CODEN: JASMAN

The Journal of the Acoustical Society of America

ISSN: 0001-4966

Vol. 123, No. 6 June 2008

(Continued)

CONTENTS—*Continued from preceding page*

ULTRASONICS, QUANTUM ACOUSTICS, AND PHYSICAL EFFECTS OF SOUND [35]

TRANSDUCTION [38]

STRUCTURAL ACOUSTICS AND VIBRATION [40]

NOISE: ITS EFFECTS AND CONTROL [50]

ARCHITECTURAL ACOUSTICS [55]

ACOUSTIC SIGNAL PROCESSING [60]

(Continued)

CONTENTS—*Continued from preceding page*

(Continued)

CONTENTS—*Continued from preceding page*

SPEECH PROCESSING AND COMMUNICATION SYSTEMS [72]

BIOACOUSTICS [80]

JASA EXPRESS LETTERS

INDEX TO VOLUME 123

Measurement of the resonance frequency of single bubbles using a laser Doppler vibrometer

Theodore F. Argo IV and Preston S. Wilson
Applied Research Laboratories, The University of Texas at Austin, P.O. Box 8029, Austin, Texas 78713-8029, and Department of Mechanical Engineering, The University of Texas at Austin, 1 University Station C2200, Austin, Texas 78712-0292
targo@mail.utexas.edu, pswilson@mail.utexas.edu

Vikrant Palan
Polytec, Inc., Tustin, California 92780
v.palan@polytec.com

Abstract: The behavior of bubbles confined in tubes and channels is important in medical and industrial applications. In these small spaces, traditional means of experimentally observing bubble dynamics are often impossible or significantly perturb the system. A laser Doppler vibrometer (LDV) requires a narrow (<1 mm diameter) line-of-sight access for the beam and illumination of the bubble does not perturb its dynamics. LDV measurements of the resonance frequency of a bubble suspended in a small tank are presented to illustrate the utility of this measurement technique. The precision of the technique is similar to the precision of traditional acoustic techniques.

PACS numbers: 43.30.Xm [GD]
Date Received: October 29, 2007 **Date Accepted:** February 28, 2008

1. Introduction

The behavior of bubbles and bubbly liquids is important in many applications ranging from underwater sound and sonar,[1,2] to industrial processes, sonochemistry, and cavitation,[3–6] to medical acoustics.[7–9] Understanding the behavior of single bubbles is fundamental to the study of collections of bubbles in any of these applications. Recently, the behavior of bubbles confined in tubes or channels and near surfaces has become important for medical and industrial applications.[10–12] In confined spaces, traditional means of experimentally observing bubble dynamics, either acoustically with hydrophones[13] or optically with Mie scattering[14] or stroboscopy,[15] are often impossible or they significantly perturb the system. A laser Doppler vibrometer (LDV)[16] only requires a narrow (less than 1 mm diameter) line-of-sight access for the beam and illumination of the bubble does not perturb its dynamics. To illustrate the utility of this measurement technique, bubble resonance frequencies, obtained from LDV measurements of the acoustically excited response of a bubble suspended in a small tank are presented and compared to theory. No absolute standard exists to assess the accuracy of bubble resonance frequency measurements, therefore the precision of the technique was considered and found to be similar to the precision of a traditional acoustic technique[17] that inferred the bubble resonance from a pair of acoustic pressure measurements.

2. Description of the apparatus and measurement procedure

A small acrylic-walled tank (35 cm × 35 cm × 13 cm, 0.625 mm wall thickness) with a tight fitting lid was filled with degassed distilled water. A single air bubble generated by a syringe and a needle was captured under a pair of parallel nylon monofilament lines (0.15 mm diam.) and positioned in the tank, as shown in Fig. 1. It has been shown that positioning a bubble of the size range used here (radii between 0.8 and 1.5 mm) with fine fibers has a negligible effect on the bubble's resonance frequency.[17] The tank was completely filled and closed so that no air remained in the tank. Acoustic excitation was provided by an electromagnetic shaker and a circular piston (2.5 cm diam.) through a hole in the tank wall and a rubber membrane. The source

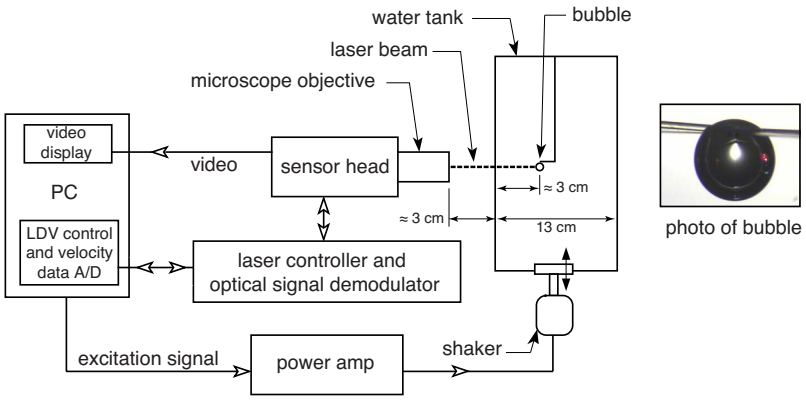


Fig. 1. (Color online) The measurement instrumentation is shown in schematic on the left. The empty arrowheads represent signal paths. A photo (obtained with the stereo microscope described in the text) of the bubble positioned underneath a pair of monofilament nylon fibers is shown above right. This pair of fibers was held in place by a wire frame that was attached to the top of the tank. The bubble was centered in the tank for the two dimensions not explicitly indicated.

signal (band limited pseudorandom noise, 1–5 kHz) was generated by the data acquisition computer and directed to a power amp and the shaker. Standing waves were thus set up inside the tank and the bubble was forced into oscillation. A pressure spectrum of the tank recorded with a miniature hydrophone (Bruel & Kjaer 8103) located near the position of the bubble, but in absence of the bubble, is shown in Fig. 2(a). Discussion of this response is deferred until Sec. 4. The normal velocity of the bubble wall was observed using a laser Doppler vibrometer (Polytec OFV-534). This procedure was repeated for four bubble sizes.

3. Description of the vibration measurement and data analysis

A LDV is based on the principle of detection of the Doppler shift of a monochromatic, directional, coherent light beam that is scattered from the surface of interest. The frequency of the scattered light (compared to the frequency of a reference beam) is used to determine the component of velocity along the axis of the incident beam.[16] A helium-neon laser with a wavelength of 633 nm is used in the vibrometer utilized in this study. A number of challenges are associated with using a LDV to measure bubble motion. The measurements must be conducted through a volume of water and the tank wall, which causes additional attenuation of the laser beam as compared to a beam path of the same length in air. If an absolute measurement of velocity is required, the optical index of refraction of the water and the tank wall must be accounted for.

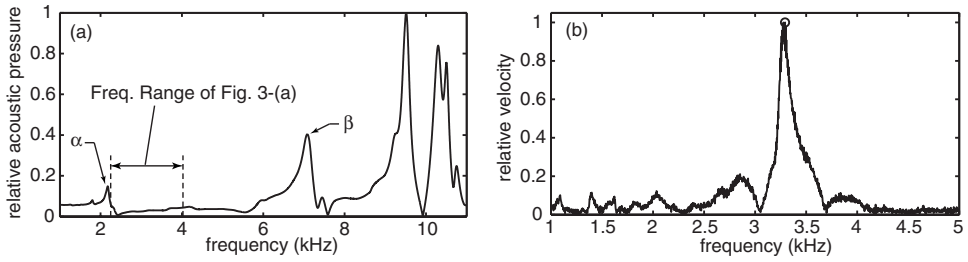


Fig. 2. The acoustic pressure spectrum measured inside the tank near the location of the bubble is shown in (a). The acoustic pressure was normalized by the maximum pressure. The frequency range of the bubble resonance measurements that appear in Fig. 3(a) is also indicated. A typical spectrum of the bubble wall velocity measured with the LDV is shown in (b). The velocity was normalized by the maximum velocity. The open circle identifies the resonance frequency.

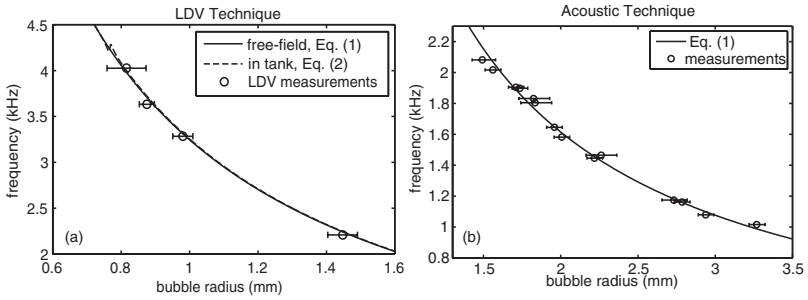


Fig. 3. In (a), bubble resonance frequencies measured with the LDV are compared to the free-field model, Eq. (1), and the rigid tank model, Eq. (2). The two curves nearly overlay one another, except for bubble radii near 0.8 mm. Results obtained with an acoustic technique (see Ref. 17) appear in (b). In both cases, the error bars represent uncertainty in the bubble size measurement.

The spherical shape of the bubble's surface further reduces the intensity of light scattered back to the interferometer, compared to that scattered from a flat surface. Finally, scattering from the tank walls must not interfere with the light scattered from the bubble.

To overcome these difficulties, an LDV sensor head (Polytec OFV-534) that projects and receives the laser beam through a microscope objective was used. A coaxial video image of the bubble was also acquired through the objective lens and displayed in real time on the data acquisition PC, which afforded precise alignment of the laser beam with the normal point on the bubble surface. The sensor head was mounted on a photographic tripod and positioned through manual manipulation of the tripod controls. The beam was focused to a diameter of approximately 2 μm, and therefore it illuminated a relatively small portion of the bubble surface. For the smallest bubble in this study, the patch illuminated by the laser deviated from a plane by at most 6 picometers, which is a negligible fraction of the optical wavelength. Hence, the LDV was effectively observing motion that was normal to the bubble surface.

An objective lens with a magnification factor of 10 was used to ensure that sufficient light was reflected back to the photodetector, as determined by near-unity coherence[18] between the input excitation signal and the measured velocity. The working distance provided by the objective allowed for unfocused light to pass through the tank wall, which in turn reduced spurious reflections to a negligible level. A time domain voltage signal that is a direct analog of the normal surface velocity of the illuminated patch on the bubble was output from the controller/demodulator (Polytec OFV-5000). A PC-based data acquisition system was used to acquire and process the velocity signals. For a given bubble size, the time domain signal was windowed and a fast Fourier transform was performed. Fifty frequency-domain averages were computed and the average spectra (resolution bandwidth=3.125 Hz) was saved. A typical spectrum is shown in Fig. 2(b). The resonance frequency of the bubble was taken to be the frequency that corresponded with the maximum amplitude of the spectra. Finally, a stereo microscope with a charge coupled device camera and diffuse white backlighting (oriented on an axis normal to the LDV axis) were used to measure the bubble size.

4. Results

The resonance frequencies extracted from the LDV velocity spectra are shown in Fig. 3(a). The error bars represent uncertainty in the measured bubble radii due to the resolution (pixelization) of the digital micrographs, which were obtained with various degrees of magnification. The solid line is the prediction of the bubble resonance frequency for free field conditions[19]

$$\omega_0 = \left[\frac{P_0}{\rho a^2} \left(\mathrm{Re}\,\Phi - \frac{2\sigma}{aP_0} \right) \right]^{1/2}, \tag{1}$$

where the hydrostatic pressure at the bubble was $P_0 = 1.03 \times 10^5$ Pa, the density of the water was $\rho = 998$ kg/m^3, a was the measured bubble radius, and the surface tension of the water was σ

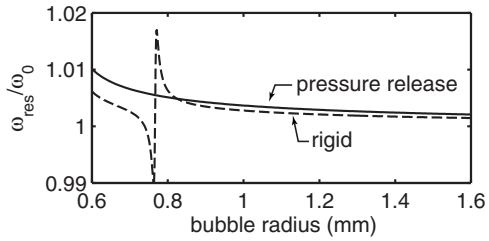


Fig. 4. The effect of tank reverberation on the resonance frequency of a bubble is shown for both rigid and pressure release tank walls. For both cases, the bubble has a slightly higher resonance frequency inside the tank, but there is less than a one percent deviation from free-field behavior, except for bubbles near 0.8 mm in radius in a rigid-walled tank.

$=0.0725$ N/m. The frequency-dependent thermal behavior of the gas is described by Φ and is given by Eq. (27) in Ref. 19. The gas properties diffusivity $D_g=2.08\times10^{-5}$ m^2/s and the ratio of specific heats $\gamma=1.4$ were used to calculate Φ.

Equation (1) is for a bubble in a free field, but the resonance frequency of a bubble can be modified by the presence of tank walls, as quantified by Eq. (35) of Ref. 13 in terms of an infinite series

$$\omega_{\text{res}} = \omega_0 + \frac{4\pi c^2 a}{\omega_0 V}\sum_{n=1}^{N} T_n, \tag{2}$$

where T_n is given by Eq. (14) of Ref. 13 and is proportional to the mode shape functions. The fully enclosed, acrylic-walled tank in this work exhibited a lowest-order resonance frequency near 2.2 kHz, as shown in Fig. 2(a), and no other resonances are present below 6 kHz. The tank used in this work can support a static pressure, hence it appears acoustically rigid at low frequencies. This assertion is supported by calculation of the lowest-order eigenfrequency of a rigid-walled tank of the same dimensions, using Eq. (9.2.7) of Ref. 20, which yields 2.1 kHz, closely matching the frequency of the peak labeled α in Fig. 2(a). At higher frequencies, the tank appears pressure release. The evidence for this is threefold. First, the higher-order resonances that are predicted for a rigid-walled tank between 2 and 6 kHz are absent in Fig. 2(a). Second, the reflection coefficient for the three medium problem given by Eq. (6.3.7) of Ref. 20, for a water-acrylic-air system between 2 and 4 kHz, is approximately -0.999, which is nearly pressure release. Third, the lowest-order resonance frequency predicted for a pressure-release tank of the same dimensions is 6.4 kHz, which is similar to the frequency of the peak labeled β in Fig. 2(a).

To bound both behavior regimes, Eq. (2) was evaluated for rigid and pressure release walls. The modal summations were taken to $N=1000$, where convergence was achieved to less than 1 part in 10^4. The quality factors of the tank's resonances are inputs to Eq. (2). For these calculations, the quality factors were all set to 100, which exceeds the actual quality factor measured inside the tank for the frequency range of the bubble measurements. The sound speed inside the tank was set to 1481 m/s. The results are shown in Figs. 3(a) and 4. Within the experimental bubble size range, the deviation between the pressure release tank model and the free field model Eq. (1) has a mean value of 0.36% and at most the deviation is 0.54%. The deviation between the rigid tank model and the free field model Eq. (1) has a mean value of 0.27% and at most the deviation is 0.64%. Since these deviations are small relative to the uncertainty in bubble size measurement, and since determining the exact effect of the tank reverberation is beyond the present scope, the free field bubble model Eq. (1) is used in Sec. 5.

For comparison to the LDV technique, bubble resonance measurements obtained with an acoustical technique[17] are shown in Fig. 3(b). The acoustic pressures inside a tank both with and without a bubble present were measured. Manipulation of these measured pressures yielded the bubble resonance frequency. The acoustic measurements are compared to Eq. (1) with P_0

$= 1.00 \times 10^5$ Pa[17] and the remaining physical parameters unchanged from the previous case.

5. Discussion

There is no accepted standard for the assessment of the absolute accuracy of a bubble resonance measurement, therefore relative measures of precision were used to characterize the two measurement techniques. The RMS error between the measurements in Fig. 3 and Eq. (1) are 1.5% for the LDV technique and 2.1% for the acoustic technique. We therefore conclude that a LDV is a viable alternative for observing bubble resonance and is capable of relative precision equal to that of a traditional acoustic technique. The noninvasive nature of the LDV technique permits use in confined spaces, such as in a narrow tube with a diameter on the order of the bubble diameter, where acoustic techniques would be difficult due to the size requirement imposed on the measurement hydrophone, and where the Mie scattering technique would be impossible, since it requires the light source and receiver axes to be separated by an angle of about 80°.

Acknowledgments

This work was supported by the University of Texas at Austin Cockrell School of Engineering, Applied Research Laboratories at the University of Texas at Austin, and Polytec, Inc.

References and links

[1]H. Medwin and C. S. Clay, *Fundamentals of Acoustical Oceanography* (Academic, Boston, 1998).
[2]T. G. Leighton, *The Acoustic Bubble* (Academic, London, 1994).
[3]K. S. Suslick and G. J. Price, "Applications of ultrasound to materials chemistry," Annu. Rev. Mater. Sci. **29**, 295–326 (1999).
[4]C. E. Brennen, *Cavitation and Bubble Dynamics* (Oxford University Press, New York, 1995).
[5]P. R. Gogate and A. B. Pandit, "A review and assessment of hydrodynamic cavitation as a technology for the future," Ultrason. Sonochem. **12**, 21–27 (2005).
[6]A. G. Chakinala, P. R. Gogate, A. E. Burgess, and D. H. Bremner, "Treatment of industrial wastewater effluents using hydrodynamic cavitation and the advanced Fenton process," Ultrason. Sonochem. **15**, 49–54 (2008).
[7]L. Hoff, *Acoustic Characterization of Contrast Agents for Medical Ultrasound Imaging* (Kluwer, Dordrecht, 2001).
[8]T. Ikeda, S. Yoshizawa, M. Tosaki, J. S. Allen, S. Takagi, N. Ohta, T. Kitamura, and Y. Matsumoto, "Cloud cavitation control for lithotripsy using high intensity focused ultrasound," Ultrasound Med. Biol. **32**, 1383–1397 (2006).
[9]X. Yang, R. A. Roy, and R. G. Holt, "Bubble dynamics and size distributions during focused ultrasound insonation," J. Acoust. Soc. Am. **116**, 3423–3431 (2004).
[10]J. Cui, M. F. Hamilton, P. S. Wilson, and E. A. Zabolotskaya, "Spherical bubble pulsation between parallel plates," J. Acoust. Soc. Am. **119**, 2067–2072 (2006).
[11]S. Qin and K. W. Ferrara, "The natural frequency of nonlinear oscillation of ultrasound contrast agents in microvessels," Ultrasound Med. Biol. **33**, 1140 (2007).
[12]R. J. Dijkink, J. P. van der Dennen, C. D. Ohl, and A. Prosperetti, "The acoustic scallop: A bubble-powered actuator," J. Micromech. Microeng. **16**, 1653–1659 (2006).
[13]T. G. Leighton, P. R. White, C. L. Morfey, J. W. L. Clarke, G. J. Heald, H. A. Dumbrell, and K. R. Holland, "The effect of reverberation on the damping of bubbles," J. Acoust. Soc. Am. **112**, 1366–1376 (2002).
[14]D. L. Kingsbury and P. L. Marston, "Mie scattering near the critical angle of bubbles in water," J. Opt. Soc. Am. **71**, 358–363 (1981).
[15]Y. Tian, J. A. Ketterling, and R. E. Apfel, "Direct observation of microbubble oscillations," J. Acoust. Soc. Am. **100**, 3976–3977 (1996).
[16]H. E. Albrecht, N. Damaschke, M. Borys, and C. Tropea, *Laser Doppler and Phase Doppler Measurement Techniques* (Springer, Berlin, 2003).
[17]V. Leroy, M. Devaud, and J.-C. Bacri, "The air bubble: Experiments on an unusual harmonic oscillator," Am. J. Phys. **70**, 1012–1019 (2002).
[18]J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Analysis*, 2nd ed. (Wiley, New York, 1993).
[19]K. W. Commander and A. Prosperetti, "Linear pressure waves in bubbly liquids: Comparison between theory and experiments," J. Acoust. Soc. Am. **85**, 732–746 (1989).
[20]L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. (Wiley, New York, 2000).

A mechanism stimulating sound production from air bubbles released from a nozzle

Grant B. Deane and Helen Czerski
Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093-0238; gdeane@ucsd.edu, hczerski@ucsd.edu

Abstract: Gas bubbles in water act as oscillators with a natural frequency inversely proportional to their radius and a quality factor determined by thermal, radiation, and viscous losses. The linear dynamics of spherical bubbles are well understood, but the excitation mechanism leading to sound production at the moment of bubble creation has been the subject of speculation. Experiments and models presented here show that sound from bubbles released from a nozzle can be excited by the rapid decrease in volume accompanying the collapse of the neck of gas which joins the bubble to its parent.

PACS numbers: 43.30.Lz, 43.30.Nb [DKW]
Date Received: January 10, 2008 **Date Accepted:** February 25, 2008

1. Introduction

It has long been known that the sounds of running water are associated with the creation of bubbles.[1] The sound produced by newly formed oceanic bubbles is of interest because these bubbles enhance the flux of greenhouse gases across the ocean surface, create aerosols, and generate underwater ambient noise.[2–5] Newly formed bubbles behave like lightly damped natural oscillators responding to nonequilibrium initial conditions, and produce a short acoustic pulse at the moment of their formation.[6] Because of their importance to a wide range of subject areas, the behavior of spherical gas and cavitation bubbles has been well studied. The acoustic behavior of spherical bubbles is governed by the Rayleigh–Plesset equation, the validity of which has been verified by numerous theoretical and laboratory studies.[6–9]

2. Bubble sound excitation mechanism

There is an extensive body of literature on the acoustical properties of gas bubbles, but comparatively little is known about the mechanism driving the production of sound when bubbles are first formed. Various mechanisms have been proposed, including the increase in the internal pressure of the bubble associated with the Laplace pressure, hydrostatic pressure effects, shape mode coupling, and a fluid jet associated with the collapsing bubble neck.[9–13] Estimates of the Laplace and hydrostatic pressure effects show that they represent a minor ($<10\%$) contribution to the noise.[9] Shape mode coupling seems to play a significant role in the damping of highly distorted bubbles[14] (e.g., bubbles fragmenting in turbulence or detaching from a nozzle), but a secondary role in bubble acoustic excitation.[15]

The idea that bubble sound production is driven by the jet of water associated with the collapse of the neck of air formed immediately after bubble pinch-off was suggested by Longuet-Higgins,[16] and has been examined by Manasseh and co-workers[12,13] for rapidly (>10 Hz) sparged bubbles. Manasseh *et al.* studied the acoustic emissions of bubbles released from a nozzle with simultaneous high-speed photographs,[12] and found that the initial fall in pressure in the fluid surrounding the bubble is associated with the neck-breaking process and the rapid retraction of the tip of the bubble once it has detached. Here we will show through experiments and an analytical model that acoustic excitation can be explained by the decrease in bubble volume that accompanies neck collapse, and is driven by surface tension forces.

Bubble detachment and concurrent acoustic emission are illustrated in Fig. 1 (see Sec. 6). The pressure pulse radiated by the bubble [Fig. 1(a)] shows that the acoustic excitation begins just before detachment, and is largely complete within a single oscillation. Image I

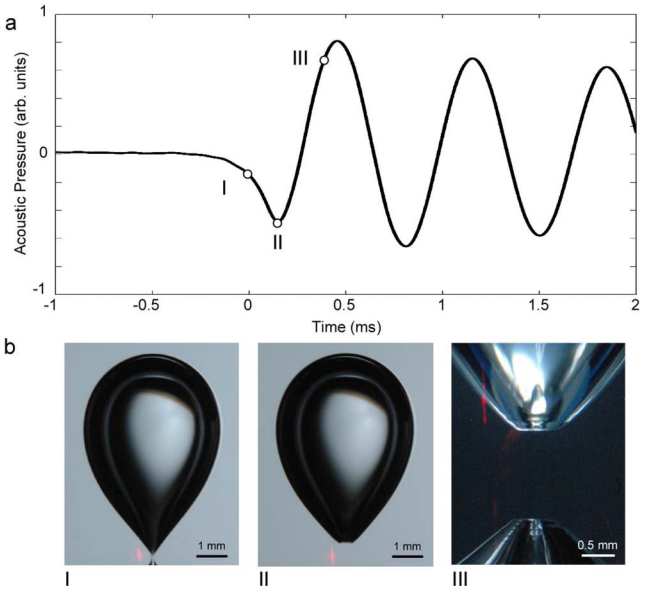


Fig. 1. (Color online) Acoustic emission and bubble release. (a) The acoustic pressure trace of a bubble released from a nozzle in the laboratory. The pressure preceding bubble release and the first few oscillations are shown. The acoustic pressure is in arbitrary units because it was measured in a small, reverberant chamber (see Sec. 6). (b) Single strobe images showing the bubble shape during acoustic excitation. The red, vertical line to the left of the bubble neck is light from the laser trigger. The release nozzle is positioned below the bottom of the images. The time of the images relative to the pressure trace is indicated by roman numerals. Image I shows the bubble immediately prior to neck rupture. Image II shows the state of neck collapse at the first pressure minimum. Image III shows a magnified view of the collapsing neck. A small, reentrant jet can be seen forming at the base of the neck and capillary waves can be seen propagating along the neck.

shows the bubble shape immediately prior to detachment. Image II, taken 160 μs after detachment, illustrates the rapid collapse of the neck remnant. Image III, taken 320 μs after detachment, shows a small reentrant jet of water forming within the collapsing neck.

3. Model for neck collapse

A simple analytical model of neck collapse can be developed as follows. Immediately preceding rupture, the bubble geometry is divided into a sphere, a cone, and a hyperbolic region (Fig. 2). The behavior of the hyperbolic region up to the point of neck rupture has been studied.[16,17] Our calculations show that it represents only a small fraction of the total volume of the collapsing neck and plays a minor role in bubble acoustic excitation. Accordingly, it is modeled as a simple cylinder of length x_0 and radius r_0. The conical region is characterized by its slope, η. For the bubble illustrated in Fig. 1, $\eta \approx 0.84$.

The radius of curvature at the neck end during collapse is small, resulting in a large Laplace pressure jump across the boundary and rapid inward acceleration. We begin by assuming that the surface tension energy in a frustrum of neck is converted into kinetic energy of the fluid within it. Following this line of reasoning yields expressions for the neck velocity in terms of distance along the neck measured from the point of rupture, x:

$$u = \begin{cases} \left(\dfrac{4\sigma}{\rho r_0}\right)^{1/2}, & x < x_0 \\ \left(\dfrac{4\sigma(1+\eta^2)^{1/2}}{\rho \eta x}\right)^{1/2}, & x \geq x_0, \end{cases} \tag{1}$$

where σ is the fluid surface tension and ρ is the fluid density. The neck collapse time is found by integrating the reciprocal of the velocity over x:

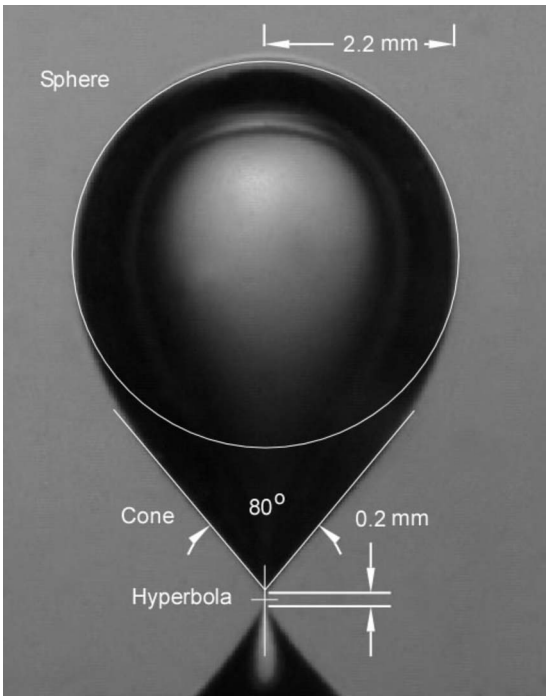


Fig. 2. Geometry for the neck collapse model.

$$\tau = \begin{cases} \left(\dfrac{\rho r_0}{4\sigma}\right)^{1/2} x, & x < x_0 \\ \left(\dfrac{\rho r_0}{4\sigma}\right)^{1/2} x_0 + \dfrac{2}{3\eta}\left(\dfrac{\rho}{4\sigma(1+\eta^2)^{1/2}}\right)^{1/2}\left[(r_0+\eta(x-x_0))^{3/2} - r_0^{3/2}\right], & x \geq x_0. \end{cases} \tag{2}$$

The velocity and collapse time are plotted in Fig. 3 as a function of x along with measurements of these quantities for the bubble illustrated in Fig. 2 (see Sec. 6). Neglecting the small volume of the hyperbolic region by setting $x_0 = r_0 = 0$, Eq. (2) can be used to calculate the decrease in neck volume with time:

$$\Delta V = -\frac{3\pi\sigma\eta(1+\eta^2)^{1/2}}{\rho} t^2. \tag{3}$$

where $t = 0$ at the moment of neck rupture.

4. Forcing term in the Rayleigh–Plesset equation

The rapidly changing neck volume drives the bubble into breathing mode oscillations. The dynamics of the collapse are inherently nonspherical, but we assume that the breathing mode response of the bubble can be described assuming spherical symmetry. To drive the bubble, we have calculated the change in external pressure that would be required to account for the change in bubble volume associated with the neck collapse. The decreasing volume results in an increase in pressure inside the bubble, which can be calculated by assuming the polytropic relationship:

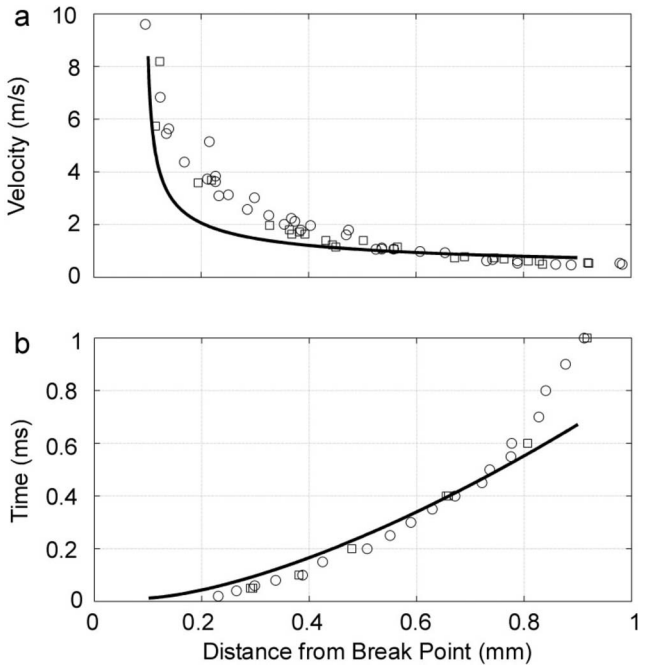


Fig. 3. Comparison of neck collapse model with experimental data. Circles and squares respectively correspond to fresh and salt water measurements for r_0=4.5 μm and x_0=100 μm. The cylinder radius was determined from the neck velocity at the end of cylinder collapse and the cylinder length was estimated from the bubble photographs. (a) The velocity of the base of the collapsing neck as a function of distance from the rupture point. (b) The collapse time versus distance from the rupture point.

$$p_{\text{in}} = p_{\text{in},0}\left(\frac{V_0}{V_0 + \Delta V}\right)^{\kappa}, \tag{4}$$

where p_{in} is the gas pressure internal to the bubble, $p_{\text{in},0}$ is the equilibrium pressure inside the bubble, V_0 is the equilibrium volume of the bubble, κ is the gas polytropic index, and we have neglected thermal losses associated with gas compression driven by the neck collapse. Thermal losses due to the natural response of the bubble are accounted for using an effective thermal viscosity, μ_{th}, as described by Prosperetti.[7] The required external pressure change is calculated from Eq. (4) by invoking continuity of normal stress across the bubble wall and neglecting surface tension and viscous forces there. The final result is a driving term on the right hand side of the linearized Rayleigh–Plesset equation:

$$\frac{\partial^2 \varepsilon}{\partial t^2} + \left[\frac{4(\mu + \mu_{\text{th}})}{\rho R_0^2} + \frac{kR_0}{1 + k^2 R_0^2}\omega\right]\frac{\partial \varepsilon}{\partial t} + \left[\frac{3\kappa p_{in,0}}{\rho R_0^2} - \frac{2\sigma}{\rho R_0^3} + \frac{k^2 R_0^2}{1 + k^2 R_0^2}\omega^2\right]\varepsilon = f(t), \tag{5}$$

where ε is the fractional increase in bubble radius, μ is the fluid viscosity, k and ω, respectively, are the wave number and angular frequency of sound at bubble resonance, R_0 is the bubble equilibrium radius, and the forcing function is given by

$$f(t) = -\frac{9\kappa\sigma\eta p_{in,0}(1 + \eta^2)^{1/2}}{4\rho^2 R_0^5} t^2; \quad t > 0. \tag{6}$$

In deriving Eq. (5), we have assumed that the velocity and displacement coefficients inside the square brackets are constant and take the values they have at the bubble's natural frequency. For the approximately 2 mm bubbles we are studying, the frequency-dependent radiation term (the k term in square brackets multiplying ε) is less than 1/5000 of the dominant pressure term and

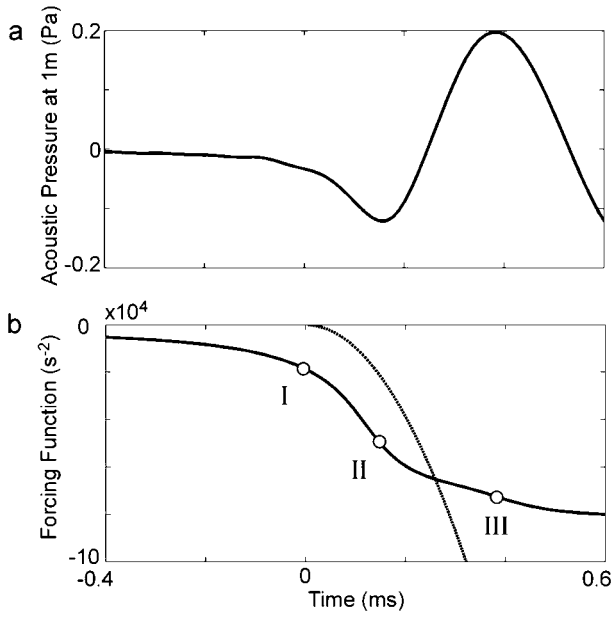


Fig. 4. Comparison of forcing function calculations with experimental data. (a) The acoustic pressure trace from a bubble released in a large pool. The bubble size and release nozzle were similar to those for the photographed bubbles. (b) The measured and theoretical forcing functions. The solid line is the measured forcing function, determined from analysis of bubble emissions. The roman numerals indicate the time of the images shown in Fig. 1. The dotted line corresponds to the forcing function calculated according to Eq. (6) with R_0=2.2 mm, σ =0.072 N/m, ρ=1000 kg/m^3, κ=1.4, and η=0.84.

can be neglected. The frequency-dependent effects of damping have been estimated (see below), and can also be ignored.

The initial shape of the radiated acoustic pulse can be analyzed to provide an estimate of the bubble forcing function, which can then be compared with Eq. (6). Although we measured the bubble acoustic signature in a small, acrylic chamber, we have not analyzed it because of the effects of reverberation. The problem with reverberation is not the level of signal which it contributes to the hydrophone output, but rather that it represents a coherent driving signal with a specific phase on the bubble wall.[18] To mitigate this problem, we have analyzed acoustic pulses recorded in a reverberation-limited test pool[19] radiated by bubbles of the same size (within 10%) and released from the same nozzle as those photographed in the small acrylic chamber but at a depth of 2 m (within 5%). These measurements were made with an International Transducer Corporation 6050C hydrophone. Because the nearest reflecting boundary is 2 m distant, the first 2.7 ms of the bubble signature is free from the effects of reverberation.

The acceleration of the bubble wall can be calculated from the acoustic radiation using $\partial^2\varepsilon/\partial t^2 = p_a/(\rho R_0^3)$, where p_a is the acoustic pressure 1 m from the bubble.[6] Note that an initial inward acceleration of the bubble wall results in the radiation of a rarefaction. The acceleration can then be integrated once to obtain $\partial\varepsilon/\partial t$ and again to obtain ε. The first three terms in Eq. (5) are then summed to obtain $f(t)$. This procedure is similar to that carried out by Pumphrey and Elmore,[20] who computed the bubble wall motion from analytical expressions for the acoustic field. Here we apply this technique to the measured field, which enables us to compute the driving term on the Rayleigh–Plesset equation. The result of the procedure is shown in Fig. 4. The top trace shows the reverberation-limited acoustic field. The bottom trace shows the mean forcing function estimated from 84 bubbles compared with Eq. (6). Approximately 1/4 of the forcing occurs before the neck detachment, presumably driven by the radial influx of the neck walls before rupture, but the main part of the excitation occurs during neck collapse. The forc-

ing function was calculated with no damping and with ten times measured damping to explore its effects (not shown). The case of no damping produced essentially no change in the forcing function, while ten times damping changed the forcing function estimate by less than 10%.

The general shape and magnitude of the forcing is adequately accounted for by the neck collapse theory once predetachment forcing is allowed for. The neck collapse model does not contain any dissipation mechanisms and does not model the collapse behavior past the first few hundred microseconds, so the theoretical curve deviates significantly from the measured response for times greater than a few hundred microseconds.

5. Concluding remarks

We have shown that the sound produced by a bubble released from a nozzle can be explained by the collapse of the neck of air formed immediately after bubble pinch-off. The collapsing neck decreases the bubble volume and excites the bubble into oscillation. The initial response of the bubble depends on the time scale of the neck collapse and the natural frequency of the bubble. This can be seen in the width of the first rarefaction trough of the acoustic pressure trace in Fig. 1, which is noticeably less than subsequent troughs. No significant differences were observed in the neck collapse process between fresh and salt water. The collapse is driven by surface tension energy in the neck, and a simple energy balance model gives a first-order description of the process. The change in bubble volume associated with the neck collapse can be incorporated into the linearized Rayleigh–Plesset equation to derive an acoustic forcing function. An estimate of this function based on the analysis of bubble acoustic emissions compares favorably with the model predictions.

We believe that the role of neck collapse in stimulating bubble sound is not limited to nozzle release. Surface tension energy has been implicated in the acoustic stimulation of fragmenting bubbles,[15,19] and it is probable that neck collapse drives sound production from fragmentation and therefore noise production by breaking waves.[21]

6. Methods summary

The bubble-release nozzle was positioned 5 cm above the base of a transparent acrylic tank with a 13-cm-square cross section. The tank was filled with either fresh or salt water to a depth of 25 cm. An International Transducer Corporation 1089D was placed alongside the nozzle tip, approximately 1 cm from the bubble. Air bubbles were released at a nominal rate of ten per minute. Bubble release was determined by interruption of a laser beam directed through the neck region, which was used to trigger data acquisition and flash lighting. Flashes from two strobe lights, each of 15 μs duration, illuminated the bubble with a preprogrammed interval. Both images were superposed on a single frame, permitting observation of the neck development over short time intervals relative to the moment of neck rupture. The neck displacements were measured by increasing the inter-flash interval, with the first flash simultaneous with neck rupture. The neck velocity was measured using a short inter-flash interval with flash pairs occurring at successively later times after neck rupture. The reverberation-limited bubble signature measurements were taken during the experiment described in Deane and Stokes.[19]

Acknowledgments

We thank Dr. Dale Stokes and Cary Humphries for assistance with the acoustic data collection, and Dr. David Farmer for discussions. We acknowledge financial support from the National Science Foundation and the Office of Naval Research. We are also grateful to two anonymous reviewers who helped us improve the manuscript.

References and links

[1]M. Minnaert, "On musical air bubbles and the sounds of running water," Philos. Mag. **16**, 235–248 (1933).
[2]E. Lamarre and W. K. Melville, "Air entrainment and dissipation in breaking waves," Nature (London) **351**, 469–472 (1991).
[3]C. D. O'Dowd and G. de Leeuw, "Marine aerosol production: A review of the current knowledge," Philos. Trans. R. Soc. London, Ser. A **365**, 1753–1774 (2007).
[4]D. M. Farmer, C. L. McNeil, and B. D. Johnson, "Evidence for the importance of bubbles in increasing air-sea

gas flux," Nature (London) **361**, 620–623 (1993).
[5]W. K. Melville, "The role of surface-wave breaking in air-sea interaction," Annu. Rev. Fluid Mech. **28**, 279–321 (1996).
[6]T. G. Leighton, *The Acoustic Bubble* (Academic, London, 1994).
[7]A. Prosperetti, "Thermal effects and damping mechanisms in the forced radial oscillations of gas bubbles in liquids," J. Acoust. Soc. Am. **61**, 17–27 (1977).
[8]M. S. Plesset and A. Prosperetti, "Bubble dynamics and cavitation," Annu. Rev. Fluid Mech. **9**, 145–185 (1977).
[9]H. C. Pumphrey and J. E. Ffowcs Williams, "Bubbles as sources of ambient noise," IEEE J. Ocean. Eng. **15**, 268–274 (1990).
[10]M. S. Longuet-Higgins, "An analytic model of sound produced by raindrops," J. Fluid Mech. **214**, 395–410 (1990).
[11]R. D. Hollet and R. M. Heitmeyer, "Noise generation by bubbles formed in breaking waves," *Sea Surface Sound* (Kluwer, Boston, 1998).
[12]R. Manasseh, S. Yoshida, and M. Rudman, "Bubble formation processes and bubble acoustic signals," *Proceedings of the Third International Conference on Multiphase Flow (ICMF'98)*, Lyon, France, 8–12 June 1998, paper no. 202 (CD-ROM).
[13]R. Manasseh, A. Bui, J. Sandercock, and A. Ooi, "Sound emission processes on bubble detachment," *Proceedings of the 14th Australasian Fluid Mechanics Conference*, Adelaide, South Australia, edited by B. B. Dally, 9–14 December 2001, Vol. **1**, pp. 857–860.
[14]M. S. Longuet-Higgins, "Nonlinear damping of bubble oscillations by resonant interaction," J. Acoust. Soc. Am. **91**, 1414–1422 (1992).
[15]G. B. Deane and M. D. Stokes, "The acoustic emissions and energetic of fragmenting bubbles," J. Acoust. Soc. Am. (in review).
[16]M. S. Longuet-Higgins, "The release of air bubbles form an underwater nozzle," J. Fluid Mech. **230**, 365–390 (1991).
[17]S. T. Thoroddsen, T. G. Etoh, and K. Takehara, "Experiments on bubble pinch-off," Phys. Fluids **19**, 042101 (2007).
[18]T. G. Leighton, P. R. White, C. L. Morfey, J. W. L Clarke, G. J. Heald, H. A. Dumbrell, and K. R. Holland, "The effect of reverberation on the damping of bubbles," J. Acoust. Soc. Am. **112**, 1366–1376 (2002).
[19]G. B. Deane and M. D. Stokes, "The acoustic signature of bubbles fragmenting in sheared flow," J. Acoust. Soc. Am. **120**, EL84–EL89 (2006).
[20]H. C. Pumphrey and P. A. Elmore, "The entrainment of bubbles by drop impacts," J. Fluid Mech. **220**, 539–567 (1990).
[21]G. B. Deane and M. D. Stokes, "Scale dependence of bubble creation mechanisms in breaking waves," Nature (London) **418**, 839–844 (2002).

Uncertainties caused by source directivity in room-acoustic investigations

Ricardo San Martín and Miguel Arana
Physics Department, Public University of Navarre, Campus de Arrosadía, 31006 Pamplona, Spain
ricardo.sanmartin@unavarra.es, marana@unavarra.es

Abstract: Although deviations in the measurement of acoustic parameters should be lower than the subjectively perceivable change in the corresponding parameter measured, this study reflects that directionality of sound sources could cause wide audience areas to break away from this criterion at high frequencies, even when using dodecahedron loudspeakers which meet the requirements of the ISO 3382 standard. The directivity of four different acoustic sources was measured and the influence of its accurate orientation spatially quantified in five enclosures for speech and music. By means of simulation software, the number of receivers affected by uncertainties greater than difference limens was established.

PACS numbers: 43.55.Mc, 43.55.Ka, 43.38.Ja [NX]
Date Received: October 18, 2007 **Date Accepted:** November 12, 2007

1. Introduction

The directionality of a sound source has a significant effect on the results obtained for acoustic parameters derived from the impulse response, especially at high frequencies. This is found to be true even for dodecahedron loudspeakers meeting the requirements of the ISO 3382 standard[1] regarding the directional patterns of sound sources. Results obtained from different sources can differ substantially, so much that it could lead to misleading evaluations of the acoustic attributes at the same source–receiver position. Even for two random–specific orientations of the same source, results may differ by more than the subjectively perceivable change—one just noticeable difference(jnd)—of the corresponding parameter.[2]

Up to now, research on uncertainties due to source orientation has been limited to, as well as by, measurement procedures. The peculiarity of the phenomenon, which appears solely at higher frequencies, seems to suggest that simulation software—whose main limitation is ray tracing at low frequencies—is not a limiting tool. It may enable us to test the "source orientation" variable spatially in a time-saving way, in sharp contrast to the use of measurements. In addition, as can be seen in this work, present display facilities implemented by developers of the programs are a great aid for interpreting results.

By means of simulation software, this study has set out to determine the effect of the directivity of sources—which meet the requirements of ISO 3382—on the results of acoustic parameters in enclosures for speech and music where such a standard is applied. Four different *omnidirectional* types of sources were surveyed and tested. Three of them, S1, S2, and S3, were different commercial dodecahedron sources while S4 was the source developed at the Institute of Technical Acoustics in Aachen, Germany.[3] This last source consists of a three-way measurement loudspeaker in which a subwoofer is used to achieve the required sound power at low frequencies and two specially designed dodecahedron speakers with different diameters are made use of to improve the omnidirectional sound radiation in comparison with the one obtained through conventional singular dodecahedron measurement devices. So as to simulate the directional pattern of the four sources, the radiated fields were measured in an anechoic chamber. By means of a computer-controlled measurement procedure, a half-sphere of measured data in 5° angular increments between adjacent points was obtained. The full sphere was built by applying symmetry rules. Odeon© room acoustics software was used during the simulation procedure.

Table 1. Basic data of the halls.

Hall, location	Volume (m^3)	Seats	$T_{30\ mid}$[a] (s)	Shape	Rays/oct
Sarasate Theatre, Pamplona (Spain)	4 000	480	1.0	Rectangular	3 600
Bretón Theatre, Logroño (Spain)	6 300	988	1.3	Rectangular	7 000
Elmia Hall, Jönköping[b] (Sweden)	11 000	1100	2.2	Fan	15 000
Baluarte Concert Hall, Pamplona (Spain)	20 000	1568	1.9	Rectangular	30 000
Mozart Concert Hall, Zaragoza (Spain)	25 800	1992	2.9	Semi-surround	18 000

[a]From measurements according to ISO 3382 for "normal coverage."
[b]Reference 4.

2. Measurement and simulation procedures

The reliability of a room acoustical simulation varies according to a range of calculation parameters: geometry data, absorption and diffusivity coefficients of surfaces, number of rays, length of the calculated impulse response, etc. For an accurate prediction of room acoustic parameters the modeling of surface reflections often needs to be polished through the aid of measurements. This has been the case of the present research work. A complete set of measurements was designed for the prior characterization of the halls to be evaluated. They were made in unoccupied conditions and with a number of measurement positions chosen under "normal coverage" based on ISO 3382 standard requirements. The measurement technique of room impulse responses based on logarithmic sweep sine excitations was resorted to.

With the aim of selecting an adequate recreation of typical conditions in enclosures for speech and music, four different halls—all of them located in northern Spanish cities—were chosen. At the same time, a hall well known by acousticians was included in the assessment. The geometrical, absorption, and diffusivity data of the Elmia Hall in Jönköping, Sweden, were forwarded to the participants at the Second International Round Robin on Room Acoustical Simulation.[4] Moreover, this hall along with calculation parameters used by Odeon was provided by the manufacturers to their users as demo data on purchasing the software. The theatres and concert halls selected are listed in Table 1 along with basic information on each. They have reverberation times within the range of 1.0–2.9 s at midfrequencies, their volumes ranging from 4000 to 25 800 m^3 and three different shapes, i.e., rectangular, slightly fan-shaped, or semi-surrounded.

The calculation parameters were picked taking into account previous work[5] and recommendations of the program. In general, the length of the impulse response should be similar to the reverberation time of the room and in our case the number of rays was established following the criteria consisting of at least one direct ray per square meter in the most distant receiver from the source. This led to values from 3600 to 30 000 rays per octave depending on both the size of the halls and the location of the audience zones. When unknown, absorption coefficients were selected from the library provided by the software used or literature. Diffusion coefficients of 0.1 and 0.7 for smooth and rough surfaces respectively were applied. These coefficients are specified for the middle frequency around 700 Hz. Then the software expands them into values for each octave band according to typical conditions for materials.

In order to determine the effect of the source orientation on the results of various acoustic parameters, 72 simulations were carried out with each source. A full 360° rotation was covered in 5° steps. For the spatial analysis, a grid of receivers — one receiver per m^2—was placed all over the audience zone excluding balconies. Over 3000 receivers, each one "measuring" seven parameters—T_{30}, EDT, C_{50}, C_{80}, T_s, G, and LF—at eight frequency bands for each simulation and source, were finally utilized. Over 50 million data had to be carefully evaluated by means of MATLAB technical computing software.

3. Results and discussion

The first question to be tackled was the reliability of the simulations. With the help of the measurements conducted for the characterization of the halls, absorption coefficients of surfaces were slightly adjusted in order to fit the values of reverberation times in several test points with those obtained through simulation. Definitive simulated values differed less than 5%—jnd of T_{30}—with respect to the measured reverberation time in each evaluated position. Furthermore, the bearing of the source orientation on measurements was also assessed by considering one source–receiver position for each hall at the central part of the audience and turning the source in 5° steps.[2]

Mm. 1. shows the results obtained for source S2 at the Baluarte Concert Hall. Sound pressure level of solely the direct sound at 2 kHz octave band is represented for each one of the 72 orientations successively. The SPL_{direct} value is subject to the lobe shape of the dodecahedron loudspeaker, i.e., whether a maximum or a minimum of the source directivity is facing the receiver. Furthermore, the corresponding octave-band directivity balloon plot of the source along with a graph containing measured and simulated values in the test receiver—which is highlighted inside the grid map—can also be observed. Both measurement and simulation values follow a similar pattern.

[Mm 1. (Color online) SPL_{direct} for S2 at 2 kHz while turning the source. (Top-left hand window) Directivity balloon plot of the source. (Bottom-left hand window) Measured and simulated values in the test receiver, highlighted. Room: Baluarte Concert Hall. This is a file of type animated "gif" (3.6 Mbytes).]

To establish a final verification of simulation reliability, variations on the C_{80} parameter are shown in Mm. 2. The same source, concert hall, and receiver test are displayed; however, in this case, 4 kHz octave band is the frequency chosen for representation. The mean value for C_{80} in the highlighted receiver is in the vicinity of −2 dB but the variation of the parameter value ranges from −3 to −1.2 dB. When we evaluated the direct sound along with the whole impulse response, we also found that the coincidence between measured and simulated values favored the simulation software and its ability to accurately predict even minor changes in measurement conditions.

[Mm 2. (Color online) C_{80} for S2 at 4 kHz while turning the source. (Top-left hand window) Directivity balloon plot of the source. (Bottom-left hand window) Measured and simulated values in the test receiver, highlighted. Room: Baluarte Concert Hall. This is a file of type animated "gif" (3.6 Mbytes).]

Once the simulations were validated for all rooms by means of measurements—except for the Elmia Hall, which was tested with the aid of the demo data provided by the manufacturers—the task was to find a way of processing the vast amount of data available. On identifying the standard deviation including the effect of the source orientation—STD_S—as the parameter that characterized the dispersion of the results observed in each receiver when carrying out the 72 corresponding simulations on each source, it was possible to draw up grid plots such as those shown in Fig. 1 covering the different halls. Relative STD_S with the jnd of the respective parameter as a reference was preferred for depiction. Thus, apart from displaying the results on a sole scale for all parameters, the value of relative STD_S was also representative of the change in subjective perception that the uncertainties could produce. The jnds Bork[4] used in the Round Robin—estimated from the data of Cox *et al.*[6] and Bradley *et al.*[7]—have been made use of. As can be seen in Fig. 1, there are several variations depending on the position within the hall. The use of a gray scale for the diagram enables us to locate the most affected zones. Central areas may be more affected by the directivity of the source. The proximity of a wall and subsequently the arrival of a reflection from others may help to compensate for the uncertainty which arises as a result of the difference of levels from the direct sound. If we use ITDG—the time gap between the arrival of the direct sound and the first sound reflected from the surfaces of the room—as an indicator of the receiver positions and plot STD_{S2} at 4 kHz for all receivers, see Fig. 2(a), we will be able to detect a tendency of higher deviations when ITDG increases.

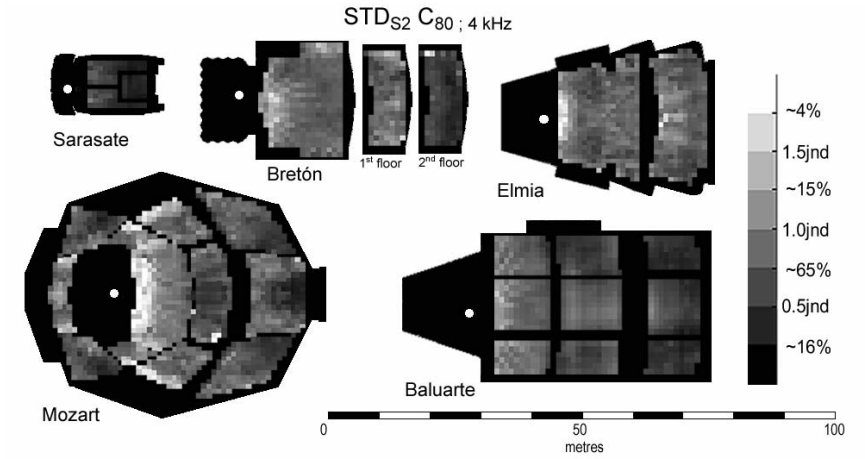


Fig. 1. Relative STDs for C_{80} parameter obtained for S2 at 4 kHz for all rooms. Reference: 1 dB—jnd for C_{80}. At colorbar, the percentage of receivers whose STDs remain within an interval of half a jnd is also displayed.

A different representation of the influence of the source orientation on acoustic parameters, particularly on clarity C_{80}, is shown in Fig. 2(b). Standard deviations STD_S within the hall are plotted again with jnd as a reference and the rate of receivers affected as well. This plot also casts light on the magnitude of the problem being analyzed. When measuring acoustic parameters, uncertainty should be lower than the subjectively perceivable change in the corresponding parameter measured. Using STD as the parameter that characterizes the deviation of the results in a measurement and stating the related 95% confidence interval, the maximum permissible difference limen should be twice the STD of the measurement apparatus. Hence, the STD should not be larger than half the jnd of the parameter under discussion. At 8 kHz octave band, this criterion is not fulfilled in over 85% of receivers in the case of S1, S2, and S3, and 18% of receivers when using such a specific source as S4.

Finally, Fig. 3(a) shows the percentage of receivers affected depending on each one of the sources in the five halls under study. For low frequencies, source directivity had no effect at all on the simulated parameter and there were similar findings for all orientations. Sources S2 and S3 delivered satisfactory results up to 1 kHz, while S1 and S4 increased this limit up to 2 and 8 kHz, respectively. At the highest frequencies, 4 and 8 kHz octave bands, the criterion of tolerance established was not met in over 80% of the receivers for the three commercial loudspeakers. Moreover, at 1 and 2 kHz octave bands, which are more common when resorting to

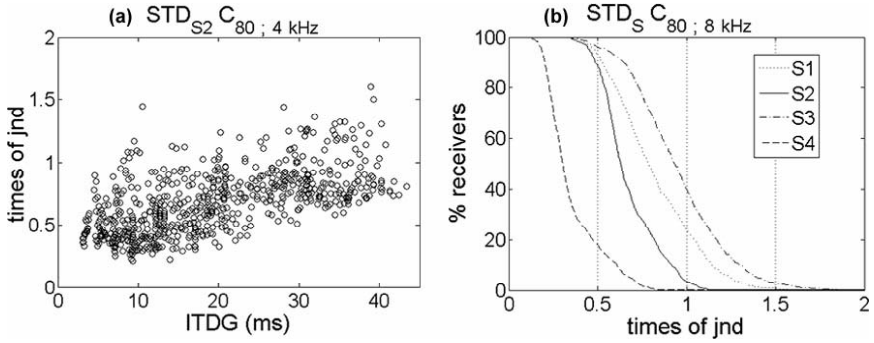


Fig. 2. Relative STDS for C_{80} parameter (a) for S2 at 4 kHz vs ITDG and (b) for all sources at 8 kHz vs percentage of receivers affected. Reference: 1 dB—jnd for C_{80}. Room: Baluarte Concert Hall.

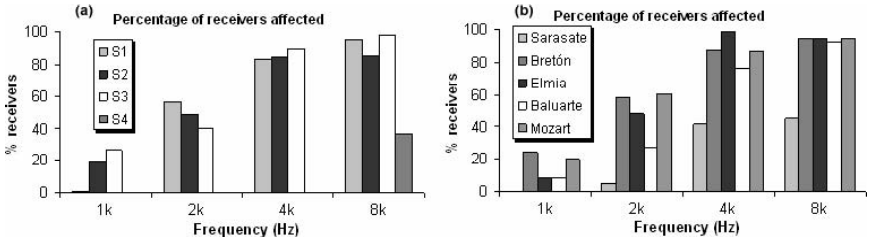


Fig. 3. Percentage of receivers with STDs for $C_{80} > 0.5$jnd (a) for all rooms and (b) for the three commercial dodecahedron loudspeakers.

average figures to define the acoustic properties of a room, a deviation any higher than half the jnd of the parameter would be expected in at least 15% and 40% of receivers, respectively, when measuring with sources S2 or S3. The described S4 measuring system has a better performance at higher frequencies and only at 8 kHz does sound radiation become perceivably directional. By setting aside that source which is only common for highly specialized measurement purposes and attempting to globally analyze the findings obtained by means of loudspeakers with common directivity patterns, we came up with the differences found in the test rooms [see Fig. 3(b)].

STD_S figures for the rest of the simulated acoustic parameters were analyzed and results similar to these reflected here for C_{80} were collected although the magnitude of uncertainty changed somewhat. Such as was stated in previous research by measurement procedures,[2] C_{50} turned out to be the most sensitive parameter, while reverberation times, whose calculation usually requires a complete integration along the time scale, were the least sensitive to the directivity of the source.

4. Conclusions

Enhanced computing power enables the acoustician to test a vast array of variables efficiently, thus making it possible to come up with configurations which would otherwise be unfeasible solely through measurement procedures. The use of simulation software could aid us in diminishing the complex procedures required when spatial distributions within halls are analyzed. In this research, a first attempt to predict uncertainties in the measurement chain of room acoustic parameters was made by means of simulation techniques. More specifically, spatial variation of uncertainty due to source orientation was analyzed. Simulated results seemed consistent with those obtained through measurement procedures.

The loudspeakers which were tested delivered satisfactory results up to 1 kHz. Below that frequency band, the bearing of source orientation is negligible for all acoustic parameters. As frequency is increased, sound radiation becomes more directional and the effect on the parameters cannot be neglected any longer. Variations on parameter figures depend on the source, the frequency band, the way those parameters are derived, as well as the position within the hall where they are going to be measured. The use of typical commercial dodecahedron sources could lead to high deviations in wide areas of audience zones of common enclosures for both speech and music. At 1 and 2 kHz octave bands, the percentage of receivers affected with uncertainties higher than the subjectively perceivable change exceeds 15% and 40%, respectively. Furthermore, at higher frequencies, a deviation greater than half the jnd of the parameter could be expected in at least 80% of receivers. Source rotations through three angular positions and the posterior averaging of subsequent measurement results—ISO 3382 requirement if the source directivity is found to have a significant effect on measured acoustic parameters—should not be discarded even in the case of dodecahedron loudspeakers having the ISO standard qualification. If the aim is to make it possible to compare measurements among different teams and equipments more accurately, the standard probably relaxes omnidirectional requirements at higher frequencies. A far-reaching aim would be to consider alternatives for assessing source omnidirectionality such as those proposed by Leishman *et al.*[8] since authors suspect that the

ISO 3382 classification method may lead to misguided conclusions about omnidirectional performance owing to its implementation by using single-plane measurement arcs.

Acknowledgment

Special thanks to Ingo Witew (Institute of Technical Acoustics, Aachen, Germany) for providing the different loudspeaker directivity data.

References and links

[1]International Organisation for Standardisation (1997). "Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters," ISO 3382, Geneva, Switzerland.
[2]R. San Martín, I. B. Witew, M. Arana, and M. Voriänder, "Influence of the source orientation on the measurement of acoustic parameters," Acta. Acust. Acust. **93**, 387–397 (2007).
[3]G. K. Behler and S. Müller (2000). "Technique for the derivation of wide band room impulse response," *Proceedings of the EAA Symposium on Architectural Acoustics*, Madrid, Spain.
[4]I. Bork, "A comparison of room simulation software – The 2nd round robin on room acoustical computer simulation," Acust. Acta Acust. **86**, 943–956 (2000).
[5]R. San Martín and M. Arana, "Predicted and experimental results of acoustic parameters in the new Symphony Hall in Pamplona, Spain," Appl. Acoust. **67**, 1–14 (2006).
[6]T. J. Cox, W. J. Davies, and Y. W. Lam, "The sensitivity of listeners to early sound field changes in auditoria," Acustica **79**, 27–41 (1993).
[7]J. S. Bradley, R. Reich, and S. G. Norcross, "A just noticeable difference in C_{50} for speech," Appl. Acoust. **58**, 99–108 (1999).
[8]T. W. Leishman, S. Rollins, and H. M. Smith, "An experimental evaluation of regular polyhedron loudspeakers as omnidirectional sources of sound," J. Acoust. Soc. Am. **120**, 1411–1422 (2006).

Determination of acoustic attenuation in the Hudson River Estuary by means of ship noise observations

Heui-Seol Roh, Alexander Sutin, and Barry Bunin
Davidson Laboratory, Stevens Institute of Technology, Hoboken, New Jersey 07030, USA
hroh@stevens.edu, asutin@stevens.edu, bbunin@stevens.edu

Abstract: Analysis of sound propagation in a complex urban estuary has application to underwater threat detection systems, underwater communication, and acoustic tomography. One of the most important acoustic parameters, sound attenuation, was analyzed in the Hudson River near Manhattan using measurements of acoustic noise generated by passing ships and recorded by a fixed hydrophone. Analysis of the ship noise level for varying distances allowed estimation of the sound attenuation in the frequency band of 10–80 kHz. The effective attenuation coefficient representing the attenuation loss above cylindrical spreading loss had only slight frequency dependence and can be estimated by the frequency independent value of 0.058 dB/m.

PACS numbers: 43.30.Nb, 43.30.Xm [WS]
Date Received: October 2, 2007 **Date Accepted:** February 24, 2008

1. Introduction

Underwater acoustic propagation in a shallow water estuary is of interest in acoustic detection of various surface and underwater threats,[1–4] acoustic tomography in the estuary,[5–7] and underwater communication.[8,9] The Maritime Security Laboratory (MSL) at Stevens Institute of Technology has recently conducted intensive investigations of acoustic wave propagation in the Hudson-Raritan Estuary. One of the important acoustic parameters is the attenuation of sound. The conventional methods of sound attenuation measurements are based on measurements of attenuation between a transmitter and a receiver.[10–13] These tests require two boats and cannot be safely conducted near navigation channels where research vessels could interfere with routine water traffic.

We applied a simpler method of attenuation calculation that is based on measurements of acoustic noise produced by passing ships and simultaneous measurements of ship locations. A similar method was considered in Ref. 13 where it was applied in the low frequency range (below 600 Hz). We are interested in a much higher frequency band (10–80 kHz) that can be used for tomography, underwater communication, and underwater threat detection in very shallow water.

2. Experimental results

In our experiments, the level of acoustic noise due to ship traffic was measured by a single hydrophone simultaneously with the distance measurements made between the hydrophone and a passing ship. This distance was measured by a video-based surface traffic tracking system developed at Stevens. This paper presents results of a test conducted on December 21, 2006 in the Hudson River Estuary near Manhattan.

Figure 1 shows the part of the Hudson River where the test was conducted adjacent to the Stevens campus. The water depth in the area of water traffic is about 15 m; the depth in the point of signal recording was 10 m. Hence there was slow depth variation along the path of the acoustic wave propagation.

Fig. 1. (Color online) Picture of the Stevens Institute of Technology campus and the test sector. The arrow shows the hydrophone position.

The wind speed at the time of measurement was about 11.3 knots and it produced waves with amplitude about 0.5 m. The surface currents were directed parallel to the coastline from north to south. Surface salinity was 17 psu and the surface temperature 10.5 °C.

The estimation of passing ships' position and the distance data from the hydrophone to the ships were obtained from ship traffic video recorded by the video-based Surface Traffic Tracking System of MSL. The prototype video surveillance system was developed at Stevens to track surface and aerial targets in a large sector of the Hudson River along the West Side of Manhattan. The system is based on an array of cameras and is designed to track, focus, and zoom in on multiple targets and events. The purpose of the tracking system is to provide real-time videos surveillance of traffic on the Hudson River, to provide for interactive video and acoustic sensor inputs, and, ultimately, to provide for automatic extraction of anomalous vessel behavior.

The noise levels of passing ships were recorded by a hydrophone (Reson TC4014) placed near the river bottom on a stand of 0.6 m height. The signal from the hydrophone was preamplified, filtered in the frequency band 5–90 kHz, transformed to digital signal, and stored in a special purpose computer on board the Stevens Research Vessel Savitsky. In signal post-processing, the calculation of spectral density in the recorded acoustic signal was conducted taking into account the hydrophone sensitivity, preamplifier gain, and transfer function of the filters.

To obtain the sound attenuation we used data from the recording of the noise of three small boats (New York Ferry, small private boats) moving along the Hudson River. The recorded noise was averaged in 10 s time window. Figure 2 shows noise levels measured by the hydrophone for a fast ferry for several distances. As the distance between hydrophone and ferry decreases, noise level increases. Speed of the passing ships is about 20 knots.

For estimation of the effective attenuation coefficient we presented the transmission loss (TL) as sum of cylindrical spreading loss and additional attenuation. This allows to express the Noise Level (NL) produced by a ship at the point of hydrophone in form,

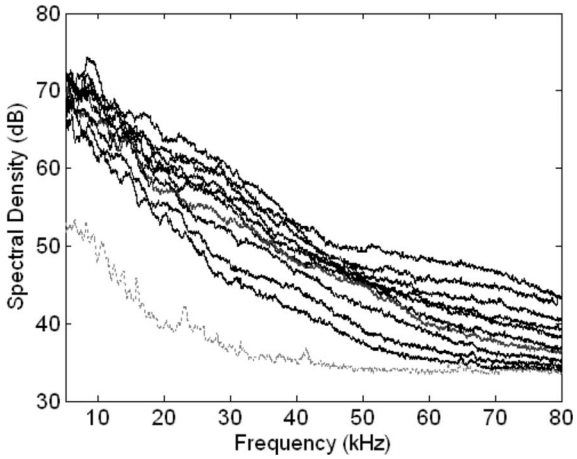


Fig. 2. Spectral noise density measured for a fast ferry for ten noise levels (dB re μPa/$\sqrt{\text{Hz}}$) for distances between 580 and 724 m with steps of 16 m (from the top to bottom, respectively). The lowest level (dotted line) represents recorded ambient noise without any water traffic.

$$NL = SL - TL = SL - 10\log(r) - \alpha r - K, \tag{1}$$

where r is the distance between a ship and the hydrophone, SL is the ship source level recalculated to 1 m from the ship, and K is parameter characterizing transition between spherical spreading near source and cylindrical spreading at great distances. This equation contains three unknown parameters of interest: (SL, K and α). Two of them, (SL, K), characterize the source of sound (i.e., a ship) and the transformation of sound to the channel, respectively. We are most interested in the relative attenuation coefficient, α, as the main acoustic parameter of the channel characterization. For estimation of the attenuation there is no need to know the values of SL and K. The attenuation coefficient was calculated by comparison of the noise level NL for distances r and $r+\Delta r$:

$$\alpha = \frac{NL(r+\Delta r) - NL(r) - 10\log(1+\Delta r/r)}{\Delta r}. \tag{2}$$

The attenuation coefficients at each frequency were calculated for three ships and relative distances between analyzing points, Δr, of 16 m. Results of average attenuation coefficient calculations are presented as a function of frequency in Fig. 3. It is seen that the attenuation coefficient has weak frequency dependence in high frequency range.

The effective attenuation coefficient with the overall average value of 0.058 dB/m with the standard deviation of 0.013 dB/m is much bigger than the volume attenuation coefficient of 0.010 dB/m at 50 kHz for sea water.[10] This large attenuation with weak frequency dependence can be associated with transmission of sound from water to the bottom and bottom scattering of sound. The detailed investigation of sound attenuation in shallow sea in the frequency band below 10 kHz[14] demonstrated that the observed weak attenuation frequency dependence can be explained by sound penetration into the bottom. Similar models can be developed for the high frequency range that may explain the observed weak frequency dependence.

3. Discussion

This paper demonstrates a method of sound attenuation measurements in a wide frequency band based on the recording of noise produced by passing ships. This method is much simpler than the standard method based on measurements of sound propagation between a transmitter and a receiver. Disadvantages of this method are connected to the possible variations of the source level that may depend on ship orientation with respect to the hydrophone. As seen from

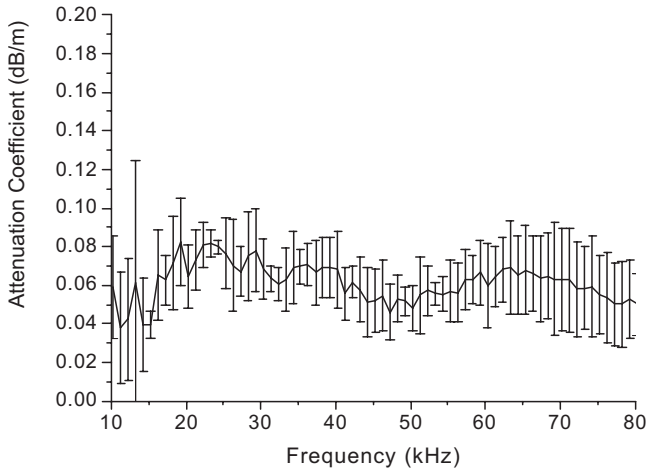


Fig. 3. Average effective attenuation coefficient and standard deviation as a function of frequency.

the presented results, the accuracy of the method is only about 22% based on this limited data. Further planned work includes additional tests aimed to investigations of ship noise radiations, including its variability among various ships and for the same ship at different times, and dependence of noise on ship orientation. The developed method is planned to be used for measurements of sound attenuation in various environmental conditions and for possible dependence on surface wave conditions. We also plan to conduct comparison of this method with standard acoustic techniques based on measurements of attenuation between a transmitter and a receiver.

Acknowledgments

The results presented in this paper were achieved through the efforts of an excellent team comprised of Stevens faculty, researchers, and students. The team included: Jeremy Turner, Peter Rogowski, Anirudh Nair, and Howie Goheen, of our maritime operations staff providing placements of hydrophone and measurements; George Kamberov and Bart Luczynski who provided the video tracking system; Nikolay Sedunov providing the hydrophone preamplifiers and filtering, Michael Tzionsky producing software for data record and preliminary signal processing.

This work was supported by ONR project No. N00014-05-1-0632: Navy Force Protection Technology Assessment Project.

References and links

[1]D. Hill and P. Nash, "Fibre-optic hydrophone array for acoustic surveillance in the littoral," Proc. SPIE **5780**, 1–10 (2005).
[2]S. Stanic, C. K. Kirkendall, A. B. Tveten, and T. Barock, "Passive swimmer detection," NRL review, http://www.nrl.navy.mil/content.php?P=04REVIEW97, 2004.
[3]R. Stolkin, A. Sutin, S. Radhakrishnan, M. Bruno, B. Fullerton, A. Ekimov, and M. Raftery, "Feature based passive acoustic detection of underwater threats," Proc. SPIE **6204**, 40–49 (2006).
[4]B. Bunin, A. Sutin, and M. Bruno, "Maritime security laboratory for maritime security research," Proc. SPIE **6540**, 65400S1–65400S8 (2007).
[5]W. Munk, P. Worcester, and C. Wunsch, *Ocean Acoustic Tomography* (Cambridge University Press, Cambridge, 1995).
[6]A. L. Matveev, D. A. Orlov, A. A. Rodionov, B. M. Salin, and V. I. Turchin, "Comparative analysis of tomographic methods for the observation of inhomogeneities in a shallow sea," Acoust. Phys. **51**(2), 218–229 (2005).
[7]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**(3), 973–986 (2000).
[8]I. F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks," ACM (Association for Computing Machinery) Sigbed Review, **1**(2), 3-8 (2004).
[9]D. B. Kilfoyle and A. B. Baggeroer, "The state of the art in underwater acoustic telemetry," IEEE J. Ocean. Eng. **25**(1), 4–27 (2000).

[10]R. J. Urick, *Principles of Underwater Sound* (McGraw–Hill, New York, 1975); C. S. Clay and H. Medwin, *Acoustical Oceanography: Principles and Applications* (Wiley, New York, 1977).
[11]F. Ingenito, "Measurement of mode attenuation coefficients in shallow water," J. Acoust. Soc. Am. **53**, 858–863 (1973).
[12]C. T. Tindle, "Attenuation parameters from normal mode measurements," J. Acoust. Soc. Am. **71**, 1145–1148 (1982).
[13]S. Lee, K. Park, J. Yoon, and P. Lee, "Measurement and analysis of broad band acoustic propagation in very shallow water," Jpn. J. Appl. Phys., Part 1 **46**(7B), 4971–4973 (2007).
[14]P. Wille, R. Thiele, and E. Schunk, "Shallow-water sound attenuation in a standard area," J. Acoust. Soc. Am. **54**, 1708–1726 (1973).

Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech

Elizabeth K. Johnson
Department of Psychology, University of Toronto, 3359 Mississauga Road, Mississauga, Ontario, Canada, L5L 1C6
elizabeth.johnson@utoronto.ca

Abstract: The Headturn Preference Paradigm was used to examine infants' use of prosodically conditioned acoustic-phonetic cues to find words in speech. Twelve-month-olds were familiarized to one passage containing an intended target (e.g., *toga* from *toga#lore*) and one passage containing an unintended target (e.g., *dogma* from *dog#maligns*). Infants were tested on the familiarized intended word (e.g., *toga*), familiarized unintended word (e.g., *dogma*), and two unfamiliar words. Infants listened longer to familiar intended words than to familiar unintended or unfamiliar words, demonstrating their use of word-level prosodically conditioned cues to segment words from speech. Implications for models of developmental speech perception are discussed.

PACS numbers: 43.71.Ft, 43.71.Sy, 43.71.Es [JH]
Date Received: December 20, 2007 **Date Accepted:** March 10, 2008

1. Introduction

One of the most fundamental questions in developmental speech perception is how infants learn to attend to the speech signal in an adult-like manner and perceive multiword utterances as strings of discrete recognizable words. This task is more complicated than common intuition suggests because fluent speech does not consist of clearly separated words that can be easily mapped onto lexical representations. On the contrary, reliable cues to word boundaries, such as silences between words, do not exist. And no single word is ever produced identically twice. Despite these difficulties, infants begin mastering the speech signal remarkably early.[1]

Within the first year of life, infants begin using many of the same bottom-up segmentation strategies as adults. For example, in English most content words begin with a stressed syllable, and adult listeners are appropriately biased to perceive stressed syllables as word onsets.[2] By 7.5 months, English learners have detected this pattern and are so focused on stress cues that they appear to perceive all stressed syllables as word onsets.[3] By the end of the first year of life infants' segmentation attempts become more accurate (i.e., more adult-like) as they expand their repertoire of segmentation strategies. For example, the use of phonotactic information (i.e., constraints on which consonant transitions are likely within versus across word boundaries) is thought to help infants overcome their earlier reliance on stress to find word boundaries.[1]

Another strategy that has been proposed to help infants refine their segmentation attempts is use of prosodically conditioned acoustic-phonetic information.[4,5] The prosodic structure of utterances can be characterized as hierarchical.[6] Roughly speaking, utterances contain intonational phrases, intonational phrases contain phonological phrases, phonological phrases contain prosodic words, and prosodic words contain syllables. Prosodic boundaries above the prosodic word level and above always coincide with word boundaries. Acoustic-phonetic cues mark the placement of speech units within the hierarchy. For example, speech units at the end of prosodic constituents tend to be lengthened,[7] and speech units along the onset of prosodic constituents tend to be more forcefully articulated than those situated in the middle.[8]

Adult listeners are sensitive to these acoustic-phonetic cues at all levels. They detect

syllable boundaries distinguishing potentially ambiguous phrases such as *known ocean* and *no notion*,[9] as well as phonological phrase boundaries that mark prosodic boundaries without breaking syllable boundaries.[10,11] Most importantly for the current study, adults also use acoustic-phonetic cues at the word level to infer speaker intent and segment words from speech. For example, the word *ham* is recognized more readily if it is produced as a monosyllabic word rather than as the first syllable of a longer word such as *hamster*.[12] Infants are similar to adults in that they have been shown to use utterance boundaries,[13] prosodic phrase boundaries,[4] and syllable boundaries[1] to find words in speech. However, infants have not yet been shown to use acoustic-phonetic cues at the prosodic word level to segment words from speech. Existing studies examining infants perception of syllable sequences straddling word boundaries (e.g., *taris* in *guitar is*[3]) were not designed to determine if infants are sensitive to acoustic-phonetic cues to word boundaries, and thus used stimuli containing multiple cues to word boundaries. Knowing whether infants use word-level acoustic-phonetic cues to segment words from speech would impact our understanding of developmental speech perception. For example, use of these cues could impose constraints on distributional theories of developmental word segmentation.[3,5,14]

In the current study we use the Headturn Preference Procedure to test 12-month-olds' use of word-level acoustic-phonetic cues to segment words from speech. The experiment familiarizes infants with two types of passages. One contained a reoccurring sequence consisting of a monosyllabic word followed by stress-final bisyllabic word (e.g., *toe#galore*); the other contained a reoccurring sequence consisting of a stress-initial bisyllabic word followed by a monosyllabic word (e.g., *dogma#lines*). Each of these target sequences is potentially ambiguous in the sense that it can be parsed as beginning with either a monosyllabic or bisyllabic word (e.g., *toe#galore* and *toga#lore*). If infants segment intended items more readily than unintended items, then this would suggest that infants use word-level acoustic-phonetic cues to locate word boundaries.

2. Method

Participants: Forty-eight American-English-learning 12-month-olds from the Baltimore-Annapolis region were tested (22 females). The infants had a mean age of 364 days (range: 351 days to 407 days). The data from nine additional infants were excluded for failing to complete the study due to fussiness (7), parental interference (1), and experimenter error (1).

Stimuli: Eight passages containing a reoccurring target syllable sequence were recorded in an infant-directed register by a female speaker naïve to the purpose of the study.

Sample passage containing stress-initial bisyllabic word: *This ruby quest will be more exciting than ever. Ruby quest weddings are frowned upon in Greenwich. The foosball pro was thrilled with that ruby quest. The station's ruby quest will end before ours. The guy who won the ruby quest wore silly goggles. I joined the singing ruby quest.*

Sample passage containing stress-final bisyllabic word: *This rue bequest will surely go down in history. Rue bequest weddings were common last May. The panda was really wild about that rue bequest. The station's rue bequest will make the cat happy. The guy who made the rue bequest wore shiny frog shoes. We sent a singing rue bequest.*

The speaker was asked to imagine that the nonsensical sequences (e.g. *gumbo#teak*) had meaning, and to produce them as an adjective preceding a noun in order to prevent the insertion of a phrase boundary between the syllables. The passages were on average 21.53 s long. The speaker also recorded four test lists (each containing 15 tokens of the same word).

In order to ensure the target syllable sequences in the passages were produced with different intended word boundaries, the duration of the onset and rime of each syllable of the familiarized sequences was measured. Past studies have shown that segments falling along prosodic boundaries tend to be lengthened relative to those not falling along a prosodic boundary. For example, the word *tune* is longer when it is realized as a monosyllabic word (*tune acquire*) than when it is embedded in a bisyllabic word (*tuna choir*).[15,16] In line with predictions from these studies, the first and third syllables or our target sequences were lengthened when produced as monosyllabic as opposed to bisyllabic words (Table 1).

Design: Infants were randomly assigned to familiarization with one of four pairs of

Table 1. Duration measurements of target sequences in the passages (onset/rime boundary in *rue* was difficult to locate). Duration (seconds).

	S1 Onset	S1 Rime	S2 Onset	S2 Rime	S3 Onset	S3 Rime
Ruby#quest	- ----	- ----	0.061 (0.011)	0.158 (0.025)	0.062 (0.011)	0.472 (0.095)
Dogma#lines	0.024 (0.004)	0.25 (0.032)	0.056 (0.014)	0.103 (0.027)	0.059 (0.01)	0.442 (0.252)
Gumbo#teak	0.020 (0.016)	0.208 (0.016)	0.083 (0.011)	0.175 (0.039)	0.083 (0.01)	0.239 (0.07)
Toga#lore	0.063 (0.029)	0.171 (0.024)	0.014 (0.003)	0.127 (0.022)	0.084 (0.02)	0.268 (0.112)
Rue#bequest	- ----	- ----	0.043 (0.009)	0.141 (0.014)	0.047 (0.008)	0.452 (0.145)
Dog#maligns	0.024 (0.01)	0.265 (0.029)	0.065 (0.016)	0.122 (0.039)	0.058 (0.014)	0.387 (0.162)
Gum#boutique	0.021 (0.003)	0.241 (0.025)	0.063 (0.009)	0.192 (0.039)	0.063 (0.01)	0.231 (0.096)
Toe#galore	0.109 (0.049)	0.185 (0.026)	0.028 (0.006)	0.137 (0.03)	0.078 (0.007)	0.282 (0.08)
Mean SW#S	0.035 (0.025)	0.210 (0.04)	0.054 (0.027)	0.141 (0.04)	0.072 (0.018)	0.355 (0.175)
Mean S#WS	0.051 (0.05)	0.230 (0.04)	0.050 (0.019)	0.148 (0.04)	0.061 (0.015)	0.338 (0.147)

passages: (1) *rue#bequest* and *dogma#lines*, (2) *toga#lore* and *gum#boutique*, (3) *ruby#quest* and *dog#maligns*, or (4) *toe#galore* and *gumbo#teak*. All infants were tested on the same four test items: *ruby*, *dogma*, *gumbo*, and *toga*. Two of these test items were familiar, in the sense that the syllable sequence occurred during the familiarization. However, one familiar syllable sequence spanned a word boundary while the other did not. The other two test items were unfamiliar.

Procedure and Apparatus: Infants were tested using a standard variant of the Headturn Preference Procedure.[1,3] The experimenter remained out of view of the infant, recording the direction and duration of the infants' orientation through the use of a button box. The randomization of stimuli and termination of trials was computer controlled. A red light and a speaker were mounted at eye level on the center of each side panel, and a green light was located at eye level on the center of the front panel. During the familiarization phase, the green light flashed at the start of each trial. Once the infant oriented toward the green center light, it stops flashing. One of the two side red lights then immediately began flashing. Once the infant oriented towards the flashing light, a sound file was presented from the speaker hidden behind the flashing light. The sound file continues to play until either the infant looks away for more than two consecutive seconds or the sound file ends. Once the infants accrued 45 s of orientation times towards each passage, the test phase began. Twelve test trials were presented during the test phase (three trials for each of the four test items). Test trials were presented in three blocks, and trial order was randomized within those blocks. Both the experimenter and the caregiver listened to masking music over headphones to prevent them from biasing the study.

3. Results

Mean orientation times to each of the three types of test items (familiar intended, familiar unintended, unfamiliar) were calculated for each of the 48 subjects (see Fig. 1). Thirty-three out of 48 subjects had longer average orientation times to familiar intended test items than unfamiliar test items. In contrast, only 22 out of 48 infants had longer orientation times to familiar unintended test items than unfamiliar test items. And most importantly, 31 out of 48 infants had longer average orientation times to familiar intended test items (*ruby* from *ruby#quest*) than the familiar unintended test items (*ruby* from *rue#bequest*). A mixed design analysis of variance, 3 (test item: intended familiar, unintended familiar, unfamiliar) $\times 4$ (familiarization condition), revealed a significant effect of test item, $F(2,44)=3.83$, $p<0.05$. There was also an effect of familiarization condition, $F(3,44)=31$, $p<0.05$, but importantly, there was no significant interaction between test item and familiarization condition, $F(6,44)=1.48$, $p>0.10$. Planned comparisons revealed a significant difference in orientation times to intended familiar and unfamiliar test items, $F(1,47)=6.12$, $p<0.05$. In addition, there was a significant difference in orientation times to intended and unintended familiar test items, $F(1,47)=4.36$, $p<0.05$. How-

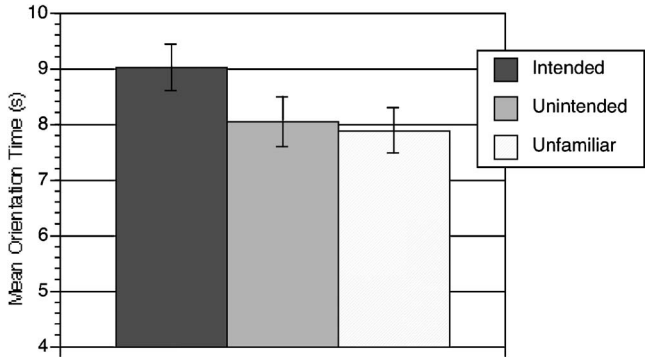


Fig. 1. Mean orientation times in seconds to intended (e.g., *ruby* from *ruby#quest*), unintended (e.g., *ruby* from *rue#bequest*), and unfamiliar test items by 12-month-olds.

ever, there was no significant difference in orientation times to familiar unintended and unfamiliar test items, $F(1,47)=0.11$, $p>0.10$. As Fig. 1 illustrates, these effects were attributable to longer orientation times to familiar intended test items ($M=9.02$ s, $SD=2.9$) than to either familiar unintended test items ($M=8.04$ s, $SD=3.1$) or to unfamiliar test items ($M=7.89$ s, $SD=2.7$).

4. Discussion

Our results demonstrate that 12-month-old infants are like adults in that they use prosodically conditioned acoustic-phonetic cues to segment words from running speech. Note that these cues constrained infants' segmentation behavior in the face of misleading syllable distribution cues to word boundaries, i.e., despite the fact that the conditional probabilities marking the transition between the two syllables of the intended and unintended words were equal. These findings suggest that sensitivity to the prosodic structure of utterances may be fundamental to early word recognition, perhaps even more so than sensitivity to syllable distribution cues. It may also be the case that simple exemplar models cannot explain the results reported in this study. Instead, infants may be like adults in that they compute the prosodic structure of utterances online and use this information to infer word boundaries.[17] These findings contribute to a growing body of evidence demonstrating that acoustic-phonetic variation plays an important role in word recognition.[10,12,17–20]

One limitation of the current study is that we have not identified which acoustic-phonetic cues infants use to detect boundaries at the prosodic word level. Recent studies have suggested that duration cues alone are sufficient for adults to perceive the boundary between prosodic words.[12,17] However, these studies did not rule out the possibility that adults are also sensitive to other prosodically conditioned acoustic-phonetic cues to word boundaries, such as degree of coarticulation or changes in consonant realization due to initial strengthening.[10,11] Thus, there remains the possibility that during the course of early language acquisition infants must learn which acoustic-phonetic cues mark boundaries in their language. It may be the case that infants learn the cues marking smaller junctures such as word boundaries by attending to large junctures such as phrase and utterance boundaries.

Another area for future research will be to identify when infants begin using word-level acoustic-phonetic cues to parse the speech stream. Doing so will be important for integrating our findings with current models of developmental word segmentation. Perhaps sensitivity to word-level acoustic-phonetic cues to word boundaries develops along with sensitivity to phonotactic and allophonic cues at around the end of the first year of life, and in combination these cues help infants overcome their over reliance on lexical stress cues to word boundaries, i.e., integration of word-level acoustic-phonetic cues with other word segmentation cues may help English learners begin extracting noninitially stressed words from speech. Another possibility

is that sensitivity to acoustic-phonetic cues begins developing even earlier, perhaps at the very earliest stages of word segmentation.[19] If this were the case, then this would have important implications for current models of developmental word segmentation. Distributional models posit that infants begin segmenting words from speech by tracking conditional probabilities between syllables, i.e., syllables pairs that are statistically likely to co-occur are likely to be words. Once infants have segmented enough words from speech by tracking conditional probabilities between syllables, then they can notice that most of the words begin with a stressed syllable.[14] Eventually, enough additional segmentation cues are learned for infants to reach an adult-like ability to extract words from speech. If infants were sensitive to prosodically conditioned acoustic-phonetic cues at the onset of word segmentation abilities, then this sensitivity would greatly constrain the use of distributional strategies to find word boundaries. Thus, it is important that future work investigate when and how infants begin detecting prosodically conditioned word boundaries in fluent speech.

Acknowledgments

I thank P. Jusczyk and the Jusczyk Lab Executive Committee for inspiring this work. I thank J. Miller, S. Shattuck-Hufnagel, A. M. Jusczyk, A. Cutler, B. Landua, and A. Seidl for invaluable help leading to the completion of this study.

References and links

[1]P. Jusczyk, *The Discovery of Spoken Language* (MIT Press, Cambridge, MA, 1997).
[2]A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception," J. Mem. Lang. **31**, 218–236 (1992).
[3]P. W. Jusczyk, D. Houston, and M. Newsome, "The beginnings of word segmentation in English-learning infants," Cogn. Psychol. **39**, 159–207 (1999).
[4]A. Gout, A. Christophe, and J. Morgan, "Phonological phrase boundaries constrain lexical access II. Infant data," J. Mem. Lang. **51**, 548–567 (2004).
[5]M. Shukla, M. Nespor, and J. Mehler, "An interaction between prosody and statistics in the segmentation of fluent speech," Cogn. Psychol. **54**, 1–32 (2007).
[6]M. Nespor and I. Vogel, *Prosodic Phonology* (Foris, Dordrecht, 1986).
[7]C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**, 1707–1717 (1992).
[8]C. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740 (1997).
[9]L. Nakatani and K. Dukes, "Locus of segmental cues for word juncture," J. Acoust. Soc. Am. **62**, 714–719 (1977).
[10]T. Cho, J. McQueen, and E. Cox, "Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English," J. Phonetics **35**, 210–243 (2007).
[11]A. Christophe, S. Peperkamp, C. Pallier, E. Block, and J. Mehler, "Phonological phrase boundaries constrain lexical access I: Adult data," J. Mem. Lang. **51**, 523–547 (2004).
[12]A. P. Salverda, D. Dahan, and J. McQueen, "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," Cognition **90**, 51–89 (2003).
[13]A. Seidl and E. K. Johnson, "Infant word segmentation revisited: Edge alignment facilitates target extraction," Dev. Sci. **9**, 566–574 (2006).
[14]E. Thiessen and J. Saffran, "When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants," Dev. Psychol. **39**, 706–716 (2003).
[15]A. Turk and S. Shattuck-Hufnagel, "Word-boundary-related durational patterns in English," J. Phonetics **28**, 397–440 (2000).
[16]T. Cho and E. K. Johnson, "Acoustic correlates of phrase-internal lexical boundaries in Dutch," In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, edited by S. H. Kin and M. J. Bae (Sunjin, Jeju, Korea, 2004), pp. 1297–1300.
[17]K. Shatzman and J. McQueen, "Prosodic knowledge affects recognition of newly acquired words," Psychol. Sci. **17**, 372–377 (2006).
[18]B. McMurray and R. N. Aslin, "Infants are sensitive to within-category variation in speech perception," Cognition **95**, B15–B26 (2005).
[19]E. K. Johnson, "Speaker intent influences infants' segmentation of potentially ambiguous utterances, " Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS, 2003), pp. 1995–1998, Barcelona, Causal Productions.
[20]P. Luce and C. McLennan "Spoken word recognition: The challenge of variation," In *Handbook of Speech Perception*, edited by D. B., Pisoni and R. E. Remez, Wiley-Blackwell, Oxford, pp. 591–609 (2005).

Human motion analyses using footstep ultrasound and Doppler ultrasound[a)]

Alexander Ekimov and James M. Sabatier
National Center for Physical Acoustics, The University of Mississippi, 1 Coliseum Drive, University, Mississippi 38677
aekimov@olemiss.edu, sabatier@olemiss.edu

Abstract: Human footsteps generate periodic broadband frequency envelopes of sound due to dynamic friction forces. Also, human body motion when walking is a cyclic temporal process. The individual body parts have different acoustic cross sections and velocities that form unique human Doppler signatures. The paper introduces an approach to analyze this motion using passive and active ultrasound. The passive method employs a narrowband microphone that is sensitive to the sound from footsteps. The active method utilizes continuous-wave ultrasound to measure the Doppler shifted signal from the body appendages. These two methods show time synchronization between Doppler and ultrasonic human footstep signatures.

PACS numbers: 43.35.Yb, 43.20.Ye [NX]
Date Received: February 6, 2008 **Date Accepted:** March 17, 2008

1. Introduction

Walking people create unique footstep acoustic[1–3] and Doppler signatures[4–7] that can be applied for human motion analyses. Footstep acoustic signatures are characterized by broadband frequency responses from a few Hertz up to ultrasonic frequencies. The dynamic forces between the foot and the supported surface generate the vibrations and sound. At ultrasonic frequencies, the sound in air propagates over greater distances than vibrations in the ground,[3] because the sound absorption in air is significantly less than vibration absorption in the ground.[8,9] The sound measurements at ultrasonic frequencies allow the possibility of passive footstep signal detection and analysis in the presence of high background noise levels expected at low frequencies in urban areas.[3,9] Typically, noise levels are highest at low frequencies and roll-off with increasing frequency.[10]

For the active method, the reflected ultrasonic signal from the human body is frequency modulated by the motion of the human body appendage (leg, arm, torso, etc.). The acoustic cross section of the appendage influences the amplitude of the received Doppler signal. These human Doppler signatures have been extensively studied using pulsed and continuous-wave (cw) radar[4–7,11] and ultrasonic sonar systems.[12–14] These Doppler signatures can be used to recognize and discriminate human motion among other moving and stationary objects for many practical applications including human intrusion security systems, human locomotion health monitoring, and computer animation of human motion.

This letter introduces human motion experimental results using both passive and active Doppler ultrasonic methods. Experiments were conducted using a tentative measurement system that was assembled from two sets of narrowband ultrasonic transducers. These transducers had resonant frequencies of 25.5 and 40 kHz for the passive and active methods, respectively, that reduced electrical and acoustical interference between them. The results from human motion measurements and analyses of persons walking on a 20 m straight-line track are presented and discussed.

[a)] A portion of this work has been presented in Alexander Ekimov and James M. Sabatier, "Ultrasonic signatures of human motion," J. Acoust. Soc. Am. **121**(5), 3115 (2007).

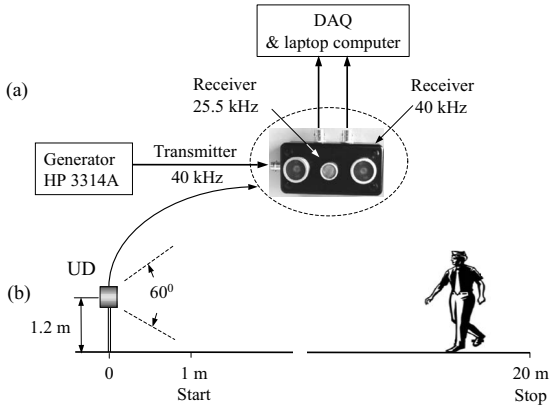


Fig. 1. Measurement of human motion: (a) is the block diagram of the measuring system and (b) is the setup for the measurements of human ultrasonic signatures in a building hallway.

2. Two-frequency band ultrasonic detector

To measure the human motion acoustic signals, a two-frequency band ultrasonic detector (UD) that uses commercially available, low-cost ultrasonic ceramic transducers was developed. A block diagram of the UD is presented in Fig. 1(a), where the UD combines three narrowband ultrasonic transducers that were assembled into a small plastic enclosure.

An ultrasonic receiver (UR) (250SR) was used for passive measurements of footstep sound. This UR had a resonance frequency of 25.5 kHz and a typical bandwidth (at −6 dB) of 1 kHz. The directivity (at −6 dB) was 60°. A receiving low-noise preamplifier (Stanford Research Systems SR 560 [not shown in Fig. 1(a)]) amplified signals from the UR with a gain of 60 dB. The 25.5 kHz transducer provided sound detection in air at greater distances than would a higher-frequency transducer. For reference, the maximum energy absorption,[10] $\alpha(f)$, in air [from Eq. (1)] at 25.5 kHz is 0.825 and 1.32 dB/m at 40 kHz.

$$\alpha(f) = 0.000033 \times f\,[\mathrm{dB/m}], \tag{1}$$

where f is the frequency of sound in Hz.

An ultrasonic Doppler sonar (UDS) was assembled from two (transmitter and receiver) ultrasonic transducers (MATSU/PAN EFR-RCB40K 54). These transducers had a resonance frequency of 40 kHz, a typical bandwidth (at −6 dB) of 2 kHz, and a directivity (at −6 dB) of 55°. One of the sensors, acting as a transmitter, emitted an ultrasonic wave while the other acted as a receiver. As can be seen from Eq. (2), the 40 kHz Doppler sonar provided 1.56 times higher sensitivity to a human body motion V than would be expected from a 25.5 kHz Doppler sonar, because the Doppler shift Δf in the reflected signal is proportional to the transmitted frequency f.[15]

$$\Delta f = \frac{2V}{C} f, \tag{2}$$

where C is the sound speed.

A HP 3314A signal generator applied a cw electrical signal at 40 kHz to the transmitter. Signals from the receiver were amplified with a Stanford Research Systems SR 560 [not shown in Fig. 1(a)] with the gain of 30 dB.

Data recording and processing were conducted using a two-channel, 24 bit data acquisition board (DAQ) (Echo Indigo IO) and a laptop computer with Sound Technology software (LAB432).

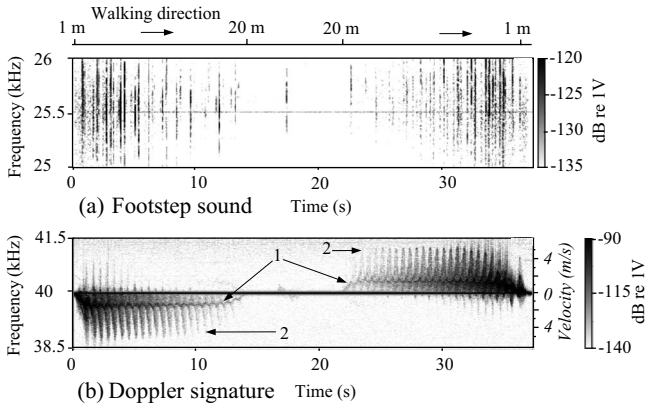


Fig. 2. The typical spectrograms of footstep passive ultrasonic (a) and active Doppler signatures (b), where 1 is a torso motion and 2 is leg and arm motion. A person started walking at 1 m from the UD location and stopped walking at 20 m from the UD. The person stood still for 5–8 s at 20 m, and then turned around, stood still for 5–8 additional seconds and then walked back to the start position. The sampling rate was 96 kHz and the FFT size was 16 384.

3. Experimental setup

The measurements of the sound from walking people using the UD were conducted on a straight-line track marked on the floor in a hallway of a modern university building. The length of the track was 20 m. In these experiments a person walked on the track, which was 0.3 m wide. The UD was placed on a tripod at a 1.2 m height and located at one end of the track. In the test configuration, the beam patterns of the ultrasonic transducers were oriented along the walking track. Each person started walking at 1 m from the UD location and stopped walking at 20 m from the UD as shown in Fig. 1(b). The person stood still for 5–8 s at 20 m, and then turned around, stood still for 5–8 additional seconds and then walked back to the starting position.

A DAQ with the sampling rate of 96 and 48 kHz antialiasing filter simultaneously acquired signals from the UDS and UR. These signals were stored on the laptop computer for each individual walker and a database for approximately 100 walkers was assembled.

4. Experimental results

The measured data were processed using the short-time Fourier Transform[16] (STFT) with a 0.17 s sliding Hanning window. This time window is approximately equal to the time duration of a single footstep, that is defined as the time interval for a single footstep from "heel strike" to "toe slap and weight transfer."[1] All the data were corrected for the amplifier gains and expressed in dB re:1V. The results are presented as spectrograms, which is the magnitude of the STFT. Three typical ultrasonic signatures of a walking human are shown below.

The most common spectrograms of footstep passive and active Doppler signals are presented in Figs. 2(a) and 2(b) in the frequency bands 25–26 kHz for the footstep passive sound signature and 38.5–41.5 kHz for the Doppler signature. The fast Fourier transform (FFT) size was 16 384, which corresponded to 5.9 Hz in the spectral line resolution and 170 ms in the time resolution. The person walked at constant speed of motion and more than 20 footsteps were required to complete the distance of 19 m. The horizontal line at 40 kHz in Fig. 2(b) is the sum of direct coupling between the transmitter and the receiver through the air and the common enclosure and the reflected signals from stationary objects. The strongest reflection from a walking person corresponded to the fluctuating line near the frequency of 39.7 kHz [marked as No. 1 in Fig. 2(b)] for walking away from the UD and 40.3 kHz for walking towards the UD. The Doppler shift Δf is proportional to the person's speed V. For the frequency shift of $|\Delta f| = 300$ Hz the speed of the walker V follows from Eq. (2):

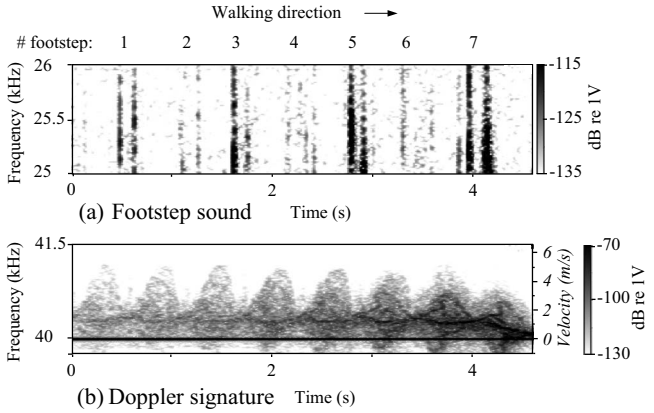


Fig. 3. The spectrograms of footstep ultrasonic (a) and the Doppler signatures (b) of a person walking toward to UD. Cotton clothing. The sampling rate was 96 kHz and the FFT size was 8192.

$$V = \frac{\Delta f}{2 \times f} \times C = \frac{300}{2 \times 40000} \times 343 = 1.29 \text{ [m/s]}, \tag{3}$$

where $C = 343$ m/s is the sound speed in air, $f = 40$ kHz, $\Delta f = 300$ Hz.

Mm. 1. Video file (3.8 Mb). This is a file of type "mpg." This video file and human ultrasonic signatures (presented in Fig. 2) were recorded at the same time. A camera was co-located with the UD and placed on a tripod at a 1m height.

The direct estimation of V from data presented in Fig. 2(b) gives the average value of

$$V = \frac{D}{t} = \frac{19}{15} = 1.27 \text{ [m/s]}, \tag{4}$$

where $D = 19$ m is the track length, $t = 15$ s is the time needed to traverse the track by the walker.

The speed V calculated from the Doppler shift [Eq. (3)] and from the experimental geometry [Eq. (4)] is approximately the same, so the frequency shift of $\Delta f = \pm 300$ Hz corresponds to the body part having the maximum cross section, that is the torso.[4] The envelopes of curves marked as No. 2 in Fig. 2(b) correspond to the motions of the legs and arms, which have smaller cross sections than the torso and therefore have lower amplitudes. Leg and arm motions had larger Doppler shift values when compared with the torso motion, which corresponded to the larger speeds of these body parts. For example, while the maximum torso velocity is 1.8 m/s, the arms and legs have maximum velocities of 5.3 m/s.

A comparison of the relative phases of curves No. 1 and the envelopes No. 2 in Fig. 2(b) shows that the minimum Doppler shift value of the torso motion corresponded to the maximum Doppler shift value of leg and arm motions. The torso motion is in the opposite phase to one set of leg and arm motions and in phase to another set of leg and arm motions. The minimum value of the Doppler shift in the envelopes No. 2 ($\Delta f = 0$ Hz) corresponds to the zero speed of a leg and opposing arm when the foot of that leg is in contact with the floor.

More detailed analyses of friction and Doppler signatures are presented in two characteristic spectrograms shown in Figs. 3(a) and 3(b) and Figs. 4(a) and 4(b). These spectrograms are zoomed portions of the data from the fully walked path. They correspond to seven footsteps of the same walker in different dress (cotton and polyester cloth, respectively) to exclude specific personal differences from the investigation. These two data sets were separated in time by a half year. The FFT size was 8192, which corresponded to 11.72 Hz in the spectral line resolution and 85.3 ms in the time resolution.

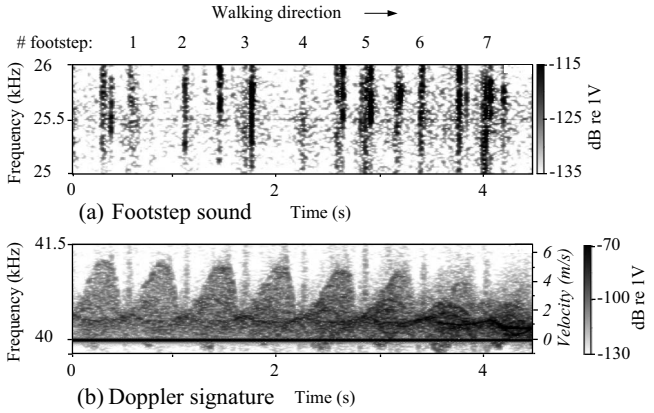


Fig. 4. The spectrograms of footstep ultrasonic (a) and the Doppler signatures (b) of a person walking toward to UD. Polyester clothing. The sampling rate was 96 kHz and the FFT size was 8192.

Mm. 2. Audio file (2.5 Mb). This is a file of type "wav." This is a raw data file of human ultrasonic signatures presented in Fig. 4.

The spectrogram in Fig. 3(a) shows seven time-frequency responses separated in time corresponding to seven individual footsteps. The time difference between pairs of vertical lines in Fig. 3(a) is about 150 ms, which corresponds to the two phases in footstep motion from "heel strike" to "toe slap and weight transfer" as described in Ref. 1. The first phase (heel strike) included the deceleration stage of the leading foot. The second phase (toe slap and weight transfer) included the toe slap resulting from the deceleration stage of the leading foot and the weight transfer resulting from the acceleration stage of the trailing foot.[1] The spectrogram in Fig. 3(a) shows qualitatively different magnitudes in footstep signatures for the left and right legs of a walker (odd and even footsteps). The high-frequency responses in Fig. 3(a) correspond in time to the minimum values of leg and arm velocity in Fig. 3(b) and were created by the friction forces between a foot and the floor.

The spectrogram in Fig. 4(a) shows additional friction responses in the footstep signature, which coincide with the maximum velocity of leg and arm motion in Fig. 4(b). These signals result from the rubbing or friction motion between the clothed arms and torso and the legs. These signals are strongly influenced by the clothing type. In Figs. 3(a) and 3(b), the walker was wearing cotton clothing and in the Figs. 4(a) and 4(b) the same walker's dress was synthetic polyester.

Figures 3(a) and 4(a) show the cyclical friction signals and Figs. 3(b) and 4(b) show Doppler shifts. Both sets of figures, Figs. 3(a) and 3(b) and also Figs. 4(a) and 4(b) show periodic friction and velocity signal correlation. The passive and active ultrasonic signatures are synchronized in time and may be used for human characterization and recognition among other moving and stationary objects.

5. Conclusions

The passive ultrasonic and active Doppler ultrasound signatures of walking persons were measured and analyzed for a building hallway at a distance range 1–20 m. The simultaneous measurement of active and passive ultrasonic signals was tested and analyzed for human motion in a building. The two methods show temporally synchronized Doppler and ultrasonic footstep signals.

Ultrasonic friction signals created by the sliding contacts between the foot and the floor corresponded in time to the minimum velocity in leg and arm motion. Ultrasonic signals due to clothing friction were detected and correlated in time to the maximum velocity in leg and arm motion. The magnitudes of these signals had a dependence on a clothing type.

Acknowledgment

This work was supported by the Department of the Army, Army Research Office, under Contract No. W911NF-04-1-0190. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the sponsor.

References and links

[1]A. Ekimov and J. M. Sabatier, "Vibration and sound signatures of human footsteps in buildings," J. Acoust. Soc. Am. **120**, 762–768 (2006).
[2]A. Ekimov and J. M. Sabatier, "Broad frequency acoustic response of ground/floor to human footsteps," Proc. SPIE **6241**, OL1–OL8 (2006).
[3]A. Ekimov and J. M. Sabatier, "Ultrasonic wave generation due to human footsteps on the ground," J. Acoust. Soc. Am. **121**, EL114–EL119 (2007).
[4]J. L. Geisheimer, E. F. Greneker, and W. S. Marshall, "A high-resolution Doppler model of human gait," Proc. SPIE **4744**, 8–18 (2002).
[5]P. van Dorp and F. C. A. Groen, "Human walking estimation with radar," IEE Proc., Radar Sonar Navig. **150**, 356–365 (2003).
[6]M. Otero, "Application of a continuous wave radar for human gait recognition," Proc. SPIE **5809**, 538–548 (2005).
[7]P. Cory, H. R. Everett, and T. H. Pastore, "Radar-based intruder detection for a robotic security system," Proc. SPIE **3525**, 62–72 (1998).
[8]L. B. Evans and H. E. Bass, "Tables of absorption and velocity of sound in still air at 68° F," Report No. WR72-2, Wyle Laboratories, Huntsville, AL (1972).
[9]A. Ekimov and J. M. Sabatier, "Passive ultrasonic methods for human footsteps detection," Proc. SPIE **6562**, 656203-1–656203-10 (2007).
[10]H. E. Bass and L. N Bolen, "Ultrasonic background noise in industrial environments," J. Acoust. Soc. Am. **78**, 2013–2016 (1985).
[11]V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: phenomenon, model, and simulation study," IEEE Trans. Antennas Propag. **42**, 2–21 (2006).
[12]Z. Zhang, P. O. Pouliquen, A. Waxman, and A. G. Andreou, "Acoustic micro-Doppler radar for human gait imaging," J. Acoust. Soc. Am. **121**, EL 110–113 (2007).
[13]A. Ekimov and J. M. Sabatier, "Ultrasonic signatures of human motion," J. Acoust. Soc. Am. **121**, 3115 (2007).
[14]A. E. Ekimov and J. M. Sabatier, "Passive and active ultrasonic methods for human motion detection," *Proc. MSS-BAMS*, 1–8 (2006).
[15]R. J. Urick, *Principles of Underwater Sound* (Peninsula, Los Altos, CA, 1983).
[16]*Applications of Digital Signal Processing to Audio and Acoustics*, edited by M. Kahrs and K. Brandenburg (Kluwer, New York, 2002).

Bayesian geoacoustic inversion in a dynamic shallow water environment

Yong-Min Jiang and N. Ross Chapman
School of Earth and Ocean Sciences, University of Victoria, PO Box 3055 Victoria, British Columbia, V8W 3P6, Canada
minj@uvic.ca, chapman@uvic.ca

Abstract: This paper presents results for matched field Bayesian geoacoustic inversion of multitonal continuous wave data collected on the New Jersey continental shelf. To account for effects of significant spatial and temporal variation of the water column sound speed, the sound speed profile was represented by empirical orthogonal functions. Data error information for the inversion was estimated from multiple time windows of the data. Inversion results for the sediment sound speeds at three ranges are in excellent agreement with the ground truth.

PACS numbers: 43.30.Pc, 43.60.Pt [JL]
Date Received: September 13, 2007 **Date Accepted:** February 25, 2008

1. Introduction

Bayesian geoacoustic inversion by matched field (MF) processing combines measured data and prior information of geoacoustic model parameters to infer estimates of the model parameters and their uncertainties. The inversion requires information about the statistics of the data errors, which is provided in terms of a data error covariance matrix. For MF geoacoustic inversion, data error is defined in terms of the difference between the complex pressures of the measured and modeled acoustic fields on an array of hydrophones. The error consists of both measurement errors, and theory errors such as incorrect geoacoustic parameterization and inaccurate propagation modeling. Theory errors are usually much more significant in MF inversions.[1–3]

A source of error that has generally been neglected in most geoacoustic inversions is error from unknown variation in the water sound speed profile (SSP).[4,5] This issue was one of the subjects of investigation in the recent shallow water experiment (SW06) sponsored by the Office of Naval Research. SW06 was a multidisciplinary, multi-institution, and multinational shallow water experiment that was carried out near the shelf break on the New Jersey continental shelf from mid-July to mid-September, 2006.[6] The geoacoustic component of SW06 included experiments designed by several research groups to acquire data over a broad frequency band from 50 to 20 000 Hz for geoacoustic characterization of the ocean bottom. An important objective of the geoacoustic experiments was to establish an experimental benchmark for comparing the performance of state of the art geoacoustic inversion methods. This paper and the companion paper by Huang *et al.*[7] present the first results of inversions of data from the experimental benchmark site.

The SSP in the water is critical in MF inversions for generating an accurate calculation of the replica field. Errors in the SSP will cause mismatch with the measured field that can lead to erroneous estimates of the geoacoustic model parameters in the inversion. Normally, it is assumed that the water column SSP is known from measurements in the experiment or otherwise, and is spatially invariant over the propagation range. However, it was evident from preliminary inversion tests with the SW06 data that the conventional practice of using individual SSPs measured at specific sites and times was not effective. A different approach was required to obtain reliable solutions for the inverse problem.

This paper presents an effective approach to account for the effects of theory errors due to unknown variations in the water sound speed in MF geoacoustic inversions. Along with the geoacoustic and geometric parameters, the SSP is also considered as unknown, and the MF

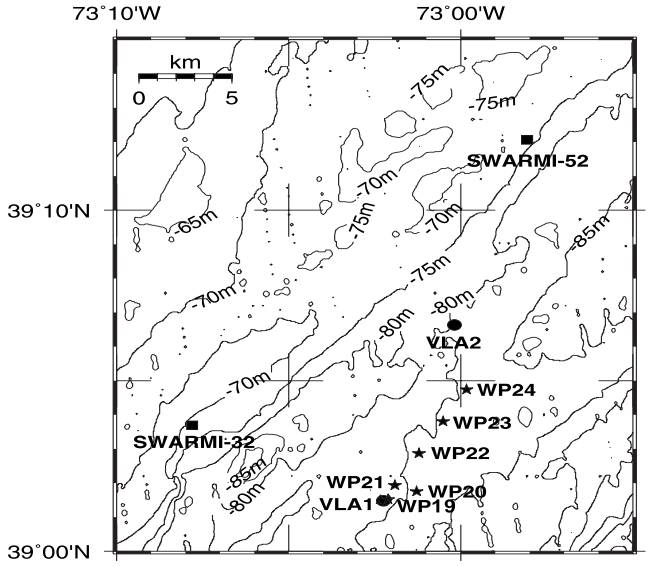


Fig. 1. Experimental site with bathymetry source and receiver positions.

inversion estimates the parameters of an approximate profile that generates the best fit of the replica field at the array. The data error covariance matrix is estimated using a set of independent data samples observed at different times in the experiment. The dynamically (spatially and temporally) varying ocean waveguide provides the means to include information about the theory errors, by evaluating the data errors under different propagation conditions.

2. Geoacoustic inversion experiments in SW06

In this paper, our focus is on the low frequency data, from 50 to 1000 Hz, that were collected at Scripps Institution of Oceanography's (SIO) vertical line array, shown as VLA1 in Fig. 1. The location near the shelf break was well chosen to magnify the effect of SSP variations. The data were acquired during source transmissions from stations WP21 (1 km), WP22 (3 km), and WP23 (5 km) along a well surveyed experimental track. The water depth was weakly range dependent, varying from 79.0 m at VLA1, to 78.8 m at WP21, 79.6 m at WP22, and 80.6 m at WP23, according to 12 kHz echo sounder data on the source ship. Sixteen hydrophones were equally spaced at 3.75 m on VLA1, with the lowest at 8.2 m above the sea floor. However, mechanical noise from sea surface motion contaminated the data on the top four phones, so that only the bottommost 12 were used in the inversions.

The acoustic data were recorded on Julian Day 239 between 2150 and 2345 Greenwich Mean Time. Two groups of multitonal continuous wave (cw) signals were transmitted alternately for 5 min at each station, from a source deployed at 30 m. The low frequency band contained cw signals at 53, 103, 203, and 253 Hz, and the mid frequency band contained signals at 303, 403, 503, 703, and 953 Hz. The time series for each frequency band were segmented into 2.62 s windows, and the separation of any consecutive segmented windows was larger than the estimated correlation time of background noise. The quality of the data was examined at each frequency by checking the signal to noise ratio and the Bartlett mismatch of the spectral components with respect to a reference time segment. On the basis of this analysis, seven of the transmitted tones were selected for the inversions: 53, 103, 203, 253, 303, 403, and 503 Hz.

The water column SSP was highly variable in space and time near the shelf break site due to internal waves, fronts, eddies, and diurnal tides. SSPs derived from conductivity-temperature-depth (CTD) measurements at each station, and at two other locations, are shown

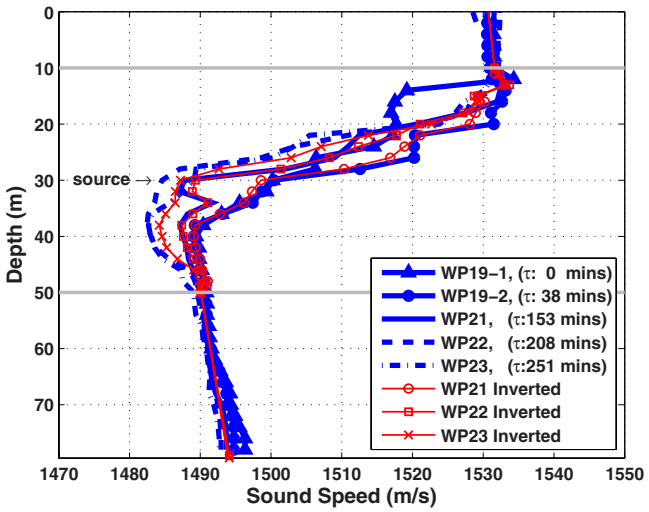


Fig. 2. (Color online) Sound speed profiles at each station.

in Fig. 2. The profiles have almost identical values from 0 to 10 m, and from 50 m to the sea bottom, but vary in the middle part from 10 to 50 m. The maximum difference of the sound speeds at the source depth of 30 m (right pointing arrow in the figure) is about 16 m/s. The measurement delay times of SSPs with respect to the SSP at WP19-1 are shown in the legend of Fig. 2.

3. Bayesian MF geoacoustic inversion of the multitonal data

The inversion estimated the values for three different sets of parameters: geometric parameters of the experimental arrangement (water depth, source-receiver range, source depth, and array orientation); parameters of the water column SSP; and the geoacoustic model parameters. The environment was assumed to be range independent, and the replica fields were calculated using the normal mode propagation model ORCA.[8] Although the experimental geometry was known accurately from measurements in the experiment, the geometrical parameters were included to provide a qualitative check on the inversion performance. Accurate estimation of the geometric parameters provided confidence that the estimated water column SSP was generating an adequate prediction of the replica field at the array.

3.1 Data error covariance matrix estimation

The formal development of Bayesian MF inversion and its implementation with Markov Chain Monte Carlo sampling algorithm is given in Refs. 1–3 and will not be repeated here. Instead we outline briefly the method used to estimate the data error covariance matrix that contains the information about theory errors. The data error covariance matrix, C_f, was estimated from multiple 2.62 s time segments from the data observed at different times in the experiments at each range. The number of data segments was 72, 68, and 66 for the inversions at 1, 3, and 5 km, respectively. The theoretical basis of this approach is similar to that presented in Ref. 3. The inversions were carried out using a 2.62 s data sample for each station, and the appropriate covariance matrix estimated from the multidata windows. The consistency of the optimal geoacoustic inversion[9] results from all the data segments mentioned above suggests that any 2.62 s data sample could be used in the Bayesian geoacoustic inversion.

Although the data were collected in a nominally fixed range geometry, the slowly varying conditions in the water column and the source motion due to sea state forcing provided different measures of the data/model misfit information. The approximate estimate of C_f was obtained from the average of this information over the sets of data segments. As noted previ-

ously, statistical tests were applied to the data residuals to test the assumptions of Gaussian distribution, spatial correlation, stationarity, and frequency independence,[1,3] and the results showed no evidence against the assumptions.

3.2 Water column sound speed profile representation

In the early stage of this study, the inversion tests indicated significant variation in the parameter estimates for inversions that used only the specific SSP measured at each source station. The estimated values of the geometric parameters shifted away from the "true" measured values, and the sediment sound speed values estimated at the three waypoints were not consistent. The inaccurate values of the geometric parameters indicated poor performance of the inversion, and so, reduced the confidence in the estimates of the geoacoustic parameters.

To account for the effect of the SSP in the inversion, the strongly varying part of the water column SSP in the thermocline (10–50 m) was represented in terms of a linear combination of a set of empirical orthogonal functions (EOFs). The samples shown in Fig. 2 that were measured around the time of the experiment were used to construct the EOFs. Only the coefficients of the first four EOFs were included in the inversion since they accounted for over 99% of the fit for the profile shape. Since the environment was assumed range independent, the estimated SSP was interpreted as a range averaged profile over the propagation path. All of the SSPs measured during the transit along the track line were used in the analysis, for the purpose of isolating the conditions during the experiment.

3.3 Geoacoustic parameterization

Extensive geological and geophysical surveys such as shallow cores, deep drills, *in situ* sediment probes, grab samples, and high resolution chirp sonar sub-bottom surveys were carried out in the vicinity of the experimental site prior to SW06. The surveys showed the presence of an interface known as the "R" reflector pervasively in the region at a depth around 20 m below the sea floor, with sand or mud layers on top of it. The sound speed decreases in these layers.[3] Based on the previous survey results, a single sediment layer over a half space geoacoustic model was used to approximate the ocean bottom in this work.

The geoacoustic parameters being inverted were sediment depth, p-wave sound speed and attenuation of the sediment and the basement, and density of the sediment. The attenuation and the density of the basement were held fixed at 0.3 dB/m at 1 kHz and 2.15 g/cm^3, respectively, according to the canonical geoacoustic model[3] for this region. Sediment attenuation and density were considered homogeneous (no depth gradient) for the model at all ranges. The frequency dependence of the attenuation was considered explicitly in the study by inverting the constant factor α and the exponent β according to $\alpha_p(f) = \alpha(f/f_0)^{\beta}$, where the unit of frequency f is kHz, f_0 is 1 kHz, and α_p is in dB/m. The attenuation was then converted into a modal attenuation coefficient and introduced into the wave equation solution as the imaginary part of the wave-number.

The sediment sound speed was treated differently in the geoacoustic model at different ranges. At the closest range of 1 km the sound speeds at both the top and the bottom of the sediment were inverted, but at the greater ranges of 3 and 5 km sediment sound speed was modeled as iso-speed in the layer.

3.4 Inversion results

The geometric parameters were sensitive at all ranges, and the estimated values were in very close agreement with the known values from the experiment. Water depth, range, source depth and the distance of bottommost sensor to the sea floor were the most sensitive of these parameters; array tilt was less sensitive and the value indicated a nearly vertical array ($\sim 2°$).

The coefficients of the EOFs that approximated the water column sound speed profile were very sensitive as well, and there was significant variation in the effective profiles at the different ranges, as shown in Fig. 2. The inverted SSP for 1 km was in good agreement with the SSP measured at WP19-2 (near VLA1); the inverted SSP for 3 km was in good agreement with

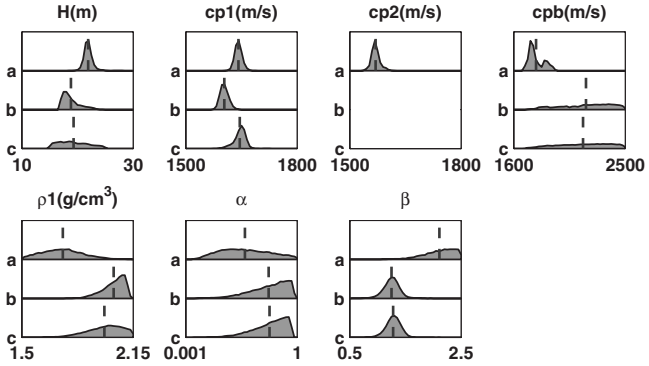


Fig. 3. Marginal probability distributions of geoacoustic inversion results. Dashed lines are means of the estimates. Rows a, b, and c are the inversion results for 1, 3, and 5 km data, respectively.

the SSP measured at the previous station, WP21 (i.e., between the source and the receiver); and the SSP estimate for 5 km was close to the average of the SSPs measured at WP22 and WP23 (3 km station and the source).

The marginal probability distributions of the geoacoustic parameter estimates are shown in Fig. 3, and the relative *maximum a posteriori* (MAP) estimates and 95% highest probability density (HPD) are given in Table 1. In Fig. 3, the vertical dashed lines indicate the mean values of the estimates, and the limits in each panel are the lower and upper search bounds for the parameter. At the range of 1 km, the inversion is able to resolve the parameters of the sediment layer. The sediment thickness H is 21 m, with sound speed c_{p1} of around 1637 m/s at the top of the layer, decreasing to a low speed c_{p2} around 1573 m/s at the sediment/basement interface (the mean value of sediment sound speed is 1604 m/s), and a sound speed jump to 1740 m/s in the basement. These estimates are consistent with the values from previous surveys for the depth to the *R* reflector, and with *in situ* measurements of the sea floor sound speed (1620–1660 m/s).[10] In particular, the decrease in sound speed below the seafloor indicates the presence of different, slow-speed sediment material, consistent with results from 3–5 m vibracores.[11] The estimated values for the sediment and basement sound speed and the sediment thickness are also consistent with the estimates from SIO's approach,[7] which are 1599.7 m/s for the mean sediment sound speed (sound speed at the top of the sediment is 1589.5 m/s, at the bottom is 1609.9 m/s), 1739.1 m/s for the basement sound speed and 24.2 m for the sediment thickness. The sediment sound speed is also very well recovered at the longer ranges of 3 and 5 km, and the values are consistent with the average value in the sediment from the 1 km inversion.

Table 1. Summary of MAP estimates and 95% HPD interval of geoacoustic model estimates in the form of [left bound MAP right bound].

Parameters	Range 1 km	3 km	5 km
H (m)	[19.7 21.1 23.0]	[21.9 22.6 25.1]	[10.7 16.1 17.6]
c_{p1} (m/s)	[1621.6 1636.8 1654.7]	[1564.9 1573.8 1589.7]	[1570.9 1599.1 1608.6]
c_{p2} (m/s)	[1557.4 1572.8 1591.3]	…	…
c_{pb} (m/s)	[1696.2 1740.5 1765.3]	[1748.2 1767.7 1902.6]	[1746.6 1945.2 1971.6]
ρ_1 (g/cm^3)	[1.56 1.68 1.80]	[1.92 2.10 2.15]	[1.88 2.12 2.15]
α	[0.141 0.456 0.953]	[0.278 0.651 0.999]	[0.266 0.618 0.998]
β	[1.747 2.342 2.498]	[1.544 2.062 2.334]	[1.041 1.391 1.729]

Sediment density ρ_1 and attenuation α_{p1} are generally not as well recovered as the sound speeds. The estimated values of the exponent of attenuation reveal a nonlinear power law of frequency dependence. The trend of increasing sensitivity of the attenuation with range makes physical sense since attenuation effects will accumulate over the range. At 5 km, the estimated value of β is 1.39 with the constant factor α (the attenuation in dB/m at 1 kHz) equal to 0.6. The decrease in the estimated value of β with increasing range can be attributed to the shallower penetration depth of signal at larger ranges, assuming that the attenuation is greater near the sea floor.

Although the estimates of the frequency dependence of the attenuation differ from other reported results,[12] it is important to understand that the attenuation estimated from MF inversions is an effective loss parameter. In addition to the contribution of the intrinsic attenuation of the sediment material, the estimate also includes contributions from other energy loss mechanisms over the long range propagation path, such as scattering due to seabed roughness and inhomogeneities in the water column.

4. Summary

This paper applies MF Bayesian geoacoustic inversion to the data collected from a site near the shelf break on the New Jersey continental shelf, where the water column SSP was highly variable in space and time. Data/model misfit errors were taken into account by estimating a data error covariance matrix from multiple data samples observed at different times. The unknown variations in the water column SSP were included by using a set of EOFs that were generated from CTD measurements during the time of the experiment to parameterize the profile. The geoacoustic inversion results presented here and the companion paper by Huang *et al.*[7] form the basis of an experimental benchmark to assess the performance of present day inversion techniques. Although the performance of various techniques has been evaluated with synthetic data sets, the comparative performance against experimental data from a well surveyed site is not known. Full evaluation involves analysis of reasons for differences of inversion results to determine limitations of specific approaches.

Acknowledgments

This work is supported by the Office of Naval Research Ocean Acoustic Team. The authors would like to thank Dr. William S. Hodgkiss and Dr. Peter Gerstoft from Scripps Institution of Oceanography for providing the data. The support of the crews of the R.V. KNORR and the other participants in the scientific team are greatly appreciated.

References and links

[1]S. E. Dosso, P. Nielsen, and M. J. Wilmut, "Data error covariance in matched-field geoacoustic inversion," J. Acoust. Soc. Am. **119**, 208–219 (2006).
[2]C.-F. Huang, P. Gerstoft, and W. S. Hodgkiss, "On the effect of error correlation on matched-field geoacoustic inversion," J. Acoust. Soc. Am. **121**, EL64–69 (2007).
[3]Y.-M. Jiang, N. R. Chapman, and M. Badiey, "Quantifying the uncertainty of geoacoustic model parameters for the New Jersey shelf by inverting air gun data," J. Acoust. Soc. Am. **121**, 1879–1895 (2007).
[4]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**, 769–781 (2001).
[5]Y.-T. Lin, C.-F. Chen, and J. F. Lynch, "An equivalent transform method for evaluating the effect of water column mismatch on geoacoustic inversion," IEEE J. Ocean. Eng. **31**, 284–298 (2006).
[6]D. J. Tang, J. Moum, J. Lynch, P. Abbot, R. Chapman, P. Dahl, T. Duda, G. Gawarkiewicz, S. Glenn, J. Goff, H. Graber, J. Kemp, A. Maffei, J. Nash, and A. Newhall, "Shallow Water '06 - A Joint Acoustic Propagation / Nonlinear Internal Wave Physics Experiment," Oceanogr. **20**, 156–167 (2007).
[7]C.-F. Huang, P. Gerstoft, and W. S. Hodgkiss, "Effect of ocean sound speed uncertainty on matched-field geoacoustic inversion," J. Acoust. Soc. Am. **123**(6), EL162–EL168 (2008).
[8]E. K. Westwood, C. T. Tindle, and N. R. Chapman, "A normal mode model for acousto-elastic ocean environments," J. Acoust. Soc. Am. **100**, 3631–3645 (1996).
[9]N. R. Chapman, Y.-M. Jiang, W. S. Hodgkiss, and P. Gerstoft, "Geoacoustic inversion in the SW06 shallow water experiments," *Proceeding of Underwater Acoustic Measurements Conference*, Crete, June, 2007, pp. 163–170.
[10]Personal communications with Dr. D. J. Tang.
[11]A. Turgut, D. Lavoie, D. J. Walter, and W. B. Sawyer, "Measurements of bottom variability during SWAT New

Jersey Shelf experiments," *Impact of Littoral Environmental Variability on Acoustic Predications and Sonar Performance* (Kluwer, Dordrecht, 2000), 91–98.
[12]S. M. Dediu, W. L. Siegmann, and W. M. Carey, "Statistical analysis of sound transmission results obtained on the New Jersey continental shelf," J. Acoust. Soc. Am. **122**, EL23–EL28 (2007).

Effect of ocean sound speed uncertainty on matched-field geoacoustic inversion

Chen-Fen Huang
Department of Marine Environmental Informatics, National Taiwan Ocean University, Taiwan 202
chenfen@mail.ntou.edu.tw

Peter Gerstoft and William S. Hodgkiss
Marine Physical Laboratory, Scripps Institution of Oceanography, La Jolla, California, USA
gerstoft@ucsd.edu, whodgkiss@ucsd.edu

Abstract: The effect of ocean sound speed uncertainty on matched-field geoacoustic inversion is investigated using data from the SW06 experiment along a nearly range-independent bathymetric track. Significant sound speed differences were observed at the source and receiving array and several environmental parameterizations were investigated for the inversion including representing the ocean sound speed at both source and receivers with empirical orthogonal function (EOF) coefficients. A genetic algorithm-based global optimization method was applied to the candidate environmental models. Then, a Bayesian inversion technique was used to quantify uncertainty in the environmental parameters for the best environmental model, which included an EOF description of the ocean sound speed.

PACS numbers: 43.30.Pc, 43.60.Pt [JL]
Date Received: September 11, 2007 **Date Accepted:** February 25, 2008

1. Introduction

Uncertainty in ocean sound speed profiles has significant impact on matched-field geoacoustic inversion.[1–4] Although the goal of inversions is to infer the geoacoustic properties of the sea floor based on acoustic field observations received on an array, uncertainty resulting from temporal and spatial variability of the ocean sound speed plays an import role in the estimation of geoacoustic parameters, especially for higher frequencies.

The purpose of this paper and a companion paper[5] is to report the geoacoustic inversion results from stationary source data obtained during the Shallow Water 2006 experiment (SW06). We focus on the parametrization of the ocean environment, while the companion paper focuses on inverting data at several ranges. However, the two papers both invert data for a source station at 1 km range from the vertical line array. This allows for a detailed discussion of the inversion method employed in the two papers.

The experimental site is on the outer continental shelf of the western North Atlantic and is located roughly 100 miles east of Atlantic City, New Jersey. During the experiment, both acoustic and oceanographic data were collected. Significant ocean sound speed variations were observed from the conductivity-temperature-depth (CTD) measurements made at the source and receiving array. To mitigate the ocean sound speed mismatch, several environmental parameterizations were investigated, including three range-independent models and two range-dependent models. Of interest is which model can better represent the geoacoustic environment sensed by the acoustic transmissions. During this experiment, the values of the geometric parameters are reasonably well known through direct measurements.

2. The experiment

On JD239, acoustic transmissions were made from a J-15 source at 30 m depth deployed from the R/V Knorr. The experimental geometry is illustrated schematically in Fig. 1(a). A 16-element 56.25 m aperture autonomous recording vertical line array (VLA) was moored at lo-

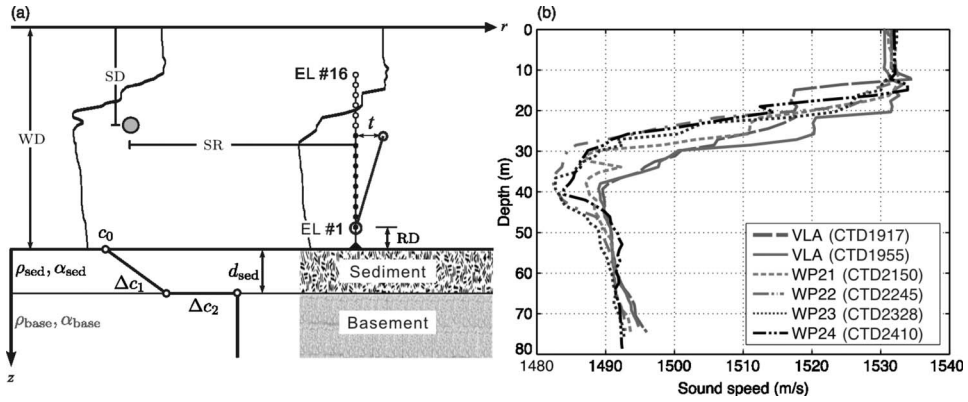


Fig. 1. Range-dependent parameterization of the SW06 environment. (b) Measured sound-speed profiles during the acoustic transmissions. The measurement time of each CTD cast is indicated as a suffix.

cation 39°1.477′N, 73°2.256′E where the water depth was 79 m. The lowermost element was 8.2 m above the bottom. Acoustic transmissions were carried out from four stations (at distances of 1, 3, 5, and 7 km from the VLA). At each station, two sets of 5 min multitonal combs were transmitted, low frequencies at 53, 103, 203 and 253 Hz, and mid frequencies at 303, 403, 503 and 703 Hz. Only the multitonal transmissions at the 1 km range (wp21; 2155-2205 universal time clock (UTC)) are discussed here.

The sound-speed profile in the water column was measured by CTD casts. Figure 1(b) shows sound-speed profiles during the acoustic transmissions. The CTD measurements at the VLA (CTD 1917 and 1955) exhibit small variations above 10 m and below 35 m. However, noticeable time-evolving sound-speed fluctuations on the order of 10 m/s are observed in the thermocline (where the sound source was located). The CTD casts were carried out at each station immediately before the start of the acoustic transmissions. Of note are the lower sound speed values in the thermocline observed in the CTDs taken at the source stations (CTD 2150, 2245, 2328 and 2410). These variations in ocean sound speed structure have an affect on the observed acoustic fields.

A geophysical survey in the study area was conducted using a high-resolution chirp sonar and GeoProbe system.[6] The chirp sonar survey, along the acoustic transmission track, showed the prominent shallow subsurface *R* reflector at 22.4 m (the *R* reflector represents an erosional surface formed during the last low stand of sea level). Preliminary GeoProbe results using the frequency band 20–50 kHz indicate that the surficial sediment sound speed is around 1620–1660 m/s near the VLA site.

3. Matched-field geoacoustic inversion

Each 5 min time series was sampled at 50 kHz and processed using 2^{17}-point fast Fourier transforms, corresponding to a snapshot of 2.6 s. The data cross-spectral density matrices (CSDMs) were computed as the average of outer products of four snapshots representing a time epoch of about 3.8 s. Due to mechanical strum contamination in the upper few array elements, only the data recorded on the lower ten elements are used.

3.1 Environmental parameterization

The base line model parameters are divided into three subsets: geoacoustic, geometrical, and ocean sound-speed parameters. Table 1 lists each environmental parameter and their search bounds. These values were selected based on *a priori* environmental knowledge.

The geoacoustic model is assumed to be range independent with a sediment layer overlying a basement, Fig. 1(a). The sediment sound speed varies linearly with depth, whereas the basement sound speed is constant. The sediment attenuation is assumed to have linear fre-

Table 1. Parameters with search bounds and Bayesian inversion results for the RD-2 model. $SAGA_{powell}$ is the best fit model using GA and the Powell method. $SAGA_{mean}$ and σ are the mean and standard deviation estimated from the Markov chain Monte Carlo derived posterior probability densities.

Model parameter		Search bounds		Inversion results	
Description	Symbol	Upper	Lower	$SAGA_{powell}$	$SAGA_{mean} \pm \sigma$
Geometric					
Source range (m)	SB	1000	1050	1040	1040 ± 3
Source depth (m)	SD	29	31	30.4	30.1 ± 0.5
Water depth at src (m)	WD_{src}	77	81	78.6	78.1 ± 0.7
Water depth at rcv (m)	WD_{rcv}	77	81	80.6	80.3 ± 0.5
1st receiver depth (m)	RD	7.5	9	8.3	8.2 ± 0.4
Array tilt (m)	t	0.5	1.5	1.15	1.08 ± 0.2
Geoacoustic					
Comp. speed (m/s)	c_0	1580	1700	1590	1596 ± 11
Incr. comp. speed (m/s)	Δc_1	0	100	20.4	23.9 ± 18
Attenuation (dB/λ)	α_{sed}	0.001	1	0.05	0.20 ± 0.2
Density (g/cm^3)	ρ_{sed}	1	2	1.8	1.7 ± 0.1
Layer thickness (m)	d_{sed}	10	30	24.2	24.7 ± 3
Incr. comp. speed (m/s)	Δc_2	50	150	129	108 ± 24
Ocean sound speed					
EOF 1_{src}/EOF 1_{rcv}		−65	−20	−65/−59	$-56 \pm 7/-52 \pm 10$
EOF 2_{src}/EOF 2_{rcv}		−25	20	−11.3/12.3	$-8.2 \pm 7/5.1 \pm 0.8$
EOF 3_{src}/EOF 3_{rcv}		−10	20	1.0/11.4	$10.0 \pm 7/7.1 \pm 8$

quency dependence, since at short range the sensitivity to nonlinear dependence is negligible. Separate simulations suggest that the inversion results are insensitive to density and attenuation in the sub-bottom. Therefore, they were set at nominal values[7] of density 2.1 g/cm^3 and attenuation 0.2 dB/λ.

The geometric parameters included in the inversions are the source range, source depth, and water depth. The array configuration is defined by estimating an effective horizontal offset between the 10th and first elements for tilt and estimating the distance of the first array element from the sea floor.

The ocean sound speed profile is parametrized by the first three EOF coefficients. Because the sound-speed difference in the thermocline layer was significant between the measured CTDs, Fig. 1(b), an empirical orthogonal function (EOF) analysis of the sound-speed measurements was carried out using all CTD measurements (16 casts) along the 80 m ISO-bath track during the period JD239 1917 to JD243 2106 UTC. Figure 2 summarizes the EOF analysis for the sound-speed profile measurements. The first three EOFs shown in Fig. 2(c) account for 95% of the variance.

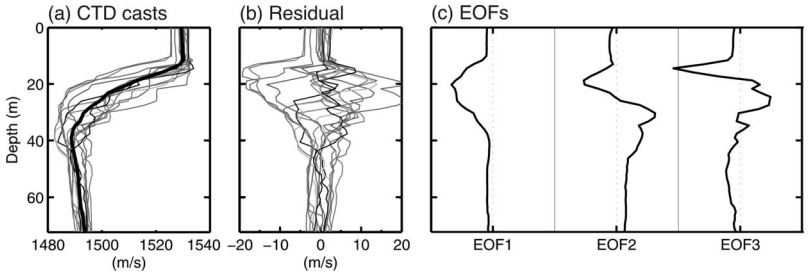


Fig. 2. EOF analysis for the SW06 CTD casts. (a) Sound-speed profiles measured from R/V Knorr and the average sound-speed profile (thick line). (b) Residual sound-speed profiles. (c) First three EOFs.

3.2 Inversion models

In the following three range-independent (RI) and two range-dependent (RD) inversion models, all five models use the same geoacoustic and geometric parameters, except for WD, as shown in Fig. 1(a) and Table 1. First, the ocean environment is assumed range independent. A typical inversion would be done using the sound speed profile corresponding to the CTD taken at the time closest to the execution of the experiment. The first model, RI-1, uses the sound speed profile measured 5 min prior to the transmission (CTD2150). The second model, RI-2, uses the sound speed profile measured at the VLA (CTD1955). The third model, RI-3, inverts for the first three EOF coefficients in order to model and estimate the sound-speed profile.

Second, a range-dependent water column is considered. The water depths at the source (WD_{src}) and VLA (WD_{rcv}) were both included in the inversion for model RD-1 and the sound-speed profiles measured at these two sites were used. In model RD-2, a total of six (three at each end) EOF coefficients were used to model the sound-speed profiles at the source and the VLA along with estimating the corresponding water depths as well.

3.3 Objective function and optimization method

The objective function measures the discrepancy between the measured acoustic and replica fields calculated for likely values of the unknown parameters. The data misfit objective function is based on the incoherent multifrequency Bartlett processor:

$$\phi(\mathbf{m}) = 1 - \frac{1}{L}\sum_{l=1}^{L} \mathbf{d}_l(\mathbf{m})^{\dagger}\mathbf{R}_l\mathbf{d}_l(\mathbf{m}), \tag{1}$$

where $\mathbf{d}(\mathbf{m})$ is the replica field generated for the vector of unknown parameters $\mathbf{m}$ normalized to have unit length, $\mathbf{R}$ is an estimated CSDM normalized to have unit trace, L the number of frequencies, and the superscript † denotes complex conjugate transpose. For the range-independent models, the normal-mode propagation code SNAP[8] was used to compute the replica fields. As for the range-dependent parameterizations, the replica fields were calculated by SNAPRD[8] based on adiabatic normal modes.

A global optimization method based on genetic algorithms (GAs) is used for the optimization. The values of the GA parameters are as follows: the population size 64; reproduction size 0.5; crossover probability 0.8; mutation probability 0.05; and the number of forward model runs for each population was 4000. For the best environmental model, the posterior probability was sampled by a Metropolis–Hastings algorithm.[8,9]

3.4 Inversion results

Matched-field geoacoustic inversion using all transmitted frequencies (eight frequencies and ten phones) was carried out for each of the above described environmental models. Table 2 tabulates the minimum values of the objective function ϕ and the water depth (WD) from the inversions. The minimum value is $SAGA_{powell}$,[8] which is found at the end of each GA run by

Table 2. Minimum values of the objective function ϕ and the parameter WD for different environmental parameterizations. For the range-dependent models, the values are for the parameters WD_{src}/WD_{rcv}.

	Range independent			Range dependent	
	RI-1	RI-2	RI-3	RD-1	RD-2
ϕ	0.301	0.212	0.196	0.304	0.187
WD	77.0	79.8	79.1	77.0 /81.0	78.6 / 80.6

carrying out a local optimization using the Powell method. Only the inversion results for the water depth are reported in the table, but a similar comparison could be made using any of the geometric parameters.

From the direct measurements, the source and VLA water depths are 79 and 78.8 m, respectively, with a standard deviation of 0.4 m. Hence, the search bound of water depth for the RI/RD cases is from 77 to 81 m. Comparing the inversion results from the range-independent models, we found that RI-1 (using the source sound speed profile) gives a rather poor inversion, the WD estimation is unreasonable (at the lower bound), and the misfit objective function has relatively high value. The inversion from RI-2 (using the VLA sound speed profile) is improved: a 30% reduction of the misfit value and the inverted WD appears near the center of the search interval. RI-3 (inverting 3 EOF coefficients) gives the best inversion results for this class. For the range-dependent models, RD-1 produces a poor inversion similar to that of RI-1. This may indicate that the source sound speed profile is not sufficiently representative of the source region of the track. RD-2 (including three sound speed EOFs and WD at both the source and VLA; an additional four more parameters than estimated in RI-3) outperforms all of the other environmental models. The smallest misfit value is achieved and an excellent agreement is obtained between the estimated and measured water depth values.

Model RD-2 is chosen for further parameter uncertainty analysis using a Markov chain Monte Carlo method based on the Metropolis–Hastings algorithm.[9] Table 1 summarizes the estimated value, $SAGA_{powell}$, and the parameter uncertainty estimate, $SAGA_{mean} \pm \sigma$, for each of the model parameters in RD-2. We see that some parameters are estimated very accurately, for example, the mean value of the upper sediment sound speed c_0 is 1596 m/s with a standard deviation of only 11.5 m/s. The inversion result is in very good agreement with the sandy bottom geoacoustic properties indicated by *in situ* measurements.[6]

The estimated sound speed profiles [Fig. 3(a)] at the source (thick dashed) and the VLA (thin dashed) are very similar to each other and resemble the measured profile at the VLA

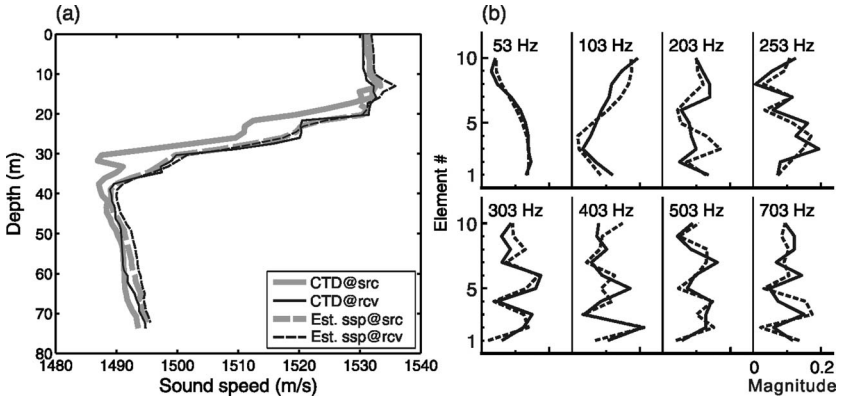


Fig. 3. Inversion results for RD-2. (a) Estimated sound speed profiles. (b) Comparison of the measured (solid) and modeled (dashed) sound fields from $SAGA_{powell}$ on the vertical array for each of the frequencies used in the inversion.

(thin solid). This might explain that the consistent misfit values obtained for both RI-3 and RD-1 models. Figure 3(b) shows good agreement between the measured (solid) and modeled (dashed) acoustic fields for the frequencies used in the inversion. At short range, it appears that the acoustic field is not particularly sensitive to the range-dependent ocean sound speed structure, and an equivalent range-independent sound speed model may be sufficient for describing the environment.

4. Discussion

The experiment took place in a very complicated ocean environment at the edge of the Gulf stream, which potentially creates strong internal waves. In a companion paper Jiang and Chapman[5] (UVic) inverts similar data along the same track with the data at 1 km range common to both papers. In the following, our paper is referred to as SIO. It is useful to discuss the difference in results, but first the two approaches are outlined:

1. Data: UVic uses one 2.6 s interval, whereas SIO uses four snapshots covering 3.8 s. More snapshots would have been beneficial.
2. UVic uses 72 2.6 s intervals (about 5 min) to estimate the data-error covariance matrix, SIO assumes a constant data error estimated from the data. This is expected to mainly influence the uncertainty analysis.
3. SIO estimates the EOFs from 16 CTD casts. However, since the sound speed profiles only vary close to the thermocline, UVic estimates the EOFs between 10 and 50 m depth and bases the estimates on just five sound speed profiles. More sound speed observations would have been preferable. This is expected to have little impact on the results.
4. Objective function (both use a Bartlett power metric), forward model (UVic uses ORCA, whereas SIO uses SNAPRD), and optimization method (both use Markov chain Monte Carlo) are not expected to result in significant differences.
5. Parameterization:

(a) Both include source range and depth, and array tilt and depth as best results are obtained if the geometric parameters are given some modest range of uncertainty.

(b) Attenuation and density are not important and will not be discussed.

(c) SIO constrains the sediment to have a positive gradient, whereas UVic also allows for negative gradients. Often marine sediments have positive gradients. In the present case, Fig. 5 of Ref. 10 from a nearby site shows cores having a negative gradient. In hindsight, the SIO sediment sound speed gradient should have been unconstrained.

(d) UVic does not allow for any ocean sound-speed range dependence. SIO allows for range dependence by inverting for ocean sound speeds at both source and VLA. This could represent an improvement if the sound speed was mildly range dependent and likely more important for the larger source ranges.

Comparing the obtained ocean bottom sound speed profile UVic obtains 1636, 1572, 1740 m/s for sediment top and bottom, and bottom half space, respectively, with a 21.1 m sediment layer. Corresponding values from SIO are 1589, 1609, and 1739 m/s with a 24 m sediment layer. Thus, the bottom half-space sound speed agrees perfectly. Based on the results of Turgut's[6] chirp profile, the sediment depth is 22.4 m, in reasonable agreement with both UVic (21.1 m) and SIO (24.2 m). The top sediment has a difference of 40 m/s between the two methods. However, the average sound speed in the sediment is probably more important for wave propagation. For this average, UVic obtains 1604 m/s, whereas SIO obtains 1599.4 m/s, also an excellent agreement. For the ocean sound speed, both obtain a sound speed profile that is similar to the measured profile at the VLA.

Uncertainty in ocean sound speed structure affects matched field geoacoustic inversion results. In the SW06 experiment, spatio-temporal variability of the ocean sound speed

structure was observed during the acoustic transmissions. The uncertainty in the ocean sound speed profiles was mitigated by including EOF coefficients for either a range-independent or range-dependent ocean sound speed environment in the inversion. The inversion results show that including sound speed EOF coefficients in the inversion yields a significantly better estimate of the geoacoustic parameters. While spatial sound speed variations were observed at the source and VLA, it was found that a single estimated sound speed profile obtained from the inversion was sufficient to represent this mildly range-dependent environment.

Acknowledgments

The SW06 was a large experiment sponsored the Office of Naval Research and involving about 50 PIs collecting both acoustic and oceanographic data. This work was supported by the Office of Naval Research Grant No. N00014-05-1-0264. In addition, C.F.H. was supported by the National Science Council and the Asian-Pacific Ocean Research Center at National Sun Yat-sen University sponsored by the Ministry of Education of Taiwan under the Projects Nos. NSC96-2218-E-019-003 and APORC-95C100303. The support of the ARL:UT team providing the acoustic source transmissions and the R/V KNORR crew is appreciated.

References and links

[1]P. Gerstoft and D. F. Gingras, "Parameter estimation using multifrequency range-dependent acoustic data in shallow water," J. Acoust. Soc. Am. **99**, 2839–2850 (1996).
[2]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**, 769–781 (2001).
[3]Y.-T. Lin, C.-F. Chen, and J. F. Lynch, "An equivalent transform method for evaluating the effect of water-column mismatch on geoacoustic inversion," IEEE J. Ocean. Eng. **31**, 284–298 (2006).
[4]C. F. Huang, P. Gerstoft, and W. S. Hodgkiss, "On the effect error correlation on matched-field geoacoustic inversion," J. Acoust. Soc. Am. **121**(2), EL64–EL69 (2007).
[5]Y.-M. Jiang and N. R. Chapman, "Bayesian geoacoustic inversion in a range dependent shallow water environment," J. Acoust. Soc. Am. **123**(6), EL155–EL161 (2008).
[6]A. Turgut, "SW06 bottom characterization by using chirp sonar and GeoProbe data," in *Shallow Water 2006 Experiment San Diego Workshop* (2007).
[7]W. M. Carey, J. Doutt, R. B. Evans, and L. M. Dillman, "Shallow-water sound transmission measurements on the New Jersey continental shelf," IEEE J. Ocean. Eng. **20**, 321–336 (1995).
[8]P. Gerstoft, SAGA Users guide 5.0, an inversion software package. An updated version of "SAGA 2.0," SACLANT Undersea Research Centre, SM-333, La Spezia, Italy (1997).
[9]S. E. Dosso, "Quantifying uncertainty in geoacoustic inversion I: A fast Gibbs sampler approach," J. Acoust. Soc. Am. **111**, 129–142 (2002).
[10]Y.-M. Jiang, N. R. Chapman, and M. Badiey, "Quantifying the uncertainty of geoacoustic parameter estimates for the New Jersey shelf by inverting air gun data," J. Acoust. Soc. Am. **121**, 1879–1894 (2007).

ACOUSTICAL NEWS—USA

E. Moran
Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747–4502

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

New Fellows of the Acoustical Society of America

John C. Osler—For contributions to seabed geoacoustics.

Subramaniam D. Rajan—For development of inverse methods in ocean acoustics.

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.

2008

29 June–4 July Joint Meeting of the Acoustical Society of America, European Acoustics Association and the Acoustical Society of France, Paris, France [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; E-mail: asa@aip.org; WWW: http://asa.aip.org].

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. E-mail icben2008@att.net].

10–14 Nov 156th Meeting of the Acoustical Society of America, Miami, FL [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; E-mail: asa@aip.org; WWW: http://asa.aip.org].

2009

18–22 May 157th Meeting of the Acoustical Society of America, Portland, OR [Acoustical Society of America, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502; Tel.: 516-576-2360; Fax: 516-576-2377; E-mail: asa@aip.org; WWW: http://asa.aip.org].

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

 0001-4966/2008/123(6)/4023/2/$23.00

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)

ACOUSTICAL NEWS—INTERNATIONAL

Walter G. Mayer
Physics Department, Georgetown University, Washington, DC 20057

International Meetings Calendar

Below are announcements of meetings and conferences to be held abroad. Entries preceded by an * are new or updated listings.

June 2008

4–6 **5th International Styrian Noise, Vibration and Harshness Congress 2008**, Graz, Austria (www.accgraz.com).

18–27 *__Summer School in Underwater Acoustics 2008__, Heraklion, Greece (ssua08.iacm.forth.gr).

29–4 **Acoustics'08 Paris: 155th ASA Meeting + 5th Forum Acusticum (EAA) + 9th Congrès Français d'Acoustique (SFA)**, Paris, France (www.acoustics08-paris.org).

July 2008

6–10 **15th International Congress on Sound and Vibration, Daejeon**, Korea (www.icsv15.org).

7–10 **18th International Symposium on Nonlinear Acoustics (ISNA18)**, Stockholm, Sweden (www.congrex.com/18th_isna).

27–31 **10th Mechanics of Hearing Workshop**, Keele University, UK (www.mechanicsofhearing.com).

August 2008

25–28 **1st International Conference on Water Side Security**, Lyngby, Denmark (www.wss2008.org).

25–29 **10th International Conference on Music Perception and Cognition (ICMPC 10)**, Sapporo, Japan (www.icmpc10.typepad.jp).

September 2008

8–12 **International Symposium on Underwater Reverberation and Clutter**, Lerici, Italy (isurc2008.org).

9–11 **6th International Symposium on Ultrasonic Doppler Methods for Fluid Mechanics and Fluid Engineering**, Prague, Czech Republic (isud6.fsv.cvut.cz).

10–12 **Autumn Meeting of the Acoustical Society of Japan**, Fukuoka, Japan (www.asj.gr.jp/index-en.html).

15–17 **International Conference on Noise and Vibration Engineering (ISMA2008)**, Leuven, Belgium (www.isma-isaac.be).

22–26 **INTERSPEECH 2008- 10th ICSLP**, Brisbane, Australia (www.interspeech2008.org).

October 2008

3–5 *__Seventh International Conference on Auditorium Acoustics__, Oslo, Norway (ioa.org.uk).

6–8 *__Acoustics Week in Canada__, Vancouver, B.C., Canada (www.caa-aca.ca/vancouver2008)

14–15 *__Underwater Noise Measurement__, Southampton, UK (underwaternoise2008.lboro.ac.uk).

21–22 *__Institute of Acoustics (UK) Autumn Conference 2008__, Oxford, UK (ioa.org.uk).

21–23 **International Conference on Low Frequency Noise and Vibration**, Tokyo, Japan (www.lowfrequency2008.org).

21–24 **acústica 2008**, Coimbra, Portugal (www.spacustica.pt).

26–29 **inter-noise 2008**, Shanghai, China (www.internoise2008.org).

November 2008

2–5 **IEEE International Ultrasonics Symposium**, Beijing, China (ewh.ieee.org/conf/ius_2008).

5–7 **Iberoamerican Acoustics Congress (FIA 2008)**, Buenos Aires, Argentina (www.adaa.org.ar).

14–18 **20th Session of the Russian Acoustical Society**, Moscow, Russia (www.akin.m).

20–21 *__Reproduced Sound 24__, Brighton, UK (www.ioa.org.uk).

24–26 **Australian Acoustical Society National Conference**, Geelong, Vic., Australia (www.acoustics.asn.au).

December 2008

17–19 *__Symposium on the Acoustics of Poro-Elastic Materials (sapem 2008)__, Bradford, UK (sapem2008.mateleys.com).

April 2009

5–9 **Noise and Vibration: Emerging Methods (NOVEM 2009)**, Oxford, UK (www.isvr.soton.ac.uk/NOVEM2009).

13–17 **2nd International Conference on Shallow Water Acoustics**, Shanghai, China (www.apl.washington.edu).

9–24 **International Conference on Acoustics, Speech, and Signal Processing**, Taipei. R.O.C. (www.icassp09.com).

July 2009

5–9 *__16th International Congress on Sound and Vibration__, Krakow, Poland (www.icsv16.org).

August 2009

23–28 **Inter-noise 2009**, Ottawa, Ont., Canada (Contact: TBA).

September 2009

6–10 **InterSpeech 2009**, Brighton, UK (www.interspcech2009.org).

19–23 *__IEEE 2009 Ultrasonics Symposium__, Rome, Italy (Email: pappalar@uniroma3.it).

October 2009

26–28 **Euronoise 2009**, Edinburgh, UK (www.euronoise2009.org.uk).

August 2010

23–27 **20th International Congress on Acoustics (ICA2010)**, Sydney, Australia (www.ica2010sydney.org).

September 2010

26–30 **Interspeech 2010**, Makuhari, Japan (www.interspeech2010.org).

 0001-4966/2008/123(6)/4025/1/$23.00

BOOK REVIEWS

P. L. Marston
Physics Department, Washington State University, Pullman, Washington 99164

These reviews of books and other forms of information express the opinions of the individual reviewers and are not necessarily endorsed by the Editorial Board of this Journal.

Self-Consistent Methods for Composites, Volume 1 – Static Problems

S. K. Kanaun and V. M. Levin

Springer, The Netherlands, 2008, 376 pp. $169, 978-1-4020-6663-4.

This book is the first of two volumes on the topic of self-consistent methods in composites. The current volume covers quasistatic problems in elasticity, thermal and electric fields, as well as some interactions between thermal/electric fields and elastic fields in composites—although this reviewer did not find any discussion of piezoelectric media here. According to the Preface to Volume 1, the second volume will be devoted to wave propagation problems in composites and other types of heterogeneous materials. The authors have tried to make the two volumes as independent as possible, but presumably the second volume will nevertheless make good use of the quasistatic results of the first volume for waves at low frequencies and long wavelengths.

The book is written at a fairly high mathematical level, treating elastic compliances and stiffnesses as fourth-rank tensors, as is entirely appropriate. The authors include appendices to remind the reader of various identities that may be useful when operating with fourth-rank tensors. On the other hand, they do not include any discussion of the Voigt notation that is well known in acoustics and seismology whereby the second rank stress and strain tensors are treated as 6-vectors, and the fourth rank compliance or stiffness tensors as 6×6 matrices. Of course, choice of the style of treatment is to some extent up to the scientist or engineer studying the problem. But I think the book would have become more useful to a broader audience if the authors had included some discussion of the translation between these two distinct (but nevertheless equivalent) ways of presenting the same information.

The main ideas that are treated in the text have to do with two approaches to averaged equations for composite, or otherwise heterogeneous (say random earth), media. The two ideas that they stress are effective field approximations and effective medium approximations. The effective field method (or EFM) assumes that inclusions in the heterogeneous medium behave as if isolated in the original matrix material, while the exciting or effective field acting on this inclusion is a certain sum of the external applied field and the perturbations due to all the surrounding inclusions. In contrast the effective medium method (or EMM) assumes that each inclusion is an isolated one imbedded in the overall composite medium with the applied field being just the actual overall applied field. Well-known examples of these two approaches for dielectrics are those of Maxwell Garnet (1904) for EFM and Bruggeman (1935) for EMM. The authors' point of view is interesting because it differs somewhat from most other presentations in the recent years; i.e., most writers on this topic consider the EMM to be the first truly self-consistent approach, and the EFM to be one of the non-self-consistent variety. However, the authors do make a valid point here in that when you are discussing self-consistency: just what is it anyway that is really self-consistent in a given approach? This reviewer has indeed heard arguments made for still other quite distinct choices of the "self-consistent" terminology, and putting personal prejudices aside, this is again clearly a choice for the individual scientist/engineer to make. We just need to be careful not to assume that our personal choices are somehow universally held, because they probably are not.

One part of the book of special note is Section 9.5 on "Cross-property relations." The particular properties studied in the section are those of thermoelastic media, and the connections are between effective thermoelastic constants of such composites and their effective overall elastic behavior. Indeed, the second author of the book (Levin) published some of the earliest work on this topic (in the 1960s), surely before the concept of "cross-property relations" was very widely known. So it is a special treat to see this topic presented here at the end of this volume, as one way of showing the audience how all the preceding work is most modern and topical.

The text as a whole is certainly appropriate for a graduate level course in mechanics and/or materials engineering. The book will also be of interest to researchers in the general area of composites, quasistatic or not, as many of the ideas in the literature are explored at a goodly length here, and so it may help to clarify certain technical issues that are not always given sufficient space in a typical journal article. The second volume (which I have not seen) is presumably the one that will be of most interest to those in the acoustics/seismology community. But the present volume does have merit on its own, as mentioned already, since it will be directly useful for studies of both lower frequency waves and also quasistatic behavior.

JAMES G. BERRYMAN
Earth Sciences Division,
Geophysics Department,
Lawrence Berkeley National Laboratory,
1 Cyclotron Road,
MS 90R1116 Berkeley, CA 94720

Musimathics Volume 2

Gareth Loy

The MIT Press, Cambridge, Massachusetts 2007 562 pp. $50 (hardcover). ISBN: 978-0-262-12285-6

Musimathics is an ambitious two-volume, thousand-page tome by Gareth Loy. Trained as a musician and composer, he was formerly a graduate researcher at Stanford's Center for Research in Music and Acoustics (CCRMA) and lecturer in the Music Department at the University of California, San Diego. In recent years, Loy has plied his services as an independent consultant in the digital audio field.

With *Musimathics*, Loy has, in effect, written "Everything I always wanted to know about mathematics related to music that no one ever taught me." As a musician, I can identify with this. The motivation for writing *Musimathics* may seem a puzzle to the electrical engineer or acoustical physicist. Wasn't this all taught in college? The answer is: not really, and especially not to musicians trained in the conservatory tradition.

Technology has changed the landscape, and today a music education need no longer follow the traditional conservatory model. Many composers, music theorists, and musicologists develop software, and all aspects of music are legitimate fields for scientific study. Dozens of music research centers have been set up around the world, competing for both industrial and governmental agency support. More and more music students (especially composers) seek interdisciplinary technical training. Major universities expect all faculty to pursue research, obtain grants, and publish in peer-reviewed journals. For these musicians, *Musimathics* will serve as a wellspring of valuable insight.

This review concerns volume 2 of *Musimathics*, covering the mathematical framework for analysis and synthesis of sounds. The chapter contents of the book are as follows:

 0001-4966/2008/123(6)/4027/1/$23.00

Digital signals and sampling
Musical signals
Spectral analysis and synthesis
Convolution
Filtering
Resonance
The wave equation
Acoustical systems
Sound synthesis
Dynamic spectra

Added to this are 11 brief appendices (30 pages in total) that range from elementary mathematical topics to esoterica such as the Walsh-Hadamard transform.

Loy has targeted a specific niche with this volume: those left behind by traditional textbooks on signal processing. In order to keep engineering texts concise, Loy observes, "most of the common sense has been removed on purpose." His formula is: "to 'rehydrate' the common sense back into it," i.e., injecting its relevance to topics of musical interest.

In terms of pace and style, Loy demands much from the reader, as he moves from elementary to advanced material quickly. For example, his path to the fast Fourier transform takes 43 pages. Compare the charming but glacially paced *Who Is Fourier*? (Translational College of LEX, 1995), which takes 429 pages of baby steps to explain the FFT.

I do not see the brisk pace as a problem. However, to support such a pace, one should prepare the reader. A problem of the exposition of *Musimathics* is a lack of foreshadowing. The introductions are brief, then the narrative simply dives in. The reader is not informed at the outset where the argument is going, or what stops it is going to make along the way. I would have liked to see a notice at the beginning of each section stating the goal: what readers can expect to learn and the motivation for the intermediate steps. I wonder how the experience of teaching a course based on this book might have changed it. In the broader context of this informative tome, however, this is merely a quibble.

As every author knows, technical figures, particularly in a tutorial volume, pose a daunting challenge. Designing a proper figure with the correct proportions to teach a particular concept is a time-consuming craft. *Musimathics* is beautifully illustrated using native Framemaker graphics for geometrical figures, Mathematica for waveforms, and composites of the two approaches as necessary.

Loy adds value with a web site to accompany the book. Readers can find a detailed list of errata discovered since its publication, a useful service. They can also download a programming language called Musimat, which was designed for the musical examples in his book.

To a current graduate student in music composition, these two volumes are at least as important as the *Harvard Dictionary of Music*. In summary, *Musimathics* Volume 2 is a major achievement in the literature of musical acoustics.

CURTIS ROADS
Media Arts and Technology, joint appointment in Music
University of California
Santa Barbara CA 93106-6065

Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms, and Acoustic Virtual Reality

Michael Vorländer

Springer-Verlag, Berlin, 2008. 335 pp. $129 (hardcover). ISBN: 978-3-540-48829-3

Drawing from his work as both educator and researcher at RWTH Aachen and conceived of as a textbook bringing together "all work done in acoustic simulation, auralization and acoustic virtual reality systems," this new book by Michael Vorländer is unique within the literature. While previous monographs have addressed portions of the topics covered, no other source provides a similar breadth of coverage. This is particularly true of the in-depth treatments of geometrical-acoustics simulation methods and methods for the auralization of airborne sound insulation and structure-born sound. Likewise this book is unique in that, being developed from course notes, it comprises, in roughly equal parts, comprehensive introductory material and up-to-date descriptions of current areas of research.

As it is used here, the term "auralization" was first coined by Kleiner[1] to mean "the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space." Vorländer broadens this definition to include any process that renders audible numerical data arrived at through simulation, measurement, or synthesis and notes that, in contemporary parlance, both the process and the result of the process are termed (an) auralization. Under this definition, auralization, as a field, consists of a broad variety of techniques that follow a three-step process of "sound field modeling, processing with an arbitrary sound signal and sound reproduction."

Consistent with its beginnings as course notes, the first five chapters of the book provide an overview of linear acoustics in fluids and structures. Though the coverage is comprehensive, including full chapters on sound sources and on sound fields in enclosures, it is sometimes schematic. As such, these chapters are best supplemented by lecture and would not be well suited to self-learning by an individual not familiar with acoustics. The coverage is also limited in scope in that it is aimed primarily at motivating the simulation methods discussed later in the book. For example, the treatment of scattering is clearly oriented toward application in geometrical-acoustics models and does not address the wave theory of surface scattering.

In a similar manner, Chapter 6 provides an overview of psychoacoustics that is directed toward application. The physiology and function of the auditory periphery are introduced, leading to a discussion of psychoacoustical metrics, mostly derived from the work of Zwicker.[2] Curiously, neither the notion of critical bands nor the concept of the Bark scale is formally introduced, though use of them is implied. Next, spatial hearing is introduced, as is the use of dummy heads, which includes a concise, informative history of their development. The chapter concludes with a substantial and thorough discussion of metrics used to describe perceptual aspects of sound fields in rooms, based largely on those described in the ISO 3382-1 standard.[3]

Chapters 7 through 9 constitute what Vorländer terms the "core" of the book, and it is here that the concept of auralization is introduced. The process is represented as a three-stage block diagram comprising sound generation, sound transmission, and sound reproduction, and this processing chain provides the framework around which the remainder of the book is structured.

Following the introduction of auralization as a concept, Chapter 7 serves primarily as a targeted introduction to relevant aspects of linear-systems theory and digital signal processing. Chapter 8 addresses the first block of the auralization processing chain providing a review of the issues involved in simulation of sound sources and a synopsis of the readily available anechoic source material. This segment of the book concludes by addressing in Chapter 9 a variety of implementational issues primarily related to convolution and synthesis of binaural signals.

Chapters 10 through 13 address the second block of the auralization processing chain, with Chapter 10 giving an outline of available techniques for numerical simulation of sound fields. Computational geometrical-acoustics methods are then given an extensive treatment in Chapter 11 with particular attention paid to issues of uncertainty and computation time. The general focus is on addressing those techniques that have been implemented in commercial software. Much of the material in this chapter is unique or not easily obtained elsewhere. However, as a consequence of the focus on commercial methods, topics at the forefront of research, such as simulation of diffraction, are described only in summary.

Complementing this treatment of sound fields in rooms, Chapters 12 and 13 develop from basic concepts of sound insulation and structural vibration models suitable for auralization. These models are described with sufficient detail to enable implementation. Being based in part on work conducted at RWTH Aachen, much of the material in these chapters is unique to the book, particularly with respect to use of such simulations for auralization.

Expanding the coverage of the book beyond the typical techniques of auralization, Chapter 14 gives an overview of the techniques of binaural-transfer-path analysis and synthesis, first developed at HEAD acoustics for analysis and synthesis of sound fields in automobiles.

With methods for simulation of sound fields and the synthesis of binaural signals having been addressed, Chapter 15 describes the implementation of these methods for real-time interactive auralization, particularly with

 0001-4966/2008/123(6)/4028/2/$23.00

respect to auralization of sound fields in rooms. The associated perceptual requirements resulting from interactivity are also covered.

The book concludes with a chapter on sound reproduction for auralization, which also includes a brief overview of virtual-reality systems. The coverage is quite comprehensive, addressing both loudspeaker-based techniques and binaural techniques using either headphones or loudspeaker systems. The author's clear preference is for binaural systems, which attempt to reproduce the sound field at the ears, and loudspeaker systems based on wave-field synthesis, which attempt to reproduce the sound field in a volume in space, as opposed to methods such as vector-based amplitude panning and first-order ambisonics (B-format), which achieve spatial effects through injection of localization cues.

An annex containing tables of material properties (absorption coefficients, scattering coefficients, and sound reduction indices) compiled from various sources supplements the book. Particularly useful are the tables of random-incidence scattering coefficients, which compile measurements of this parameter for which data are rarely available.

In summary, I find much to recommend in this new book, particularly because of its unique place in the literature and the qualifications of the author to speak to the topic at hand. Given the extensive experience of the author, when reading through Chapters 10 through 16, I often found myself wishing he had written more extensively and in greater technical detail in his areas of expertise. Much more, for example, could have been written regarding technical and practical details of wave-field synthesis. The same is true for the modeling of diffraction in room-acoustic simulation, evaluation of auralization accuracy, and wave-based methods for sound-field simulation. The author addresses this, to some extent, by providing an extensive set of references to many recent papers, but with the consequence that the book is no longer self-contained. Nonetheless, it stands as the current reference volume in this field of growing interest. Ironically, its greatest shortcoming is not scientific, technical, or even structural, but rather proofreading. In what I assume is no fault of the author, the text is beset by spelling errors, incorrectly formatted fonts, and missing punctuation.

JASON E. SUMMERS
Naval Research Laboratory
Acoustics Division, Code 7142
Washington, District of Columbia 20375

[1]M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization – An overview," J. Audio Eng. Soc. **41**, 861–874 (1993).
[2]E. Zwicker and H. Fastl, Psychoacoustics – Facts and Models, 2nd ed. (Springer, New York, 1999).
[3]ISO 3382-1, "Acoustics – Measurement of the reverberation time of rooms with reference to other parameters" (1997).

 0001-4966/2008/123(6)/4028/2/$23.00

REVIEWS OF ACOUSTICAL PATENTS

Sean A. Fulop
Dept. of Linguistics, PB92
California State University Fresno
5245 N. Backer Ave., Fresno, California 93740

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*
JEROME A. HELFFRICH, *Southwest Research Institute, San Antonio, Texas 78228*
DAVID PREVES, *Starkey Laboratories, 6600 Washington Ave. S., Eden Prairie, Minnesota 55344*
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, *Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York 14627*

7,316,523

43.20.Tb ACOUSTICALLY MATCHED METHOD AND APPARATUS FOR SCREEDING CONCRETE

J. Dewayne Allen and Richard P. Bishop, assignors to Allen Engineering Corporation
8 January 2008 (Class 404/114); filed 25 January 2005

The process of vibratory screeding of freshly placed concrete is enhanced by matching the acoustic impedance of portions of the screeding equipment that are in contact with the plastic concrete to the acoustic impedance of the concrete. This matching, which improves the transfer of energy to the concrete, is accomplished by placing between the vibration source and the concrete a quarter-wavelength thick plate of a material whose acoustic impedance approximates that of the concrete.—EEU

7,313,481

43.30.Ma METHODS AND DEVICES FOR ANALYZING AND CONTROLLING THE PROPAGATION OF WAVES IN A BOREHOLE GENERATED BY WATER HAMMER

Daniel Moos and Youli Quan, assignors to GeoMechanics International, Incorporated
25 December 2007 (Class 702/12); filed 25 May 2006

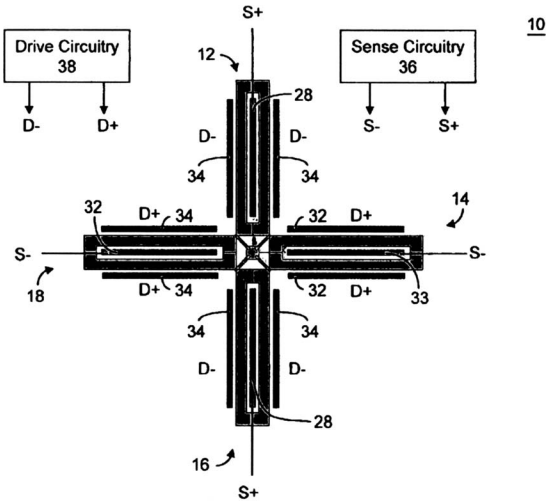
This patent discloses the use of borehole waves generated by the closing of a valve (the classical "water hammer" effect) to probe the local porosity and density of the borehole surroundings. Working from a layered model of the surrounding medium, the authors show how the wave amplitude and shape evolution can be used to determine the presence of a porous section to the borehole, and discontinuities in the density of the surroundings. It appears to be quite uncomplicated once the computer model is set up—and the description of that is largely missing from the patent.—JAH

7,319,372

43.38.Hz IN-PLANE MECHANICALLY COUPLED MICROELECTROMECHANICAL TUNING FORK RESONATORS

Zhiyu Pan *et al.*, assignors to Board of Trustees of the Leland Standford Junior University
15 January 2008 (Class 333/197); filed 15 July 2005

This patent discloses an interesting arrangement of coupled tuning fork resonators that have characteristics optimally tailored for MEMS fabrication. As shown in the figure, the arrangement of the tuning forks is radial, such that they are all supported on a central stem and by a combination of the four tips, depending on the type of coupling desired. The drive and sense is electrostatic, via electrodes on the sides of the tines of each fork and on the substrate next to them. The layout is said to provide for full differential drive and pickoff of the vibrations, and it looks well suited to that. The



authors also claim that the design lowers electrode-to-substrate parasitic capacitances. It is not stated what the arrangement would be used for, though gyroscopes and accelerometers come to mind. The patent discloses only the manufacturing process, not any of the performance details.—JAH

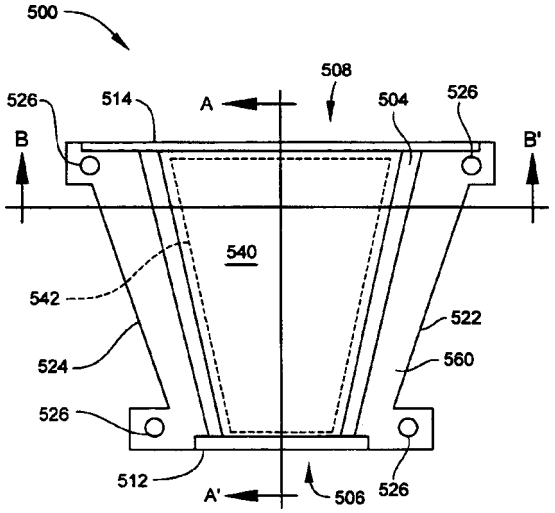
 0001-4966/2008/123(6)/4031/9/$23.00

7,315,627

43.38.Ja SOUND-DAMPING LAMINATE FOR LOUDSPEAKER STRUCTURE

David H. Cox *et al.*, assignors to Harman International Industries, Incorporated
1 January 2008 (Class 381/353); filed 16 August 2004

The walls of high quality horns and waveguides should be rigid and nonresonant, yet a lightweight structure is equally important. This patent describes a three-layer damping panel that can be incorporated into the structure of a horn. The core layer embedded between the inner and outer



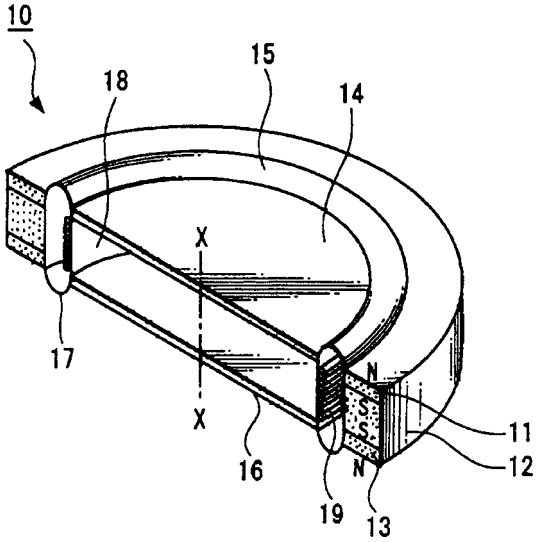
horn surfaces may consist of any suitable damping material, including silicon rubber, resilient foam, corrugated fiber, and balsa wood.—GLA

7,317,810

43.38.Ja MAGNETIC CIRCUIT AND SPEAKER

Yoshio Ohashi, assignor to Sony Corporation
8 January 2008 (Class 381/421); filed in Japan 6 January 2004

The invention is a thin, lightweight speaker with two, symmetrically mounted diaphragms **14** and **16**. It belongs to the general category of open-gap moving coil speakers in that voice coil **19** is adjacent to magnetic assembly **13** rather than in a gap between two pole pieces. An unusual three-layer magnetic structure is used. The top and bottom magnets are magnetized vertically, with poles (from top to bottom) north, south, south, north as indicated. The central magnet is magnetized radially, with its south pole adjacent to the voice coil. A relatively uniform magnetic field is generated in close proximity to the coil, and at nearly 90° to the conductors.—GLA

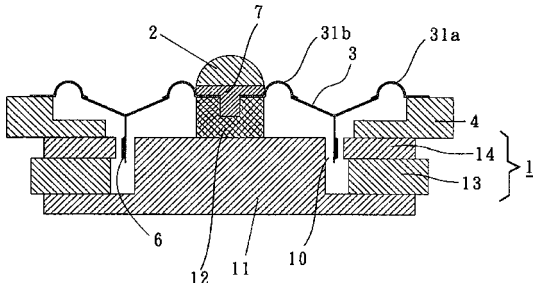


7,319,772

43.38.Ja SPEAKER DEVICE FOR IMPROVING MID/HIGH-RANGE FREQUENCIES

George Chang and Kaohsiung Hsien, Taiwan
Taiwan 15 January 2008 (Class 381/343); filed 7 January 2005

Many loudspeakers incorporate a central projecting bullet (often called a phasing plug) to smooth response at higher frequencies. This patent argues that better results can be obtained if bullet **2** is made of sound absorbing



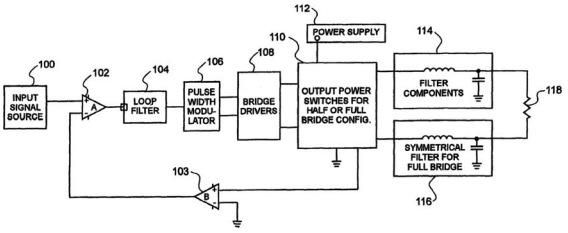
material. Several frequency response curves are included in the patent document, but some of them may be mislabeled because they seem to contradict the benefits set forth.—GLA

7,319,763

43.38.Lc POWER AMPLIFICATION FOR PARAMETRIC LOUDSPEAKERS

Jeevan G. Bank and James J. Croft, III, assignors to American Technology Corporation
15 January 2008 (Class 381/77); filed 11 July 2001

A parametric loudspeaker modulates an ultrasonic beam to produce sound from thin air. The process is inherently inefficient, and the reactive load presented by most ultrasonic transducers further burdens the power



amplifier. This patent describes the design of an interesting Class D (switching) amplifier optimized for this service.—GLA

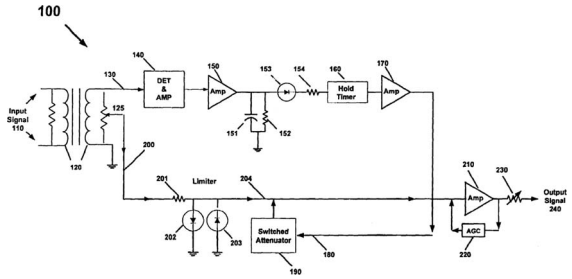
7,310,420

43.38.Si TELEPHONE HEADSET AMPLIFIER WITH NOISE BLANKING CAPABILITY

Robert E. Lucey and Michael B. Lasky, assignors to GN Netcom A/S
18 December 2007 (Class 379/392); filed 28 April 2004

Wireless telephone handsets are susceptible to bursts of noise from electrical interference. This patent describes a sophisticated squelch circuit that attenuates such noise before it reaches the listener's ear. Several variants are disclosed but they all split the incoming audio signal into two electronic

paths. One path detects excessive peaks and then triggers an attenuator in the second path. Some variants include a delay in the main signal path. The



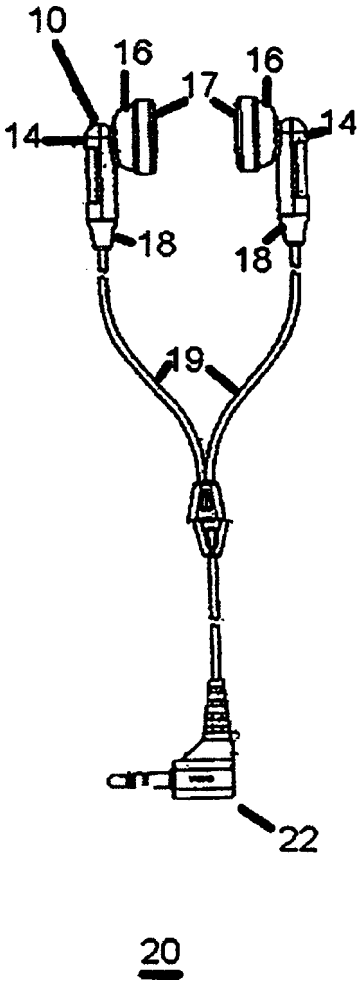
techniques are similar to those employed in high-quality audio limiters, which may be why all 17 patent claims are clearly restricted to telephone communications.—GLA

7,319,762

43.38.Si HEADSET WITH FLASHING LIGHT EMITTING DIODES

Douglas J. Andrea and Qunsheng Liu, assignors to Andrea Electronics Corporation
15 January 2008 (Class 381/74); field 5 July 2006

Near the beginning of the patent document we are told, "A user can, and often does, wear the headset when listening to an audio signal, and even when not." Thus the need for a combination headset and color organ that



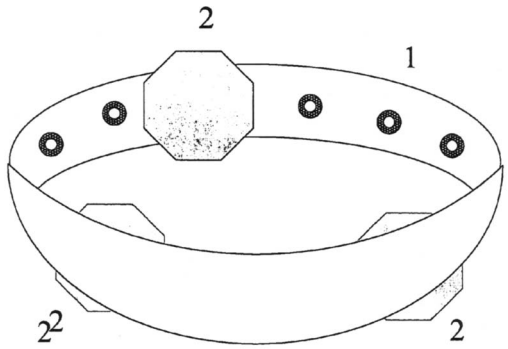
notifies others in the vicinity when it is safe to interrupt the user's audio reverie.—GLA

7,310,427

43.38.Si RECREATIONAL BONE CONDUCTION AUDIO DEVICE, SYSTEM

Sheldon M. Retchin and Martin Lenhardt, assignors to Virginia Commonwealth University
18 December 2007 (Class 381/380); filed 29 July 2003

This patent describes a bone conduction audio sweatband. A preferred embodiment utilizes two or more transducers (vibrational woofers and tweeters), each covered by a waterproof polymeric material. "The device is



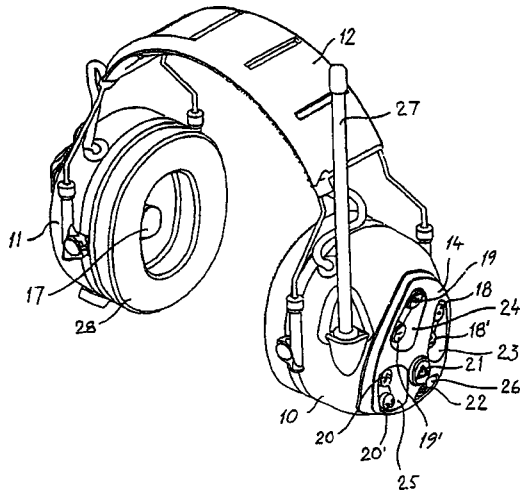
tunable for sound quality and comfort by adjusting and moving the sound transmitting transducers around the head of the user."—GLA

7,317,809

43.38.Si ARRANGEMENT IN ACOUSTIC HEADSETS

Christer Almqvist, assignor to Peltor AB
8 January 2008 (Class 381/371); filed in Sweden 15 August 1997

This is an unusual patent for three reasons: First, its history goes back ten years to August 1998 when the first application was filed. Second, there is only one patent claim. Finally, the novelty of the invention appears to lie in a single ancillary feature rather than the device itself. What is pictured is



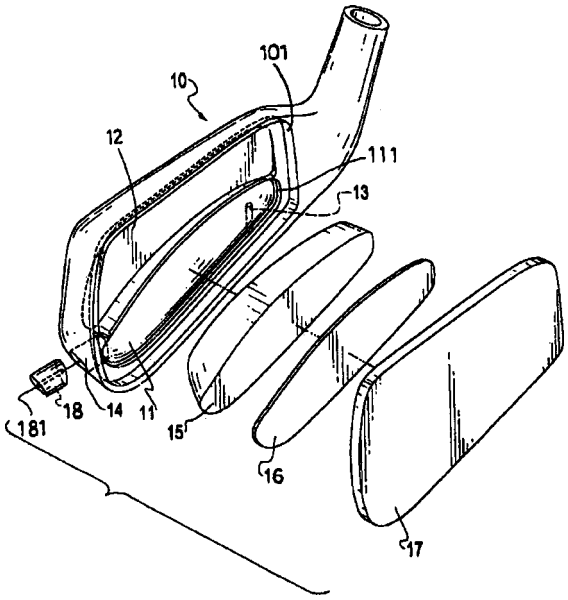
a combination radio and noise canceling headset. "User-friendly" control is provided by several groups of buttons, each group arranged in a separate recess.—GLA

7,303,485

43.40.Jc SHOCK-ABSORBING GOLF CLUB HEAD

Wen-Cheng Tseng, Chienchen Dist., Kaohsiung, Taiwan
4 December 2007 (Class 473/332); field in Taiwan 31 December 2003

Head body **10**, which is partly hollow, has shock absorbing core **15** and inner cap **16** located behind faceplate **17**. Optional plug **18** can have a "ventilation hole" **181**. When the club face strikes the ball, the cavity **12**



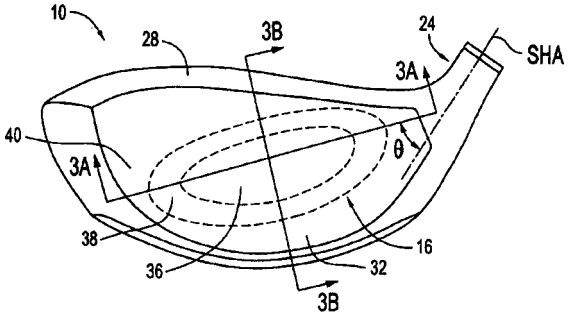
produces the sound effect of an impact while the "remaining shock" is absorbed by core **15**.—NAS

7,297,072

43.40.Kd COMPOSITE METAL WOOD CLUB

Jeffrey W. Meyer *et al.*, assignors to Acushnet Company
20 November 2007 (Class 473/332); field 25 August 2006

Hitting face **16** of golf club head **10** has an inner elliptical area **36** with major axis aligned as shown and having a higher stiffness, defined as the product of the Young's modulus with the cube of the face thickness, than does the elliptical donut **38** surrounding it. This changes the coefficient of restitution in the direction of the high toe to low heel (along cut line **3A**-**3A**) to coincide with the ball impact pattern, therefore increasing the elastic deformation of the club face and reducing the rate of deformation of the



ball, the deformation of which is said to be parasitic. A means of adding dampening to the club head is also described.—NAS

7,317,994

43.40.Le METHOD AND APPARATUS FOR SIGNAL SIGNATURE ANALYSIS FOR EVENT DETECTION IN ROTATING MACHINERY

Naresh Sundaram Iyer *et al.*, assignors to General Electric Company
8 January 2008 (Class 702/56); filed 10 August 2005

The method delineated in this patent in essence makes use of the facts that an impact on a mechanical component causes that component to "ring" at its natural frequency, and that this ringing decays exponentially. The event detection system described here thus seeks to isolate events in acquired vibration signals and to determine the frequencies and damping associated with the events of interest. It then uses this frequency and damping information to detect the occurrences of given events in other vibration signals.—EEU

7,318,007

43.40.Le REAL TIME GEAR BOX HEALTH MANAGEMENT SYSTEM AND METHOD OF USING THE SAME

Sarkis Barkhoudarian, assignor to United Technologies Corporation
8 January 2008 (Class 702/184); filed 31 December 2003

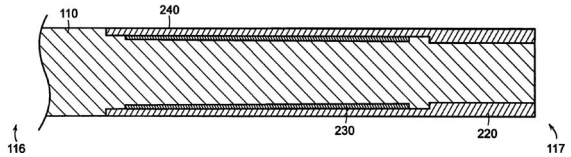
This system for tracking the progression of wear in gears makes use of a plurality of proximity sensors positioned around a rotating gear. The times between signals resulting from individual gear tooth passages are determined, along with amplitude data, and are used to determine the natural frequencies associated with deflections of the gear teeth and the presence of chatter. This information then may be used to predict the imminence of failure in real time.—EEU

7,297,068

43.40.Tm VIBRATION DAMPING FOR A CUE STICK

Paul D. Costain, Beverly, Massachusetts, and Bill Stroud, Ruidoso Downs, New Mexico
20 November 2007 (Class 473/44); filed 22 October 2004

Pool cue shaft body **110** is formed so that sleeve **220** can be fitted over the body. Damping material **230** fills the void in the shaft body. This is all at the butt end of the cue, as damping the free end may change the vibration characteristic players are accustomed to when striking the ball. The treat-



ment at the butt end is said to reduce the vibrations at this location and improve the shot.—NAS

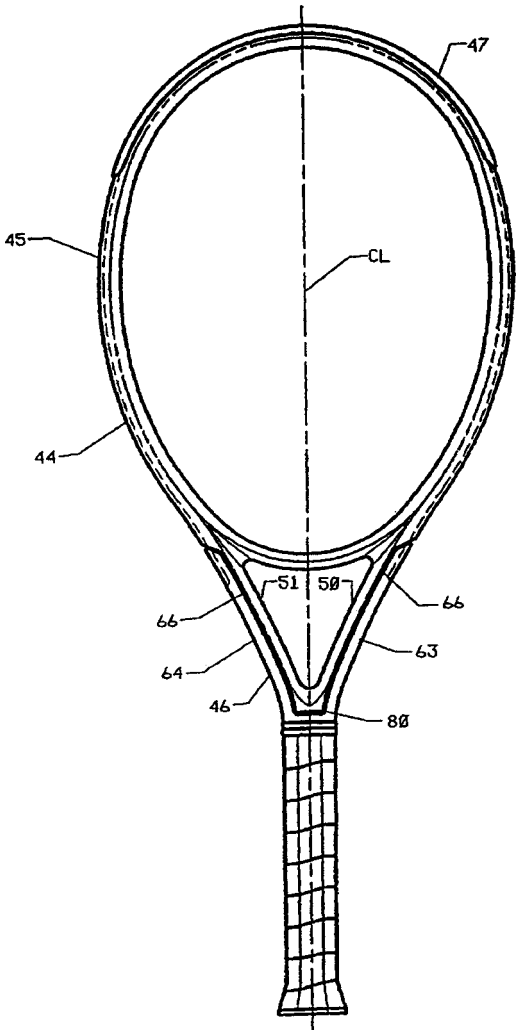
7,297,080

43.40.Tm GAME RACQUET WITH SEPARATE HEAD AND HANDLE PORTIONS FOR REDUCING VIBRATION

William D. Severa *et al.*, assignors to Wilson Sporting Goods Company
20 November 2007 (Class 473/536); filed 6 January 2005

Shock on a typical tennis racquet can last about 5 ms and vibration can

last about 1000 ms after ball impact. This can cause physical problems such as tendonitis, or tennis elbow. A vibration absorbing material **66**, which can be urethane, natural rubber, butyl rubber, or synthetic rubber, is placed between the handle portion **46** and the head portion **45** of racquet **44**.—NAS

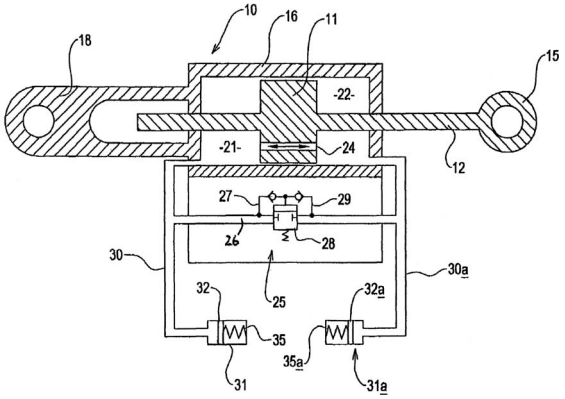


7,314,124

43.40.Tm VIBRATION DAMPING APPARATUS

Gerald Henry Martyn *et al.*, assignors to Westland Helicopter Limited
1 January 2008 (Class 188/318); filed in United Kingdom 18 November 2004

This damper, intended for reducing the vibrations of helicopter blades and the like, consists of a piston **11** that moves within a cylinder **16** with energy dissipation resulting as the fluid that fills the spaces **21** and **22** on the two sides of the cylinder is pushed through opening **24**. Additional control of the damping force is obtained by the fluid being channeled through side branches to a relief and check valve arrangement **28** and to relief chambers **31** and **31a** in which the pressure-induced motions of pistons are opposed by springs.—EEU



7,315,774

43.40.Vn JERK MANAGEMENT USING MULTIVARIABLE ACTIVE DRIVELINE DAMPING

Robert L. Morris, assignor to GM Global Technology Operations Incorporated
1 January 2008 (Class 701/53); filed 22 March 2006

In vehicles with hybrid propulsion systems there tend to occur sudden changes in acceleration of the driveline components as alternate power sources are turned on and off, as energy storage devices are activated, and as the operator changes gears. When the direction of the torque on an axle changes there can result gear lash and jerks as slack is taken out of the driveline and driveline components impact one another. A multivariable feedback control system is provided that restricts the axle torque when a torque reversal occurs and that controls the driveline component speeds to minimize the effect of lash take-up.—EEU

7,316,525

43.40.Vn VORTEX INDUCED VIBRATION OPTIMIZING SYSTEM

Donald Wayne Allen *et al.*, assignors to Shell Oil Company
8 January 2008 (Class 405/211); filed 6 January 2006

Cylindrical bodies that anchor floating items, such as buoys, to the bottom of a body of water are subject to vortex-induced vibrations due to currents. These vibrations lead to undesirable axial oscillatory stresses in these cylindrical bodies. This patent describes systems for measuring these stresses and controlling the tension in the cylindrical bodies automatically so as to reduce these stresses.—EEU

7,311,175

43.50.Gf ACOUSTIC LINER WITH BYPASS COOLING

William Proscia and Christopher D. Jones, assignors to United Technologies Corporation
25 December 2007 (Class 181/290); filed 10 August 2005

Two panels are spaced apart, and a resonator chamber is aligned between them to enhance the performance of a noise attenuation acoustic liner for gas turbine aircraft engines or the like.—CJR

7,311,957

43.55.Ev SOUND ABSORBING MATERIAL AND PROCESS FOR MAKING

Matthew Bargo, II, assignor to CTA Acoustics, Incorporated
25 December 2007 (Class 428/74); filed 12 September 2003

Mix together a blended matrix of man-made fibers, a co-binder (like a phenolic resin), and fibrous cellulose, heat or cure gently, and you end up with a nonflammable, heat-resistant sound absorbing material, suitable perhaps to line various compartments of an automobile and reduce noise transfer from an engine.—CJR

7,318,498

43.55.Ev DECORATIVE INTERIOR SOUND ABSORBING PANEL

Daniel Scott Woodman *et al.*, assignors to Azdel, Incorporated
15 January 2008 (Class 181/290); filed 6 April 2004

This is another embodiment of a sandwich panel to absorb sound inside an automobile, described as a multilayered fiber reinforced thermoplastic sound absorbing panel. The patent claims this system is less expensive to manufacture.—CJR

7,310,558

43.66.Ts PEAK-DERIVED TIMING STIMULATION STRATEGY FOR A MULTICHANNEL COCHLEAR IMPLANT

Richard Van Hoesel, assignor to Hearworks PTY, Limited
18 December 2007 (Class 607/57); filed in Australia 24 May 2001

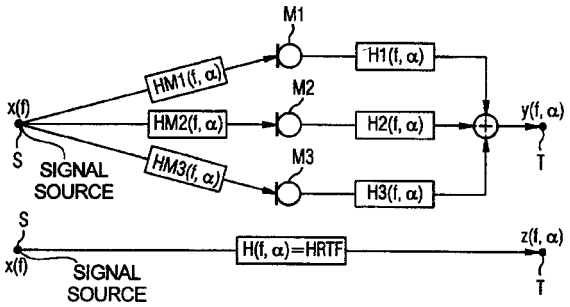
Electrode stimuli are generated for binaural cochlear implant fittings so as to preserve interaural time delays between ears by controlling the time and intensity of the multichannel signal peaks for each ear. Peak magnitude pulses are prioritized and are presented at the times they actually occur instead of being presented at zero crossings. The advantage is said to be that the listener may be able to better separate harmonics from multiple sound sources that are mixed together.—DAP

7,313,241

43.66.Ts HEARING AID DEVICE, AND OPERATING AND ADJUSTMENT METHODS THEREFOR, WITH MICROPHONE DISPOSED OUTSIDE OF THE AUDITORY CANAL

Volkmar Hamacher and Torsten Niederdränk, assignors to Siemens Audiologische Technik GmbH
25 December 2007 (Class 381/60); filed in Germany 23 October 2002

To enhance localization ability, the difference between the input signal at an in-the-canal microphone position and the signal external to the auditory canal at which the hearing aid microphone system is located is applied as a compensation factor to the hearing aid transfer function.—DAP

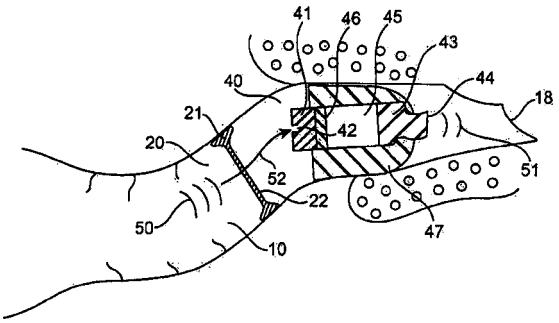


7,313,245

43.66.Ts INTRACANAL CAP FOR CANAL HEARING DEVICES

Adnan Shennib, assignor to InSound Medical, Incorporated
25 December 2007 (Class 381/325); filed 22 November 2000

An acoustically permeable cap protects a hearing device seated deep within the ear canal from fluids and solid debris. The cap is made with a porous membrane and has a retention ring that prevents foreign materials from entering the ear canal by conforming to the walls of the ear canal.—DAP



7,315,626

43.66.Ts HEARING AID WITH PERFORMANCE-OPTIMIZED POWER CONSUMPTION FOR VARIABLE CLOCK, SUPPLY VOLTAGE AND DSP PROCESSING PARAMETERS

Søren Louis Pedersen, assignor to Microsound A/S
1 January 2008 (Class 381/323); filed in Denmark 21 September 2001

Current consumption for the digital signal processor in a hearing aid is proportional to clock frequency and supply voltage. At higher signal-to-noise ratios in the input signal, less signal processing is required, thereby allowing the clock frequency and supply battery voltage to be lowered for increased battery life.—DAP

7,319,768

43.66.Ts HEARING AID AND METHOD FOR THE DETECTION AND AUTOMATIC SELECTION OF AN INPUT SIGNAL

Gerard van Oerle, assignor to Phonak AG
15 January 2008 (Class 381/312); filed 16 March 2004

Two or more analog input signals are analyzed and selectively routed

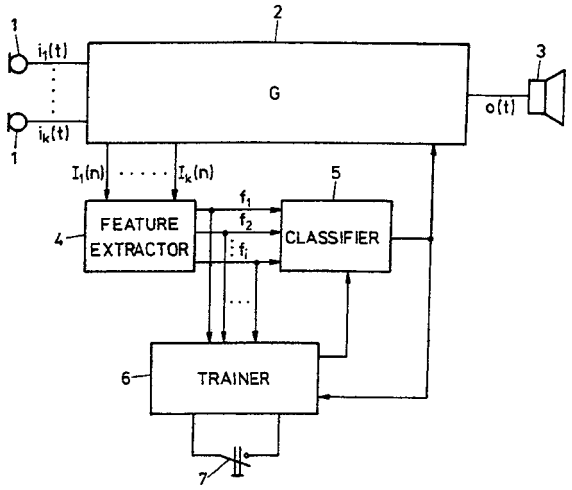
to the appropriate signal processing paths. The audio signal processing circuitry is kept in a reduced power state until a relevant signal is detected by determining if it exceeds a minimum amplitude threshold for a predetermined time.—DAP

7,319,769

43.66.Ts METHOD TO ADJUST PARAMETERS OF A TRANSFER FUNCTION OF A HEARING DEVICE AS WELL AS HEARING DEVICE

Silvia Allegro-Baumann *et al.*, assignors to Phonak AG
15 January 2008 (Class 381/312); filed 9 December 2004

Features from the input signal are extracted and classified. Operating parameters of the hearing aid are adjusted in accordance with the class most nearly corresponding to the momentary acoustic scene. Thereafter, the hearing device is further trained using the new acoustic scene data to improve classification accuracy and speed.—DAP

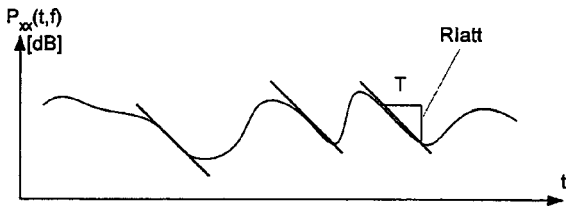


7,319,770

43.66.Ts METHOD OF PROCESSING AN ACOUSTIC SIGNAL, AND A HEARING INSTRUMENT

Hans-Ueli Roeck and Manuela Feilner, assignors to Phonak AG
15 January 2008 (Class 381/321); filed 30 April 2004

Methodology is provided for suppressing room reverberation by calculating hearing aid gain based on a room impulse attenuation measure (the maximum negative slope of the logarithm of the input signal power as a function of time in multiple frequency bands) and signal-to-reverberation-noise ratio. Gain monotonically increases with signal-to-reverberation-noise ratio. Maximum gain is applied if the difference between the acoustic input signal power and the acoustic input signal power delayed is positive.—DAP



7,319,771

43.66.Ts VIBRATOR FOR BONE CONDUCTED HEARING AIDS

Kristian Åsnes, assignor to P & B Research AB
15 January 2008 (Class 381/326); filed in Sweden 2 June 2000

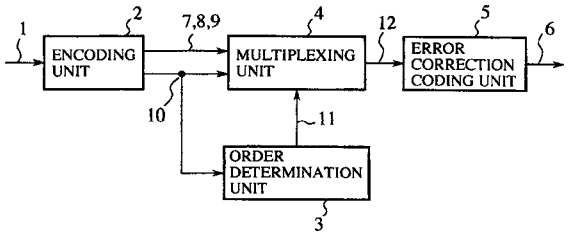
The goal is to provide a more powerful vibrator device that is both smaller physically and takes less energy compared to currently available devices. Static and dynamic magnetic fields are separated enough so that the dynamic field does not pass through the permanent magnets. Two permanent magnets work independently of each other such that the static field is confined to a portion of the magnetic path, but both the static and dynamic fields coincide in the air gaps.—DAP

7,315,871

43.72.Gy SOUND ENCODER AND SOUND DECODER

Hirohisa Tasaki, assignor to Mitsubishi Denki Kabushiki Kaisha
1 January 2008 (Class 704/229); filed in Japan 25 July 2001

Frame-to-frame bit error variations are taken into account in this error protection scheme for codes representing a sound signal. A sound encoder determines the order in which codes are to be multiplexed on a frame-by-frame basis and calculates and embeds an error correction code for the multiplexed code. The result is said to provide sound encoding having better immunity to bit errors.—DAP



7,318,035

43.72.Gy AUDIO CODING SYSTEMS AND METHODS USING SPECTRAL COMPONENT COUPLING AND SPECTRAL COMPONENT REGENERATION

Robert Loring Andersen *et al.*, assignors to Dolby Laboratories Licensing Corporation
8 January 2008 (Class 704/500); filed 8 May 2003

To reduce the amount of information transmitted, only a baseband portion of the original audio signal is encoded and a residual portion is discarded. Coupled channel signals having spectral components related to a composite of two or more audio input signals are formulated. Energy measures are obtained of the spectral components of the audio inputs, residual and coupled signals and are used to calculate coupling scale factors and scale factors. From this information, a synthesized signal is generated during decoding to substitute for the missing residual portion.—DAP

7,319,756

43.72.Gy AUDIO CODING

Arnoldus Werner Johannes Oomen and Leon Maria Van De Kerkhof, assignors to Koninklijke Philips Electronics N.V.
15 January 2008 (Class 380/236); filed in the European Patent Office 18 April 2001

An audio signal is scrambled by identifying and analyzing separately transient, sinusoidal, and noise components. At least one of the three components is encrypted prior to quantization of the audio signal, which results

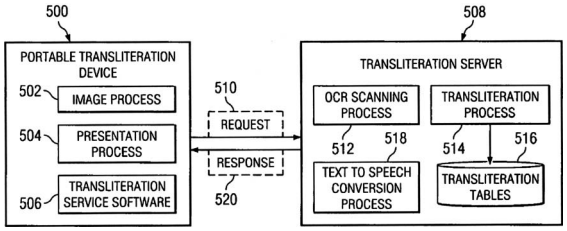
in a signal having lower quality than the original. Encryption occurs on parameters such as absolute frequency and differential amplitude between linked signal segments.—DAP

7,310,605

43.72.Ne METHOD AND APPARATUS TO TRANSLITERATE TEXT USING A PORTABLE DEVICE

Janani Janakiraman and David Bruce Kumhyr, assignors to International Business Machines Corporation
18 December 2007 (Class 704/277); filed 25 November 2003

A camera in a portable device captures an image of text to be transliterated. The image is transmitted via wireless link, with source and target languages identified, to an Internet transliteration service. The desired transliterated result returned contains a phonetic pronunciation of the text in the



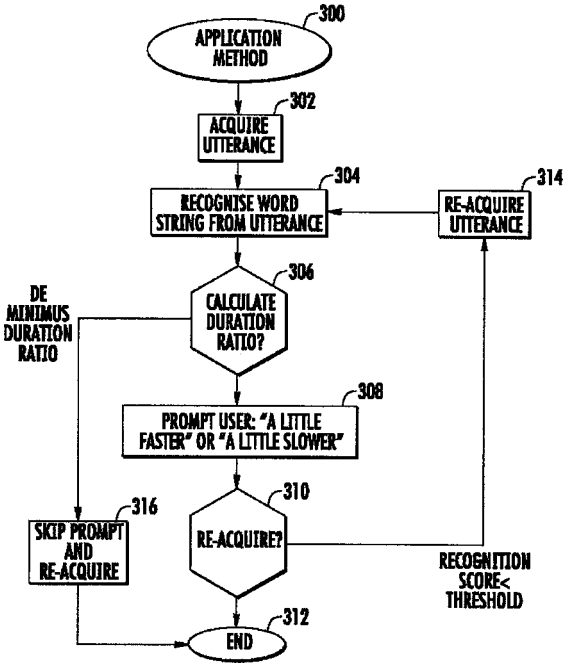
source language and is presented on the portable device in text form on a display or in audible form with a text-to-speech conversion.—DAP

7,318,029

43.72.Ne METHOD AND APPARATUS FOR AN INTERACTIVE VOICE RESPONSE SYSTEM

Wendy-Ann Coyle and Stephen James Haskey, assignors to International Business Machines Corporation
8 January 2008 (Class 704/231); filed in United Kingdom 24 October 2002

During speech recognition, an average of the ratios of actual delivery duration to the ideal delivery duration of the words in an utterance is calculated. Feedback prompting is then provided to the speaker to speed up or



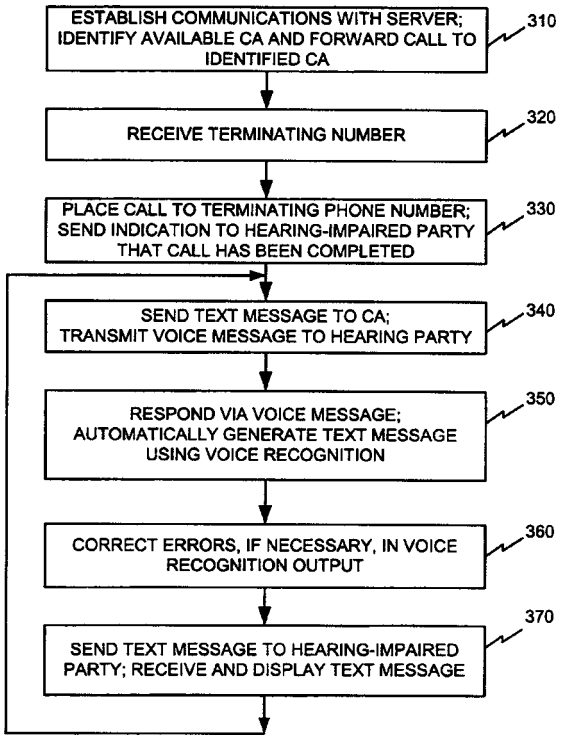
slow down the utterance if the duration is not ideal for the speech recognizer.—DAP

7,315,612

43.72.Ne SYSTEMS AND METHODS FOR FACILITATING COMMUNICATIONS INVOLVING HEARING-IMPAIRED PARTIES

William A. McClelland, assignor to Verizon Business Global LLC
1 January 2008 (Class 379/52); filed 2 November 2004

Voice recognition replaces the traditional TDD/TTY devices required for hearing impaired or deaf persons to communicate with hearing persons over telephone lines. The hearing party is identified based on telephone number, and a file of previously stored words is accessed. Text messages are generated from voice messages received from the hearing party via voice recognition using the stored word file and a voice link established by a



server forwarding the call to an identified communication assistant. The text messages corresponding to the voice messages are transmitted to the hearing impaired or deaf person.—DAP

7,317,787

43.72.Ne VOICE ENHANCING FOR ADVANCE INTELLIGENT NETWORK SERVICES

Susanne Marie Crockett *et al.*, assignors to AT&T Knowledge Ventures, L.P.
8 January 2008 (Class 379/88.03); filed 30 April 2002

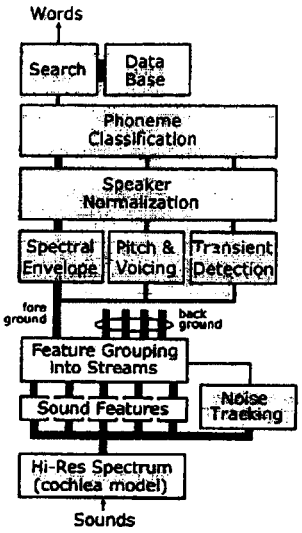
This patent relates to a method for subscribers to verbally alter, in real time, services such as call forwarding, call blocking and caller ID, for example when they are traveling. In response to voice announcements from an intelligent peripheral to a subscriber of parameters available for changing, voice instructions are received from the subscriber, recognized, and translated into digital commands that change stored parameters for that subscriber.—DAP

7,319,959

43.72.Ne MULTISOURCE PHONEME CLASSIFICATION FOR NOISE-ROBUST AUTOMATIC SPEECH RECOGNITION

Lloyd Watts, assignor to Audience, Incorporated
15 January 2008 (Class 704/254); filed 14 May 2003

For automatic voice dialing applications, 600 spectral values are separated into several streams that group sounds from the same source prior to analysis for phoneme classification. Thereafter, using a high resolution cochlear model, phoneme-level classification accuracy in noisy environments is said to improve by providing a spectral envelope, pitch and voicing information, normalized speaker characteristics, and detected transients as in-



puts to the classifier. For further accuracy improvement, the system is trained with a full phoneme target set in which syllable stress is taken into account.—DAP

7,313,260

43.80.Vj CONTROLLING THICK-SLICE VIEWING OF BREAST ULTRASOUND DATA

Shih-Ping Wang and Fangyi Rao, assignors to U-Systems, Incorporated
25 December 2007 (Class 382/128); filed 18 July 2006

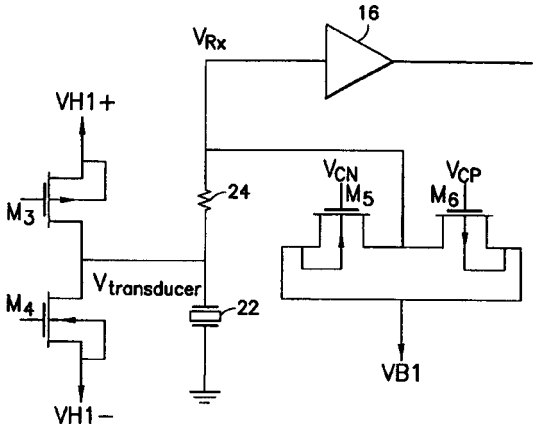
An ultrasound image representing a thick-slice or slab-like portion of breast volume essentially parallel to the standard x-ray mammographic view is registered with the x-ray image to improve the viewer's perception of breast structures being displayed in both images.—RCW

7,314,445

43.80.Vj INTEGRATED LOW-VOLTAGE TRANSMIT/RECEIVE SWITCH FOR ULTRASOUND IMAGING SYSTEM

Robert G. Wodnicki and Rayette A. Fisher, assignors to General Electric Company
1 January 2008 (Class 600/437); filed 30 December 2003

A low-voltage transmit-receive switch placed between the output of a high-power pulser and the input of a receiver preamplifier is comprised of a



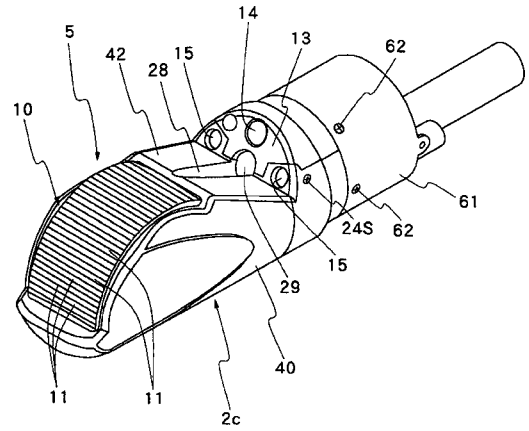
series resister and a parallel MOSFET pair.—RCW

7,318,806

43.80.Vj ULTRASOUND ENDOSCOPE

Shinichi Kohno, assignor to Fujinon Corporation
15 January 2008 (Class 600/463); filed in Japan 18 October 2002

A housing at the distal end of this endoscope can be split into a casing and head block to facilitate maintenance of the components. An ultrasound transducer is mounted on the front portion of the head block. An illuminator and an optical image pick-up are located in an inclined wall rising obliquely on the rear side of the ultrasound transducer. A passage for a biopsy needle



is located between the ultrasound transducer and the parts for visual observation.—RCW

LETTERS TO THE EDITOR

This Letters section is for publishing (a) brief acoustical research or applied acoustical reports, (b) comments on articles or letters previously published in this Journal, and (c) a reply by the article author to criticism by the Letter author in (b). Extensive reports should be submitted as articles, not in a letter series. Letters are peer-reviewed on the same basis as articles, but usually require less review time before acceptance. Letters cannot exceed four printed pages (approximately 3000–4000 words) including figures, tables, references, and a required abstract of about 100 words.

Introducing atmospheric attenuation within a diffusion model for room-acoustic predictions (L)

Alexis Billon
University of Liège, Sart-Tilman B28, B-4000 Liège, Belgium

Judicaël Picaut[a)]
Laboratoire Central des Ponts et Chaussées, Section Acoustique Routière et Urbaine, Route de Bouaye, B.P. 4129, 44341 Bouguenais Cedex, France

Cédric Foy
CEBTP-SOLEN, 12 Avenue Gay Lussac, ZAC La Clef Saint Pierre, 78990 Elancourt, France

Vincent Valeau
Université de Poitiers, LEA UMR CNRS 6609, 40 Avenue du Recteur Pineau, 86022 Poitiers Cedex, France

Anas Sakout
Université de La Rochelle, LEPTIAB, Avenue Michel Crépeau, 17042 La Rochelle Cedex 01, France

(Received 30 October 2007; revised 14 February 2008; accepted 10 March 2008)

This paper presents an extension of a diffusion model for room acoustics to handle the atmospheric attenuation. This phenomenon is critical at high frequencies and in large rooms to obtain correct acoustic predictions. An additional term is introduced in the diffusion equation as well as in the diffusion constant, in order to take the atmospheric attenuation into account. The modified diffusion model is then compared with the statistical theory and a cone-tracing software. Three typical room-acoustic configurations are investigated: a proportionate room, a long room and a flat room. The modified diffusion model agrees well with the statistical theory (when applicable, as in proportionate rooms) and with the cone-tracing software, both in terms of sound pressure levels and reverberation times. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2903872]

PACS number(s): 43.55.Br, 43.55.Ka [NX] Pages: 4040–4043

I. INTRODUCTION

Recently, Valeau *et al.*[1] have proposed a generalization and a numerical implementation of a so-called diffusion model,[2,3] first proposed by Ollendorf,[4] to describe the reverberant sound field in enclosures. In room acoustics, this model has been applied successfully to single-space enclosures,[1] coupled-rooms systems,[5] fitted rooms,[6] both for low[1] and high[7–9] wall absorption coefficients. However, in this model, the atmospheric attenuation is not taken into account, although this phenomenon can be significant at high frequencies, particularly for large enclosures.

In the present paper, a modification of the diffusion model is proposed to deal with the atmospheric attenuation (Sec. II). In Sec. III, predicted reverberation times (RTs) and sound pressure levels (SPLs) obtained with the diffusion model are compared with the statistical theory and a cone-tracing software for a quasi-cubic room, for various atmospheric attenuations. For long and flat rooms, the diffusion model is compared to the cone-tracing software only. Section IV concludes the paper.

II. MODIFIED ROOM ACOUSTIC DIFFUSION MODEL

A. Diffusion model

The diffusion model is based on the concept of sound particles traveling at a velocity c along straight lines and striking walls or scattering objects, in a room of volume V and surface S. The room-acoustic diffusion model, proposed first by Ollendorff,[4] and further developed by Picaut *et al.*[2] for modeling the acoustics of enclosed spaces, is based on an analogy with the diffusion of particles in a medium containing scattering obstacles, as presented by Morse and Feshbach.[10] Following the same approach, the atmospheric

[a)]Author to whom correspondence should be addressed. Electronic mail: judicael.picaut@lcpc.fr

 0001-4966/2008/123(6)/4040/4/$23.00

attenuation term can be introduced in the diffusion model, leading to a modified expression of the diffusion constant:

$$D' = D \times \frac{1}{1 + m\lambda}, \tag{1}$$

where m is the coefficient of atmospheric attenuation (in m^{-1}). D is the diffusion constant without atmospheric attenuation,

$$D = \frac{\lambda c}{3}, \tag{2}$$

where $\lambda = 4V/S$ is the classical mean free path for a perfectly diffuse room. This last relation shows that the diffusion of the sound energy density in the room is modified by the atmospheric attenuation. However, the sound absorption being usually very small over a mean free path (i.e., $m\lambda \ll 1$), one can consider thereafter that $D' \approx D$. Similarly, by introducing the atmospheric attenuation term in the derivation of the diffusion model, the diffusion equation for the energy density w in the room, with a sound source term $P(\mathbf{r},t)$, becomes:

$$\frac{\partial}{\partial t} w(\mathbf{r},t) - D'\nabla^2 w(\mathbf{r},t) + mcw(\mathbf{r},t) = P(\mathbf{r},t) \quad \text{in } V. \tag{3}$$

In this equation, the attenuation term has the typical form of an absorption term within the diffusion framework. In order to take into account the sound absorption occurring at the room surface, a mixed boundary condition must also be introduced:[1]

$$-D' \frac{\partial w(\mathbf{r},t)}{\partial \mathbf{n}} = hw(\mathbf{r},t) \quad \text{on } S, \tag{4}$$

where h is the local exchange coefficient, $\mathbf{n}$ the outgoing normal vector, and $\partial/\partial\mathbf{n}$ the normal derivative. The exchange coefficient is expressed by using the absorption coefficient α of the wall. Several expressions have been derived.[1,3,7–9] In this paper, the Eyring's expression of the exchange coefficient is considered:

$$h = -\frac{c \ln(1-\alpha)}{4}. \tag{5}$$

B. Room energy balance

The energy balance in the room can be obtained by integrating Eq. (3) over the volume V (with $D' \simeq D$):

$$\int_V \frac{\partial w(\mathbf{r},t)}{\partial t} \mathrm{d}V_r - \int_V D\nabla^2 w(\mathbf{r},t) \mathrm{d}V_r + \int_V mcw(\mathbf{r},t) \mathrm{d}V_r = \int_V P(\mathbf{r},t) \mathrm{d}V_r. \tag{6}$$

Using the Gauss' theorem and the boundary condition of Eq. (4), it follows:

$$\int_V \frac{\partial w(\mathbf{r},t)}{\partial t} \mathrm{d}V_r + \int_S hw(\mathbf{r},t) dS + \int_V mcw(\mathbf{r},t) \mathrm{d}V_r = P(t). \tag{7}$$

Last, by considering the case of a diffuse sound field in the room [i.e., $w(\mathbf{r},t) = w(t)$ is supposed to be uniform] with uniform absorption coefficient at walls, Eq. (7) gives the well-known energy balance of the statistical theory:[11]

$$V\frac{\mathrm{d}w(t)}{\mathrm{d}t} - \frac{\ln(1-\alpha)cS}{4} w(t) + Vmcw(t) = P(t), \tag{8}$$

The diffusion model can then be seen as an extension of statistical theory to spatially varying reverberant sound fields.

III. NUMERICAL VALIDATION

Three typical geometrical configurations met in room acoustics have been investigated, considering varying atmospheric attenuation. The wall absorption coefficient is homogeneous within the rooms. For the room with proportionate dimensions, the diffusion model is compared to the statistical theory and a cone-tracing software (CATT-Acoustics). For flat and long rooms, the sound field is known to be nondiffuse and the diffusion model is compared to a cone-tracing software only. The diffusion model is implemented in a finite-elements software with 3000 mesh elements for the three rooms. Equation (3) along with Eq. (4) are solved with the following initial conditions:

$$w(\mathbf{r},0) = 0 \quad \text{in } V, \tag{9}$$

$$w(\mathbf{r},0) = \frac{W}{V_s} \quad \text{in } V_s, \tag{10}$$

in order to obtain the reverberation time (RT), with V_s and W the volume and the sound power of the source, respectively. The decay of sound pressure level at location $\mathbf{r}$ can be expressed as:[12]

$$\mathrm{SPL}(\mathbf{r},t) = 10 \log\left[\frac{\rho c^2}{P_{\mathrm{ref}}^2} w(\mathbf{r},t)\right], \tag{11}$$

where P_{ref} is equal to 2×10^{-5} Pa. The reverberation times are then estimated over the decay range from -5 to -35 dB.[13]

To calculate the steady sound field, Eq. (3) along with Eq. (4) are solved for a source with a sound power level of 100 dB. The sound pressure level including the direct sound field can be expressed as:

$$\mathrm{SPL}_{\mathrm{tot}}(\mathbf{r}) = 10 \log\left[\frac{\rho c}{P_{\mathrm{ref}}^2}\left(\frac{W}{4\pi r^2} \exp(-mr) + cw(\mathbf{r})\right)\right], \tag{12}$$

where r is the distance from the source center ($r = |\mathbf{r}|$), and $w(\mathbf{r})$ the stationary solution of the diffusion equation.

For the cone-tracing software, the number of rays and the ray truncation time are chosen as 5×10^5 and 150% of the reverberation time, respectively. Since the diffusion

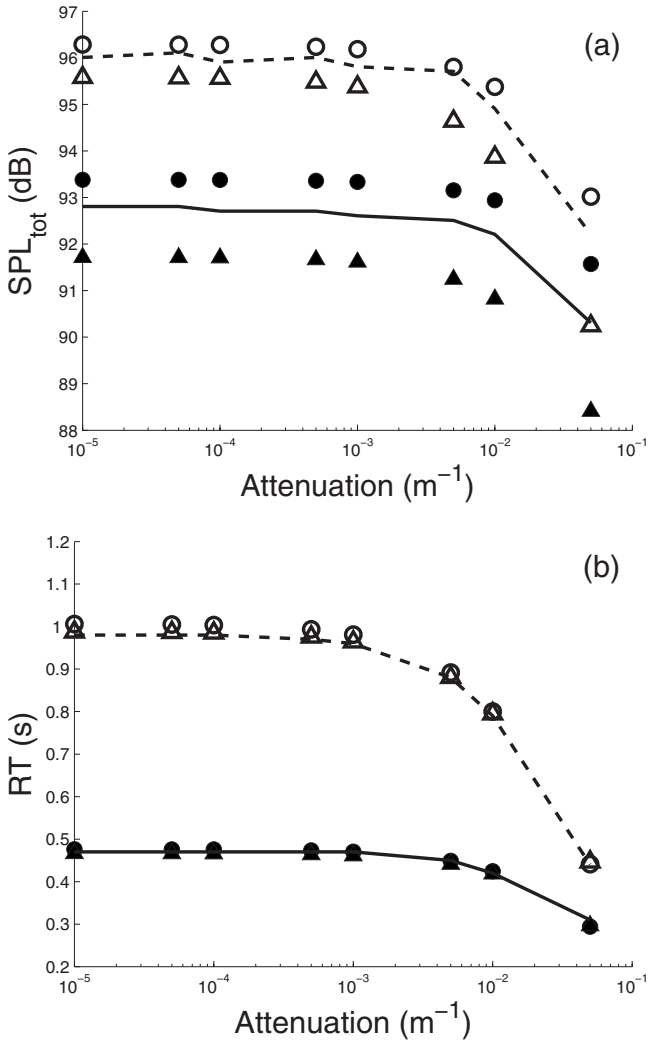


FIG. 1. (a) Sound pressure level and (b) reverberation time at (1,1,1) m as a function of the atmospheric attenuation, in a proportionate room of size $(5\times4\times3)$ m^3: (○) and (●) diffusion model, (– –) and (—) cone-tracing software, (△) and (▲) statistical theory, respectively, for α=0.1 and 0.2.

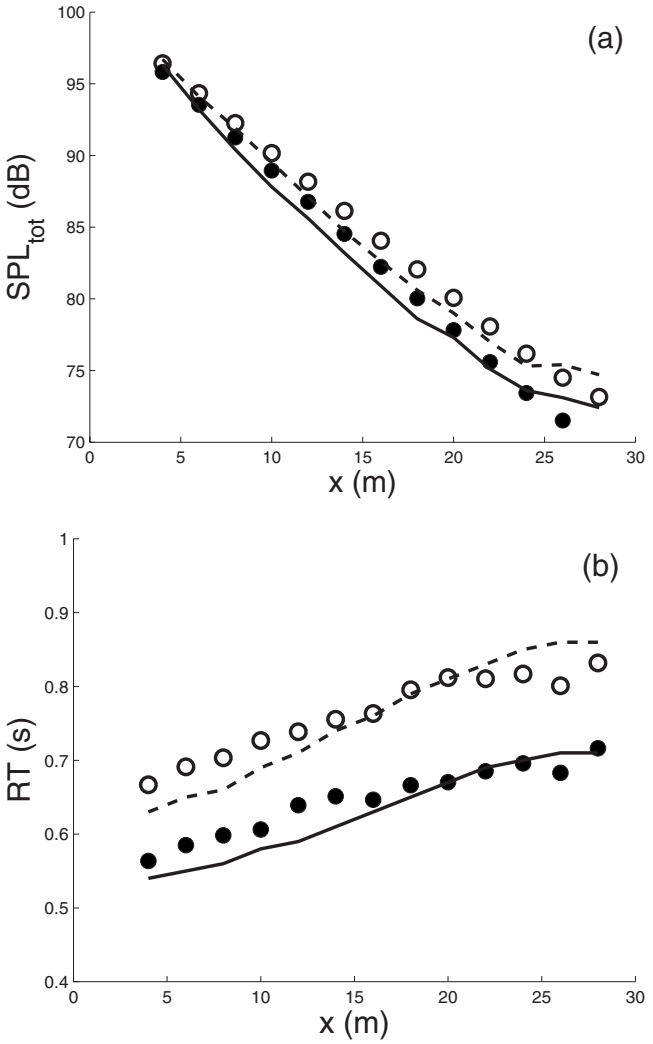


FIG. 2. (a) Sound pressure level and (b) reverberation time as function of the receiver location in a center axis of long room of size $(30\times2\times2)$ m^3: (○) and (●) diffusion model, (– –) and (—) cone-tracing software, without and with atmospheric attenuation, respectively.

model assumes that the room is perfectly diffuse, the wall reflections are set to be also completely diffusive in the cone-tracing software.

A. Proportionate room

Let us consider a proportionate room of size $(5\times4\times3)$ m^3. The sound source is located in the middle of the room at coordinate (2.5,2,1.5). The wall absorption coefficient is uniform and its value is equal to 0.1 and 0.2. The atmospheric attenuation m is varied from 1×10^{-5} to 0.05 m^{-1}, corresponding to a [50, 16000] Hz acoustic wave propagating in an air at 20 °C and 50% of relative humidity.[14]

For diffusion model and the cone-tracing software, the SPL and the RT are calculated at point (1,1,1). For the sound pressure level, as well as for the reverberation time (Fig. 1), the three models are in very good agreement for both absorption coefficient values. The diffusion model agrees better with the cone-tracing software; the maximal discrepancy for the SPL is then less than 1 dB and 0.02 s for the RT. It can be noted that the diffusion model does not give exactly the same results as the statistical model. According to Sec. II B, both models are in agreement only if the sound energy density is perfectly uniform in the room. However, for the diffusion model, even for a cubic room, the energy density is not perfectly uniform, mainly close to the sound source. It explains why the diffusion model and the statistical theory give a slightly different result in Fig. 1.

B. Long room

A sound source is now located at (2,1,1) in a long room of size $(30\times2\times2)$ m^3. The wall absorption coefficient is uniform and equal to 0.1. The atmospheric attenuation m is equal to 0 (no atmospheric attenuation) and 0.01 m^{-1} corresponding to a 7.2 kHz acoustic wave propagating in the air at 20 °C and with 50% of relative humidity.

The SPL and the RT are evaluated every 2 m along a line following the length of the room and passing through the sound source (Fig. 2). The diffusion model is in good agreement with the cone-tracing software both with and without atmospheric attenuation. In terms of sound level, the maximal discrepancy is less than 2 dB. Moreover, for the sound

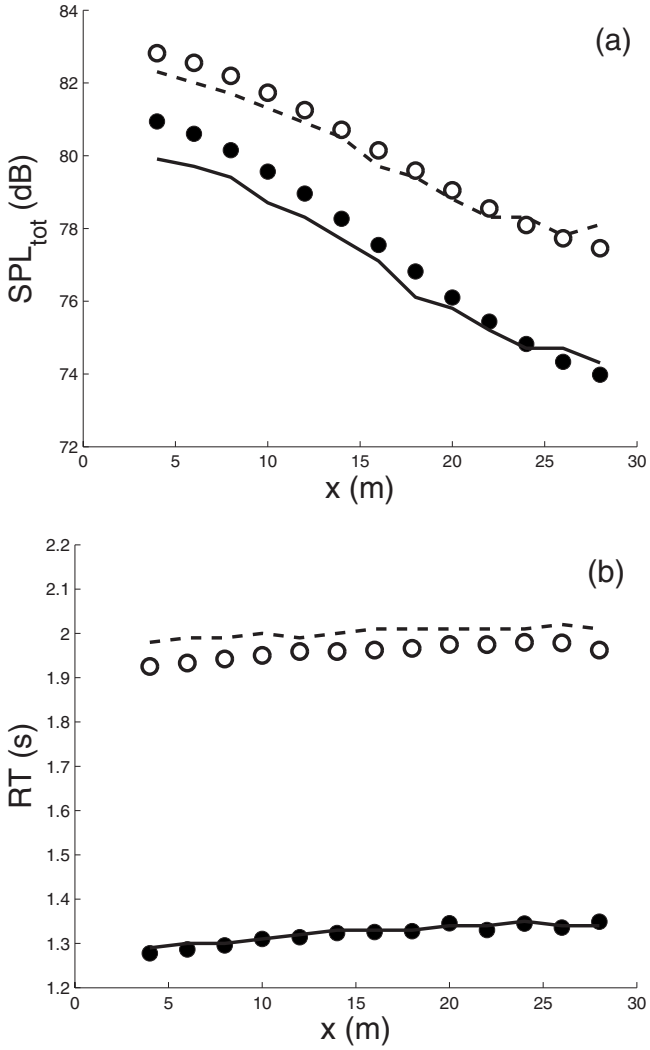


FIG. 3. (a) Sound pressure level and (b) reverberation time as function of the receiver location in a flat room of size $(30\times30\times3)$ m^3: (○) and (●) diffusion model, (– –) and (—) cone-tracing software, without and with atmospheric attenuation, respectively.

decay, the diffusion model shows the typical increase of reverberation time with the source-receiver distance[15] and the error relative to the cone-tracing software is less than 10% at any receiver location. In this configuration, the atmospheric attenuation has an effect mainly on the sound decay.

C. Flat room

A sound source is located at (2,2,1) in a flat room of size $(30\times30\times3)$ m^3. Again, the wall absorption coefficient is uniform and equal to 0.1. The atmospheric attenuation is set to 0 and 0.01 m^{-1}.

The SPL and the RT are measured every 2 m along a line passing through the room center at 1 m high (Fig. 3). The agreement between the diffusion model and the cone-tracing software is here again very good. In terms of sound level, the maximum discrepancy is equal to 1.2 dB. For the reverberation time, the error relative to the cone-tracing software is inferior to 5% at any receiver location. In this configuration, the atmospheric attenuation has an effect both on the spatial attenuation and on the temporal sound decay of the sound energy.

IV. CONCLUSION

This work introduces a modification into the room-acoustic diffusion model to account for the atmospheric attenuation, whose effect is particularly important for acoustical predictions at high frequencies and in large rooms. An additional absorption term is thus introduced within the diffusion equation, as well as in the expression of the diffusion constant. The room energy balance shows that the modified diffusion model is consistent with the statistical theory. To validate this additional absorption term in the diffusion equation, three typical room-acoustic geometries are investigated. In the proportionate room, the atmospheric attenuation is varied over a wide range of values: the modified diffusion model shows a very good agreement with the statistical theory and a cone-tracing software both in terms of sound pressure and reverberation time. For the flat and long rooms, the modified diffusion model agrees very well also with the cone-tracing software, for the spatial variations of the sound pressure level, as well as for the temporal sound decay.

ACKNOWLEDGMENT

The authors wish to thank the Agence de l'Environnement et de la Maîtrise de l'Énergie (ADEME) for providing financial support of this work.

[1] V. Valeau, J. Picaut, and M. Hodgson, "On the use of a diffusion equation for room-acoustic prediction," J. Acoust. Soc. Am. **119**, 1504–1513 (2006).
[2] 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).
[3] J. Picaut, L. Simon, and J.-D. Polack, "Sound field in long rooms with diffusely reflecting boundaries," Appl. Acoust. **56**, 217–240 (1999).
[4] F. Ollendorff, "Statistical room acoustics as a problem of diffusion—a proposal," Acustica **21**, 236–245 (1969), in German.
[5] A. Billon, V. Valeau, A. Sakout, and J. Picaut, "On the use of a diffusion model for acoustically coupled rooms," J. Acoust. Soc. Am. **120**, 2043–2054 (2006).
[6] V. Valeau, M. Hodgson, and J. Picaut, "A diffusion-based analogy for the prediction of sound in fitted rooms," Acust. Acta Acust. **93**, 94–105 (2007).
[7] Y. Jing and N. Xiang, "A modified diffusion equation for room-acoustic predication," J. Acoust. Soc. Am. **121**, 3284–3287 (2007).
[8] A. Billon, J. Picaut, and A. Sakout, "Prediction of the reverberation time in high absorbent room using a modified-diffusion model," Appl. Acoust. **69**, 68–74 (2008).
[9] Y. Jing and N. Xiang, "On boundary conditions for the diffusion equation in room-acoustic prediction: Theory, simulations, and experiments," J. Acoust. Soc. Am. **123**, 145–153 (2008).
[10] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw–Hill, New York, 1953).
[11] H. Kuttruff, *Room Acoustics*, 3rd ed. (Applied Science, London, 1991).
[12] A. D. Pierce, *Acoustics: An Introduction to its Physical Principles and Applications* (Acoustical Society of America, New York, 1981), Vol. **1**.
[13] ISO 3382, "Acoustics—Measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization (1997).
[14] H. Bass, H.-J. Bauer, and L. Evans, "Atmospheric absorption of sound: Analytical expressions," J. Acoust. Soc. Am. **52**, 821–825 (1972).
[15] J. Kang, *Acoustics of Long Spaces: Theory Design and Practice* (Thomas Telford Ltd., London, 2002).

A method of measuring the Green's function in an enclosure (L)

Yu Luan[a] and Finn Jacobsen
Acoustic Technology, Department of Electrical Engineering, Technical University of Denmark, Building 352, Ørsteds Plads, DK-2800 Kgs. Lyngby, Denmark

(Received 10 January 2008; revised 8 April 2008; accepted 8 April 2008)

The acoustic Green's function can be measured using a device with two matched microphones mounted in a tube driven by a loudspeaker combined with another microphone that represents the observation point. Good agreement is obtained between the measured and theoretical Green's function in a rectangular room below 320 Hz. At higher frequencies the agreement is less good because of the imperfect geometry of the room.
 [DOI: 10.1121/1.2917804]

PACS number(s): 43.55.Mc, 43.20.Rz, 43.20.Ks, 43.58.Fm [AJZ] Pages: 4044–4046

I. INTRODUCTION

There is a growing need, e.g., in the automotive industry for experimental determination of acoustic transfer functions in connection with analysis of complicated sources. The transfer function of concern, henceforth called the Green's function, is the complex ratio of the sound pressure at a given position to the volume acceleration of a monopole at another position. The purpose of this letter is to examine a simple technique for experimental determination of the Green's function based on pressure microphones mounted in tube driven by a loudspeaker. The technology is commercially available and has been validated in an anechoic room.[1] In this letter the performance of the device is evaluated in a lightly damped room.

II. OUTLINE OF THEORY

The acoustic Green's function is the solution to the inhomogeneous wave equation with a mass injection source term and given boundary conditions,[2]

$$G(\mathbf{r},\mathbf{r}_0)=\frac{p(\mathbf{r})}{j\omega\rho Q}, \tag{1}$$

where $p(\mathbf{r})$ is the sound pressure at $\mathbf{r}$, Q is the volume velocity of a harmonic monopole at $\mathbf{r}_0$, ω is the radian frequency, and ρ is the density of air. Experimental determination of this quantity obviously involves measuring the frequency response between the volume velocity (or acceleration) of a small source and the signal from a pressure microphone. However, ordinary loudspeakers do not provide a signal proportional to their volume velocity.

Figure 1 shows an arrangement for determining the volume velocity at the opening of the tube, Q, driven by a loudspeaker at the other end. The sound pressure signals p_A and p_B are measured at a distance of $l+\Delta l$ and l from the opening with matched quarter-inch microphones. The tube section has a diameter of 4 cm, from which it follows that the nonaxisymmetric (1, 0) and (2, 0) modes can propagate above 5.0 and 8.3 kHz, respectively.[2] However, since the microphones measure the sound pressure on the axis of the tube these modes, which do not contribute to the volume velocity at the opening, are not detected. The cut-on frequency of the first axisymmetric mode is 10.5 kHz.

In the frequency range where it can be assumed that only plane waves are measured, the entire sound field can be determined if the sound field is sampled at two positions (unless they are spaced a multiple of half a wavelength),[3] and the volume velocity at the opening of the tube can easily be shown to be

$$Q=\frac{S}{\rho c}\cdot\frac{p_A\cos kl-p_B\cos[k(l+\Delta l)]}{j\sin k\Delta l}, \tag{2}$$

where S is the cross-sectional area of the tube, c is the speed of sound, and k is the wave number.[1] It now follows that the Green's function is

$$G(\mathbf{r},\mathbf{r}_0)=\frac{p_C}{j\omega\rho Q}=\frac{Q^*p_C}{j\omega\rho|Q|^2} \tag{3}$$

[where $p_C=p(\mathbf{r})$]. Expressed in terms of frequency responses between the three pressure signals the Green's function becomes

$$G(\mathbf{r},\mathbf{r}_0)=\frac{\sin k\Delta l}{kS}\times\frac{H_{AC}\cos kl-|H_{AB}|^2H_{BC}\cos[k(l+\Delta l)]}{\cos^2kl-2\,\mathrm{Re}\{H_{AB}\}\cos kl\cos[k(l+\Delta l)]+|H_{AB}|^2\cos^2[k(l+\Delta l)]}, \tag{4}$$

where the frequency responses H_{AB}, H_{AC}, and H_{BC} are estimated in the usual manner from cross and auto spectra,[4] e.g., the frequency response between microphone signals A and B is the ratio of the cross spectrum S_{AB} to the auto spectrum S_{AA},

$$H_{AB}=\frac{S_{AB}}{S_{AA}}. \tag{5}$$

Equation (4) can be simplified if the three signals can be assumed to be perfectly coherent. However, in measurements in enclosures a poor signal-to-noise ratio of the signal from microphone C may occur at antiresonance frequencies.

[a]Author to whom correspondence should be addressed; electronic mail: yl@oersted.dtu.dk.

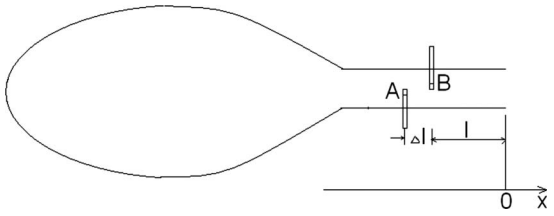


FIG. 1. Device for measuring the output volume velocity.

III. EXPERIMENTAL RESULTS

To test the method with a complicated Green's function a simple experiment has been carried out in a lightly damped rectangular room with dimensions $3.29 \times 4.38 \times 3.29$ m and a reverberation time of 5 s in the frequency range of concern, corresponding to a Schroeder frequency[5] of 650 Hz. The tube, and thus the room, was driven by a Brüel & Kjær "OmniSource" (B&K 4295), a loudspeaker mounted in an "inverted horn" of hard plastic, and radiating through a small opening.[6] The volume of the inverted horn is about 2.3 l. On the source a "Volume velocity adaptor" of type B&K 4299 was mounted. This device is a 10-cm-long tube of hard plastic that fits the OmniSource and makes it possible to measure the sound pressure at two positions in the tube with a matched set of B&K 4178 (1/4 in.) microphones (microphones A and B in the foregoing). The dimensions l and Δl are 3 and 2 cm, respectively. Finally, the sound pressure at the observation point was measured with a 1/2 in. microphone of type B&K 4192 (microphone C in the foregoing). The three frequency responses were measured using a B&K "PULSE" analyzer using a frequency span of 200 Hz, center frequencies between 160 and 600 Hz, and 3200 spectral lines corresponding to a resolution of 62.5 mHz. The PULSE analyzer was also used for measuring the reverberation time of the room using the conventional interrupted noise method. No correction for possible phase and amplitude mismatch between the matched microphones was attempted, but the agreement between their responses was ascertained by measuring one time with the microphones interchanged.

In what follows measured Green's functions are compared with the theoretical function for an enclosure[2,5]:

$$G(\mathbf{r},\mathbf{r}_0) = -\frac{1}{V}\sum_{m=0}^{\infty}\frac{\psi_m(\mathbf{r})\psi_m(\mathbf{r}_0)}{k^2 - k_m^2 - jk/(\tau_m c)}, \tag{6}$$

where V is the volume of the room, ψ_m is the mode shape (a simple product of three cosines in a rectangular room), k_m is the wave number corresponding to the natural frequency of the mth mode, and τ_m is the time constant of the mode. The corresponding reverberation time is 13.8 times larger.[5]

Figure 2 shows a typical example of a measured and predicted Green's function in the frequency range between 60 and 260 Hz for a source position of (1.65, 2.19, 0.01) m and a receiver position of (1.65, 2.19, 1.04) m. As can be seen the agreement is generally quite good, although deviations occur around 160 Hz. It can also be seen that some peaks are underestimated and some troughs are overestimated; and some peaks and troughs are shifted a little. The underestimation of a few peaks and overestimation of a few troughs are probably due to the fact that the broadband reverberation time has been measured in the frequency range of concern, whereas the actual modal time constants of different modes may vary. (To determine the individual time constants would require measuring, say, 3 dB bandwidths of all the modes; and this is only possible when the modes are well separated.) The small frequency shifts that occur, e.g., at 175 Hz are in all probability due to the fact that the room was not perfectly rectangular.

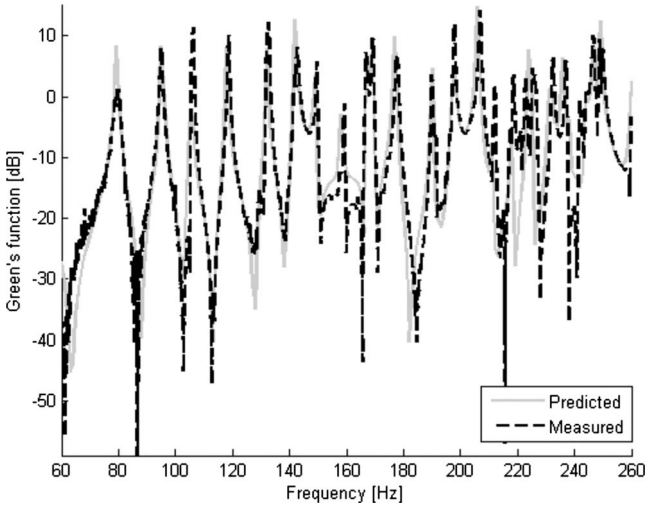


FIG. 2. Measured and predicted Green's function in a rectangular enclosure of 47 m^3.

Figure 3 shows an example of a measured and predicted Green's function in the frequency range between 260 and 460 Hz. The agreement is clearly less good, in particular between 320 and 350 Hz. This is undoubtedly due to the irregular geometry of the room (the floor of the room was, in fact, neither completely horizontal nor completely flat). Above 460 Hz the agreement (not shown) is in general no longer acceptable, but there is no reason to expect this to be due to a failure of the experimental arrangement.

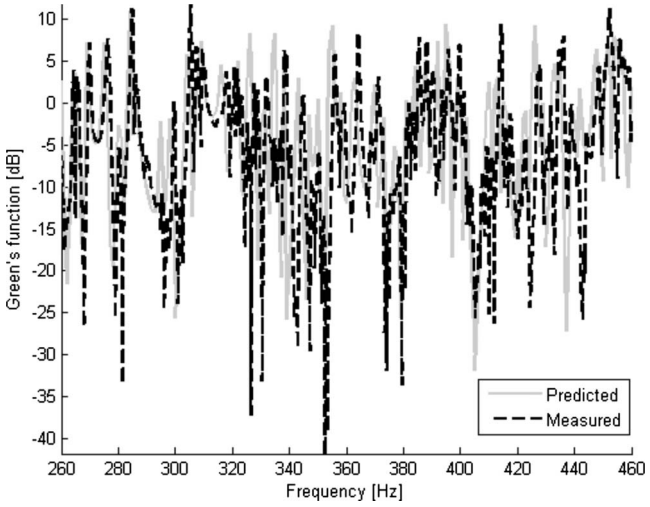


FIG. 3. Measured and predicted Green's function in a rectangular enclosure of 47 m^3.

IV. CONCLUSIONS

A method of measuring the Green's function has been examined in a rectangular enclosure and found to give results in good agreement with theoretical predictions in the frequency range below 320 Hz. The poor agreement at higher frequencies is due to deviations between the assumed perfectly rectangular geometry and the actual, slightly irregular geometry of the room.

[1]S. Gade, N. Møller, J. Hald, and L. Alkestrup, "The use of volume velocity source in tranfer measurements," in *Proceedings of Inter-Noise 2004*, Prague, Czech Republic, 2004.
[2]P. M. Morse and K. U Ingard, *Theoretical Acoustics* (McGraw–Hill, New York, 1968).
[3]J. Y. Chung and D. A. Blaser, "Transfer function method of measuring in-duct acoustic properties. I. Theory," J. Acoust. Soc. Am. **68**, 907–913 (1980).
[4]J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Processing*, 2nd ed. (Wiley, New York, 1993).
[5]H. Kuttruff, *Room Acoustics*, 4th ed. (E & FN Spon, London, 2000).
[6]J.-D. Polack, L. S. Christensen, and P. M. Juhl, "An innovative design for omnidirectional sound sources," Acta. Acust. Acust. **87**, 505–512 (2001).

Comment on "Three-dimensional finite element modeling of guided ultrasound wave propagation in intact and healing long bones," [J. Acoust. Soc. Am. 121(6), 3907–3921 (2007)]

Xiasheng Guo, Dong Zhang, Di Yang, and Xiufen Gong
Institute of Acoustics, Key Laboratory of Modern Acoustics (Nanjing University), Ministry of Education, Nanjing University, Nanjing 210093, China

Junru Wu[a)]
Department of Physics, The University of Vermont, Burlington, Vermont 05405, USA

(Received 27 December 2007; revised 11 March 2008; accepted 13 March 2008)

Protopappas *et al.* performed finite element (FE) studies on the propagation of guided ultrasound waves in intact and healing long bones, and found that the dispersion of guided modes was significantly influenced by the irregularity and anisotropy of the bone. A time-frequency (t-f) method was applied to the obtained signals and several wave modes were identified. However, this technique was unable to quantify their observations and provide monitoring capabilities. One possible reason of this shortcoming may come from the inherent disadvantage of the t-f method. The objective of this comment is to demonstrate that it is necessary to combine other techniques with FE simulations for the extraction of significant quantitative ultrasonic features. Individual guided modes in an isotropic pipe have been theoretically examined using the normal mode expansion (NME) method, and many modes that are missed by the t-f analysis have been identified. It is concluded that in order to extract quantitative ultrasonic features, FE simulations should be supplemented by other techniques such as the NME. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2904929]

PACS number(s): 43.80.Ev, 43.80.Qf, 43.80.Jz [CCC] Pages: 4047–4050

I. INTRODUCTION

Guided ultrasound (US) waves have recently been proposed to evaluate the mechanical properties of long bones.[1,2] Since the guided US wave can propagate across the bone thickness, its propagation dispersion may provide useful information related to the mechanical property and architecture of bone. Therefore, measurement of the dispersion relation of guided waves may be used to monitor the health of a bone. Experimentally, a broadband signal is usually used to excite several modes of guided waves in long bones. Some signal processing methods, such as the two-dimensional fast Fourier transform (2D-FFT) method,[3] the time-frequency (t-f) method[1,4] and spectrum estimation methods,[5] have been utilized to analyze the obtained broadband multi-mode signals.

In a paper by Protopappas *et al.* [J. Acoust. Soc. Am. 121(6), 3907–3921 (2007)][4] the propagation of guided ultrasound waves in intact and healing bones was modeled by using a finite element (FE) method. A hollow circular cylinder model and a more realistic model of a pipe incorporating the callus tissue with its cross-section dimensions extracted from images obtained from computed tomography scans were studied. The authors applied a t-f technique, the so-called reassigned smoothed-pseudo Wigner–Ville (RSPWV) energy distribution, to analyze signals obtained from the simulation. The calculated amplitudes (in dB) of the energy distribution of different modes were used in t-f representation[1,3] to examine the mode-resolving capability of the t-f method. In the t-f representation, the color of a point represented the amplitude of the energy distribution. In the case of an intact isotropic hollow circular cylinder [Fig. 10(b) in their paper], L(0,5) mode in the frequency range of 0.55–0.8 MHz, L(0,4) mode in 0.7–0.85 MHz, L(0,8) mode at its cutoff frequency (1.05 MHz) and the fundamental L(0,1) and F(1,1) modes in 0.05–0.15 MHz, were identified. Other guided wave modes could not be clearly identified from the t-f diagram.

As indicated by Protopappas *et al.*, RSPWV was unable to give sufficient ultrasonic features to monitor the bone healing process.[4] In our opinion, failure of this technique in this regard may be due to the inherent disadvantage of the t-f method itself. Although t-f analysis is capable of representing the dispersion of individual wave modes, it is unable to discern modes when dispersion relations of several modes are very similar or even overlapping. The main problem for not providing sufficient ultrasonic features was that with t-f analysis it is generally difficult to quantify the dispersion characteristics of a mode. Mode quantification in the t-f domain requires image analysis and mode-tracking techniques to be applied to the t-f image. On the contrary, the nature of the NME technique makes it possible to recognize each individual guided mode separately. Thus, the combination of normal mode expansion (NME) with FE simulations may make t-f analysis more useful in practice.

In this study, we calculated the time domain signals in terms of the longitudinal and the first-order flexural modes for an isotropic circular cylinder by using the NME method.

[a)]Author to whom correspondence should be addressed. Electronic mail: jun-ru.wu@uvm.edu.

The excitation was a three-cycle Gaussian modulated 1 MHz sinusoidal wave with a −6 dB bandwidth of 0.55 MHz. The energy distribution spectra of the calculated modes were then obtained in using two different manners to examine the mode recognition capabilities of the t-f method. The objective of this comment is to demonstrate that in t-f analysis FE simulation alone is not sufficient to give system complete and correct ultrasonic features. One should either apply image analysis and mode-tracking techniques to the t-f images or combine the FE simulations with other techniques, such as the NME.

II. METHODS

For a hollow circular cylinder, the time-dependent displacement field can be expressed in terms of the generalized Fourier series as[6]

$$\mathbf{u}(\mathbf{r},t) = \sum_m \xi_m(t)\mathbf{u}_m(\mathbf{r},\omega_m), \tag{1}$$

where $\xi_m(t)$ is the mth mode generalized Fourier coefficient and $u_m(r,\omega_m)$ is the eigenfunction of the mth mode displacement.

In our calculation, **s** was a spatially uniform force applied perpendicularly to the outer surface of the pipe within the range of a 5-mm-diam circular zone. Temporally, **s** was a three-cycle Gaussian-modulated 1 MHz sinusoidal wave, with a −6 dB bandwidth of 0.55 MHz; the excitation was exactly the same as that used by Protopappas *et al.* Since there was no body force, the generalized Fourier coefficients $\xi_m(t)$ can be determined by[6]

$$\xi_m(t) = \omega_m^{-1}\int_0^t d\tau\left[\sin\omega_m(t-\tau)\cdot\int_S \mathbf{s}\cdot\mathbf{u}_m(\mathbf{r},\omega_m)dS\right], \tag{2}$$

where S represented the outer surface of the pipe. The eigenfunction of normal modes $\mathbf{u}_m(\mathbf{r},\omega_m)$ was determined by the free boundary conditions and the Navier equation,

$$(\lambda+2\mu)\nabla(\nabla\cdot\mathbf{u}_m)+\mu\nabla\times(\nabla\times\mathbf{u}_m)=-\rho\omega_m^2\mathbf{u}_m, \tag{3}$$

where λ and μ are the Lame constants which could be deduced from the Young's modulus and the Poisson ratio, ρ is the density, and ω_m is the angular frequency.

The same parameters of bone used by Protopappas *et al.* were adopted in simulation, i.e., the inner and outer diameters were 4.53 and 8.61 mm, respectively; the Young's modulus 14 GPa; the Poisson's ratio 0.37, and the density 1500 $\mathrm{kg/m^3}$.

The normal displacement waveforms of all the longitudinal modes and first-order flexural modes that exist in the frequency range 0–1.5 MHz were calculated at a distance of 3.6 cm from the source, where a receiver was located. A time step of 0.05 μs, corresponding to a sampling frequency of 20 MHz, was adopted for the calculation of a 50 μs time period.

As for the t-f method, the method (the RSPWV distribution) developed by Protopappas *et al.* using the frequency- and time-smoothing $W/10$ point Hamming windows was followed, where W denotes the number of points of the signals.

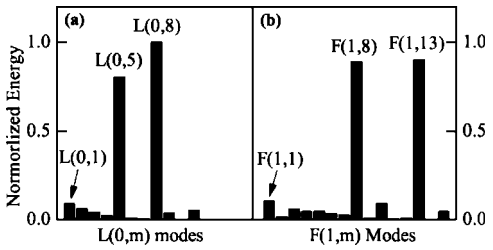


FIG. 1. Normalized energy of (a) longitudinal modes and (b) first-order flexural modes.

III. RESULTS

The energy distribution was first calculated for different excited modes. As shown in Fig. 1 the normalized energy is presented for 13 longitudinal modes [Fig. 1(a)] and 15 first-order flexural modes [Fig. 1(b)] that might be excited. As was reported by Protopappas *et al.*, the L(0,5) and L(0,8) modes had greater energy than other longitudinal modes. Data shown in Fig. 1(a) confirmed their results, but nontrivial energy spectra for additional modes as shown in Fig. 1(b) also existed. Particularly, the energy spectra for the F(1,8) and F(1,13) modes appeared to be comparable with those of L(0,5) and L(0,8). It is noted that the energy spectra of the fundamental modes L(0,1) and F(1,1) did not appear to be as strong as those described by Protopappas *et al.*

The calculated amplitude (in dB, with a lower limit of −24 dB for all the t-f diagrams) of the energy distribution of longitudinal modes was used in t-f representation[1,3] to examine the mode-resolving capability of the t-f method. In the t-f presentation, the color of a point represented the amplitude of the energy distribution. The time domain signal of each longitudinal mode was first calculated by using Eq. (2); then the t-f method was applied to these signals, respectively; finally the results were superimposed with (t-f) dispersion curves of various modes and presented in Fig. 2(a), which shows the t-f characteristics of each mode separately. The calculated time domain signals of each mode were then added and the t-f method was applied to the summed signals with the results presented in Fig. 2(b), in which constructive and destructive interference affect each mode differently.

As expected, L(0,5) and L(0,8) modes were clearly identified in both cases. However, other modes were not detected in the case shown in Fig. 2(b); the fundamental mode L(0,1) was not observable in 0.05–0.15 MHz in Fig. 2(b).

Figure 3(a) illustrates the superimposed time-frequency diagram of L(0,5) and L(0,8), while Fig. 3(b) illustrates that of F(1,8) and F(1,13). These two figures seem to suggest that L(0,5) and L(0,8) shared similar dispersive characteristics with F(1,8) and F(1,13), respectively. The energy distributions presented by the time-frequency distributions of the above-mentioned pairs of modes had a similar feature.

IV. DISCUSSION

For the hollow circular cylinder case, Protopappas *et al.*[4] found several characteristics from the t-f diagrams: the fundamental L(0,1) and F(1,1) modes in 0.05–0.15 MHz, the

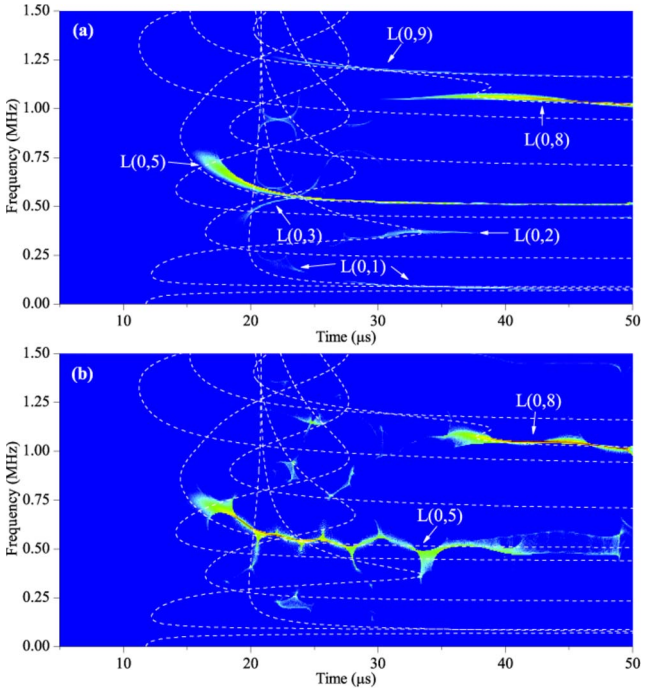


FIG. 2. (Color online) (a) The superimposed time-frequency diagram of longitudinal mode signals. (b) The time-frequency diagram of superimposed longitudinal mode signals.

L(0,4) mode in the range of 0.7–0.85 MHz, the L(0,5) mode in the range of 0.55–0.85 MHz, and the L(0,8) mode at its cutoff frequency can be clearly identified.

We calculated the theoretical signals of the longitudinal modes and first-order flexural modes in the same circular pipe. Of course, it was not possible to calculate all the modes, because there were infinite numbers of modes as the circumferential order varies from 1 to infinity.

It should be pointed out that results from Figs. 1 and 2 indicate that the strong energy components observed by Protopappas *et al.* in the low frequency range of 0.05–0.15 MHz might not correspond only to the L(0,1) and F(1,1) but also to some other higher-order fundamental flexural modes.

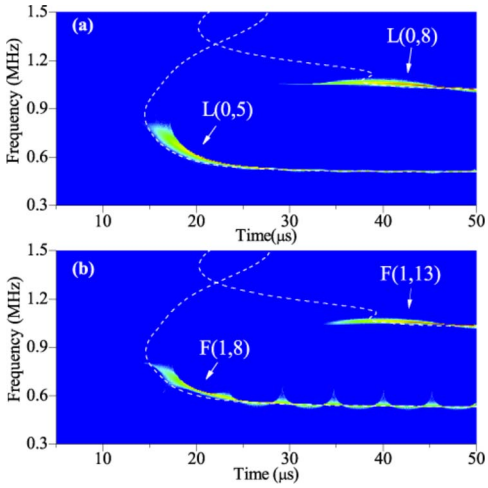


FIG. 3. (Color online) Superimposed time-frequency diagrams of (a) L(0,5) and L(0,8), and (b) F(1,8) and F(1,13).

From Fig. 1 and Fig. 2(a), we can see that other modes (including longitudinal modes and flexural modes) were also excited, but they did not appear in our Fig. 2(b) and Fig. 10(b) of the original paper by Protopappas *et al.*[4] This was because the dispersion curves of these modes are either too close to or even overlapping with that of other mode(s), e.g., L(0,3) is too close to L(0,5) while F(1,13) is overlapping with L(0,8).

Additionally, as illustrated in Fig. 3, the L(0,5) and F(1,8) modes, and the modes L(0,8) and F(1,13), possessed not only similar dispersion characteristics, but also similar t-f distributions. Consequently, the strong signals described as the L(0,5) and L(0,8) in the paper of Protopappas *et al.* might not only belong to these two modes, but also be shared by flexural modes, such as the F(1,8) and F(1,13), and other higher-order flexural modes.

It is worth mentioning that, as was indicated by Protopappas *et al.*, a realistic bone pipe model should consist of transversely isotropic material and nonidealized cross section. For the former, similar results could be obtained with the NME method; for the latter, NME analysis should be numerically carried out based on the dispersion characteristics calculated by FE or boundary element methods.

Actually, the combination of FE and NME techniques could make it possible to control the generated modes. The so-called coupled finite element-normal modes expansion method has been applied to the study of Lamb waves in a plate, and allows the determination of the amplitude of each wave mode.[7] We believe that with this technique applied in a bone model, one is capable of extracting ultrasound features quantitatively from obtained signals.

In conclusion, the inability to quantify the observed experimental results might come from the inherent disadvantages of the time-frequency method. To extract the detailed ultrasonic features, one could either apply image analysis and mode-tracking techniques to the t-f image or combine the FE simulations with the NME technique. We hope this comment would be helpful to the authors of the original paper and other researchers in the field.

ACKNOWLEDGMENTS

This work is supported by the Program for New Century Excellent Talents in University (06-0450) and the Scientific Research Foundation of Graduate School of Nanjing University (No. 2006CL07).

[1]Protopappas, V. C., Fotiadis, D. I., and Malizos, K. N., "Guided ultrasound wave propagation in intact and healing long bones," Ultrasound Med. Biol. **32**(5), 693–708 (2006).
[2]Bossy, E., Talmant, M., and Laugier, P., "Three-dimensional simulations of ultrasonic axial transmission velocity measurement on cortical bone models," J. Acoust. Soc. Am. **115**(5), 2314–2324 (2004).
[3]Alleyne, D., and Cawley, P., "A two-dimensional Fourier transform method for the measurement of propagation multimode signals," J. Acoust. Soc. Am. **89**(3), 1159–1168 (1991).
[4]Protopappas, V. C., Kourtis, I. C., Kourtis, L. C., Malizos, K. N., Massalas, C. V., and Fotiadis, D. I., "Three-dimensional finite element modeling of guided ultrasound wave propagation in intact and healing long bones," J. Acoust. Soc. Am. **121**(6), 3907–3921 (2007).
[5]Vollmann, J., Breu, R., and Dual, J., "High-resolution analysis of the complex wave spectrum in a cylindrical shell containing a viscoelastic me-

dium," J. Acoust. Soc. Am. **102**(2), 896–920 (1997).
[6]Tang, L. G., and Cheng, J. C., "Numerical analysis on laser-generated guided elastic waves in a hollow cylinder," J. Nondestruct. Eval. **21**(2), 45–53 (2002).
[7]Emmanuel, M., Jamal, A., and Christophe, D., "Modeling of Lamb waves generated by integrated transducers in composite plates using a coupled finite element-normal modes expansion method," J. Acoust. Soc. Am. **107**(1), 87–94 (2000).

Computing the far field scattered or radiated by objects inside layered fluid media using approximate Green's functions

Mario Zampolli,[a] Alessandra Tesei, and Gaetano Canepa
NATO Undersea Research Centre, Viale San Bartolomeo 400, 19126 La Spezia, Italy

Oleg A. Godin
CIRES, University of Colorado at Boulder and NOAA/Earth System Research Laboratory, 325 Broadway, Boulder, Colorado 80305-3328, USA

(Received 11 October 2007; accepted 27 February 2008)

A numerically efficient technique is presented for computing the field radiated or scattered from three-dimensional objects embedded within layered acoustic media. The distance between the receivers and the object of interest is supposed to be large compared to the acoustic wavelength. The method requires the pressure and normal particle displacement on the surface of the object or on an arbitrary circumscribing surface, as an input, together with a knowledge of the layered medium Green's functions. The numerical integration of the full wave number spectral representation of the Green's functions is avoided by employing approximate formulas which are available in terms of elementary functions. The pressure and normal particle displacement on the surface of the object of interest, on the other hand, may be known by analytical or numerical means or from experiments. No restrictions are placed on the location of the object, which may lie above, below, or across the interface between the fluid media. The proposed technique is verified through numerical examples, for which the near field pressure and the particle displacement are computed via a finite-element method. The results are compared to validated reference models, which are based on the full wave number spectral integral Green's function. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2902139]

PACS number(s): 43.20.Mv, 43.20.El, 43.40.Rj, 43.30.Jx [SFW] Pages: 4051–4058

I. INTRODUCTION

The computation of the field surrounding a closed surface, such as the wet surface of a transducer or of a scatterer, via the Helmholtz-Kirchhoff integral,[1] is a common practice in various fields of acoustics. The quantities, which one must know in order to accomplish the task, are the acoustic pressure p and the normal particle displacement u_n on the closed surface, as well as the Green's function for the background medium. Usually, p and u_n can be obtained from analytical or numerical modeling techniques, or from experimental data, while the Green's function is known as a closed form expression for the simpler cases, such as an infinite homogeneous background medium, or via some form of integral representation for more complex configurations. For the case of layered acoustic media, the Green's functions can be expressed by wave number spectral integrals,[2–6] which are well known but are not straightforward to evaluate.

In practical applications, the Helmholtz-Kirchhoff integral is often replaced by a discrete sum,[7–9] requiring the evaluation of a large number of Green's functions for the various source-receiver configurations and at a large number of different frequencies. The number of Green's functions needed for the computation rapidly increases with frequency, which can quickly lead to impractical computation times if one performs the evaluation of the Green's functions by numerical integration. This paper describes an efficient technique for computing the acoustic field generated by a radiating or scattering object inside a layered acoustic medium, at distances large compared to the wavelength, using approximations of the relevant Green's functions written in terms of elementary functions. The approximate Green's functions, which can also be applied to source-receiver configurations where the source and the receiver are close to the fluid-fluid interface, are based on the asymptotic evaluation of the wave number integral by the method of steepest descent.[5,6] No restrictions are imposed on the shape of the object of interest or on its location with respect to the interface between the two background media.

An alternative and potentially attractive technique for computing the layered medium Green's functions is the method of complex images.[10–15] The complex image solution is based on the approximation of the reflection and transmission coefficients in the Green's function integrand via a discrete sum of image point sources, with complex source point coordinates. The two main advantages of this method are its validity in the near field and the existence of a simple recursion relation in frequency[12] for computing the source parameters. The fitting of the transmission coefficient, which is needed in those configurations where the source is located in one of the two media, and the receiver is located in the other medium, is affected by convergence problems which have been reported in the literature.[13] For this reason, the present work is focused on the steepest descent approximations of the Green's functions.

[a]Electronic mail: zampolli@nurc.nato.int.

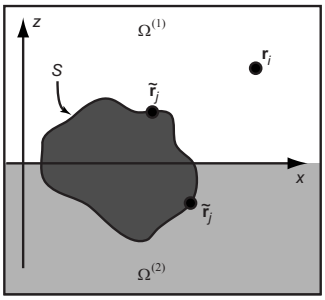


FIG. 1. A scattering or radiating object is included in a two-layered acoustic medium. The Green's function sources $\tilde{\mathbf{r}}_j$ can be in two different configurations with respect to the background acoustic media, $\Omega^{(1)}$ and $\Omega^{(2)}$, and with respect to the receiver $\mathbf{r}_i$.

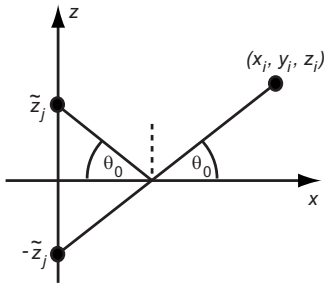


FIG. 2. A source is located above the fluid-fluid interface, at $(0,0,\tilde{z}_j)$, and the receiver is in the same medium.

Section II describes the technique for obtaining the far field via the Helmholtz-Kirchhoff integral. The approximate Green's functions[5,6] for the source and the receiver in the same medium, and for the source and the receiver in different media, are presented for a two-layered medium with refractive index $n>1$ and $n<1$. The method is applied in Sec. III to scattering from an elastic spherical shell buried and partially buried with respect to the interface between the two media. Results for a proud sphere have been presented in an earlier article.[9] The numerical examples are restricted to spherical shapes because of the availability of validated reference solutions, which are specialized to such cases.[16,17] This circumstance does not imply any loss of generality and does not preclude the applicability of the technique to arbitrary three-dimensional objects.

II. ACOUSTIC FAR FIELD FOR AN OBJECT IN A TWO-LAYERED MEDIUM

Throughout the entire paper, the attention is restricted to harmonic oscillations of constant frequency f. The complex time dependence $e^{-i\omega t}$, with t representing time, $i=\sqrt{-1}$, and $\omega=2\pi f$, is factored out of the equations. The two-layered medium consists of an upper fluid, having density $\rho_f^{(1)}$ and sound speed $c_f^{(1)}$ and occupying the half-space $z>0$, and a lower fluid, having density $\rho_f^{(2)}$ and sound speed $c_f^{(2)}$ and occupying the half-space $z<0$. The results presented here are also applicable to media with weak attenuation, which can be accounted for by introducing complex sound speeds.

A general case, representing the possible configurations of the object with respect to the two-layered medium and with respect to the receiver, is the situation depicted in Fig. 1. Those circumstances, where the object lies only within the upper fluid or the lower fluid, are special cases of the general scenario shown in the figure. The complex acoustic pressure p and the normal particle displacement u_n on a surface S enclosing the object may be known from experimental measurements or from numerical results. The volume $\Omega^{(1)}$ represents the difference between the $z>0$ half-space and the volume enclosed by S, and similarly $\Omega^{(2)}$ is the $z<0$ half-space minus the volume enclosed by S.

The acoustic near field, sampled at the points $\tilde{\mathbf{r}}_j=(\tilde{x}_j,\tilde{y}_j,\tilde{z}_j)\in S$, with $1\leqslant j\leqslant N_{\text{src}}$, can be propagated to a number of receivers located at $\mathbf{r}_i=(x_i,y_i,z_i)\in\Omega^{(1)}$, with $1\leqslant i\leqslant N_{\text{rec}}$, using the discrete approximation of the Helmholtz-Kirchhoff integral:[9]

$$p(\mathbf{r}_i)=\sum_j\left(\frac{\partial G_{ij}^{(12)}}{\partial n_j}p_j-\rho_f(\tilde{z}_j)\omega^2 G_{ij}^{(12)}(u_n)_j\right)dA_j, \tag{1}$$

where $p_j=p(\tilde{\mathbf{r}}_j)$, $(u_n)_j=u_n(\tilde{\mathbf{r}}_j)$, and dA_j is the portion of the area on S associated with each point. The operator $\partial/\partial n_j$ denotes the derivative in the direction of the outward pointing normal vector at the point $\tilde{\mathbf{r}}_j\in S$. The function $\rho_f(\tilde{z}_j)=\rho_f^{(1)}$ for $\tilde{z}_j>0$ and $\rho_f(\tilde{z}_j)=\rho_f^{(2)}$ for $\tilde{z}_j<0$. The notation $G_{ij}^{(12)}=G^{(12)}(\mathbf{r}_i,\tilde{\mathbf{r}}_j)$ represents the Green's function for the two-layered medium, with the source at $\tilde{\mathbf{r}}_j$ and the receiver at $\mathbf{r}_i$. The full wave number spectral integral representations of the reflected and transmitted (or refracted) fields caused by a point source in a two-layered fluid medium are well known from the literature[2,4–6] and can be evaluated using a number of numerical techniques. This can lead to large computation times, particularly if the evaluation of the wave number integrals is required for a large number of sources N_{src}. Numerical experiments[9] indicate that the far field computed with Eq. (1) converges if the source locations $\tilde{\mathbf{r}}_j$ are spaced by less than $\lambda_{\min}/3$, where $\lambda_{\min}$ represents the shortest wavelength in the frequency band of interest. For many examples of practical interest, where $k_{\max}a=O(10)$, with $k_{\max}=2\pi/\lambda_{\min}$, and where a is a representative dimension of the object, the implication is that $N_{\text{src}}=O(10^3)$. The computation times resulting from the direct integration of the wave number kernels in such cases can be significant, particularly if the field needs to be calculated over a broad range of densely spaced frequencies.

If $k|\mathbf{r}_i-\tilde{\mathbf{r}}_j|\gg 1$, with $k=\omega/c_f^{(1)}$, substantial improvements in the computational efficiency can be obtained by employing approximate representations of $G^{(12)}$ based on the steepest descent evaluation of the wave number spectral integral, which are obtained by neglecting second order terms compared to unity in the far field asymptotic expansions of the reflected and transmitted acoustic fields.[5,6]

A. Source above the interface: The reflected field contribution

The Green's function describing the pressure field generated by a point source at $\tilde{\mathbf{r}}_j=(0,0,\tilde{z}_j)$, $\tilde{z}_j>0$, received at $\mathbf{r}_i=(x_i,y_i,z_i)$ (Fig. 2) is[4,5]

$$G^{(12)}(\mathbf{r}_i,\tilde{\mathbf{r}}_j) = \frac{\exp(ikR)}{R} + i\int_0^\infty V(\xi)J_0(\xi r)e^{i\mu(z_i+\tilde{z}_j)}\frac{\xi}{\mu}d\xi, \tag{2}$$

where J_0 is the zero order Bessel function, $\mu=\sqrt{k^2-\xi^2}$, $r=\sqrt{x_i^2+y_i^2}$, and $R=\sqrt{r^2+(z_i-\tilde{z}_j)^2}$. The plane-wave reflection coefficient V is expressed as a function of $\xi=k\cos\theta_0=k_2\cos\theta_1$,

$$V(\xi) = \frac{m\mu-\mu_1}{m\mu+\mu_1}, \tag{3}$$

using the ratio of densities $m=\rho_f^{(2)}/\rho_f^{(1)}$ and $\mu_1=\sqrt{k_2^2-\xi^2}$, with $k_2=\omega/c_f^{(2)}$.

The first order accurate steepest descent approximation of the integral in Eq. (2) yields[5,6]

$$G^{(12)}(\mathbf{r},\tilde{\mathbf{r}}_j) = \frac{\exp(ikR)}{R} + \frac{\exp(ikR_1)}{R_1}\left[V(\xi) - \frac{iN}{kR_1}\right], \tag{4}$$

where $R_1=\sqrt{r^2+(z_i+\tilde{z}_j)^2}$ and θ_0 is the grazing angle on the interface between the two fluid media. The coefficient N, which introduces a correction to the zero order accurate ray approximation, is given by

$$N = \frac{1}{2}\left(\frac{\partial^2 V}{\partial\theta^2} - \frac{\partial V}{\partial\theta}\tan\theta\right)_{\theta=\theta_0}. \tag{5}$$

Defining the refractive index as $n=c_f^{(1)}/c_f^{(2)}$ and the critical angle of reflection $\delta=\arccos n$, the lateral wave component

$$2i\cos\delta\exp[ikR_1\cos(\theta_0-\delta)] \times\{mkR_1^2[\cos\theta_0\sin\delta\sin^3(\delta-\theta_0)]^{1/2}\}^{-1} \tag{6}$$

must be added to the right hand side of Eq. (4) in those cases, where $|\mathrm{Im}(n)|\ll 1$ and $\mathrm{Im}(k/|k|)\ll 1$ (moderate damping), in the region $\theta_0<\mathrm{Re}(\delta)$, if $\mathrm{Re}(n)<1$, or in the region $\theta_0<\arccos[1/\mathrm{Re}(n)]$, if $\mathrm{Re}(n)>1$. While the proportional to N correction in Eq. (4) originates from an accurate evaluation of the contribution due to the stationary point $\xi=k\cos\theta_0$ of the exponential in the integrand in Eq. (2), the lateral wave (6) originates mathematically as a contribution associated with the branch point $\xi=k_2$ of the integrand.

B. Source below the interface: The transmitted (or refracted) field contribution

For a source at $\tilde{\mathbf{r}}_j=(0,0,-\tilde{z}_j)$, $\tilde{z}_j>0$ and for a receiver located at $\mathbf{r}_i=(x_i,y_i,z_i)$, $z_i>0$, the Green's function is

$$G^{(12)}(\mathbf{r}_i,\tilde{\mathbf{r}}_j) = i\int_0^\infty W(\xi_2)J_0(\xi_2 r)e^{i(\mu_2\tilde{z}_j-\mu_{2,1}z_i)}\frac{\xi_2}{\mu_2}d\xi_2, \tag{7}$$

with the plane-wave transmission coefficient being W, $\xi_2=k_2\cos\theta_{0,2}=k\cos\theta_{1,2}$, $\mu_2=\sqrt{k_2^2-\xi_2^2}$ and $\mu_{2,1}=\sqrt{k^2-\xi_2^2}$. The angle $\theta_{0,2}$ is the incident grazing angle, and $\theta_{1,2}$ is the transmission grazing angle, from the lower fluid into the upper fluid, defined according to Figs. 3 and 4.

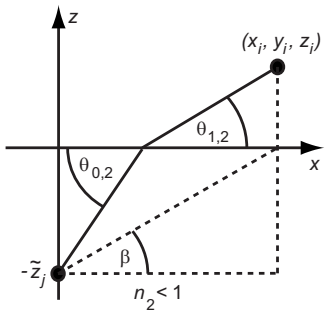


FIG. 3. The source is below the fluid-fluid interface, at $(0,0,-\tilde{z}_j)$, and the receiver is in the upper medium. The refractive index $n_2<1$.

The steepest descent evaluation of the wave number integral yields, upon neglecting of terms $O(k^{-2}R^{-2})$,[6]

$$G^{(12)}(\mathbf{r}_i,\tilde{\mathbf{r}}_j) = \sqrt{\frac{\cos\theta_{0,2}\sin\theta_{0,2}}{r(\tilde{z}_j+z_i\sin^3\theta_{0,2}/(n_2\sin^3\theta_{1,2}))}} \times\exp\left[ik_2\left(\frac{\tilde{z}_j}{\sin\theta_{0,2}}+\frac{n_2 z_i}{\sin\theta_{1,2}}\right)\right] \times\left[W(\cos\theta_{0,2}) + \frac{iN_t\sin^3\theta_{0,2}}{2k_2(\tilde{z}_j+z_i\sin^3\theta_{0,2}/(n_2\sin^3\theta_{1,2}))}\right], \tag{8}$$

where $n_2=c_f^{(2)}/c_f^{(1)}$. In this case, the plane-wave transmission coefficient is written as a function of $q=\cos\theta_{0,2}$,

$$W(q) = \frac{2m_2\sqrt{1-q^2}}{m_2\sqrt{1-q^2}+\sqrt{n_2^2-q^2}}, \tag{9}$$

with $m_2=\rho_f^{(1)}/\rho_f^{(2)}$, and

$$N_t = \frac{3\cos\theta_{0,2}}{\sin^2\theta_{0,2}}\left(\frac{\tilde{z}_j+z_i\sin^5\theta_{0,2}/(n_2^3\sin^5\theta_{1,2})}{\tilde{z}_j+z_i\sin^3\theta_{0,2}/(n_2\sin^3\theta_{1,2})} - \frac{1+\cos^2\theta_{0,2}}{3\cos^2\theta_{0,2}}\right)\frac{\partial W}{\partial q} - \frac{\partial^2 W}{\partial q^2}. \tag{10}$$

The stationary point in the integrand in Eq. (7), which gives the field (8), corresponds to the refracted ray which obeys Snell's law.

As $z_i\to 0$, the angle $\theta_{1,2}\to 0$, and the terms containing inverse powers of $\sin\theta_{1,2}$ in Eq. (8) diverge. This implies that, for small z_i, the numerical evaluation of $G^{(12)}$ can be-

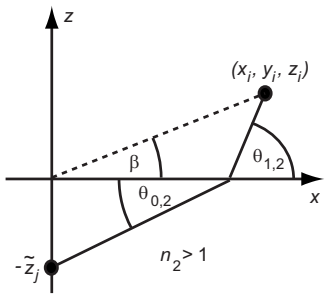


FIG. 4. The source is below the fluid-fluid interface, at $(0,0,-\tilde{z}_j)$, and the receiver is in the upper medium. The refractive index $n_2>1$.

come unstable, which makes it necessary to resort to more convenient limiting forms of Eq. (8). For the case $n_2<1$ and $r>\tilde{z}_j \cot\beta$, where β is defined in Fig. 3, one obtains in the limit of small z_i,

$$G^{(12)}(\mathbf{r}_i,\tilde{\mathbf{r}}_j)=\frac{2[z_i+i/(k_2 m_2 \sin\delta)]\cot\delta}{\sqrt{r(r-\tilde{z}_j\cot\delta)^3}} \times \exp\left[ik_2\left(r\cos\delta+\tilde{z}_j\sin\delta+\frac{z_i^2\cos\delta}{2(r-\tilde{z}_j\cot\delta)}\right)\right]. \tag{11}$$

If $\beta<\delta$ and $k_2 z_i^2 \ll R(\cos\beta-\cos\delta)$, the contribution of an additional stationary point in the asymptotic evaluation of the wave number integral yields

$$(r^2+\tilde{z}_j^2)^{-1/2}\exp(ik_2\sqrt{r^2+\tilde{z}_j^2}-k_2 z_i\sqrt{\cos^2\beta-n_2^2}) \times\left[\frac{2m_2\sin\beta}{m_2\sin\beta+i\sqrt{\cos^2\beta-n_2^2}}-\frac{iN_d}{k_2\sqrt{r^2+\tilde{z}_j^2}}\right], \tag{12}$$

which must be added to Eq. (11). The factor N_d is given by the expression

$$N_d=N+A_d+k_2 z_i m_2\sin\beta(B_d+C_d), \tag{13}$$

with

$$A_d=\frac{2ik_2 z_i(1-n_2^2)\cos^2\beta\sin\beta}{(\cos^2\beta-n_2^2)(m_2\sin\beta+i\sqrt{\cos^2\beta-n_2^2})^2}, \tag{14}$$

$$B_d=\frac{2n_2^2-\cos^2\beta(3n_2^2+1-2\cos^2\beta)}{(\cos^2\beta-n_2^2)^{3/2}(m_2\sin\beta+i\sqrt{\cos^2\beta-n_2^2})}, \tag{15}$$

$$C_d=\frac{k_2 z_i\cos^2\beta\sin^2\beta\sqrt{\cos^2\beta-n_2^2}}{(\cos^2\beta-n_2^2)^{3/2}(m_2\sin\beta+i\sqrt{\cos^2\beta-n_2^2})}, \tag{16}$$

and where N is obtained from Eq. (5) by replacing n with n_2, m with m_2 in the reflection coefficient $V(\xi)$, which in this case is computed for the angle $\theta_{0,2}$. Mathematically, the component (12) of the transmitted field arises as a contribution of an additional, complex stationary point. Physically, it corresponds to a diffracted arrival, which propagates along a path distinct from the refracted ray.

When $n_2>1$, the interface is upward refracting (Fig. 4). In this case, if $r>z_i/\sqrt{n_2^2-1}$ and $\tilde{z}_j\to 0$, so that $\theta_{0,2}\to 0$, the transmission coefficient $W(\cos\theta_{0,2})\to 0$, and instabilities arise in the numerical evaluation of Eq. (8). The asymptotic form, which makes it possible to circumvent these difficulties for small z_i, is

$$G^{(12)}(\mathbf{r}_i,\tilde{\mathbf{r}}_j)=\left(\frac{2m_2\tilde{z}_j}{\sqrt{n_2^2-1}}+\frac{2im_2^2}{k_2(n_2^2-1)}\right) \times[r(r-z_i/\sqrt{n_2^2-1})^3]^{-1/2} \times\exp\left[ik_2\left(r+z_i\sqrt{n_2^2-1}+\frac{\tilde{z}_j^2}{2(r-z_i/\sqrt{n_2^2-1})}\right)\right]. \tag{17}$$

If $\cos\beta>1/n_2$, then the term

$$m_2/\sqrt{r^2+z_i^2}\times\exp(ik_2 n_2\sqrt{r^2+z_i^2}-k_2\tilde{z}_j\sqrt{n_2^2\cos^2\beta-1}) \times\left[\frac{2n_2\sin\beta}{n_2\sin\beta+im_2\sqrt{n_2^2\cos^2\beta-1}}+\frac{N_e}{k_2\sqrt{r^2+z_i^2}}\right] \tag{18}$$

must be added to the right hand side of Eq. (17), where

$$N_e=m_2(n_2^2-1)\frac{A_e+B_e}{d_{e,1}^{3/2}d_{e,2}^3}+\frac{C_e}{d_{e,1}d_{e,2}^2}-ik_2\tilde{z}_j n_2^2\sin\beta\frac{D_e+E_e}{d_{e,1}^{3/2}d_{e,2}}, \tag{19}$$

using

$$A_e=n_2^3\sin^2\beta(3-\sin^2\beta)-2n_2(n_2^2-1), \tag{20}$$

$$B_e=im_2\sin\beta(2+n_2^2\cos^2\beta)\sqrt{n_2^2\cos^2\beta-1}, \tag{21}$$

$$C_e=2k_2\tilde{z}_j n_2 m_2^2(n_2^2-1)\cos^2\beta\sin\beta, \tag{22}$$

$$D_e=2-\cos^2\beta(3+n_2^2-2n_2^2\cos^2\beta), \tag{23}$$

$$E_e=k_2\tilde{z}_j n_2^2\cos^2\beta\sin^2\beta\sqrt{n_2^2\cos^2\beta-1}, \tag{24}$$

$$d_{e,1}=n_2^2\cos^2\beta-1, \tag{25}$$

and

$$d_{e,2}=n_2\sin\beta+im_2\sqrt{n_2^2\cos^2\beta-1}. \tag{26}$$

The components (17) and (18) of the transmitted wave originate as contributions due to a real and a complex stationary point and correspond to regular (ray) and diffracted arrivals, respectively.

III. NUMERICAL RESULTS

The technique presented in Sec. II is illustrated here by considering a spherical scatterer of outer radius $a=0.5$ m in two different configurations with respect to the planar interface separating two fluid half-spaces (Fig. 5): (i) The sphere is contained in the lower medium, with a distance of 20 cm between the north pole and the interface $z=0$ (buried sphere), and (ii) one hemisphere is contained in the $z>0$ half-space and the other hemisphere is contained in the $z<0$ half-space (half buried sphere). Far field results for scattering from a spherical target lying proud on the interface between two fluid media have already been presented in an earlier paper.[9]

The upper fluid has the properties $\rho_f^{(1)}=1000$ kg/m^3, $c_f^{(1)}=1500$ m/s and the lower fluid has $\rho_f^{(2)}=1800$ kg/m^3, $c_f^{(2)}=1600$ m/s with a dissipation of $\alpha=0.5$ dB/λ. The critical angle of reflection is around $\delta=20°$. The scatterer is assumed to be a void elastic spherical shell of 1 cm thickness, with density $\rho_s^{\text{shell}}=3000$ kg/m^3, longitudinal sound speed $c_L^{\text{shell}}=3500$ m/s, and shear sound speed $c_T^{\text{shell}}=1400$ m/s. Two different plane-wave insonification angles are considered for each of the test cases: supercritical insonification at $\theta_{\text{inc}}=25°$ and subcritical insonification at $\theta_{\text{inc}}=15°$. The two

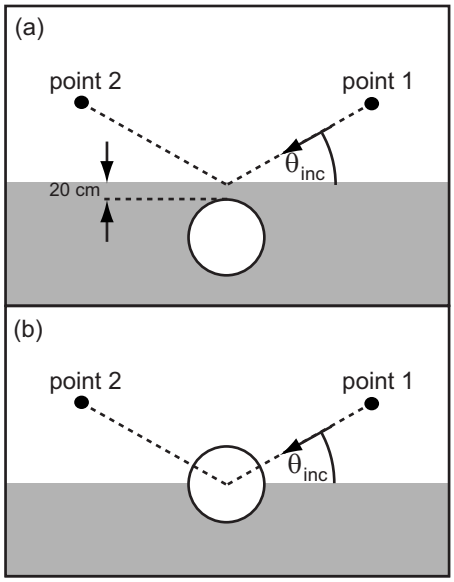


FIG. 5. A void spherical shell is buried 20 cm below the fluid-fluid interface (a) and half buried with respect to the interface (b). The direction of the incoming plane wave is denoted by the arrow. The receivers are located at point 1 (backscatter direction) and point 2 (forward scatter direction).

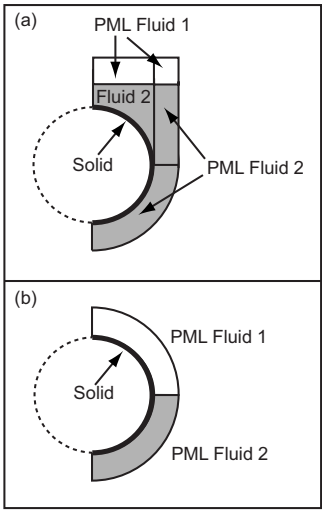


FIG. 6. Schematic representation of the FE domains used for computing the near field scattered pressure and displacement for the (a) buried sphere and for the (b) half buried sphere.

receivers are chosen to lie one in the backscattering direction (labeled "point 1"), and the other one in the forward scattering direction (labeled "point 2"), at 50 m distance from the projection of the centroid of the sphere on the $z=0$ plane, as shown in Fig. 5.

The scattered acoustic pressure p^{scat} and the corresponding normal displacement u_n^{scat} on the surface of the sphere are computed over the frequency band $f=100$ Hz–10 kHz at increments of 5 Hz, using the finite-element (FE) scattering model described in Ref. 9. The FE tool treats axially symmetric geometries with nonsymmetric incident field forcing via an azimuthal Fourier mode decomposition of the problem around the axis of symmetry. This results in a finite number of two-dimensional (2D) models (one for each azimuthal mode), each of which is independently solved from the others. The 2D computational domains used to define the geometry of the FE model are schematically represented in Fig. 6. The perfectly matched layers[18] (PMLs), with which the open physical domain is artificially truncated in the model, can also be seen in the figure. Details on the emulation of the Sommerfeld radiation condition using PMLs and on the criteria for subdividing the computational domain into a mesh of finite elements are given in Ref. 9. For the buried sphere, it is necessary to include a physical layer of fluid between the target and the fluid-fluid interface [Fig. 6(a)]. This requires using triangular mesh elements, yielding a total number of 110 000 complex FE degrees of freedom (DOFs) for each azimuthal Fourier mode. The computational domain for the half buried sphere can be constructed with rectangular elements. In this case, the PMLs can be applied in direct contact with the spherical target, and the number of DOF is considerably smaller (30 000 DOFs). The computations are carried out on an AMD64 Opteron 285 platform, with a CPU clock speed of 2.6 GHz.

The incident acoustic field is expressed in terms of incoming, boundary-reflected and transmitted plane waves. The quantity of interest, computed with the numerical model, is the far field target strength

$$\text{TS}(\mathbf{r}_i) = 20 \log_{10}(\mathbf{r}_i |p^{\text{scat}}(\mathbf{r}_i)|/|p^{\text{inc}}|), \tag{27}$$

where the scattered field $p^{\text{scat}}(\mathbf{r}_i)$ at the receiver is computed using Eq. (1) and the approximations of $G_{ij}^{(12)}$, presented in Sec. II, together with the scattered pressure p^{scat} and the normal displacement u_n^{scat} sampled on the wet surface of the sphere. The amplitude of the incoming plane wave is $|p^{\text{inc}}|=1$ for the cases considered here. The normal derivative of the Green's functions in Eq. (1) is computed by a first order finite difference.

The sampling points of p^{scat} and u_n^{scat} are approximately equidistributed on the wet surface of the sphere, with a spacing of $\lambda_{\min}/3.8$, which yields a total of 2188 source points for the Green's functions. To evaluate $\partial G_{ij}^{(12)}/\partial n_j$, it is also necessary to compute $G_{ij}^{(12)}$ at a second set of source points, obtained by moving the points on the sphere surface out by a small distance along the unit normal. The evaluations of $G_{ij}^{(12)}$ must be repeated for each of the 1981 frequencies in the frequency band and for each of the two receiver locations, which implies a total of 17×10^6 computations of $G_{ij}^{(12)}$ required to obtain the target strength plots of Secs. III A and III B. The CPU times for the TS computation at both receiver points and for all frequencies are reported in detail in Secs. III A and III B, and do not exceed 125 s for the buried case and 85 s for the half buried case. These times are essentially negligible compared to the CPU time required to solve the near field FE problem.

In what follows below, the results obtained with the technique presented in this paper are labeled "FE-HK" (for "finite element with Helmholtz-Kirchhoff postprocessing"). Results computed with the transition matrix (T-matrix) approach,[16,17] which has been validated in the past by comparison with experimental data[19,20] and with other numerical techniques,[21] are used as reference solutions. The compari-

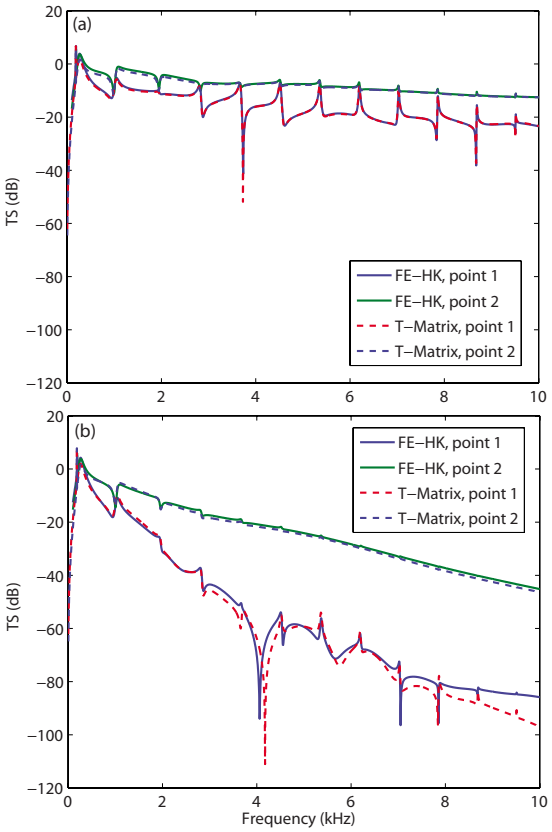


FIG. 7. (Color online) Scattering from the buried spherical shell at (a) supercritical and (b) subcritical insonifications.

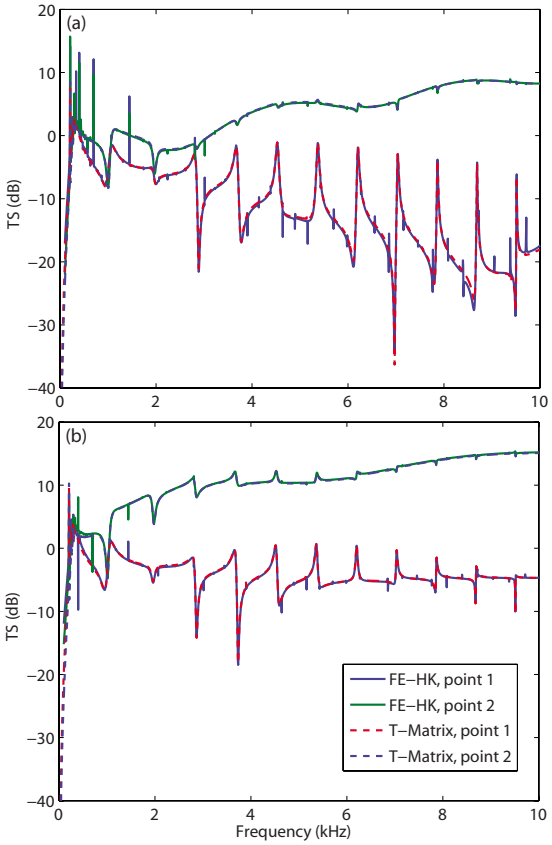


FIG. 8. (Color online) Scattering from the half buried spherical shell at (a) supercritical and (b) subcritical insonifications.

son with the *T*-matrix results, which are obtained using the full wave number integral solution, is particularly appropriate in the context of this paper, since it substantiates the validity of the asymptotic approximations of the Green's functions employed in the FE-HK computations.

A. Buried sphere

For the supercritical incidence case, the approximations given in Eq. (8) are used to propagate the scattered field to the receiver points via Eq. (1). The CPU time for the entire far field computation, using the FE solution as an input, is 64 s for this case. Figure 7(a) shows the excellent agreement between the FE-HK result and the *T*-matrix reference solution.

At subcritical incidence, the backscattering and forward scattering receiver points move closer to the surface, and it becomes necessary to use Eq. (17). The correction term given in Eq. (18) must be added to the right hand side of Eq. (17) for those source-receiver configurations, for which $\cos\beta > 1/n_2$. In this case, the CPU time for the evaluation of Eq. (1) for the two receivers, over the entire bandwidth of the sweep, increases to 125 s. The agreement between the present result and the *T*-matrix solution, shown in Fig. 7(b), is overall very good, with some discrepancies appearing above 3 kHz in the backscattering direction. The comparison with alternative solutions for the same problem, computed using other tools,[22] such as boundary element methods,[23–25] could not resolve the disagreement.

B. Half buried sphere

For the half buried spherical scatterer, Green's function sources above and below the $z=0$ interface must be taken into account to compute the target strength. The contributions to Eq. (1) coming from point sources for which $\tilde{z}_j < 0$ are computed according to the procedure already described in Sec. III A. The Green's functions for sources contained in the $z>0$ half space are computed using Eq. (4), together with the lateral wave correction term (6), which must be added to the right hand side of Eq. (4) for those source-receiver configurations, for which $\theta_0 < \mathrm{Re}(\delta)$.

Figure 8 shows that the overall agreement between the FE-HK results and the *T*-matrix results is very good for both supercritical and subcritical incidences. The CPU times for the supercritical and for the subcritical case are 42 and 82 s, respectively. However, the FE-HK results exhibit some spurious narrow resonance peaks, which are more evident in the supercritical case. The nonphysical resonances are a consequence of the ill-posed finite element boundary value problem, which is formulated in terms of a nonviscous fluid model: The density jump across the fluid-fluid interface gives rise to an incident acoustic field with a discontinuous tangential displacement at the $z=0$ interface. The points, where the $z=0$ interface meets the outer surface of the elastic structure, define a line along which the normal vector on the target surface is parallel to the interface between the two fluids. Along this line, the component of the total acoustic particle displacement tangential to the $z=0$ interface must be con-

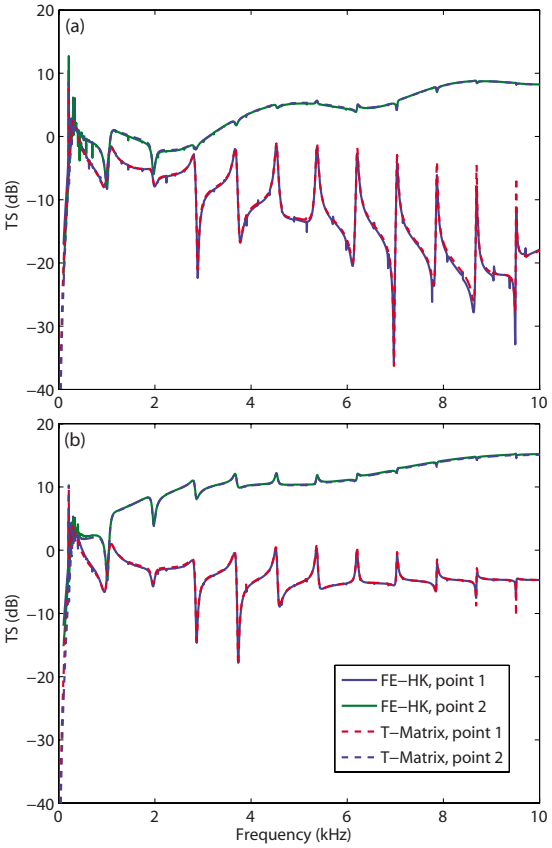


FIG. 9. (Color online) Scattering from the half buried spherical shell at (a) supercritical and (b) subcritical insonifications, with a material damping of 0.02 dB/λ in the shell.

tinuous with the normal component of the elastic displacement on the target surface. This, in turn, leads to the condition that the mechanical impedance of the elastic structure must be infinite along the same line. Such a constraint is nonphysical and causes the spurious resonances which can be seen in Fig. 8.

Similar observations have been previously reported by Fricke[26,27] in the finite difference analysis of scattering from floating elastic ice features. In Fricke's work, the numerical model was stabilized by the introduction of a small material damping. Following the same approach, a dissipation of $\alpha = 0.02$ dB/λ is introduced by adding an imaginary component to the longitudinal and transverse sound speeds of the shell material. Figure 9 shows the improved agreement with the *T*-matrix solution: The nonphysical peaks almost disappear, while the structure of the target response is mostly unaltered by the damping.

IV. CONCLUSIONS

An efficient method for computing the far field radiated or scattered from generic three-dimensional objects, located within layered acoustic fluid media, has been presented. The technique is based on the evaluation of the discretized Helmholtz-Kirchhoff integral, which takes the pressure and normal displacement on the surface of the object as input. In practice, these two quantities can be determined analytically, numerically, or by experiments, which makes the general technique widely applicable to various fields of acoustics. The efficiency comes from the availability of approximate formulas for the Green's functions of the layered medium, which are based on the asymptotic evaluation of the full wave number integrals.

The computational examples for scatterers embedded inside a two-layered medium demonstrate the effectiveness of the technique. An extension to include more general background media, such as Pekeris waveguides or stepwise stratified media, should be fairly straightforward by adapting the method of images, essentially following earlier work by Fawcett.[12]

ACKNOWLEDGMENTS

The authors are grateful to Finn Jensen of NURC for the useful discussions and inputs to this paper. The *T*-matrix results for the partially buried sphere were kindly provided by Raymond Lim from NSWC Panama City, whom the authors also thank for the fruitful discussions. Useful inputs and discussions with William Kuperman of UC San Diego and Henrik Schmidt of MIT have also contributed to the work presented here.

[1] A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications* (Acoustical Society of America, Melville, NY, 1991).
[2] P. M. Morse and K. U. Ingard, *Theoretical Acoustics* (McGraw-Hill, New York, 1968).
[3] H. Schmidt and J. Glattetre, "A fast field model for the three-dimensional wave propagation in stratified environments based on the global matrix method," J. Acoust. Soc. Am. **78**, 2105–2114 (1985).
[4] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics* (Springer-Verlag, New York, 2000).
[5] L. M. Brekhovskikh and O. A. Godin, *Acoustics of Layered Media II: Point Sources and Bounded Beams* (Springer-Verlag, Berlin, 1992).
[6] L. M. Brekhovskikh and O. A. Godin, *Akustika Neodnorodnykh Sred. I: Osnovy Teorii Otrazheniya i Rasprostraneniya Zvuka (Acoustics of Inhomogeneous Media. Vol. I: Fundamentals of Sound Reflection and Propagation Theory)* (Nauka, Moscow, 2007) (in Russian).
[7] H. Schmidt, "Virtual source approach to scattering from partially buried elastic targets," AIP Conf. Proc. **728**, 456–463 (2004).
[8] I. Lucifredi and H. Schmidt, "Subcritical scattering from buried elastic shells," J. Acoust. Soc. Am. **120**, 3566–3583 (2006).
[9] M. Zampolli, A. Tesei, F. B. Jensen, N. Malm, and J. B. Blottman, "A computationally efficient finite element model with perfectly matched layers applied to scattering from axially symmetric objects," J. Acoust. Soc. Am. **122**, 1472–1485 (2007).
[10] M. A. Nobile and S. I. Hayek, "Acoustic propagation over an impedance plane," J. Acoust. Soc. Am. **78**, 1325–1336 (1985).
[11] J. A. Fawcett, "Complex-image approximations to the half-space acousto-elastic Green's function," J. Acoust. Soc. Am. **108**, 2791–2795 (2000).
[12] J. A. Fawcett, "A method of images for a penetrable acoustic waveguide," J. Acoust. Soc. Am. **113**, 194–204 (2003).
[13] J. A. Fawcett and R. Lim, "Evaluation of the integrals of target/seabed scattering using the method of complex images," J. Acoust. Soc. Am. **114**, 1406–1415 (2003).
[14] M. Ochmann, "The complex equivalent source method for sound propagation over an impedance plane," J. Acoust. Soc. Am. **116**, 3304–3311 (2004).
[15] G. Taraldsen, "The complex image method," Wave Motion **43**, 91–97 (2005).
[16] R. Lim, J. L. Lopes, R. H. Hackman, and D. G. Todoroff, "Scattering by objects buried in underwater sediments: Theory and experiment," J. Acoust. Soc. Am. **93**, 1762–1783 (1993).
[17] R. Lim, "Acoustic scattering by a partially buried three-dimensional elastic obstacle," J. Acoust. Soc. Am. **104**, 769–782 (1998).
[18] J. P. Bérenger, "A perfectly matched layer for the absorption of electromagnetic waves," J. Comput. Phys. **114**, 185–200 (1994).
[19] A. Tesei, A. Maguer, W. L. J. Fox, R. Lim, and H. Schmidt, "Measurements and modeling of acoustic scattering from partially and completely

buried spherical shells," J. Acoust. Soc. Am. **112**, 1817–1830 (2002).
[20]H. J. Simpson, B. H. Houston, and R. Lim, "Laboratory measurements of sound scattering from a buried sphere above and below the critical angle (L)," J. Acoust. Soc. Am. **113**, 39–42 (2003).
[21]J. J. Shirron and T. E. Giddings, "A finite element model for acoustic scattering from objects near a fluid-fluid interface," Comput. Methods Appl. Mech. Eng. **196**, 279–288 (2006).
[22]F. B. Jensen, M. Zampolli, and A. Tesei, "Benchmarking scattering from spheres and cylinders near the seafloor: A numerical comparative study," Technical Report NURC-FR-2007-005, NATO Undersea Research Center, 2007.
[23]S. Amini and P. J. Harris, "A comparison between various boundary integral formulations of the exterior acoustic problem," Comput. Methods Appl. Mech. Eng. **84**, 59–75 (1990).
[24]I. Karasalo, "Exact finite elements for wave propagation in range-independent fluid-solid media," J. Sound Vib. **172**, 671–688 (1994).
[25]I. Karasalo, "On evaluation of hypersingular integrals over smooth surfaces," Comput. Mech. **40**, 617–625 (2007).
[26]J. R. Fricke, "Acoustic scattering from elemental Arctic ice features: Numerical modeling results," J. Acoust. Soc. Am. **93**, 1784–1796 (1993).
[27]J. R. Fricke, Ph.D. thesis, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution (1991).

Nonlinear radial oscillations of encapsulated microbubbles subject to ultrasound: The effect of membrane constitutive law

Kostas Tsiglifis and Nikos A. Pelekasis[a)]
Department of Mechanical and Industrial Engineering, University of Thessaly Pedion Areos, Volos, Thessaly 38334 Greece

(Received 9 June 2007; revised 17 March 2008; accepted 22 March 2008)

The nonlinear radial oscillations of bubbles that are encapsulated in an elastic shell are investigated numerically subject to three different constitutive laws describing the viscoelastic properties of the shell: the Mooney–Rivlin (MR), the Skalak (SK), and the Kelvin–Voigt (KV) models are used in order to describe strain-softening, strain-hardening and small displacement (Hookean) behavior of the shell material, respectively. Due to the isotropic nature of the acoustic disturbances, the area dilatation modulus is the important parameter. When the membrane is strain softening (MR) the resonance frequency decreases with increasing sound amplitude, whereas the opposite happens when the membrane is strain hardening (SK). As the amplitude of the acoustic disturbance increases the total scattering cross section of a microbubble with a SK membrane tends to decrease, whereas that of a KV or a MR membrane tends to increase. The importance of strain-softening behavior in the abrupt onset of volume pulsations, that is often observed with small insonated microbubbles at moderately large sound amplitudes, is discussed. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2909553]

PACS number(s): 43.25.Ba, 43.25.Yw, 43.20.Px [AJS] Pages: 4059–4070

I. INTRODUCTION

The common approach of most studies modeling contrast agents is a direct transfer of the achievements of classical physical acoustics to biological systems. In the pioneering studies of microbubble pulsations in blood flow, contrast agents are commonly described by various forms of the Rayleigh–Plesset equation (Church, 1995; Frinking and De Jong, 1998). In one of the earlier attempts to model a contrast agent, Church used a generalized Rayleigh–Plesset model that accounted for the shell thickness and viscoelastic properties. In this manner he was able to show the effect of shell properties on the resonance frequency and sound attenuation in a liquid containing microbubbles. To this end, he used the Kelvin–Voigt constitutive law, which is essentially Hooke's law for an incompressible material and predicts the stresses developing on the shell membrane for small displacements. It is valid in the limit of small amplitude acoustic disturbances. Adopting a slightly different approach, Frinking and de Jong modeled the microbubble shell as a membrane of infinitesimal thickness and used simple linear models or semiempirical laws, respectively, for the description of shell elasticity and viscosity. They contacted simulations in the linear and nonlinear regime of acoustic disturbances and thus were able to point out the importance of higher harmonics in the scattered signal, as well as the effect of viscoelastic properties of the shell on the bubble response. However, upon comparing the predictions of their model with the available scattering data obtained at higher acoustic pressures, they reported failure to predict the dependence of scattering cross section on increasing acoustic pressure. Sboros *et al.* (2002) reached a similar conclusion when they compared the same models against their own measurements.

An effort towards a more rigorous theoretical description of radial pulsations of microbubbles in blood flow was made by Khismatullin and Nadim (2002). In that study the radial motion of a microbubble that is encapsulated by a viscoelastic membrane and surrounded by a slightly compressible viscoelastic liquid was examined, assuming that the viscoelastic properties of the shell and the liquid are described by the Kelvin–Voigt (KV) and the 4-constant Oldroyd models, respectively. In this fashion they were able to calculate resonance frequencies and damping coefficients for linearly pulsating microbubbles. As was already shown elsewhere (Church, 1995; Hoff *et al.*, 2000) because of membrane elasticity resonance occurs at higher frequencies than for the case free bubbles. However, their theory is restricted to small-amplitude oscillations only, hence the effect of the appearance of higher harmonics and subharmonics was restricted to the second harmonic response. The Church–Hoff model (Hoff *et al.*, 2000), is an adaptation of the Church model (Church, 1995), taken in the limit of small shell thickness in comparison with the radius. As an alternative approach, Sarkar *et al.* (2005) modeled the effect of shell dilatational elasticity through interfacial tension and its variation with shell interfacial area, while also including the effect of dilatational shell viscosity through a Newtonian viscoelastic model for the membrane material.

As will be seen in the following, this is a recurring issue with most models of contrast agents, namely their predictive value at large acoustic pressures is limited. In particular, it

[a)]Author to whom correspondence should be addressed. Electronic mail: pel@uth.gr

should be stressed that all the above models ignore changes in the material with varying sound amplitude, adopting a type of Hooke's law for the material's mechanical behavior. However, most materials exhibit a varying apparent elasticity modulus when they are subject to external disturbances of increasing intensity or increasing frequency. In the following, apparent elasticity modulus denotes the varying slope of the stress-strain relation of a particular material. Thus, there are materials, called strain hardening, whose stress-strain relationship exhibits a larger slope as deformations increase. This essentially amounts to an increased apparent elasticity modulus. A characteristic example of this type of material is that of the lipid bilayer that forms the red blood cell membrane as well as of certain polymers that are used in the manufacturing of contrast agents. On the other hand, if the stress-strain slope is reduced as deformations increase, the material is called strain softening, e.g., rubber. Such behavior is accounted for by the constitutive law describing the membrane material. The Skalak law (Skalak *et al.*, 1973) belongs to the class of constitutive laws describing materials that are strain hardening by nature and it is widely used for describing the mechanical behavior of the red blood cell membrane, while the Mooney–Rivlin law is a often used to characterize strain softening materials. The importance of these material properties has already been recognized in the modeling of blood cells or capsules in general (Barthès-Biesel *et al.*, 2002) where by the term capsule we refer to drops surrounded by an elastic membrane.

The scope of this paper is to emphasize the flow structure interaction aspect of contrast agent dynamics by cross examining the effect of membrane viscoelastic behavior along with that of external liquid attributes such as viscosity, compressibility, and nonlinearity in the acoustic disturbance. A detailed account of the model employed for the description of the microbubble is given in Sec. II, based on the model of Keller and Miksis (1980). The encapsulating shell is modeled as a thin membrane via one of the three membrane constitutive laws that were mentioned above, i.e., the Kelvin–Voigt, Mooney–Rivlin, and Skalak laws. The fluid and structure problems are coupled at the microbubble interface where the stress balance is imposed. The numerical methodology is briefly presented in Sec. III. As a final product the resonance frequency and scattering cross section of the microbubble are calculated for a wide parameter range in Sec. IV. The impact of the constitutive law on the interpretation of certain experimental observations is also stressed.

II. PROBLEM FORMULATION

We consider an encapsulated microbubble with initial radius R_0, submerged in a Newtonian liquid of density ρ_l, dynamic viscosity μ_l and static pressure P'_{st} taken to be at 1 bar. The microbubble consists of ideal gas encapsulated in a viscoelastic membrane. The latter is taken to be volume incompressible with shear modulus G_s and viscosity μ_s. The shell thickness δ is taken to be much smaller than the initial radius. Initially the membrane is at static equilibrium where it may develop uniform residual stresses, assuming a spherosymmetric configuration:

$$R_E = R_0 - u'|_{r'=R'_0}, \tag{1}$$

where $u'|_{r'=R'_0}$, is the radial displacement that produces the residual stresses and R_E the microbubble equilibrium radius that is free of any stresses. For stress free initial conditions $R_E=R_0$; throughout this study primed letters denote dimensional variables. The gas inside the microbubble exerts at the membrane a pressure $P'_{g,0}$ the variations of which are applied instantaneously and uniformly throughout the gas due to its negligible density. We also assume that the microbubble executes adiabatic oscillations. Consequently, each moment the pressure inside the bubble is correlated with the microbubble volume as

$$P'_g V'^{\gamma} = P'_{g,0} V'^{\gamma}_0, \tag{2}$$

with V'_0 denoting the initial microbubble volume and $\gamma=1.4$ the polytropic constant for an adiabatic process.

The bubble is insonated by a sinusoidal pressure disturbance in the far field

$$P'_{\infty}(t) = P'_{\mathrm{st}} + P'_{\mathrm{Ac}}(t') = P'_{\mathrm{st}}[1 + \varepsilon \sin(\omega_f t')], \tag{3}$$

with $v_f=1-10$ MHz ($\omega_f=2\pi v_f$) the forcing frequency lying in the ultrasound range, $P'_{\mathrm{Ac}}(t')$ the far field pressure disturbance, and ε the amplitude of the acoustic disturbance. The n-harmonic component of the scattering cross section is given (Hilgenfeldt *et al.*, 1998) by

$$\sigma' = 4\pi \frac{\int_0^{t'_f} r'^2 P'^2_{\mathrm{Sc}} dt'}{\int_0^{t'_f} P'^2_{\mathrm{Ac}} dt'}, \tag{4a}$$

$$\sigma'_{\mathrm{Sc},n} = 4\pi \frac{\int_0^{t'_f} r'^2 P'^2_{\mathrm{Sc},n} dt'}{\int_0^{t'_f} P'^2_{\mathrm{Ac}} dt'}, \tag{4b}$$

$$P'_{\mathrm{Sc}}(r',t') = P'_l(r',t') - P'_{\mathrm{st}} - P'_{\mathrm{Ac}}(t'); \tag{4c}$$

P'_{Sc} is the scattered pressure from the microbubble, registered in the host fluid at a distance r' from the microbubble's center of mass, while subscript n denotes the n-harmonic component of the scattering cross section. In the present study σ'_{Sc} is evaluated at the interface, in which case $r'=R'$ is the instantaneous external microbubble radius.

The initial external radius of the microbubble, R_0, is assigned as the characteristic length of the problem. Since the time scale of microbubble oscillations is determined by the external forcing frequency, ω_f, the characteristic time of the problem is $1/(\omega_f)$ and subsequently the characteristic velocity, $\omega_f R_0$. Finally, the characteristic pressure is defined via the characteristic velocity as $\rho_l \omega_f^2 R_0^2$.

A. Governing equations of the external liquid

The pulsating motion of the microbubble may exhibit very large velocities, especially as the amplitude of the acoustic disturbance increases as is the case when high me-

chanical index ultrasonic bursts are employed. As a result of its low viscosity, viscous effects may be neglected in the bulk of the host fluid, taken to be either water or blood, and a velocity potential, ϕ', may be introduced which simplifies the analysis significantly. In addition, inclusion of liquid compressibility is required in the model in order to account for fluid motion in the far field. When the Mach number of the flow is small but not negligible, $M=\omega_f R_0/(c)\ll 1$, based on the radial velocity of the microbubble interface, $\dot{R}' \sim \omega_f R_0$, the far field flow is compressible and is described by the wave equation. Near the bubble-host fluid interface the flow field can be described by the Laplacian to leading order (Prosperetti and Lezzi, 1986). In this fashion, and utilizing the known wave form for the pressure disturbance that is applied in the far field, the nonlinear ordinary differential equation describing spherosymmetric oscillations of a microbubble in a compressible liquid reads

$$(1-M\dot{R})R\ddot{R}+\left(\frac{3}{2}-\frac{M\dot{R}}{2}\right)\dot{R}^2=(1+M\dot{R})(P_l|_{r=R}-P_{\mathrm{st}}-P_{\mathrm{Ac}})+MR\frac{d}{dt}(P_l|_{r=R}-P_{\mathrm{Ac}}), \tag{5}$$

where R is the dimensionless external microbubble radius at time t, $\dot{R}=dR(t)/(dt)$, $\ddot{R}=d^2R(t)/(dt^2)$ and $P_l|_{r=R}$ is the dimensionless pressure of the host liquid calculated at the microbubble's interface. Equation (5) provides the instantaneous location of the bubble's interface once the liquid pressure is known. It is essentially the Keller–Miksis model (1980) describing moderate, fast, or even very fast radial oscillations of a bubble by properly accounting for compressibility effects when M is small but not negligible. The normal component of the viscous stress exerted on the microbubble reads in spherical coordinates as

$$\mathbf{n}\cdot\mathbf{X}_l'\cdot\mathbf{n}\big|_{r'=R'}=\mu_l\left(2\frac{\partial u_r'}{\partial r'}-\frac{2}{3}\boldsymbol{\nabla}'\cdot\mathbf{u}'\right), \tag{6}$$

where the second term on the right hand side arises as a result of compressibility even though it is not very important for small M; see Prosperetti and Lezzi (1986) for more details on the asymptotic validity of the Keller–Miksis model for small Mach numbers.

B. Modeling of the mechanical behavior of the membrane constitutive laws

The liquid pressure exerted on the membrane can be calculated via a stress balance that is applied on the membrane itself. In this fashion the microbubble model can be completed by correlating the pressure of the external liquid, $P_l|_{r=R}$, calculated on the interface of the bubble, with the instantaneous pressure inside the bubble, P_g, the viscous stresses in the liquid, and the viscoelastic stresses that develop on the membrane due to its radial deformation and velocity. Subsequent substitution in Eq. (5) provides a nonlinear ordinary differential equation that can be solved for the radial position and velocity of the interface.

When the shell thickness is infinitesimally small, a single force balance can be written for the gas–liquid interface,

$$[P_G'\mathbf{I}-P_l'\mathbf{I}+\mathbf{X}_l']\cdot\mathbf{n}=\sigma(\boldsymbol{\nabla}_s'\cdot\mathbf{n})\cdot\mathbf{n}-\nabla_s'\cdot\mathbf{X}_M', \tag{7}$$

where $\mathbf{I}$ denotes the unitary stress tensor, $\mathbf{n}$ the normal vector at the interface pointing towards the host fluid, $\boldsymbol{\nabla}_s'$ the surface gradient operator, σ the interfacial tension between the gas in the microbubble and the host liquid in the presence of the membrane, P_l', $\mathbf{X}_l'$ the pressure and viscous stress tensor, respectively, in the liquid, and $\mathbf{X}_M'$ the two-dimensional stress tensor containing the stresses that develop on the membrane surface as a result of its mechanical properties such as elasticity and viscosity, Pozrikidis, 1992; lower and upper case symbols in bold denote vectorial and tensorial quantities, respectively, throughout this study. A detailed presentation of the stresses that develop on the membrane, depending on the constitutive law that describes the mechanical behavior of the material that forms the membrane, is provided in the following.

C. Kelvin–Voigt model

One of the earlier used constitutive laws (Church, 1995) governing the mechanical behavior of the membrane is the Kelvin–Voigt law (KV) that relates the viscoelastic stresses to the strain and rate of strain tensors, $\boldsymbol{\Gamma}'$, $\dot{\boldsymbol{\Gamma}}'$, in a linear fashion,

$$\mathbf{X}_M'=2(G_s\boldsymbol{\Gamma}'+\mu_s\dot{\boldsymbol{\Gamma}}'),\quad \boldsymbol{\Gamma}'=\frac{1}{2}[\boldsymbol{\nabla}'\mathbf{u}'+(\boldsymbol{\nabla}'\mathbf{u}')^T],$$

$$\dot{\boldsymbol{\Gamma}}'=\frac{1}{2}[\boldsymbol{\nabla}'\mathbf{w}'+(\nabla'\mathbf{w}')^T], \tag{8}$$

where $\mathbf{u}'$ and $\mathbf{w}'$ are ascribed as the dimensional displacement and velocity vectors inside the membrane, respectively, G_s, μ_s signify the shell shear modulus and viscosity expressed in kg/(m s^2) and kg/(m s), respectively, and superscript T denotes the transpose of a tensor. We consider radial pulsations and neglect inertia effects in, and shape oscillations of, the shell, which is taken to be at equilibrium at all times. It should also be stressed that the above model is essentially Hooke's law, with the addition of a viscous term, and therefore is strictly valid for small membrane displacements. Nevertheless, different variations of it that are valid either for finite (Church, 1995; Khismatullin and Nadim, 2002) or infinitesimal (Frinking and De Jong, 1998; Hoff *et al.*, 2000; Sarkar *et al.*, 2005) shell thickness, are extensively used in the literature over a very wide range of pressure amplitudes and viscoelastic parameter values. Consequently, we also make use of it in the present study, for the purpose of comparing its validity range against other more relevant constitutive laws that account for changes in the apparent material properties, such as the shear modulus, with increasing pressure amplitude or frequency of sound.

Following Church (1995) and Khismatullin and Nadim (2002) we consider the r component of the momentum and continuity equations for the shell, integrate in the radial direction across the shell, and take the stress equilibrium be-

tween the membrane, the external liquid, and the internal gas. Thus, we relate the liquid pressure exerted on the membrane, $P_l|_{r=R}$ with the instantaneous pressure inside the bubble. Owing to the small membrane thickness, in comparison with the microbubble radius, we proceed by taking the ratio between the shell thickness, δ, and bubble external radius, R', to be negligibly small throughout the bubble pulsation. In this fashion and neglecting any initial strain, the liquid pressure $P_l|_{r=R}$ reads in dimensionless form:

$$P_l|_{r=R} = \left[\frac{2}{We} + P_{\mathrm{st}}\right]\left(\frac{1}{R}\right)^{3\gamma} - \frac{2}{WeR} - \frac{4}{\mathrm{Re}_l}\frac{\dot{R}}{R} - \frac{4m}{\mathrm{Re}_l}\frac{\dot{R}}{R^2} - 2\frac{3G}{R}\left(\frac{R^2}{(1-u)^2} - 1\right), \tag{9}$$

where, $We = \rho_l \omega_f^2 R_0^3/(\sigma)$ denotes the Weber number comparing inertia forces in the liquid due to the external forcing with surface tension, $\mathrm{Re}_l = \rho_l \omega_f R_0^2/(\mu_l)$ and $m = 3\mu_s\delta/(\mu_l R_0)$ the Reynolds number of the external liquid, comparing forces of inertia with viscous dissipation, and the relative fluid to membrane viscosity, respectively, and $G = \delta G_s/(\rho_l \omega_f^2 R_0^3)$ the dimensionless shear stress modulus that compares elastic with inertia forces. The above equation holds when the membrane remains very thin while undergoing small displacements during the microbubble pulsation, and is essentially the Church–Hoff model for viscoelastic membranes (Hoff *et al.* 2000; Sarkar *et al.* 2005). It assumes an incompressible shell with a simplified expression for the shell displacements, $u' \approx R'^2(R' - R_0)/r'^2$.

Upon replacing Eq. (9) in Ref. 5 we obtain an ordinary nonlinear ordinary differential equation (ODE) with dimensionless time t as the only independent variable, and the external microbubble radius, R, as the only unknown. Coupled with the appropriate initial conditions, it can be integrated to provide the radial position and velocity of the membrane, and through them the rest of the important dependent variables of the flow. We allow for residual stresses at the onset of bubble vibration via the initial displacement u_0:

$$R(t=0) = 1, \quad \dot{R}(t=0) = 0, \quad u|_{r=1} = u(r=1, t=0) = u_0. \tag{10}$$

D. Strain hardening and strain softening materials

Most materials do not respond to external forces through a constant apparent elasticity modulus. Rather, they exhibit a varying slope in their stress strain relation at large deformations or at very abrupt changes of pressure, as is the case with ultrasound. Two very common families of materials characterized by this kind of response are strain softening and strain hardening materials. In the former case the membrane material is such that its shear modulus is reduced as strain grows, whereas the opposite is true for the latter type of membrane materials. Most polymer shelled air filled particles are probably strain softening, e.g., Sonazoid, see also Sarkar *et al.* (2005). Consequently taking into consideration the specific material behavior will enhance the predictive capabilities of the model. In the following we present the governing equations for the mechanical behavior of a viscoelastic membrane at equilibrium, taken to be infinitesimally thin in comparison with the radius as is normally the case with contrast agents used in ultrasound diagnostic imaging, for different types of nonlinear response.

We associate the elastic tension tensor, $\mathbf{X}'_{\mathbf{M}}$, on a deformed two-dimensional surface with the Green–Lagrange surface deformation tensor via the strain energy function $w(I_1, I_2)$, where I_1, I_2 denote the 2d strain invariants. The strain energy $w(I_1, I_2)$ depends on the nature of the membrane material and assumes different forms as the mechanical behavior of the membrane changes. A typical strain energy describing a very thin sheet of an isotropic, volume-incompressible, rubber-like material with strain-softening behavior, is the one provided by the two-dimensional Mooney–Rivlin (MR) law (Barthès-Biesel *et al.* 2002),

$$w^{\mathrm{MR}} = \frac{G_{\mathrm{MR}}}{2}\left[(1-b)\left(I_1 + 2 + \frac{1}{I_2+1}\right)' + b\left(\frac{I_1+2}{I_2+1} + I_2 + 1\right)\right], \tag{11a}$$

$$X'^{\mathrm{MR}}_{M11} = \frac{G_{\mathrm{MR}}}{\lambda_1\lambda_2}\left(\lambda_1^2 - \frac{1}{(\lambda_1\lambda_2)^2}\right)[1 + b(\lambda_2^2 - 1)], \tag{11b}$$

with G_{MR} the Mooney–Rivlin surface shear modulus expressed in $\mathrm{kg/s^2}$ and λ_1, λ_2 the principal extension ratios. When the indices are exchanged in Eq. (11b) the stress component along principal direction 2 is obtained, whereas for spherosymmetric pulsations $\lambda_1 = \lambda_2$. The case with $b=0$ corresponds to a neo-Hookean membrane, whereas as b tends to zero the membrane becomes softer; b ranges between 0 and 1. It should also be noted that the Mooney–Rivlin constitutive law allows for unrestricted area dilatation that is compensated by progressive thinning of the membrane, whereas the case with $b=0$ (neo-Hookean membrane) represents the appropriate linear stress strain relationship that accounts for the change in metric properties during deformation.

One of the most widely used constitutive laws pertaining to strain-hardening membranes is the one developed by Skalak *et al.* (1973) in order to model the lipid bilayer structure surrounding the red blood cell,

$$w^{\mathrm{SK}} = \frac{G_{\mathrm{SK}}}{2}(I_1^2 + 2I_1 - 2I_2 + CI_2^3), \tag{12a}$$

$$X'^{\mathrm{SK}}_{M11} = \frac{G_{\mathrm{SK}}}{\lambda_1\lambda_2}\{\lambda_1^2(\lambda_1^2 - 1) + C(\lambda_1\lambda_2)^2[(\lambda_1\lambda_2)^2 - 1]\}, \tag{12b}$$

with G_{SK} denoting the Skalak (SK) surface shear modulus expressed in $\mathrm{kg/s^2}$. Parameter C in the above equations is always positive and controls the extend of area incompressibility of the membrane. In the case of red blood cells $C \gg 1$ in order to accommodate the almost incompressible nature of the membrane area. Nevertheless, this is quite a general law that is used for strain-hardening membranes.

Membrane viscosity can also be accounted for via a linear Newtonian term that is added to the elastic stresses and

involves the membrane velocity, $\mu_{2d}2/(\lambda_i)\partial\lambda_i/(\partial t')$; $1/(\lambda_1)\lambda_1/(\partial t')$ is the first principal component of the surface rate of strain tensor (Barthès-Biesel *et al.* (2002)) and μ_{2d} the two-dimensional membrane viscosity expressed in kg/s. In the absence of material characterization data we take both shear and dilatational viscosities of the membrane to be μ_{2d}.

For a two-dimensional membrane the viscoelastic contributions to the force balance Eq. (7) enter through the surface divergence of the surface stress tensor. In the context of spherosymmetric oscillations only the radial component of the divergence has a nonzero contribution, while the membrane principal extension ratio due to its radial displacement reads

$$\lambda(t') = \lambda_1 = \lambda_2 = \frac{R'(t')}{R_E} = \frac{R'(t')}{R_0 - u'|_{r'=R_0'}}. \tag{13}$$

It should be noted that, in view of the isotropy in the deformation that is assumed in the present study, the main elastic effect that is assessed here is that of area dilatation due to pulsation. This is reflected in the area dilatation modulus K that is defined as the ratio between the isotropic elastic tension and the relative area change (λ^2-1) for small deformations. It turns out that a Hookean material with Poisson ratio $\nu_s=1/2$, a MR material and a SK material with $C=1$ all are characterized by area dilatation modulus K equal to $3G_{2d}$, where G_{2d} denotes the 2d shear modulus of the membrane. In the case of a KV membrane $G_{2d}=G_s\delta'$. In the following we will use this parameter in order to compare the behavior of membranes with the same area dilatation modulus that obey different constitutive laws. Thus, for a MR membrane we will use $G_{\mathrm{MR}}=G_s\delta'$, and similarly for a SK membrane we will use $G_{\mathrm{SK}}=G_s\delta'$. In the latter type of membrane it can be seen that $K=G_{\mathrm{SK}}(2C+1)$. Consequently, in the following we will ensure that the area dilatation modulus is the same when we compare strain hardening with strain softening and Hookean (i.e., KV) membranes. In addition we will present a separate set of results in order to assess the effect of parameter C on SK membranes, whichessentially represents the effect of increasing the area dilatation modulus for a given SK membrane.

Upon introduction into the stress equilibrium equation that holds on the membrane, Eq. (7), of the MR constitutive law and reverting to dimensionless formulation in a manner analogous to the Kelvin–Voigt model, we obtain the following expression for the liquid pressure on the membrane:

$$P_l|_{r=R} = \left(\frac{1}{R}\right)^{3\gamma}\left[P_{\mathrm{st}} + \frac{2}{We} + 2G\left[1-\left(1-u|_{r=1}\right)^6\right]\left[1 + b\left[\left(\frac{1}{1-u|_{r=1}}\right)^2 - 1\right]\right]\right] + \frac{2}{WeR} - \frac{4\dot{R}}{\mathrm{Re}_l R} - \frac{2G}{R}\left[1-\left(\frac{1-u|_{r=1}}{R}\right)^6\right]\left[1 + b\left[\left(\frac{R}{1-u|_{r=1}}\right)^2 - 1\right]\right] - \frac{4\dot{R}}{\mathrm{Re}_l R^2}m, \tag{14}$$

where $G=G_{\mathrm{MR}}/(\rho_l\omega_f^2R_0^3)$, $m=\mu_{\mathrm{MR}}/(\mu_l R_0)$, are the dimensionless numbers that arise and $u|_{r=1}=u_0$ the initial membrane displacement that determines the residual stresses inside the membrane. In a similar fashion, for a SK membrane the following expression for the dimensionless liquid pressure, $P_l|_{r=R}$, is derived:

$$P_l|_{r=R} = \left(\frac{1}{R}\right)^{3\gamma}\left[P_{\mathrm{st}} + \frac{2}{We} + 2G\left[\left(\frac{1}{1-u_0}\right)^2(1-C) + C\left(\frac{1}{1-u_0}\right)^6 - 1\right]\right] + - \frac{2}{WeR} - \frac{4\dot{R}}{\mathrm{Re}_l R} - \frac{2G}{R}\left[\left(\frac{R}{1-u_0}\right)^2(1-C) + C\left(\frac{R}{1-u_0}\right)^6 - 1\right] - \frac{4\dot{R}}{\mathrm{Re}_l R^2}m, \tag{15}$$

where $G=G_{\mathrm{SK}}/(\rho_l\omega_f^2R_0^3)$, $m=\mu_{\mathrm{SK}}/(\mu_l R_0)$, are the dimensionless numbers that arise. Finally, substituting the above expressions in Eq. (5) we obtain a nonlinear ODE describing the time variation of the radial position and velocity of a MR or a SK membrane with Eq. (10) providing the initial conditions.

E. Linear theory

Starting with the Kelvin–Voigt model we apply infinitesimal perturbations to the basic solution, which is the microbubble equilibrium, assuming that the membrane is free of residual stresses at $t=0$, i.e., $u_0=0$. Applying small disturbances on the external radius as well as the far field pressure,

$$R = 1 + \varepsilon R_d, \quad P'_\infty = P_{\mathrm{st}} + \varepsilon P_{\mathrm{st}}\sin(t), \quad \varepsilon \ll 1, \tag{16}$$

introducing the above expansions in the governing equations and retaining terms of order ε only, we obtain

$$\left[1 + \frac{4M}{\mathrm{Re}_l} + \frac{4Mm}{\mathrm{Re}_l}\right]\ddot{R}_d + \left[-\frac{2M}{We} + \frac{4}{\mathrm{Re}_l} + \frac{4m}{\mathrm{Re}_l} + 3\gamma M\left(\frac{2}{We} + P_{\mathrm{st}}\right) + 12GM\right]\dot{R}_d + \left[3\gamma\left(\frac{2}{We} + P_{\mathrm{st}}\right) - \frac{2}{We} + 12G\right]R_d = -\varepsilon P_{\mathrm{st}}\sin(t) - \varepsilon P_{\mathrm{st}}M\cos(t). \tag{17}$$

The above linear equation furnishes the dimensionless resonance frequency, $\omega_I=\omega_{\mathrm{Res}}/\omega_f$, and damping, s, of the microbubble via the roots of its characteristic polynomial, $\omega = s+i\omega_I$. Following the same procedure for the linear dynamics of MR and SK membranes we recover the same ω provided $\mu_{\mathrm{MR}}=3\delta'\mu_s$, $G_{\mathrm{MR}}=G_S\delta'$, and $\mu_{\mathrm{SK}}=3\delta'\mu_s$, $G_{\mathrm{SK}}=G_S\delta'$. We conclude that the microbubble behavior is independent of the membrane constitutive law if membrane displacements are small. For the SK membrane in addition to the above conditions it is required that $C=1$. For any other C value SK membranes behave differently from MR or KV membranes even in the linear regime. In all other cases the microbubble behavior is heavily dependent on the constitutive law and this is an effect that will be demonstrated in the following sections. In the same manner, the effect of the initial residual stresses of the membrane on the microbubble

scattering cross section is an additional issue that must be investigated in connection with the membrane material law.

For small external perturbations and when the microbubble has reached the phase of steady oscillations, we can neglect any transient effects and calculate the scattering cross section from Eq. (4) by employing the solution of the linearized problem Eq. (17)

$$\frac{\sigma'_{\mathrm{Sc}}}{4\pi R_0^2} = \frac{1}{\left[\left(\frac{F_3}{F_1}\right)^2 - 1\right]^2 + \delta_t^2} \sqrt{\frac{1+M^2}{F_1^2}}, \quad \delta_t = \frac{F_2}{F_1}, \tag{18}$$

where F_1, F_2, and F_3 are the factors multiplying radial acceleration, radial velocity, and radial position, respectively, in Eq. (17). In Eq. (18) the scattering cross section is evaluated on the undisturbed microbubble interface.

III. NUMERICAL IMPLEMENTATION

We use the fourth order Runge–Kutta (RK) integrator in order to solve the second order nonlinear ordinary differential equation governing the motion of the membrane. The time step of the numerical integration is fixed and is selected so that enough time steps are afforded within one period of the forced or the natural radial pulsations. Eventually results are tested for convergence with respect to time step and agreement with linear theory is established, whenever this is possible. The same approach has been successfully employed in the past for simulating large amplitude oscillations of free bubbles (Pelekasis *et al.* 2004) near the Blake threshold. In order to compute the integral $\int_0^{t_f}(RP_{\mathrm{Sc}})_n^2 dt$ in Eq. (4) we implement Parseval's identity

$$\int_0^{t_f} f(t)^2 dt = \frac{t_f}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2), \tag{19}$$

where t_f is the duration of the time integration and a_n, b_n $n=1,2,\ldots\infty$ are the Fourier coefficients of $f(t)$ which are calculated through the fast Fourier transform (FFT) algorithm. The zeroth order coefficient is not included in the right hand side of Eq. 20 since it corresponds to the time average of $f(t)$, which will be zero in view of Eq. (4).

The validity of the above numerical implementation was investigated in the case of small pressure disturbances, where numerical results are compared against the predictions of linear theory. The dimensionless scattering cross section was calculated numerically, Eqs. (5) and (14) or (15), and theoretically, Eq. (18), as a function of the forcing frequency for small sound amplitudes and a standard set of parameter values provided in the following section. Agreement between computations and linear theory was always achieved. It was also reaffirmed that when $\varepsilon \ll 1$ and $C=1$ the three constitutive laws predict the same dynamic behavior for the microbubble. However, when nonlinear perturbations are applied the three constitutive laws can exhibit quite different dynamic behavior as will be seen in the following.

It should be stressed that in graphs depicting dimensionless scattering cross section shown in the following, the external frequency, ν_f, will be scaled with resonance frequency, ν_{Res}, obtained from the characteristic polynomial of Eq. (17); ω_I goes like $1/\omega_f$ hence ω_{Res} is appropriately independent of the forcing frequency. They are both in the MHz regime, which is also the frequency range of diagnostic imaging. Scattering cross section is scaled with the microbubble interfacial area $4\pi R_0^2$.

IV. RESULTS AND DISCUSSION

In this section a detailed parametric study is presented on the effect of the microbubble properties, e.g., size and mechanical properties of the membrane, and the ultrasound characteristics, i.e., amplitude and frequency, on the response of an insonated contrast agent. The resonance frequency as well as the scattering cross section of the fundamental and higher harmonics are monitored when moderate or large acoustic disturbances are applied. The parametric study is conducted for the KV, the MR, and the SK membrane constitutive laws while the effect of residual stresses on backscatter is also investigated. Results with the KV mode are only marginally presented in order to show the limitations of this model for large sound amplitudes. The parameters of the problem are based on those provided from experimental and theoretical studies available in the literature of contrast agent research. In particular, an estimate for the membrane stiffness and friction was obtained for Sonovue™ (Gorce *et al.*, 2000), Sonazoid (Sarkar *et al.* 2005), Albunex (De Jong and Hoff, 1993), and different polymer encapsulated air bubbles (Hoff *et al.* 2000) by fitting the Church–Hoff model to their experimental recordings of scattering cross section and sound attenuation. Based on the same set of experimental data we use $\delta \approx 15$ nm as a characteristic membrane thickness and take $R_0 \approx 3$ μm as an indicative microbubble radius. Based on the same studies we allow for variation of G_s between 35 and 105 MPa and μ_s between 0.6 and 1.6 kg/(m s) and, unless otherwise specified, we use the former values for G_s and μ_s as characteristic. Nevertheless, as will be seen in the following, the area dilatation modulus, K, is the determining factor in the case of spherosymmetric bubble pulsations. Consequently, fitting data obtained in the regime of low acoustic disturbances can provide $K=3G_{\mathrm{MR}}$ for MR membranes or $K=G_{\mathrm{SK}}(1+2C)$ for SK membranes. We also set parameters b and C to 0 and 1 for the MR and the SK constitutive laws, respectively, and provide the framework for estimating these parameters based on measurements.

The physical properties of water are used for the host liquid; $\rho_l=998$ kg/m^3, $\mu_l=0.001$ kg/(m s), $C_l=1500$ m/s. In the absence of any reliable data on membrane porocity, the interfacial tension σ is set to the average of the gas-membrane and liquid-membrane tensions (Church, 1995; Khismatullin and Nadim, 2002). It is almost the same as the gas-host fluid interfacial tension, 0.072 kg/s^2, for the case of a shell with very small thickness. In any case it does not significantly affect microbubble response.

Experimental measurements, Hoff *et al.* (2000) among others, indicate that the scattering cross section from encapsulated bubbles is weaker than the one obtained from free bubbles. It was argued by Khismatullin and Nadim that this is basically due to membrane viscosity rather than elasticity.

Even though this is a valid argument it will be seen in the following that if the proper constitutive law for the membrane is not known, predictions of the resonance frequency based on simplified models may be in significant error. Resonance frequency is heavily dependent on the membrane area dilatation modulus and it could be the case that the scattered signal from a microbubble is relatively weak simply because the ultrasonic beam is out of resonance. The total scattering cross section as a function of forcing frequency, subject to increasing sound amplitude and for the four types of bubble behavior examined in the present study, i.e., a free bubble and a microbubble with a KV, a MR or a SK membrane, is presented in Fig. 1. The response refers to the state of pulsation at which the microbubble performs steady oscillations, i.e., after the initial transient period has elapsed. For reference we note that the dimensional scattering cross section for the Kelvin–Voigt model in Fig. 1(b) at resonance, becomes 1244 μm^2 in dimensional form upon multiplication by $4\pi R_0^2$ with $R_0=3$ μm.

Increasing the sound amplitude affects encapsulated bubbles by varying the effective area dilatation modulus of their membrane, i.e., $X'_M/(\lambda^2-1)$ for isotropic tension. In an average sense over one period of volume pulsation after initial transients have elapsed, λ^2-1 essentially represents the relative area dilatation $\Delta A/A$ and for a material obeying Hooke's law the ratio $X'_M/(\lambda^2-1)$ is a constant that equals the area dilatation modulus (Barthès-Biesel *et al.*, 2002). Any deviation from this behavior is identified as nonlinearity of the material. This is true for membranes that obey the MR as well as the SK constitutive laws. It results in an increase in the total scattering cross section for MR membranes and a corresponding decrease for SK membranes for reasons to be explained in detail in the following. Encapsulated bubbles tend to scatter a smaller amount of radiated energy due to the additional damping of the shell, as indicated by comparing peaks among graphs corresponding to free and encapsulated bubbles in Fig. 1. On the other hand, due to the elasticity of the encapsulating shell they can store energy, which they can then scatter back to the surrounding fluid at resonance. For microbubbles of the size relevant to our study and for large sound amplitudes, scattered energy primarily depends on the bubble radius and velocity and is called active scatter (Hilgenfeldt *et al.*, 1998). Nevertheless, especially for larger bubbles, viscous damping due to the encapsulating shell dominates, hence the decreased scatter from encapsulated microbubbles versus free bubbles. The combined result of these effects is clearly illustrated in Fig. 1, based on which it can be surmised that the attenuation or intensification of volume oscillations, that is mostly evident in high acoustic amplitudes, determines the energy scatter from the microbubble. This is corroborated by Figs. 2 and 3 where the radial displacement, velocity, and scattered pressure on an encapsulated bubble are plotted at resonance, for increasing sound amplitude. Due to the effective hardening of SK materials, i.e., the effective area dilatation modulus increases, the membrane displacement and velocity at resonance increases very mildly as the amplitude of the disturbance increases, Figs. 3(a) and 3(b). Consequently, when the sound amplitude increases the total scattering cross section decreases due to the

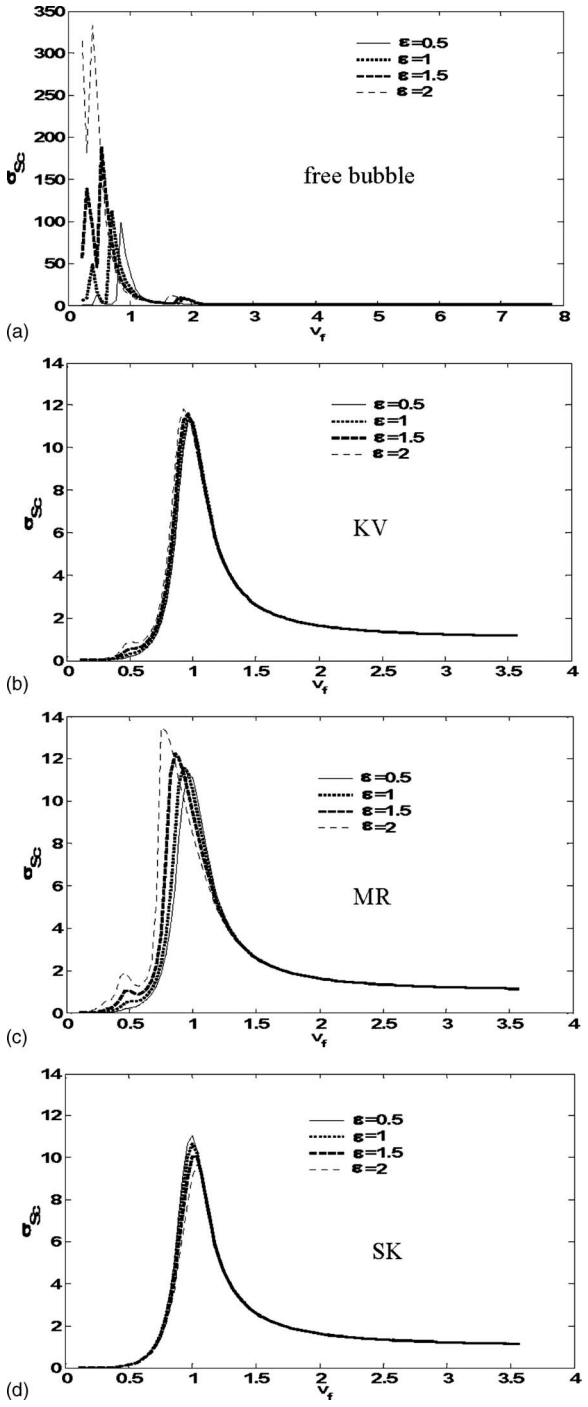


FIG. 1. Total scattering cross section vs scaled forcing frequency, when $\varepsilon=0.5$, 1, 1.5, and 2, for (a) a free bubble, (b) a KV, (c) a MR ($b=0$), and (d) a SK ($C=1$) membrane; $v_{res}=1.28$ and 2.8 MHz for a free bubble and an encapsulated microbubble, respectively.

disproportionately small increase of the microbubble's active scatter in comparison with the external disturbance. When MR membranes are subject to a sound field of increasing amplitude their effective area dilatation modulus decreases, which leads to enhancement of radial displacement and velocity, Figs. 2(a) and 2(b), that is larger than expected based solely on the amplitude of the acoustic disturbance. This is clearly manifested in the amplified total scattering cross section in Fig. 1(c). Comparing the level of total scatter between the three constitutive laws under examination, a material

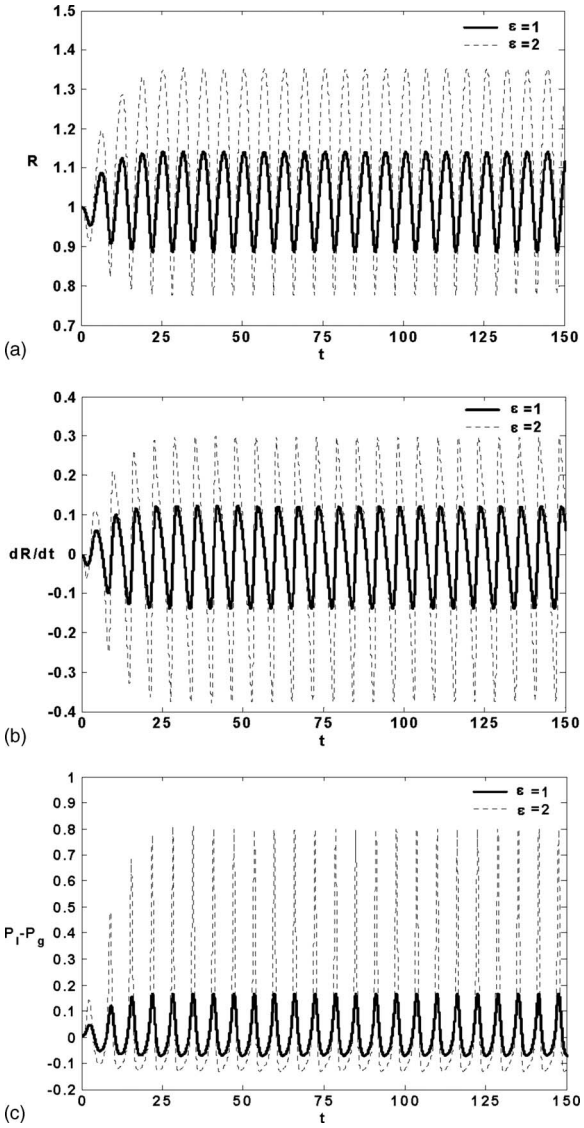


FIG. 2. Time evolution of the (a) external microbubble radius, (b) interfacial velocity, and (c) pressure load ($P_l - P_g$), for a MR membrane on resonance when $\varepsilon = 1$ and 2 ($v_f = 2.7$ and 2.4 MHz, respectively); $b = 0$.

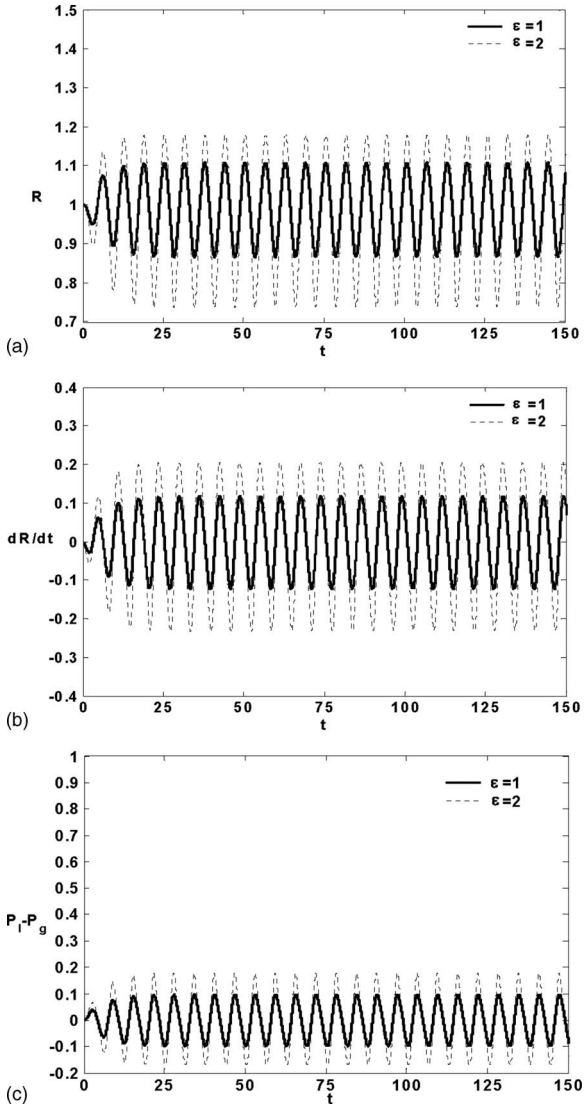


FIG. 3. Time evolution of the (a) external microbubble radius, (b) interfacial velocity, and (c) pressure load ($P_l - P_g$), for a SK membrane on resonance when $\varepsilon = 1$ and 2 ($v_f \approx 3.3$ MHz for both cases); $C = 2$.

obeying the KV law is only moderately affected by the amplitude of sound exhibiting a slight increase in the total scatter. Overall it can be argued that MR membranes permit larger deformations than SK membranes and consequently tend to scatter more echo through changes in the microbubble volume (active scatter).

As can be gleaned from Figs. 2(c) and 3(c), SK membranes develop larger extensional pressure loads for the same amplitude of the external disturbance after the initial transient has elapsed, when compared against MR membranes. In fact, this effect is intensified as their strain hardening nature is accentuated by increasing parameter C. This may explain experimental observations of shell cracking (Bloch *et al.*, 2004), at sonication before any significant area dilatation takes place. Most likely the membrane exhibits small defects at regions where excessive in plate tensions develop as a result of the large extensional load, eventually tearing the membrane apart. The particular polymer shelled contrast agent, BG1135, reportedly does not exhibit any significant harmonic or subharmonic content until the moment of cracking, in a fashion similar to strain hardening membranes as will be seen in the following. On the contrary, strain softening membranes exhibit very large compressive loads, Fig. 2(c), that may cause severe deformation and buckling of the shell (Dollet *et al.* 2008).

Figures 4(a) and 4(b) depict average area dilatation, $\Delta A/A$, as a function of sound amplitude, ε, during the phase of steady pulsation for a MR and a SK membrane. Such plots can be generated from optical measurements of contrast agent pulsation, with ε viewed as a measure of the load felt by the shell due to changes in the liquid pressure as a result of the acoustic excitation. In fact, it has been observed using the lipid shelled contrast agent BR14 (Emmer *et al.*, 2007), that there is an amplitude threshold for the onset of microbubble pulsation beyond which there is an abrupt increase in the area dilatation during steady pulsation. As the size of the microbubbles decreases, this type of threshold behavior was observed at smaller sound amplitudes.

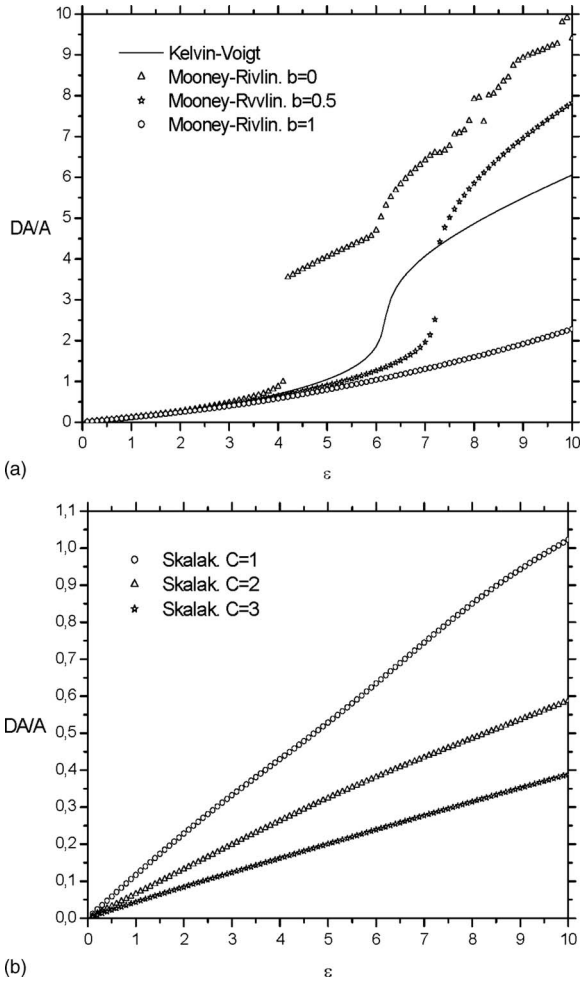


FIG. 4. Area dilatation vs sound amplitude for (a) a MR membrane with different b values, $b=0$, 0.5, and 1 (the behavior of a KV membrane is also shown for reference) and (b) a SK membrane with different C values, $C=1$, 2, and 3; $v_f=1.7$ MHz.

Indeed, calculations of the area dilatation at steady pulsation for a MR microbubble that is insonated by an increasing sound amplitude, reveal that as the membrane becomes softer, b approaches zero, the response becomes more and more abrupt, and appears at a lower amplitude threshold, Fig. 4(a). Varying the bubble size shows that the response of smaller microbubbles deviates from linearity at lower amplitudes. The reason for this abrupt increase in area dilatation for small bubbles lies in the change in resonance frequency with sound amplitude. The microbubbles used in the simulations shown in Fig. 4(a) are driven below resonance. As the amplitude of sound increases their resonance frequency decreases, owing to the strain-softening nature of the shell, until it hits the forcing frequency in which case an intense signal is obtained. The change in resonance frequency is faster for softer membranes hence the steep rise in area dilatation when $b=0$. This effect is present when the Kelvin–Voigt model is employed, despite the fact that it ignores material nonlinearity, and is attributed solely to the increasing effect of inertia with nonlinearity which also decreases resonance frequency. This is a well known result from nonlinear bubble dynamics that can be, however, significantly accentuated when material behavior is also taken into consideration,

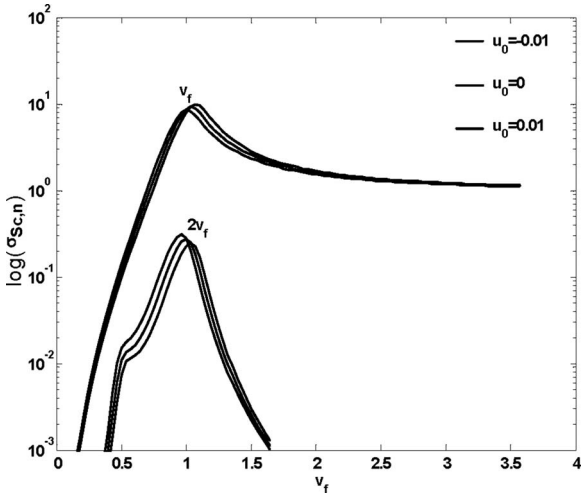


FIG. 5. Fundamental and second harmonic components of the scattering cross section vs forcing frequency for a SK membrane with $C=1$ and nonzero residual stresses at $t=0$, $u_0=u(t=0)=-0.01$, 0 and 0.01; $\varepsilon=2$.

as illustrated in Fig. 4(a). In fact, such plots can be used, along with optical measurements, as a means to characterize the membrane by estimating b and consequently determine the degree of softness. Larger microbubbles have smaller linear resonance frequencies, which may already be below the forcing frequency in which case vibration onset is not observed, hence an increase in amplitude will instigate nonlinear resonance sooner, i.e., at a lower amplitude. Such bubbles exhibit a slow increase in the gradient of the dilatation versus amplitude curve that is characteristic of strain softening membranes and determines the extent of nonlinearity in the material behavior. A similar parameter estimation can be employed for SK membranes through a plot like Fig. 5(b) illustrating the effect of membrane hardness, C, on area dilatation versus sound amplitude curves. In this case the microbubble is again driven below resonance. However, the resonance frequency of a strain-hardening material increases with increasing amplitude and the possibility for resonance is eliminated. The response of the area dilatation versus amplitude curves is typical of strain-hardening materials with a decreasing gradient as ε increases.

The effect of varying shear modulus G_s on the resonance frequency is shown in Table I as a function of the constitutive law and the sound amplitude. The resonance frequency for a given set of parameters corresponds to the maximum in the first harmonic components of the scattering cross section over the range of applied forcing frequencies. As was already known from previous studies (Church 1995; Khismatullin and Nadim 2002), for all three constitutive laws the resonance frequency increases with increasing membrane elasticity. As the amplitude of the acoustic disturbance increases, increasing ε, the extent of nonlinearity increases and as a first effect one notices a decrease in resonance frequency, slight for KV and more intense for MR membranes in the above, this is a well-known result from weakly nonlinear theory of free bubble dynamics. However, when the SK constitutive law is used the resonance frequency exhibits a slow increase, indicating a progressive stiffening of the membrane and a concomitant loss of effective system inertia. These

TABLE I. Dimensional resonance frequency of the first harmonic component as a function of area dilatation modulus and acoustic amplitude, recovered from numerical simulations.

	Free bubble		
ε	ν_{res} (MHz)		
0.5	1.1		
1	1		
1.5	0.8		
2	0.6		
	Mooney–Rivlin – ν_{res} (MHz) ($b=0$)		
	G_s (MPa)		
ε	35	70	105
0.5	2.7	3.7	4.4
1	2.7	3.6	4.4
1.5	2.6	3.5	4.3
2	2.4	3.4	4.2
	Skalak – ν_{res} (MHz) ($C=1$)		
	G_s (MPa)		
ε	35	70	105
0.5	2.8	3.7	4.4
1	2.8	3.7	4.4
1.5	2.9	3.7	4.4
2	2.9	3.7	4.4
5	3.4	4.1	4.6
	Skalak – ν_{res} (MHz) ($C=2$)		
	G_s (MPa)		
ε	21	42	63
0.5	2.8	3.7	4.4
1	2.8	3.7	4.4
1.5	2.8	3.7	4.4
2	2.9	3.7	4.4
5	3.4	4	4.5

TABLE II. Harmonic content of the scattering cross section for (a) a KV, (b) a MR ($b=0$), and (c) a SK ($C=1$) membrane, when $\varepsilon=0.5, 1, 1.5$ *and* 2. To obtain the actual dimensions one has to multiply the harmonic content by $4\pi R_0^2=113\ \mu m^2$, where $R_0=3\ \mu m$ in these simulations.

	KELVIN–VOIGT			
	$\sigma_{Sc,n}$			
ε	1st Harmonic	2nd Harmonic	3rd Harmonic	4th Harmonic
0.5	11.02	0.14	⋯	⋯
1.0	10.58	0.66	0.044	⋯
1.5	10.07	1.16	0.146	0.019
2.0	9.43	1.5	0.28	0.057
	MOONEY–RIVLIN ($b=0$)			
	$\sigma_{Sc,n}$			
ε	1st Harmonic	2nd Harmonic	3rd Harmonic	4th Harmonic
0.5	10.72	0.34	0.011	⋯
1.0	10.06	0.86	0.085	0.01
1.5	9.1	1.4	0.27	0.058
2.0	8.13	1.91	0.62	0.22
	SKALAK ($C=1$)			
	$\sigma_{Sc,n}$			
ε	1st Harmonic	2nd Harmonic	3rd Harmonic	4th Harmonic
0.5	10.88	0.02	⋯	⋯
1.0	10.45	0.085	⋯	⋯
1.5	9.76	0.145	⋯	⋯
2.0	9.26	0.25	0.01	⋯

effects can also be gleaned from the evolution of total scattering cross section for different values of ε, shown in Fig. 1. The deviation in resonance frequency between the predictions of Hooke's law, as manifested in the KV model, and those from the MR and SK constitutive laws is on the order of a few tenths of a MHz which is not negligible, given the sensitivity of modern imaging techniques, and keeps increasing with increasing amplitude of sound. It should also be stressed that it is the area dilatation modulus that determines the microbubble response. In fact, increasing C but varying G_s so that the product $G_{SK}(2C+1)$ remains constant very closely reproduces the values of resonance frequency; note that $G_{SK}=G_s\delta$, for amplitudes ε as large as 5; see also Table I. The same is true for the harmonic scatter.

Emphasis should also be placed on the harmonic content of the scattered signal since this finds extensive use in modern techniques of nonlinear signal processing (Burns *et al.*, 2000; Sarkar *et al.*, 2005). Table II shows the harmonic content of the scattering cross section of a MR and a SK membrane with varying dilatation modulus, in response to an acoustic disturbance of increasing amplitude. As was seen from Figs. 2 and 3, soft membranes exhibit larger displacements and velocities in comparison with hard membranes. As a result the amount of energy that is returned to the host fluid is scattered at lower frequencies with respect to KV and SK membranes. In addition, the content of the scattered signal in harmonic components, for given amplitude of the acoustic disturbance, is also increased which makes strain-softening membranes exceptionally useful for diagnostic tools where harmonic imaging is a preferred modality. In fact, as the amplitude of sound increases, the scattered signal from the fundamental harmonic becomes weaker at resonance, compared to that from a strain-hardening membrane, due to the appearance of higher harmonics. Another important aspect of the microbubble response at large amplitudes is the appearance of a subharmonic, $\nu_f/2$, signal in the backscatter that is especially evident for MR membranes and that can be quite useful for nonlinear image processing (Sarkar *et al.*, 2005); see also Fig. 6. In general, a rich harmonic content from a certain contrast agent is a clear indication of a strain-softening membrane.

The effect of the residual stresses on the scattering cross section is almost nonexistent for MR or KV membranes. This is not the case when the membrane material obeys the Skalak law. The more strain hardening the material is, the more intense is the shift of the resonance frequency to higher values as well as the scatter of the fundamental harmonic,

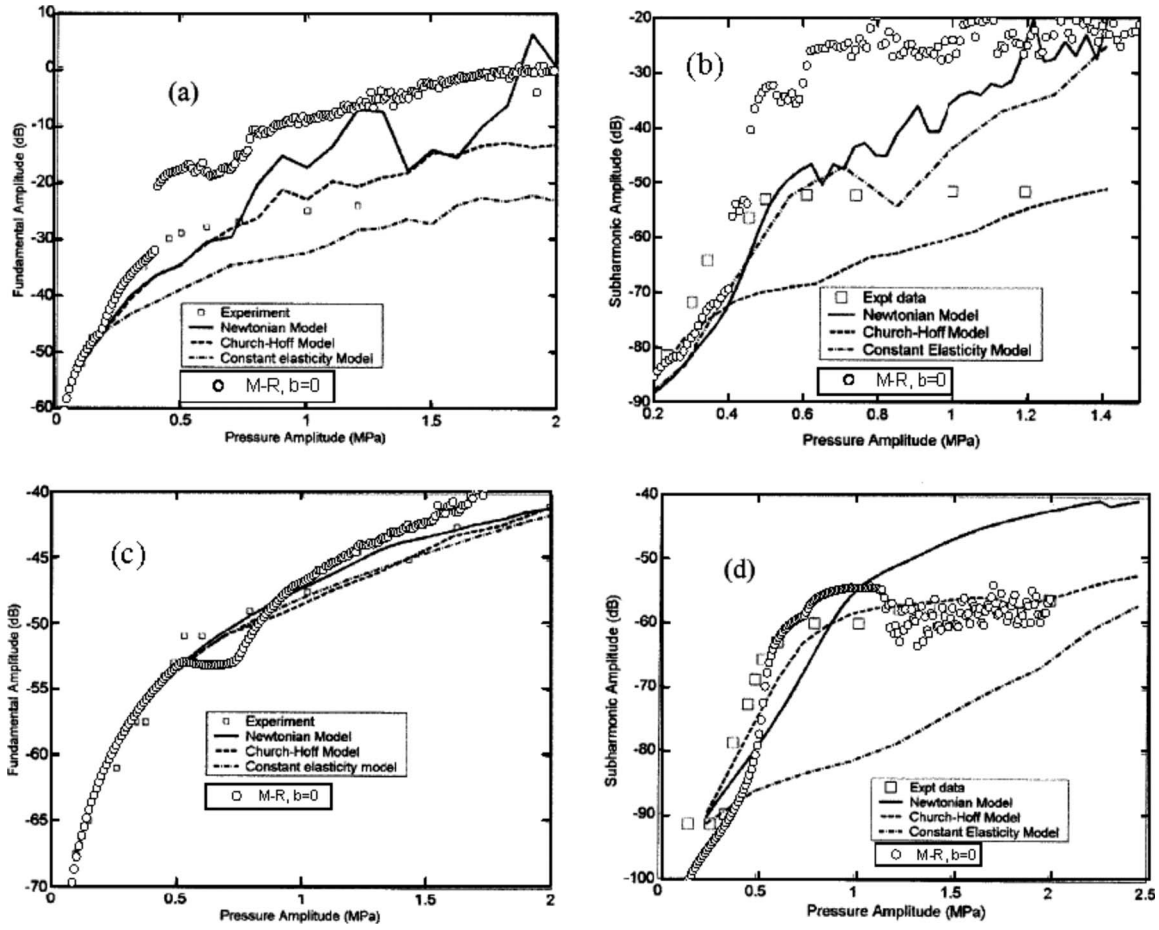


FIG. 6. Comparison between experimental measurements (Sarkar *et al.*, 2005) and predictions based on the Church–Hoff, the constant elasticity, the Newtonian, and the strain softening model, of the fundamental and subharmonic signals when (a), (b) $\omega_f=2\pi$ 2 MHz and (c), (d) $\omega_f=2\pi$ 4.4 MHz, for Sonazoid microbubbles.

Tables I and II. For the same reason the residual stresses play an increasingly important role in the microbubble response, altering both the resonance frequency and the scattering cross section. The compressive or expansive nature of initial displacements that cause the residual stresses plays a pivotal role in determining these aspects of microbubble behavior. In particular, compressive initial stresses tend to decrease the microbubble resonance frequency while decreasing the amount of scatter at resonance. The opposite is true for positive initial displacements corresponding to volume expansion. In the latter situation the relative increase in the contrast agent's area amounts to increasing its effective shear modulus as well. Consequently, the microbubble exhibits more intense scatter and larger resonance frequencies for a given amount of residual stresses at $t=0$, see also Fig. 5.

Nonlinear membrane behavior may explain certain findings of experimental investigations available in the literature. Sarkar *et al.* (2005) employed a number of the available contrast agent models in order to match experimental measurements of the fundamental and the subharmonic scatter from a Sonazoid solution. In that study the constant elasticity model fails to capture the plateau in the subharmonic scatter exhibited by the measurements at very large sound amplitudes, $\varepsilon \sim 10$, whereas the Church–Hoff model underpredicts the subharmonic measurements and cannot satisfactorily capture the plateau in the two signals. The authors attribute part of the failure to the softening of the membrane leading to higher amounts of scatter than expected based on the Church–Hoff model, which is not strictly valid when large membrane displacements are present. We carried out a number of scattering calculations using the values for $G_s \approx 52 \times 10^6$ kg/(m s^2) and $\mu_s \approx 0.99$ kg/(m s) obtained in the above study for Sonazoid by fitting the Church–Hoff model to low amplitude sound attenuation data; $R_0=1.6$ μm. Figures 7(a), (d) and 8(a), (d) from the above study are reproduced, Figs. 6(a)–6(d) in the present study, with the addition of the curve corresponding to the strain-softening membrane model presented here with b=0. The latter model predicts the fundamental and subharmonic signals quite well and for the entire range of sound amplitudes, for relatively large forcing frequencies, $\omega_f \geq 2\pi 4.4$ MHz, Figs. 6(c) and 6(d). The latter is the resonance frequency for the bubble size used in the sample under examination. For lower values of the forcing frequency the model with the strain softening membrane is qualitatively correct but tends to over-predict the two signals; see Figs. 6(a) and 6(b). This failure may be attributed to errors in the estimation of G_s, perhaps due to the fact that the attenuation data were not acquired at low enough amplitudes for thefitting to be valid, to bubble size

distribution effects or, more importantly, to variations in membrane viscosity as a result of the high frequency of the acoustic disturbance. Namely, a large number of polymeric materials exhibit a shear thinning behavior when subjected to high frequency disturbances due to disentanglement of the polymer chains. Consequently, the membrane material is expected to exhibit a higher viscosity as low frequencies, which may account for the systematic over-prediction of the two signals at low frequencies by the model presented here.

Finally, it should be stressed that for spherosymmetric pulsations it is the area dilatation modulus that determines microbubble dynamics, i.e., parameter $K=3G_{MR}$ or $G_{SK}(2C+1)$ for MR and SK membranes, respectively. Nevertheless, in the absence of shear, the set of measurements that is typically carried out in order to estimate G_s and μ_s, i.e., sound attenuation and scattering cross section measurements, is enough to fit K for either type of membrane under isotropic tension, as long as it is made at the appropriate range of low sound amplitudes and for fixed thickness δ. Parameters b and C, characterizing the degree of softness or hardness for MR and SK membranes, can be estimated by carrying out measurements of area dilatation versus sound amplitude, as the later increases beyond the range of validity of Hooke's law. They essentially determine deviations from linearity in the slope of the stress-area dilatation curve for the particular membrane material. In the case of strain softening membranes the effect of vibration onset or thresholding, Emmer *et al.* (2007), provide a type of measurement that is quite sensitive in parameter b, Fig. 4(a).

ACKNOWLEDGMENTS

Dr. Kostas Tsiglifis wishes to acknowledge scholarships "HRAKLEITOS" and "PYTHAGORAS" for financial support. The authors also wish to acknowledge P. Dallas for performing some of the simulations presented in the article as part of his Diploma Thesis.

Barthès-Biesel, D., Diaz, A., and Dhenin, E. (**2002**). "Effect of constitutive laws for two-dimensional membranes on flow-induced capsule deformation," J. Fluid Mech. **460**, 211–222.

Bloch, S. H., Wan, M., Dayton, P. A., and Ferrara, K. W. (**2004**). "Optical observation of lipid and polymer-shelled ultrasound microbubble contrast agents," Appl. Phys. Lett. **84**(4), 631–633.

Burns, P. N., Simpson, D. H., and Averkiou, M. (**2000**). "Nonlinear imaging," Ultrasound Med. Biol. **26**, S19.

Church, C. C. (**1995**). "The effects of an elastic solid surface layer on the radial pulsations of gas bubbles," J. Acoust. Soc. Am. **97**, 1510–1521.

De Jong, N., and Hoff, L. (**1993**). "Ultrasound scattering of Albunex® microspheres," Ultrasonics **31**, 175–181.

Dollet, B., van der Meer, S. M., de Jong, N., Versluis, M., and Lohse, D. (**2008**). "Nonspherical oscillations of ultrasound contrast agent microbubbles," Ultrasound Med. Biol. (to be published).

Emmer, M., van Wamel, A., Goertz, D. E., and de Jong, N. (**2007**). "The onset of microbubble vibration," Ultrasound Med. Biol. **33**(5) 941–949.

Frinking, P. J. A., and De Jong, N. (**1998**). "Acoustic modeling of shell-encapsulated gas bubbles," Ultrasound Med. Biol. **24** 523–533.

Gorce, J. M., Arditi, M., and Schneider, M. (**2000**). "Influence of bubble size distribution on the echogenicity of ultrasound contrast agents. A Study of SonoVue™," Invest. Radiol. **35**, 661–671.

Hilgenfeldt, S., Lohse, D., and Zomack, M. (**1998**). "Response of bubbles to diagnostic ultrasound: A unifying theoretical approach," Eur. Phys. J. B **4**, 247–255.

Hoff, L., Sontum, P. C., and Hovem, J. M. (**2000**). "Oscillations of polymeric microbubbles: Effect of the encapsulated shell," J. Acoust. Soc. Am. **107**(4), 2272–2280.

Keller, J. B. and Miksis, M. (**1980**). "Bubble oscillations of large amplitude," J. Acoust. Soc. Am. **68**, 628–633.

Khismatullin, D. B. and Nadim, A. (**2002**). "Radial oscillations of encapsulated microbubbles," Phys. Fluids **14**, 3534–3556.

Pelekasis, N. A., Gaki, A., Doinikov, A., and Tsamopoulos, J. A. (**2004**). "Secondary Bjerknes forces between two bubbles and the phenomenon of acoustic streamers," J. Fluid Mech. **500**, 313–347.

Pozrikidis, C. (**1992**). *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, (Cambridge University Press, Cambridge).

Prosperetti, A., and Lezzi, A. (**1986**). "Bubble dynamics in a compressible liquid, Part 1. First-order theory," J. Fluid Mech. **168**, 457–478.

Sarkar, K., Shi, W. T., Chatterjee, D., and Forsberg, F. (**2005**). "Characterization of ultrasound contrast microbubbles using *in vitro* experiments and viscous and viscoelastic interface models for encapsulation," J. Acoust. Soc. Am. **118**(1), 539–550.

Sboros, V., MacDonald, C. A., Pye, S. D., Moran, C. M., Gomatam, J., and McDicken, W. N. (**2002**). "The dependence of ultrasound contrast agents backscatter on acoustic pressure: Theory versus experiment," Ultrasonics **40**, 579–583.

Skalak, R., Tozeren, A., Zarda, R. P., and Chien, S. (**1973**). "Strain energy function of red blood cell membranes," Biophys. J. **13**, 245–280.

Focusing of shock waves induced by optical breakdown in water

Georgy N. Sankin,[a] Yufeng Zhou, and Pei Zhong
Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina 27708

(Received 19 June 2007; revised 5 March 2008; accepted 7 March 2008)

The focusing of laser-generated shock waves by a truncated ellipsoidal reflector was experimentally and numerically investigated. Pressure waveform and distribution around the first (F_1) and second foci (F_2) of the ellipsoidal reflector were measured. A neodymium doped yttrium aluminum garnet laser of 1046 nm wavelength and 5 ns pulse duration was used to create an optical breakdown at F_1, which generates a spherically diverging shock wave with a peak pressure of 2.1–5.9 MPa at 1.1 mm stand-off distance and a pulse width at half maximum of 36–65 ns. Upon reflection, a converging shock wave is produced which, upon arriving at F_2, has a leading compressive wave with a peak pressure of 26 MPa and a zero-crossing pulse duration of 0.1 μs, followed by a trailing tensile wave of −3.3 MPa peak pressure and 0.2 μs pulse duration. The −6 dB beam size of the focused shock wave field is 1.6×0.2 mm^2 along and transverse to the shock wave propagation direction. Formation of elongated plasmas at high laser energy levels limits the increase in the peak pressure at F_2. General features in the waveform profile of the converging shock wave are in qualitative agreement with numerical simulations based on the Hamilton model.
 [DOI: 10.1121/1.2903865]

PACS number(s): 43.25.Cb, 43.25.Jh [AJS] Pages: 4071–4081

I. INTRODUCTION

Electrohydraulic (EH) shock wave lithotripters, including the first generation Dornier HM-3, have been widely used in clinic for the treatment of kidney stones for over two decades.[1–3] In a typical EH lithotripter, an underwater spark discharge is used to produce a spherically divergent shock wave at the first focus (F_1) of a truncated brass ellipsoidal reflector. Upon reflection, the shock wave is converged to the second focus (F_2) of the ellipsoidal reflector where the kidney stone inside the patient is aligned under fluoroscopic or ultrasound imaging guidance. Numerous clinical studies have demonstrated that the HM-3 produces better stone comminution with higher stone-free rate compared to the second and third generation electrohydraulic (EH), electromagnetic (EM) and piezoelectric (PE) lithotripters.[4–6] The underlying mechanisms, however, have not been well understood. More recently, methods to upgrade the HM-3 lithotripter, which is still widely regarded as the golden standard in shock wave lithotripsy (SWL), for improved performance and safety have been proposed and tested *in vitro*.[7–10] The general principle and techniques developed for upgrading the HM-3 may also be applied to improve the design of EM lithotripters which have been widely used in clinical SWL.

For design optimization of a lithotripter, development of numerical models that can simulate accurately the propagation and focusing of lithotripter shock waves (LSWs) will be valuable. In recent years, several different models have been developed to describe the general characteristics of linear and nonlinear wave propagations in the original and upgraded HM-3 lithotripters. Hamilton developed a linear model which depictures the propagation of different LSW components in space and time.[11] Averkiou and Cleveland used the two-dimensional Khokhlov–Zabolotskaya–Kuznetsov (KZK) equation with initial conditions at the reflector aperture determined from geometrical acoustics.[12] Zhou and Zhong extended this approach with initial conditions taken from measurements near F_1 and introduced the concept of an equivalent reflector.[13] Tanguay and Colonius used the Euler equations to model shock wave propagation in two-phase flow.[14,15] Szeri *et al.* implemented a density jump technique to model the reflector as an interface in the fluid media.[16,17] To validate the model calculation, comparison with reliable experimental data produced by the lithotripter is critical. However, because of the inherent instability in electrical spark discharge, repeatable pressure measurements in an EH lithotripter are problematic. Therefore, for model validation it is highly desirable to develop means for generating stable focused shock waves using a reflector configuration similar to that used in an EH lithotripter.

Optical breakdown induced by a focused Q-switched laser pulse in water produces a shock wave of microsecond duration.[18–22] The focused laser initially vaporizes the water, giving rise to a bubble that expands rapidly and generates concomitantly a divergent shock wave.[22,21] Upon reaching maximum expansion, the bubble collapses violently emitting a secondary shock wave at its minimal volume.[19] This method provides a stable source for the generation of spherical shock wave with about 3% of the incident optical energy being converted into acoustic emission.[20]

In this work, the focusing of laser-generated shock waves in water by a truncated ellipsoidal brass reflector was investigated. The pressure waveform and distribution at both foci of the ellipsoidal reflector were measured. Further, the Hamilton model[11] was implemented to simulate qualitatively

[a]Author to whom correspondence should be addressed. Tel.: (919) 660-5416. Fax: (919) 660-8963. Electronic mail. gns@duke.edu

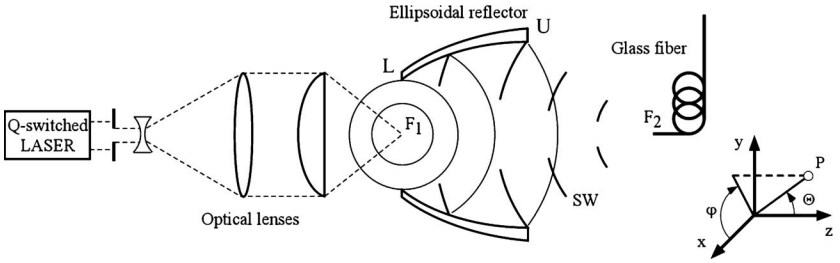


FIG. 1. A schematic diagram of the experimental setup, F_1 is the first and F_2 is the second focus of the ellipsoidal reflector.

the profile of the pressure waveform and its evolution along the reflector axis toward F_2. The simulation results were compared to the experimental measurements. In addition, this experimental system was used to simulate the effect of jitters in electric spark discharge on the resultant pressure distribution at the lithotripter focus, i.e., the variation in peak pressure at F_2 as a result of the departure of the shock wave emitter from F_1. Altogether, these experimental results may be useful for validation of three dimensional (3D) numerical models of shock wave focusing in water.

II. MATERIALS AND METHODS

A. Experimental setup

A schematic diagram of the experimental setup is shown in Fig. 1. A Q-switched neodymium doped yttrium aluminum garnet laser with a wavelength of 1064 nm and a pulse duration of ~5 ns (Tempest 10, New Wave Research, Fremont, CA) was collimated and focused by a combination of lenses ($F=30$ mm and NA=0.6 for the last focusing lens) in water to generate a single cavitation bubble via optical breakdown. The laser was operated at 9.3%, 21%, 33%, or 65% of its maximum output energy of 200 mJ (yet the optical attenuation in the focusing lens was unknown). For conciseness, we will omit the inclusion of the maximum energy in describing the laser output henceforth. The laser was aligned horizontally with its beam focus coinciding with F_1 of a truncated ellipsoidal reflector insert described previously.[7] The reflector insert has a major semiaxis $a'=132.45$ mm, a minor semiaxis $b'=71.5$ mm, and a half focal length $c'=111.5$ mm. The height of the reflector insert is 106 mm and the minimal distance from F_1 to the plane coinciding with the lower edge of the reflector insert is 4 mm.With this geometry the reflector insert covers about 39% of the full solid angle around F_1. The reflector was mounted on a (3D) translational stage (Thorlabs, Newton, NJ) and immersed in a water tank (40 cm × 31 cm × 60 cm) filled with degassed water at 20 °C.

Based on linear acoustic approximation, the beam diameter (D_B) at F_2 can be estimated by using the following equation [see Ref. 23 Eq. (5.51) on p. 107]:

$$D_B = 2.44\lambda\left(\frac{L_f}{D}\right) = 2.44(2s\tau)\left(\frac{c'}{2b'}\right) = 0.16 \text{ mm}, \tag{1}$$

where $\lambda(=2s\tau)$ is the wavelength, L_f is the focal length, D is the aperture diameter of the source, $s=1482$ m/s is the sound speed in water at 20 °C, and $\tau=28\pm8$ ns is the full width at half maximum (FWHM) of the shock wave pulse measured at a stand-off distance of $r=1.1$ mm from F_1 at 9.3% of the maximum laser output energy.[24] The peak positive pressure (p^+) at F_2 of the reflected shock wave can be estimated from energy conservation consideration. Without attenuation, the total acoustic energy delivered by the focused wave at F_2 should be equal to the energy associated with the shock wave emitted from F_1, i.e.,

$$(p^+)^2\frac{\pi D_B^2}{4} = (p_P)^2 4\pi r^2(0.39), \tag{2}$$

where $p_P=4.8\pm0.3$ MPa is the peak pressure of the shock wave measured at a stand-off distance of $r=1.1$ mm from F_1 at 9.3% of the maximum laser output energy.[24] The coefficient of 0.39 accounts for the fact that only 39% of the original shock wave emitted from F_1 will be covered by the reflector insert. From Eqs. (1) and (2), one can obtain

$$p^+ = 4\frac{r}{D_B}p_P\sqrt{0.39} = 83.5 \text{ MPa}. \tag{3}$$

B. Pressure measurement near F_1

The pressure waveforms were measured by using a fiber optic probe hydrophone (FOPH-500, RP Acoustics, Leutenbach, Germany). The 100 μm probe tip of the hydrophone was aligned at a distance r from F_1 using a combination of translational and rotational stages. The hydrophone signals were first recorded on a digital oscilloscope (500 MHz Wave Runner 6050A, LeCroy, Chestnut Ridge, NY), then inverted and deconvoluted to obtain calibrated pressure waveforms using a computer program supplied by the manufacturer. Signal averaging over 10–64 shots was performed to improve the signal-to-noise ratio of the measured pressure waveforms. When multiple signal averaging was not used (e.g., in reproducibility test), a moving average was calculated at each location using data points within a window of 10 ns duration, which corresponds to the temporal resolution of the hydrophone.

C. Alignment of the reflector

Alignment of the reflector insert and pressure measurement near F_2 was carried out by scanning along each of the three orthogonal axes (x, y, and z) until the maximum peak pressure at F_2 was detected. Figure 2 shows the distribution of the peak pressure along x and y axes, as well as along the z-axis at laser energy $E=9.3\%$. It can be seen that misalignment of the reflector insert by 0.2 mm (more than the wave-

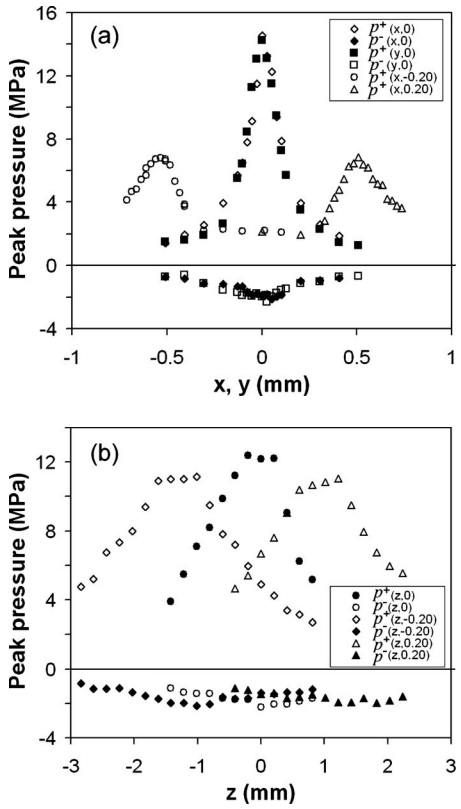


FIG. 2. The distribution of peak pressures near F_2 (p^+ for positive and p^- for negative) when reflector is displaced for +0.20, 0, and −0.20 mm (the second coordinate in the bracket) along (a) x- or y-axis and (b) z-axis. E =9.3%, PRF=5 Hz, each point represents averaged data over ten measurements.

length of the acoustic pulse) could result in a significant drop in the peak pressure. The −6 dB beam size at F_2 was estimated to be about 0.2×1.6 mm^2 in the lateral and axial directions of the reflector.

Since the beam size in the lateral direction [0.20 ± 0.03 mm, Fig. 2(a)] is only about twice the hydrophone probe diameter (d=0.1 mm), significant spatial averaging error could be introduced near F_2.[25] Assuming a Gaussian distribution for the converging LSW and an effective beam diameter $D_B^*\sim D_B-d$,[25] integration over the probe surface gives an estimated focal pressure $p^*=1.4p^+$. Away from F_2, however, the hydrophone can successfully capture the true waveform of the shock wave.

D. High-speed shadowgraph imaging

The dynamics of laser-generated shock waves and associated cavitation bubbles were captured by using a high-speed imaging system (Imacon 200, DRS Hadland, Oakland, NJ) in combination with a long-distance microscope (K2, Infinity, Boulder, CO) and a 5× objective lens. A fiber optic coupled xenon flash lamp (ML-1000, Dyna-Lite, Union, NJ) was used for illumination. A digital delay generator (DG 535, Stanford Research Systems, Sunnyvale, CA) was used to trigger the laser, flash lamp, and the high-speed camera, respectively. The laser spark measured by a photodetector (PDA50, Thorlabs, Newton, NJ) was used as the reference time for the shadowgraph images.

TABLE I. Characteristics of the pressure waveforms measured at a stand-off distance of 3 mm away from F_1 at different energy levels (PRF=5 Hz).

Laser energy E (%)	Peak pressure p_P (MPa)	Peak arrival time (ns)	Rise time (ns)	Full width at half maximum (FWHM) (ns)
9.3	2.1 ± 0.1	2019 ± 2	10 ± 3	36 ± 6
21	3.2 ± 0.1	2008 ± 1	9 ± 2	45 ± 3
33	4.0 ± 0.2	1997 ± 3	9 ± 1	53 ± 6
65	5.9 ± 0.1	1984 ± 2	10 ± 1	65 ± 4

E. Numerical analysis

Because of the extreme high gain of the reflector insert for laser-induced shock wave (G=309 for E=21% based on the data shown in Tables I and II) conventional nonlinear wave propagation models, such as the KZK equation, become invalid.[26] For modeling shock wave focusing based on the Euler equations in combination with the Tait equation, extremely large number of grid nodes will be required because of the high frequency content of the pulse, leading to unacceptably long computation time.[16] In light of these limitations, the Hamilton model of linear wave focusing in an ellipsoidal reflector[11] was chosen to reveal the general features and to facilitate the interpretation of the measured pressure waveforms in terms of various contributory components (the central wave, the wake, and the edge wave). Pressure waveforms measured near F_1 [Fig. 4(b)] were used as the source condition in the Hamilton model calculation.

III. RESULTS

A. Laser-induced shock waves and cavitation bubble at F_1

Figure 3 shows the general features of the shock wave and cavitation bubble produced by laser-induced optical breakdown at F_1. Depending on the laser energy, the maxi-

TABLE II. Characteristics of the pressure waveforms measured at F_2 at different energy levels (PRF=5 Hz, n=30) and normalized peak positive pressure determined based on theoretical calculation at z=−0.1 mm.

Laser energy E(%)	Peak pos. pressure (MPa)	Peak arrival time (ns)	Beam length Δz (mm)	Beam width Δx (mm)	Full width at half maximum (FWHM) (ns)	Theoretical peak pos. pressure (normalized)
9.3	15.0 ± 1.8	$177\,769\pm8$	1.6 ± 0.2	0.20 ± 0.03	24.5 ± 1.6	1
21	22.4 ± 1.6	$177\,747\pm5$	2.0 ± 0.2		24.1 ± 1.6	1.52
65	25.9 ± 3.9	$177\,696\pm7$			33.1 ± 5.5	2.81

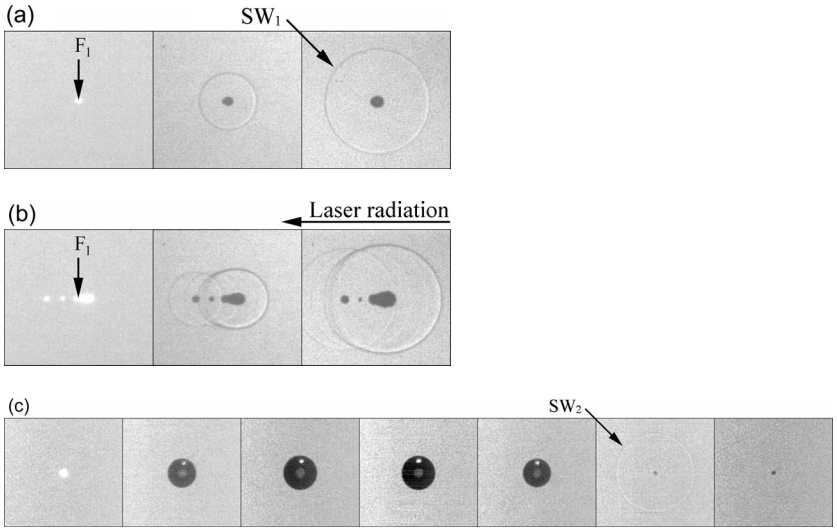


FIG. 3. High-speed images of laser-induced optical breakdown in water. SW_1: shock wave produced by the laser spark: SW_2: shock wave produced by bubble collapse. Laser energy level $E=9.3\%$ [(a) and (c)] and 65% (b), respectively; interframe time (IFT)$=0.2$ μs [(a) and (b)] and 10 μs (c), respectively; the height of each image frame is 1.7 mm.

mum bubble radius R_m varies in the range of 0.3–0.7 mm. At low energy settings, the shapes of the laser-induced plasma, bubble formation, and associated shock wave are all nearly spherical [Fig. 3(a)]. As the energy exceeds 65%, multiple spots of optical breakdown are observed with the location of the largest plasma and hypocenter of the strongest shock wave shifted toward the optical lens [Fig. 3(b)] due to peculiarities of laser-induced optical breakdown.[27]

The pressure profile of the shock wave, measured at a stand-off distance $r=3$ mm from the laser focus, shows a leading shock front followed by a tail approximated by a triangle [see Fig. 4(b)]. The pulse is essentially a compressive wave with negligible tensile component. The FWHM of the shock wave vary from 36 to 65 ns, and the rise time is about 10 ns which is close to the temporal resolution of the hydrophone [Fig. 4(a)]. As shown in Fig. 4(a) when laser energy increases the peak pressure becomes higher and pulse width widens, yet the arrival time of the shock wave is shortened as a result of nonlinear propagation. Based on the measurements, a simplified waveform at F_1 was determined [Fig. 4(b)] and used for the Hamilton model calculation. In addition, the peak pressure was found to vary linearly with pulse repetition frequency (PRF) at different laser energy levels [Fig. 4(c)] since laser output energy increases with PRF. Furthermore, the peak pressure of a spherically divergent shock wave is known to vary inversely with the propagation distance,[20–22] and this relationship is confirmed by the measurement data at two energy levels [Fig. 4(d)]. The peak pressure, arrival time, shock front rise time, and FWHM of the laser-generated shock waves are summarized in Table I.

At laser setting of $E=9.3\%$ the energy of the emitted shock wave from the optical breakdown and the potential energy of the resultant bubble at its maximum expansion can be determined as follows:

$$E_{\mathrm{SW}} = \frac{8\pi p_P^2}{3\rho c} r^2 \tau = 8\ \mu\mathrm{J} \tag{4}$$

and

$$E_b = \frac{4\pi}{3} R_m^3 p_0 = 11\ \mu\mathrm{J}. \tag{5}$$

In contrast to the spark-generated bubble (between the tips of an electrode) in an EH lithotripter, laser-induced bubble expands and collapses symmetrically in a free field, leading to the generation of a strong secondary shock wave. It has been shown that the amplitude of the secondary shock wave could be as strong as the first one when the bubble collapse time is within 50–60 μs.[24]

Figure 5 shows the angular variation of the peak pressure measured in two orthogonal planes ($\varphi=90°$ and $\Theta=90°$, respectively) at the same 3 mm stand-off distance from the laser focus. In the transverse plane ($\Theta=90°$) it was found that the pressure distribution is symmetric around the optical axis [Fig. 5(a)]. However, in the axial plane ($\varphi=90°$) the pressure distribution is not uniform, and the strongest shock wave is produced in the direction perpendicular to the optical axis [Fig. 5(b)]. This is presumably because the laser plasma is elongated along the axis (see Fig. 3 and Ref. 21) forming a cylindrically elongated bubble with resultant nonuniform pressure distribution. At two different energy levels ($E=9.3$ and 21%), the pressure amplitudes are almost doubled when Θ varies from 33° to 87°, which corresponds to the upper and the lower edges of the reflector [depictured by arrows in Fig. 5(b)].

B. Experimental measurement of the focused shock wave near F_2

Figures 6 and 7 show representative pressure profiles of the focused shock wave measured at different positions along the z-axis. At F_2 a bipolar, asymmetric acoustic pulse comprising of a stronger leading compressive wave, followed by a weaker trailing tensile wave was observed [Fig. 6(c)]. The arrival time of the shock wave was $T_a=177.77$ μs (see Table III), which corresponds to an average wave propagation speed of 1490 m/s. Due to focusing the FWHM of the shock wave near F_2 was measured to be 24–33 ns, which was sig-

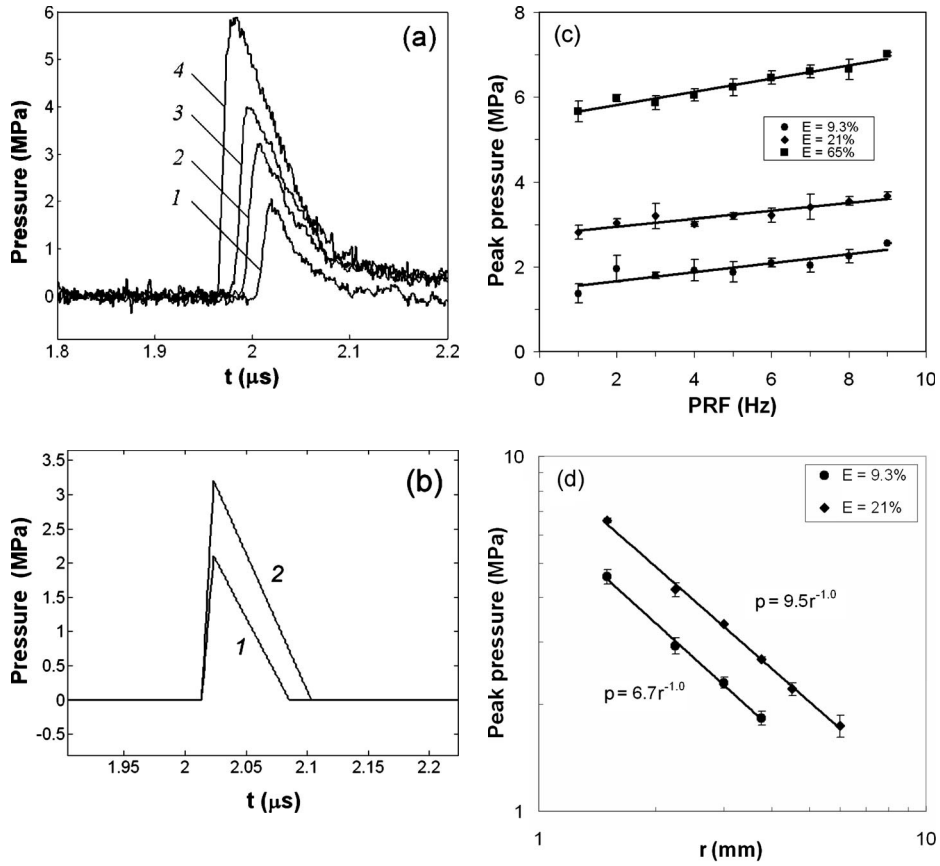


FIG. 4. (a) The pressure waveforms measured at a stand-off distance of 3 mm from F_1 produced at different laser energy levels of E=9.3% (1), 21% (2), 33% (3), and 65% (4) with PRF=5 Hz, φ=0°, and Θ=90°; (b) idealized pressure waveforms at F_1 used for theoretical calculation based on the Hamilton model for E=9.3% (1) and 21% (2); (c) calibration curve for the peak pressure at different pulse repetition rates and energy levels measured at 3 mm above the laser spark (φ=90°, Θ=90°); (d) relationship between the peak pressure p and measurement distance r above the laser spark at two different energy levels (E=9.3% and 21%, PRF=5 Hz). Each pressure waveform was obtained by signal averaging over 20 shots. In (c) and (d), error bars show standard deviation from three independent experiments.

nificantly reduced from the corresponding value near F_1 (see Tables I and II). At E=9.3% the focused shock wave measured at F_2 (waveform is averaged over 30 shots) has a peak positive pressure of 10 MPa with a zero-crossing pulse duration of 0.1 μs and a peak negative pressure of −2.6 MPa with a pulse duration of ~0.2 μs.

Away from F_2, several features in the evolution of the shock wave profile along the z-axis can be noticed. Prefocally ($z<0$), the waveform has multiple components corresponding to the central wave (C), the diffracted wave from the lower edge (E_L), the wake (W), and the inverted diffracted wave from the upper edge (E_U) of the truncated ellipsoidal reflector. A dual positive peak structure is observed [Figs. 6(a) and 6(b)]. Postfocally ($z>0$), the edge wave takes over both the center wave and the wake [Figs. 6(d) and 6(e)].

Figures 8(a) and 8(b) show the shot-to-shot variation of the peak positive pressure, FWHM, and shock front arrival time of the focused shock wave estimated using moving average at three different laser energy levels. The corresponding shock wave parameters at F_2 are summarized in Table II. Similar to the observations at F_1 when the laser energy increases, the peak pressure becomes significantly higher, while the shock wave arrival time reduces slightly [Fig. 8(a)]. In addition, when the laser energy increases from 21% to 65%, the variations in peak positive pressure and FWHM increase from 7.1% to 15% and from 6.6% to 17%, respectively. These results correlate to the instability and the associated random shift in the location of laser-induced optical breakdown at higher energy levels [see Fig. 3(b)]. The peak positive [Fig. 8(c)] and peak negative [Fig. 8(d) pressure at F_2 initially increases with the peak pressure at F_1 and PRF. However, at higher energy settings the peak pressures are saturated presumably due to the elongation in the geometry of laser-induced plasma [Fig. 3(b)]. The elongated plasma could be considered as several point sources along the major axis of the ellipsoidal reflector near F_1. As shown in Fig. 2, the superposition of these displaced shock sources could lead to a focused shock wave at F_2 with lower amplitude but longer pulse duration compared to that produced by a single shock wave induced at F_1.

In addition, jitters in the exact location of laser-induced plasma around F_1 can lead to reduced focusing gain, lower peak pressure, and shift in focal area. Since the optics are fixed on the wall of the water tank jitter around F_1 was modeled by moving both the reflector and the tip of the FOPH simultaneously by a distance of $-\mathbf{r}_P=(-x_P,-y_P,-z_P)$, which is equivalent to displacing the laser-induced plasma by $\mathbf{r}_P$. To ensure sphericity of the shock wave, this experiment was performed at a low energy level of E=9.3%. The results show that displacement of the plasma in either the x or y axis

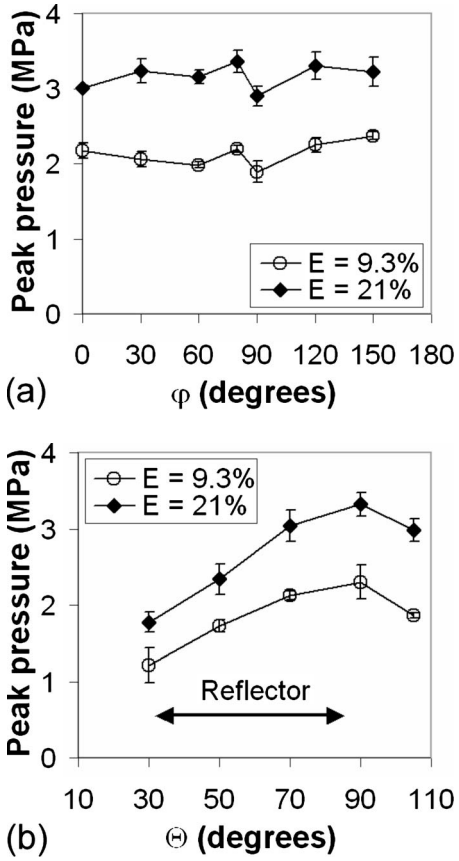


FIG. 5. Angular variation of the peak pressure of laser-generated shock waves near F_1, in the plane (a) $\Theta=90°$ and (b) $\varphi=90°$. Measurements were made at a stand-off distance of 3 mm from the laser spark, PRF=5 Hz, using signal averaging over 20 shots. Error bars represent standard deviation from three independent experiments.

leads to a much more rapid reduction of the peak pressure than in the z-axis (Fig. 9). Jitters in plasma location could also lead to different arrival times of the focused shock wave at F_2, and signal averaging from multiple pulses would reduce the measured peak pressure. These factors may contribute to the large variation in peak pressure and FWHM at high energy output levels, as shown in Fig. 8. This finding suggests that in an EH lithotripter deviation of the spark discharge in the first focal plane transverse to the lithotripter axis could significantly reduce the peak pressure at F_2 while effectively increasing the focal beam size.

Figure 10 shows the distribution of the peak pressure of the focused shock wave along z-axis at two lower energy settings (i.e., 9.3% and 21%). Although the maximum pressure was found to increase with the laser energy, the location of the positive peak remained unchanged. In contrast to lithotripter field where peak compressive pressure tends to shift postfocally and peak tensile pressure shits prefocally,[13] the measured peak pressures appear to coincide with F_2 within the measurement uncertainty (± 0.2 mm). Moreover, the length of the focal area at F_2 was found to increase from 1.6 to 2.0 mm as the laser energy increased (Table II). Interestingly, both the compressive and tensile peak pressures have a local minimum at $z=-4$ mm (Fig. 10).

High-speed images, taken by a charge coupled device video camera (GP-MF552, Panasonic, Secaucus, NJ), revealed a detectable microbubble produced at the center of the end face of the FOPH 500 fiber tip (see inset in Fig. 11). The bubble has a maximum dimension of 55×30 μm^2 at laser energy $E=50\%$ ($p^+\sim 25$ MPa). The main plot in the figure shows the hydrophone signal taken simultaneously, demonstrating a shock wave arrival time of 178 μs, bubble collapse time of ~ 8 μs and rebound of the bubble after 186 μs. Based on ten independent measurements, the collapse time of the microbubble was found to be in the range of 6.4–8.4 μs by passive cavitation detection using a 3.5 MHz focused transducer (A382S, Panametrics, Waltham, MA).

C. Theoretical modeling of the shock wave focusing near F_2

The profiles of the laser-induced focused shock wave along the z-axis of the truncated reflector were simulated by using the Hamilton model. The middle column in Fig. 6 shows the resultant waveform for the truncated reflector, while the right column shows two separate waveforms (1) from a complete insert reflector and (2) from the truncated part at the bottom normalized by the peak pressure at $z=-0.1$ mm for $E=9.3\%$. Because the Hamilton's model has a singularity at $z=0$ and the FOPH has a 100 μm core fiber diameter, $z=-0.1$ mm was chosen when calculating pressure at F_2. The primary components of the shock wave, especially the central wave (C), the edge wave (E), and the wake (W), can be clearly seen. Despite the differences in numerical values, the general profiles of the shock wave at different locations agree qualitatively with the experimental measurements [Figs. 6(a)–6(e)]. Because the truncated ellipsoidal reflector has an upper and a lower rim, two edge waves were produced. The edge wave from the lower rim (E_L) is closer to the central wave. In comparison, the edge wave from the upper rim (E_U) is further away from the central wave before reaching the focal point [Figs. 6(f) and 6(k)]. As the shock wave converges toward F_2,E_U moves close to C at a faster speed than E_L. At $z=-4$ mm, the peak of E_L merges with the peak of the wake, leading to a reduced peak negative pressure [Figs. 6(g) and 6(l)] that was also observed experimentally (see Fig. 10). Beyond the focal point, both the edge wave and the wake invert the phase and overtake the central wave [Figs. 6(j) and 6(o)]. Similar features have been observed in a previous study of the reflector insert using spark discharge from a lithotripter electrode for shock wave generation.[7]

The results for higher energy level ($E=21$ and 65%) are shown in the Fig. 7 and Table II, which reveal a linear increase of the peak pressure at $z=-0.1$ mm with initial laser-induced shock wave pressure at F_1 [see Fig. 4(b)] and qualitative agreement of the pressure waveforms along z-axis between the experimental measurement and model calculation.

IV. DISCUSSION

In this study, the focusing of laser-generated shock waves by a truncated brass ellipsoidal reflector in water was

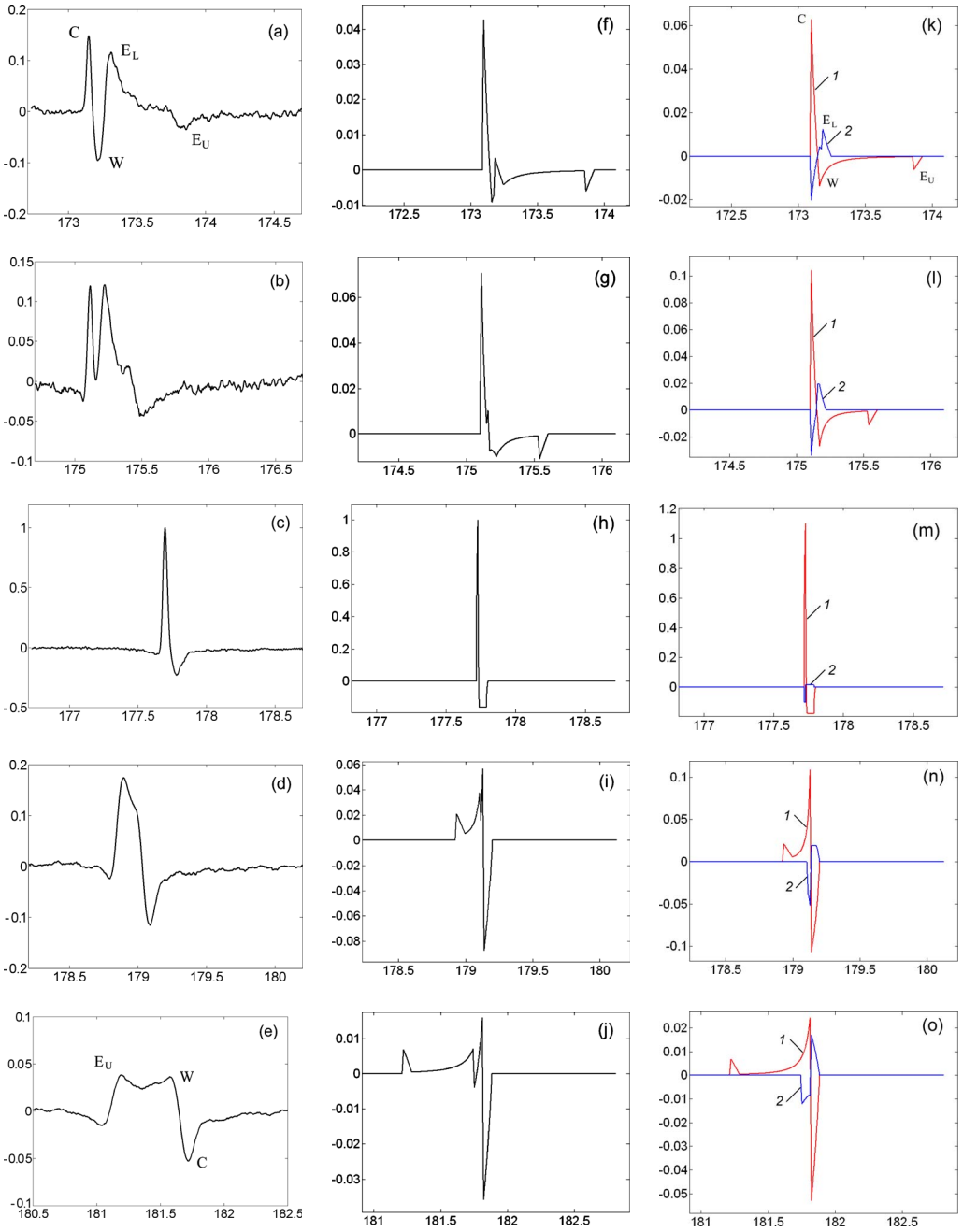


FIG. 6. (Color online) The measured (first column) and simulated (second and third columns) pressure waveforms along z-axis at z= [(a), (f), and (k)] −7, [(b), (g), and (l)] −4, (c) 0, [(h) and (m)] −0.1, [(d), (i), and (n)] 2, and [(e), (j), and (o)] 6 mm. Measurement results are normalized by the measured peak pressure at z=0, E=9.3%, and the simulated results are normalized by the simulated peak pressure at z=−0.1 mm, E=9.3%. The pressure waveform was measured at E=9.3%, PRF=5 Hz, with signal averaging over 30 [(c) and (d)] or 50 [(a), (b), and (e)] shots. The third column shows waveforms both from a complete insert reflector (1) and from the truncated part at the bottom (2).

investigated. Owing to the consistency in shock wave generation by a focused laser and the short duration of the resultant acoustic pulse (which reduces the likelihood of cavitation inception), pressure waveforms at both foci (F_1 and F_2 of the ellipsoidal reflector can be reliably measured. The shock wave at F_2 consists of a leading compressional phase ($p^+ < 32$ MPa) with a zero-crossing pulse duration $t^+ = 25-33$ ns, followed by a trailing tensile phase ($p^- > -4.3$ MPa). In comparison, the corresponding values for the spark-discharge generated shock waves in electrohydraulic shock wave lithotripters are $p^+ = 40-50$ MPa, $t^+ = 1-2$ μs, and $p^- > -10$ MPa. The beam size of laser-generated focused shock wave is also significantly smaller (1.6×0.2 mm^2 in axial and transverse directions) than its counterpart in an EH lithotripter (120×12 mm^2). It should be noted that the finite size of the hydrophone probe (0.1 mm) could lead to signal averaging and reduced peak pressure measured at F_2 when the diameter of hydrophone is comparable to the beam diameter. Therefore the actual beam diameter may be even smaller.

Although the principle of shock wave generation in EH lithotripsy and laser-induced optical breakdown is similar, the first shock wave generated by the spark discharge in an EH lithotripter is much stronger than the second shock wave

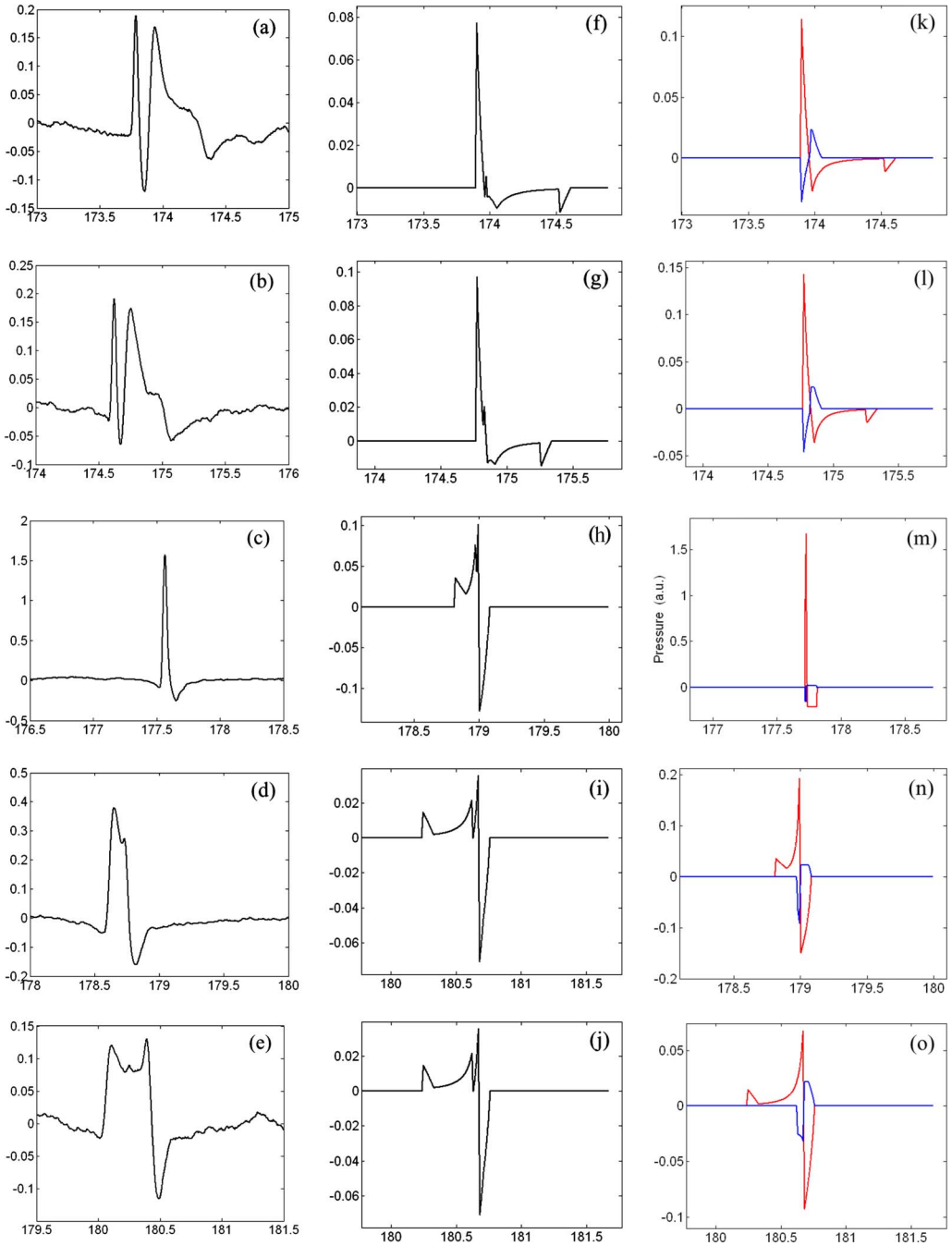


FIG. 7. (Color online) The same as Fig. 6 for E=21% at z= [(a), (f), and (k)] −5.8, [(b), (g), and (l)] −4.5, (c) 0, [(h) and (m)] −0.1, [(d), (i), and (n)] 1.8, [(e), (j), and (o)] 4.3 mm.

produced by the collapse of the bubble between the tips of the electrode.[28] In comparison, the amplitude of the second shock wave from bubble collapse in laser-based systems is similar to that of the first shock wave produced by the optical breakdown.[24] These differences may be caused by the significantly higher amount of electric energy (~16 J) that is deposited in an EH lithotripter, which yields a primary shock wave with higher peak pressure and longer pulse duration than the secondary shock wave produced by the collapse of the bubble, which is nonspherical due to the presence of the electrode tips.

TABLE III. Characteristics of the pressure waveforms measured at F_2 for different PRFs at E=9.3% (n=30).

Pulse repetition frequency (PRF) (Hz)	Peak pos. pressure (MPa)	Peak arrival time (ns)	Full width at half maximum FWHM (ns)
1	13.6 ± 1.7	177 773 ± 6	25.1 ± 2.1
5	15.0 ± 1.8	177 769 ± 8	24.5 ± 1.6
9	16.8 ± 1.4	177 765 ± 6	23.9 ± 1.5

Since the laser-generated shock wave is spherically diverging, the amplitude of the incident shock wave on the reflector surface (Θ=33°) can be estimated to be about 50 kPa with a corresponding focusing gain of ~300. For the most part, the initially divergent and subsequently focusing shock wave travels approximately at sound speed, suggesting

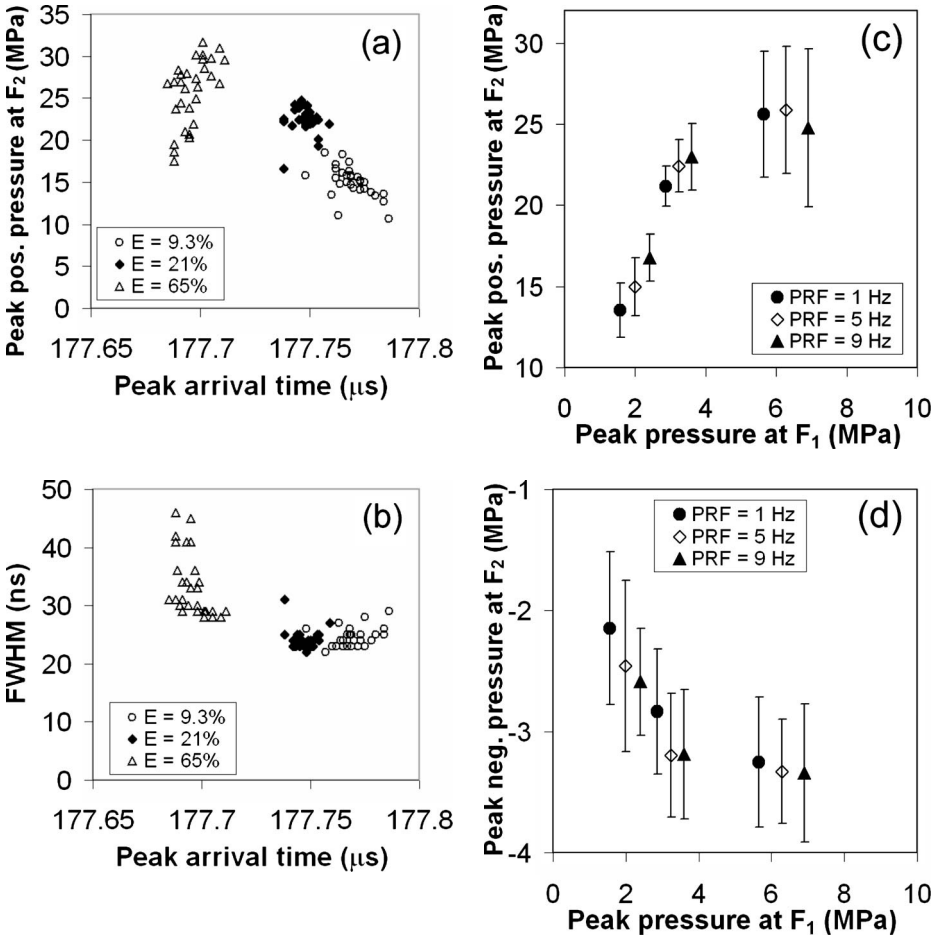


FIG. 8. [(a) and (b)] The shot-to-shot variation ($n=30$) of pressure parameters measured at F_2 at different laser energy levels. The mean with standard deviation of (c) peak positive and (d) peak negative pressure of shock wave at F_2 at different energy levels (as represented by the different peak pressures at F_1 and pulse repetition frequency (PRF) averaged over $n=30$ shots. Pressure at F_1 was measured at a stand-off distance of 3 mm away from the laser spark.

that a linear approximation can be reasonably applied to model the wave propagation, except near the focal point. Indeed, Hamilton's model correctly captures the interpulse time of various wave components although the calculated pressure of the central wave near F_2 is several times higher than the measured one (data not shown). The discrepancy is more pronounced at higher laser output energy levels (Table II). It was found that for laser energy below $E=21\%$ the peak pressure at F_2 increases with the source pressure at F_1 (Fig. 8), yet at high energy levels ($E>21\%$) the correlation between pressures at F_2 and F_1 becomes nonlinear and eventually saturated for both the compressive and tensile peak

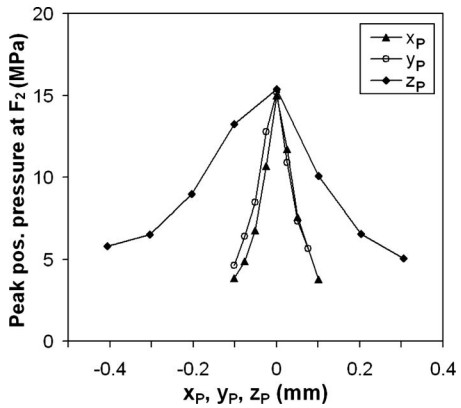


FIG. 9. Variation of the peak positive pressure measured at F_2 in relation to the jitter in the location of laser-induced plasma along either x-, y- or z-axis at F_1 ($E=9.3\%$)

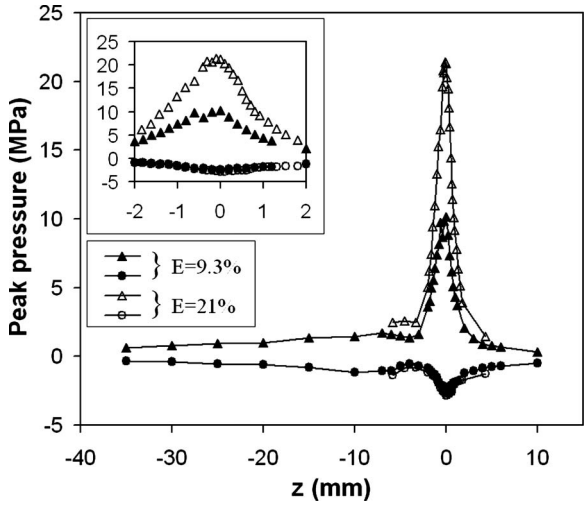


FIG. 10. The positive (triangles) and negative (circles) peak pressure distribution of the focused shock wave along z-axis near F_2. Data were taken using signal averaging over 30–64 shots at PRF=5 Hz. A zoomed view around the focal area is shown in the inset.

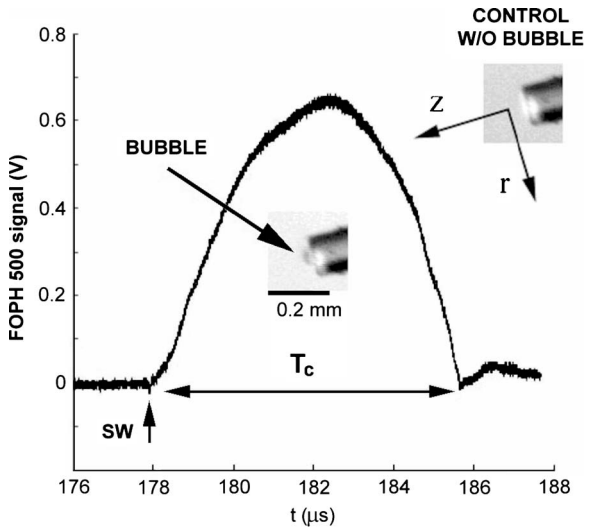


FIG. 11. Fiber optic probe hydrophone signal (140 μm fiber diameter) in response to an impinging shock wave and resultant cavitation bubble. Inset: single bubble formed on the end face of the fiber near its maximum expansion following the shock wave impact.

pressures (Fig. 8). Further increase of the laser output energy leads to increased pulse amplitude and pulse duration at F_1, which would result in larger focus size based on linear diffraction theory [Eq. (1)]. In addition, nonlinear refraction and absorption limit the pressure gain at F_2 especially for high amplitude waves. Elongation of laser plasma at high energy levels also spreads the acoustic energy to a large area, leading to a reduced pressure gain at F_2 (Fig. 9). Hence, for validation of theoretical models, a low energy setting with resultant spherical bubble in the laser-induced shockwave system should be used.

The asymmetric effect of the jitter orientation on resultant pressure reduction at F_2 (see Fig. 9) indicates that in an HM-3 where the electrode axis is tilted at 76° from the reflector axis (which is close to the most sensitive direction of 90°), the enlarged gap between the electrode tip as treatment progresses may produce a significant reduction of the peak pressure at F_2 with a concomitantly increased beam size. In comparison, the electrode in most of the newer generation electrohydraulic shock wave lithotripters is aligned with the reflector axis (which is in the least sensitive direction of 0°). Therefore, the jitter due to enlarged gap of the electrode in the newer EH lithotripters may have a less significant impact on beam size change at F_2.

Laser-generated shock waves may not be suitable for clinical lithotripsy because of the strong absorption of high frequency acoustic waves in biological tissues, which limits the penetration depth of the wave. Interestingly, fragmentation of rosin stone in water by laser-generated shock waves has been demonstrated by Musatov.[29] Their results support spallation as a potential mechanism of stone fragmentation. Because cavitation produced by laser-induced shock waves is weak[30] and the resultant maximum bubble radius is less than 0.1 mm (Fig. 11), such shock waves may be valuable for investigating the propagation and interaction of stress waves in stone phantoms by photoelastic imaging[31] with minimal interference from cavitation bubbles. The short duration of laser-induced shock wave should significantly increase the resolution of the photoelastic image. Alternatively, the microbubbles produced by laser-induced shock wave may provide a useful means for investigating bubble-cell interaction in the context of cavitation-mediated drug and gene delivery.

ACKNOWLEDGMENTS

This work was supported in part by NIH through Grant Nos. RO1-DK52985 and S10-RR16802.

[1]C. G. Chaussy and G. J. Fuchs, "Current state and future-developments of noninvasive treatment of human urinary stones with extracorporeal shock-wave lithotripsy," J. Urol. (Baltimore) **141**, 782–789 (1989).
[2]J. E. Lingeman, "Extracorporeal shock wave lithotripsy - Development, instrumentation, and current status," Urol. Clin. North Am. **24**, 185–211 (1997).
[3]J. E. Lingeman, S. C. Kim, R. L. Kuo, J. A. McAteer, and A. P. Evan, "Shockwave lithotripsy: Anecdotes and insights," J. Endourol **17**, 687–693 (2003).
[4]H. A. Fuselier, L. Prats, C. Fontenot, and A. Gauthier, "Comparison of mobile lithotripters at one institution: Healthtronics Lithotron (TM), Dornier MFL-5000, and Dornier Doli," J. Endourol **13**, 539–542 (1999).
[5]R. Gerber, U. E. Studer, and H. Danuser, "Is newer always better? A comparative study of 3 lithotrlptor generations," J. Urol. (Baltimore) **173**, 2013–2016 (2005).
[6]S. F. Graber, H. Danuser, W. W. Hochreiter, and U. E. Studer, "A prospective randomized trial comparing 2 lithotriptors for stone disintegration and induced renal trauma," J. Urol. (Baltimore) **169**, 54–57 (2003).
[7]Y. F. Zhou and P. Zhong, "Suppression of large intraluminal bubble expansion in shock wave lithotripsy without compromising stone comminution: Refinement of reflector geometry," J. Acoust. Soc. Am. **113**, 586–597 (2003).
[8]X. F. Xi and P. Zhong, "Improvement of stone fragmentation during shock-wave lithotripsy using a combined EH/PEAA shock-wave generator - In vitro experiments," Ultrasound Med. Biol. **26**, 457–467 (2000).
[9]P. Zhong and Y. F. Zhou, "Suppression of large intraluminal bubble expansion in shock wave lithotripsy without compromising stone comminution: Methodology and in vitro experiments," J. Acoust. Soc. Am. **110**, 3283–3291 (2001).
[10]D. L. Sokolov, M. R. Bailey, and L. A. Crum, "Dual-pulse lithotripter accelerates stone fragmentation and reduces cell lysis in vitro," Ultrasound Med. Biol. **29**, 1045–1052 (2003).
[11]M. F. Hamilton, "Transient axial solution for the reflection of a spherical wave from a concave ellipsoidal mirror," J. Acoust. Soc. Am. **93**, 1256–1266 (1993).
[12]M. A. Averkiou and R. O. Cleveland, "Modeling of an electrohydraulic lithotripter with the KZK equation," J. Acoust. Soc. Am. **106**, 102–112 (1999).
[13]Y. F. Zhou and P. Zhong, "The effect of reflector geometry on the acoustic field and bubble dynamics produced by an electrohydraulic shock wave lithotripter," J. Acoust. Soc. Am. **119**, 3625–3636 (2006).
[14]M. Tanguay and I. Colonius, "Numerical simulation of bubble cavitation flow in shock wave lithotripsy," in *Fourth International Symposium on Cavitation, CAV2001*, Pasadena, USA, 2001.
[15]M. Tanguay and T. Colonius, "Progress in modeling and simulation of shock wave lithotripsy (SWL)," in *Fifth International Symposium on Cavitation, CAV2003*, Osaka, Japan, 2003.
[16]A. Szeri, "Numerical modeling of shock wave focusing and bubble dynamics," personal communication (2006).
[17]J. I. Iloreta, Y. F. Zhou, G. N. Sankin, P. Zhong, and A. J. Szeri, "Assessment of shock wave lithotripters via cavitation potential," Phys. Fluids **19**, 086103 (2007).
[18]O. Lindau, "Untersuchungen zur lasererzeugten Kavitation (Investigation on laser-induced cavitation)," Ph.D. thesis, Georg-August-Universität, 2001 (in German).
[19]I. Akhatov, O. Lindau, A. Topolnikov, R. Mettín, N. Vakhitova, and W. Lauterborn, "Collapse and rebound of a laser-induced cavitation bubble," Phys. Fluids **13**, 2805–2819 (2001).
[20]A. A. Buzukov, Y. A. Popov, and V. S. Teslenko, "Experimental study of explosion caused by focusing monopulse laser radiation in water," Zh.

Prikl. Mekh. Tekh. Fiz. **10**, 17–24 (1969) (in Russian) [J. Appl. Mech. Tech. Phys. **10** 701–708 (1972)].
[21]J. Noack and A. Vogel, "Single-shot spatially resolved characterization of laser-induced shock waves in water," Appl. Opt. **37**, 4092–4099 (1998).
[22]A. Vogel, K. Nahen, D. Theisen, and J. Noack, "Plasma formation in water by picosecond and nanosecond Nd:YAC laser pulses. 1. Optical breakdown at threshold and superthreshold irradiance," IEEE J. Sel. Top. Quantum Electron. **2**, 847–860 (1996).
[23]D. A. Christensen, *Ultrasonics Bioinstrumentation* (Wiley, New York, 1988).
[24]G. N. Sankin, W. N. Simmons, S. L. Zhu, and P. Zhong, "Shock wave interaction with laser-generated single bubbles," Phys. Rev. Lett. **95**, 034501 (2005).
[25]Y. V. Sud'enkov and E. V. Ivanov, "Experimental study of the focusing of submicrosecond pressure pulses in liquids," Tech. Phys. **43**, 714–719 (1998).
[26]A. G. Musatov, O. V. Rudenko, and O. A. Sapozhnikov, "Nonlinear refraction and nonlinear absorption in the focusing of high-intensity pulses," Sov. Phys. Acoust. **38**, 274–279 (1992).
[27]J. Noack and A. Vogel, "Laser-induced plasma formation in water at nanosecond to femtosecond time scales: Calculation of thresholds, absorption coefficients, and energy density," IEEE J. Quantum Electron. **35**, 1156–1167 (1999).
[28]A. J. Coleman, M. J. Choi, J. E. Saunders, and T. G. Leighton, "Acoustic-emission and sonoluminescence due to cavitation at the beam focus of an electrohydraulic shock-wave lithotripter," Ultrasound Med. Biol. **18**, 267–281 (1992).
[29]A. G. Musatov, "Destruction of solids by powerful ultrasonic pulses," Acoust. Phys. **41**, 100–104 (1995).
[30]M. H. Niemz, C. P. Lin, C. Pitsillides, J. Cui, A. G. Doukas, and T. F. Deutsch, "Laser-induced generation of pure tensile stresses," Appl. Spectrosc. **70**, 2676–2678 (1997).
[31]X. F. Xi and P. Zhong, "Dynamic photoelastic study of the transient stress field in solids during shock wave lithotripsy," J. Acoust. Soc. Am. **109**, 1226–1239 (2001).

The role of nonlinear effects in the propagation of noise from high-power jet aircraft[a)]

Kent L. Gee[b)] and Victor W. Sparrow
Graduate Program in Acoustics, 201 Applied Science Building, The Pennsylvania State University, University Park, Pennsylvania 16802

Michael M. James,[c)] J. Micah Downing,[c)] and Christopher M. Hobbs
Wyle Laboratories, 241 18th Street South, Suite 701, Arlington, Virginia 22202

Thomas B. Gabrielson and Anthony A. Atchley
Graduate Program in Acoustics, 201 Applied Science Building, The Pennsylvania State University, University Park, Pennsylvania 16802

(Received 29 September 2007; revised 7 March 2008; accepted 10 March 2008)

To address the question of the role of nonlinear effects in the propagation of noise radiated by high-power jet aircraft, extensive measurements were made of the F-22A Raptor during static engine run-ups. Data were acquired at low-, intermediate-, and high-thrust engine settings with microphones located 23–305 m from the aircraft along several angles. Comparisons between the results of a generalized-Burgers-equation-based nonlinear propagation model and the measurements yield favorable agreement, whereas application of a linear propagation model results in spectral predictions that are much too low at high frequencies. The results and analysis show that significant nonlinear propagation effects occur for even intermediate-thrust engine conditions and at angles well away from the peak radiation angle. This suggests that these effects are likely to be common in the propagation of noise radiated by high-power aircraft. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2903871]

PACS number(s): 43.25.Cb, 43.50.Nm, 43.25.Vt [ROC] Pages: 4082–4093

I. INTRODUCTION

The role of nonlinearity in the propagation of noise radiated from high-speed jets is a question that has been the topic of investigations that span the last few decades. From the perspective of investigations on full-scale jets, Morfey and Howell[1] showed that flyover measurements made on the Concorde and other high-power aircraft exhibited anomalously low atmospheric absorption at high frequencies; an analysis of their recording equipment and the measurement environment indicated that nonlinear energy transfer to high frequencies was a possible explanation for the anomalously low absorption. More recently, analyses by Gee *et al.*[2,3] of the F/A-18E Super Hornet engine run-up data have shown evidence of nonlinear energy transfer in terms of measured versus linearly predicted spectra at multiple distances, as well as by calculations of nonlinearity indicators. Model-scale jet noise experiments that demonstrate evidence of nonlinear propagation effects include those by Gallagher and McLaughlin,[4] Petitjean *et al.*,[5] and Gee *et al.*[6]

Others have been motivated by the hypothesis that nonlinear effects are present in high-amplitude jet noise but have approached the problem indirectly by conducting controlled-source measurements or by performing numerical experiments. Pernet and Payne[7] studied the nonlinear evolution of noise signals in a plane-wave tube. Pestorius and Blackstock[8] extended the scope of Pernet and Payne's original work by propagating noise of greater bandwidth and amplitude in their duct. They also developed a nonlinear propagation algorithm that Pierce[9] later showed to be a numerical solution to the generalized Burgers equation (GBE). Blackstock[10] later used a modified version of their code to numerically propagate a noise recording from a T-38 aircraft at close range. The code predicted a nonlinear evolution of the waveform, but no measurements of the aircraft noise at greater distances were available for comparison against the prediction. Additionally, Webster and Blackstock[11] carried out free-field, high-amplitude noise experiments with an array of horn-coupled loudspeakers. They found significant evidence of nonlinear energy transfer in many of their experiments. Furthermore, when their controlled-source noise spectra were compared to a measured spectrum from a KC-135A aircraft, the controlled-source spectral levels were found to be considerably lower, suggesting that nonlinearity likely affected noise propagation from the KC-135A and other high-power aircraft. Similar arguments were made by Gee *et al.*[12] in the conclusions of a recent propagation study that employed a large horn-coupled electropneumatic driver as a source. Finally, Crighton and Bashforth,[13] Scott,[14] Lighthill,[15] Punekar,[16] and Menounou and Blackstock[17] have all analytically or numerically looked at aspects of the nonlinear propagation of noise.

[a)]Portions of this work were presented at the 2005 Joint Acoustical Society of America and Noise-Con Meeting in Minneapolis, MN and at the 12th AIAA/CEAS Aerocoustics Conference in Monterey, CA, May 2006.

[b)]Present address: Department of Physics and Astronomy, N243 ESC, Provo, UT 84602; electronic mail: kentgee@byu.edu

[c)]Present address: Blue Ridge Research and Consulting, LLC, 13 $\frac{1}{2}$ W. Walnut St., Asheville, NC 28801.

 0001-4966/2008/123(6)/4082/12/$23.00

A broad look at these prior studies tells us that there is certainly compelling evidence that nonlinear effects can influence the propagation of high-amplitude jet noise. However, despite these numerous previous investigations, there has not yet been an experiment that comprehensively addresses, from experimental and modeling standpoints, the question of the prevalence or significance of nonlinear effects in the propagation of noise radiated from a full-scale, high-power (e.g., military) jet aircraft.

In this article, we demonstrate that the propagation of noise from a high-power military jet aircraft can be highly nonlinear. We present the outcome of propagation measurements made on an F-22A Raptor during static engine run-up tests and compare the measured spectra against those predicted by two propagation models. The first model is a GBE-based nonlinear model[18] that is related to the work of Anderson[19] and Pestorius and Blackstock.[8] This model has been recently used to study the outdoor propagation of finite-amplitude periodic signals.[12] The second model is a free-field, linear propagation model that includes the effects of spherical spreading and atmospheric absorption and dispersion. The comparisons show that nonlinear effects are significant for multiple angles and engine powers and are not limited to, for example, the Mach wave (peak radiation) angle at afterburner. The scope of this article is significantly broader than that of Refs. 20 and 21, in which preliminary results were presented.

II. OVERVIEW OF PROPAGATION MODELS

A. Nonlinear model

The nonlinear propagation model is based on a formulation of the GBE that incorporates cumulative quadratic nonlinearity, atmospheric absorption and dispersion, and spherical spreading. In a retarded time frame, this formulation of the GBE may be written as

$$\frac{\partial p}{\partial r} = \frac{\beta}{2\rho_0 c_0^3}\frac{\partial p^2}{\partial \tau} + \psi_\tau\{p\} - \frac{1}{r}p, \tag{1}$$

where $p(r,\tau)$ is the acoustic pressure, r is the range variable, τ is the retarded time of propagation between the input distance and r, β is the coefficient of nonlinearity, ρ_0 is the ambient atmospheric density, c_0 is the small-signal sound speed, and ψ_τ is an operator representing atmospheric absorption and dispersion that acts on $p(r,\tau)$.

Equation (1) is solved with a hybrid time-frequency domain algorithm that is based on the work of Anderson[19] and Pestorius and Blackstock[8] and is described in detail in Ref. 18. Briefly, in the hybrid time-frequency domain solution of the GBE, an input time waveform is propagated to a greater distance via small spatial steps (in our case, one-tenth of the shock formation distance at each propagation step). Because the GBE formulation in Eq. (1) shows that the evolution of the pressure with distance is equal to the addition of three separate terms for a sufficiently small spatial step, the nonlinear and linear portions of the propagation can be treated independently over this spatial step. This allows the nonlinearity to be accounted for with the implicit Earnshaw solution in the time domain, whereas the absorption and dispersion are most conveniently handled in the frequency domain on a frequency-by-frequency basis. The spherical spreading term is a simple scaling factor in either domain but is evaluated in the frequency domain in our algorithm. A fast Fourier transform (FFT) and its inverse are used to transform the waveform to the frequency domain and back at each spatial step. Finally, in applying atmospheric absorption and dispersion to the complex pressure spectrum in the frequency domain, we are performing a FFT-based circular convolution. Because of the relatively rapid decay of the corresponding impulse response of the complex absorption transfer function, wraparound artifacts were effectively suppressed with a cosine-squared amplitude taper to the first and last 100 samples of the time waveform.

B. Linear model

It is the first term in Eq. (1) involving $\partial p^2/\partial\tau$ that produces nonlinearity; without it, Eq. (1) is simply a free-field parabolic propagation model that contains spherical spreading and atmospheric absorption and dispersion. This linearized form of Eq. (1) is the prediction model used in this study for comparisons of linear propagation versus the experiment. To obtain linearly predicted waveforms, the input waveforms are transformed to the frequency domain, where the spherical spreading and atmospheric absorption and dispersion over the propagation distances are applied. The complex pressure spectra are then inverse Fourier transformed back to the time domain to obtain the linearly predicted waveform at the comparison distance. Although this process is similar to the nonlinear model, the linear nature of these calculations allows each propagation prediction to be performed in a single spatial step.

III. MEASUREMENT SUMMARY

A. Experimental setup

Static engine run-up tests were conducted by Wyle Laboratories and Penn State for the F-22A Raptor during the early morning on 15 September 2004 at Edwards Air Force Base (EAFB). The F-22A Raptor has two Pratt and Whitney F-119 turbofan engines that are in the 160 kN (35 000 lbf) thrust class and have two-dimensional convergent-divergent nozzles capable of $\pm 20°$ thrust vectoring. (Additional information regarding the engine operating parameters is not publicly available at this time.) To measure the acoustical radiation from an engine, Bruel and Kjaer (types 4938, 4939, and 4190) and GRAS (type 40BF) condenser microphones were located at various distances along five different radials, all at a height of approximately 1.8 m. The microphone layout is shown in Fig. 1, where angles are measured relative to the jet inlet. The origin for the measurement array was located approximately 5.5 m (roughly seven to eight jet diameters) downstream from the jet nozzles. This origin reflected an attempt to locate the origin as close as possible to the dominant aeroacoustic source region downstream of the nozzle exit plane. This location, however, is only an approximation at best because not only are the exact source characteristics currently unknown, but the dominant source region is expected to vary both as a function of frequency and angle.

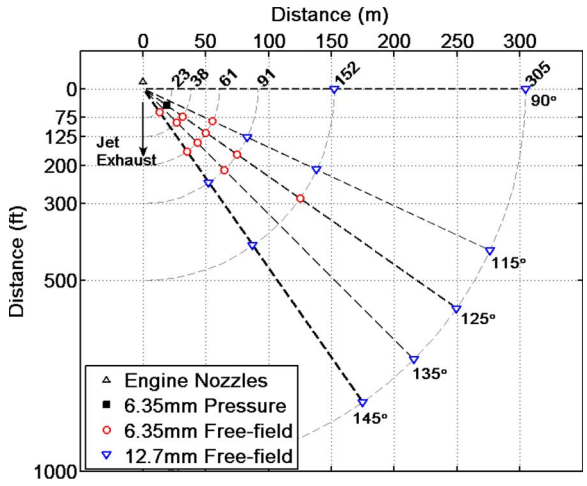


FIG. 1. (Color online) Experimental setup for the F-22A measurements at Edwards Air Force Base. Distances are shown in meters along the horizontal axis and in feet along the vertical axis.

During the tests, the engine farthest from the measurement array was held idle while the near engine's condition was varied for the run-up tests.

With the exception of the 23 m, 125° microphone, which was a 4938 pressure sensor, all microphones were free-field sensors. The 4938, 4939, and 40BF microphones have a 6.35 mm (0.25 in.) diameter diaphragm, whereas the 4190 microphones have a 12.7 mm (0.5 in.) diameter diaphragm. The 4190 microphones were located along 90° and along the 305 m arc. Because of limitations in setup time caused by security restrictions, all microphones were mounted vertically at grazing incidence, which is a nonideal configuration for the free-field sensors and affects their high-frequency response. Acquisition of the pressure waveforms was carried out using National Instruments 24 bit PXI-4472 DAQ cards with a 96 kHz sampling rate.

B. Local meteorology

The time of the tests was selected to be early morning in the hope of minimizing atmospheric effects that are usually present during the day at EAFB, namely, a significant temperature lapse and moderate winds. The run-up measurements took place between 6:30 and 8:00 a.m. Pacific daylight time (PDT), during which time atmospheric conditions were generally conducive to making propagation measurements. A meteorological station, placed at 61 m and 122.5°, monitored the local conditions during the test. The station consisted of three temperature sensors located at heights of 0.3, 1.7, and 3.3 m, two relative humidity sensors located at 0.3 and 3.3 m, and wind speed and direction gauges located at 4.3 m. Plots of the meteorological conditions during the entire testing period may be found in Ref. 18. The results of the monitoring show that relatively neutral measurement conditions occurred toward the end of the test, at approximately 7:30 a.m., when there was low wind (<0.5 m/s) and little temperature gradient (<0.3 °C/m) at the station. The particular measurements discussed hereafter were taken between 7:35 and 7:50 a.m., during which this favorable measurement environment appeared to persist. Shown in Table I are the average conditions during the three particular runs discussed in this article, which represent low engine power (idle), intermediate engine power (90% rpm), and high engine power (afterburner).

C. Expected influence of ground reflections

The jet source and microphones are both located off the ground; consequently, multipath interference effects caused by ground reflections are expected to be present in the spectra. Because both the nonlinear and linear propagation models described in Sec. II are free-field models that do not incorporate ground reflections, it is important to understand at the outset the expected effect of ground reflections on spectral behavior. A characterization of ground-induced interference effects explains the presence of "ripples" in the one-third octave spectra calculated from the measured waveforms. In addition, it illustrates that the effect of the ground cannot explain the large discrepancy that will be shown between measured spectra and predicted spectra based on the free-field linear model described in Sec. II B.

The measurements were conducted on a runway located in a dry lake bed that was several hundred meters from any buildings or other large obstructions. The terrain was extremely flat; however, the composition of the terrain varied over the measurement area, beginning with a tarmac that gave way to a lake bed loosely covered with sage brush and followed by a bare lake bed. The difference in surface hardness between the tarmac and the lake bed constitutes an impedance change along the propagation path for which an explicit accounting would normally be required. However, because the ground impedances for the various surface compositions were not measured and because lookup tables of the effective flow resistivity for various types of ground (e.g., see Ref. 22) give a rather wide range for each surface, accounting for an impedance change is not likely to be very helpful. Rather, a constant ground impedance that falls

TABLE I. Measurement times and mean ambient conditions for the F-22A engine run-up test. The temperature and relative humidity conditions have been estimated at 1.8 m via linear interpolation between measurement heights. Temperatures have been rounded to the nearest half-degree and relative humidity values to the nearest percent. Wind speeds given are at the measured height of 4.3 m.

Engine condition	Time (PDT)	Pressure (atm)	Temperature (°C)	Relative hum. (%)	Mean wind speed (m/s)
Idle	7:35	0.92	15.0	48	0.1
90% rpm	7:49	0.92	16.5	51	0.5
Afterburner	7:36	0.92	15.0	48	0.1

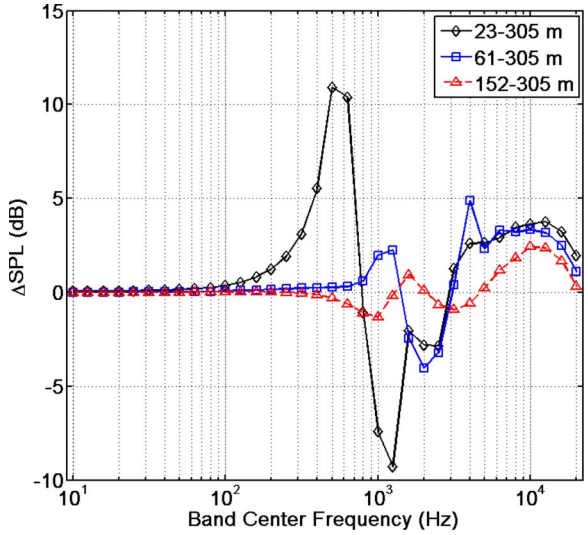


FIG. 2. (Color online) Predicted change in sound pressure level (relative to free-field predicted levels) due to ground interactions (ΔSPL) between 1.8 m high microphones for various ranges as a function of one-third octave band center frequency.

within the range of "hard ground" has been used in calculations with the recognition that the analysis is a somewhat qualitative assessment of the anticipated impact of the ground on spectral calculations.

The particular model used to make these ground reflection calculations accounts for the interaction of spherically spreading waves with a finite-impedance ground[23] as well as the effects of turbulence[24] on spectral minima and maxima. In these calculations, the point source and the receiver are both assumed to be located at a height of 1.8 m. An effective flow resistivity of 4000 kPa s/m^2 has been assumed, which nominally corresponds to exposed, rain-packed dirt.[22] The turbulence model[24] employed has as input parameters a turbulence length scale and a fluctuating index of refraction, which were set to 1.1 m and 3.0×10^{-6}, respectively.

The relative change in the free-field sound pressure level due to the ground, ΔSPL, is shown in Fig. 2 as a function of one-third octave band center frequency for the propagation ranges used subsequently in the prediction model comparisons. Calculations of ΔSPL between two measurement distances, rather than simply at a single distance, is appropriate for the current scenario because the waveform data at the model input distance also contain the effects of ground reflections. Consequently, Fig. 2 describes the differences in interference effects for the two measurement distances. The ΔSPL values for 23–305, 61–305, and 152–305 m shown in Fig. 2 demonstrate that the dominant ground interactions occur between 500 Hz and 5 kHz for these propagation ranges. It is important to note that at higher frequencies, the influence of the ground is only approximately a 2–3 dB increase in sound pressure level relative to a linear, free-field prediction made between the two microphone locations.

Before proceeding to an analysis of the measurement results in Sec. IV, it is noted that ground effects will be readily apparent in the one-third octave spectra calculated from the measured time data. However, the spectral nulls are not as deep and are significantly broader, particularly for the 90% rpm and afterburner cases, than indicated by this theoretical analysis. This may be attributed to the partially correlated nature and the spatial extent of the aeroacoustic sources in the jet plume, as opposed to the point source assumed in the ground reflection model. In addition, the frequencies at which the nulls occur also differ somewhat from the analysis, particularly for 91–305 m, which is certainly caused at least in part by the estimated values of the flow resistivity and turbulence coefficients. These discrepancies serve to make quantitative application of the theory in the form of spectral corrections inappropriate. However, the analysis is qualitatively useful in that (a) ground reflections may be identified as such in the measured spectra and that (b) the effect of the ground cannot reasonably explain the difference between measured spectra and spectra predicted using free-field, linear theory.

IV. MEASUREMENT RESULTS

In this section, one-third octave spectra calculated from the measured waveforms (hereafter referred to as measured one-third octave spectra) are shown as a function of distance for various engine powers and as a function of angle at two distances for one engine at afterburner. To calculate the spectrum, a waveform consisting of 2^{20} samples (about 10.9 s) was filtered with one-third octave filters to yield the average power in each band. The band pressure levels were then calculated from the average powers. To account for the placement of the free-field microphones at grazing incidence, manufacturer-supplied corrections have been added to the calculated spectra. For the 4190 microphones, the correction at 10 kHz is approximately 4 dB and grows to about 10 dB at 20 kHz. If these corrections are applied to one-third octave spectra, the maximum uncertainty of the correction is 2 dB for cases where the power in the 20 kHz one-third octave band is concentrated at either edge of the band. For the 4939 and 40BF microphones, the correction is relatively minor, only 2.5 dB at 20 kHz, with an uncertainty of less than 1 dB. Also presented in this section are time waveform segments for the near engine at 90% rpm and afterburner.

A. Function of engine power

1. Idle

The measured one-third octave spectra along 125° for idle, which represents the low-power case, are displayed in Fig. 3. The frequency axis has been restricted, from the 20 Hz to 4 kHz one-third octave bands, because instrumentation noise floor limits were reached outside this range at some of the distances. Because the jet mixing noise is relatively low in amplitude, the spectral shape at idle is not a characteristic "haystack" shape common in jet noise but contains other components of the overall engine noise. Further examination of the measured spectra between 23 and 305 m shows the effects of geometrical spreading and a gradual roll-off at high frequencies caused by atmospheric absorption. Also present is the influence of ground reflections, most noticeable for 23–61 m and above about 500 Hz. Finally, in the legend of Fig. 3, as well as those of subsequent figures, the overall sound pressure level (traditionally abbreviated as

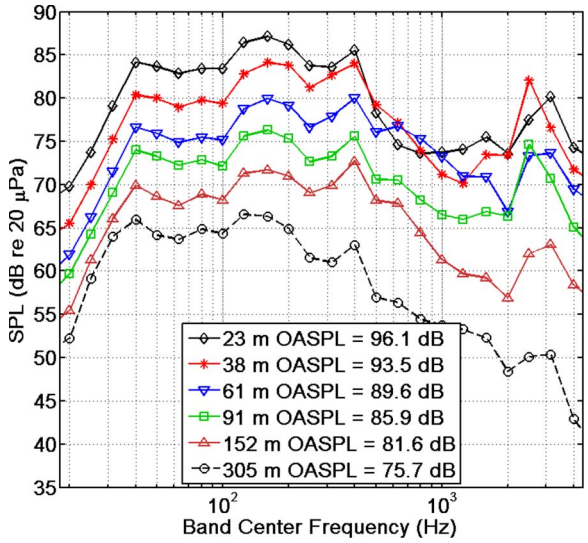


FIG. 3. (Color online) Measured one-third octave levels along 125° for idle. The measurement system noise floor limits the analysis range to the 20 Hz–4 kHz one-third octave bands. In this and in subsequent figure legends, OASPL refers to the overall sound pressure level.

OASPL within the aeroacoustics community) is given. Here, the OASPL values indicate levels that, from an occupational noise perspective, would be potentially damaging at close range if hearing protection is not worn but not high enough that nonlinear propagation effects would likely be significant.

2. 90% rpm

With the engine nearest to the measurement array at 90% rpm, the jet mixing noise is sufficient in amplitude to cause the spectra along 125° to have the characteristic haystack shape expected of jet noise close to the source. On a narrowband scale, the 23 m spectrum in Fig. 4(a) would exhibit a frequency-squared dependence at low frequencies (6 dB/octave increase) and an inverse-frequency-squared dependence at high frequencies (6 dB/octave decrease). However, because of the rate of increase of the one-third octave bandwidths, one-third octave spectra exhibit a frequency-cubed dependence at low frequencies (9 dB/octave increase) and an inverse-frequency dependence at high frequencies (3 dB/octave decrease). With the exception of the ground effects between 500 and 3 kHz, the high-frequency roll-off at 23 m obeys the expected 3 dB/octave trend. Furthermore, it is noteworthy that the slope of the roll-off does not appear to significantly change between 23 and 91 m, which could indicate nonlinear waveform steepening and shock formation during the course of propagation.

Some further discussion regarding evidence of nonlinear propagation is worthwhile at this stage. The OASPL values for 90% rpm are substantially greater than for idle and are large enough that nonlinear effects may be expected when compared to the results of past experiments with controlled sources.[11,12] The evolution of the high-frequency roll-off between 23 and 305 m in Fig. 4(a) suggests that these nonlinear effects are indeed present. According to nonlinear theory,[25] significant shock formation in a random noise waveform will cause the high-frequency spectrum to exhibit a 6 dB/octave roll-off (3 dB/octave on a one-third octave

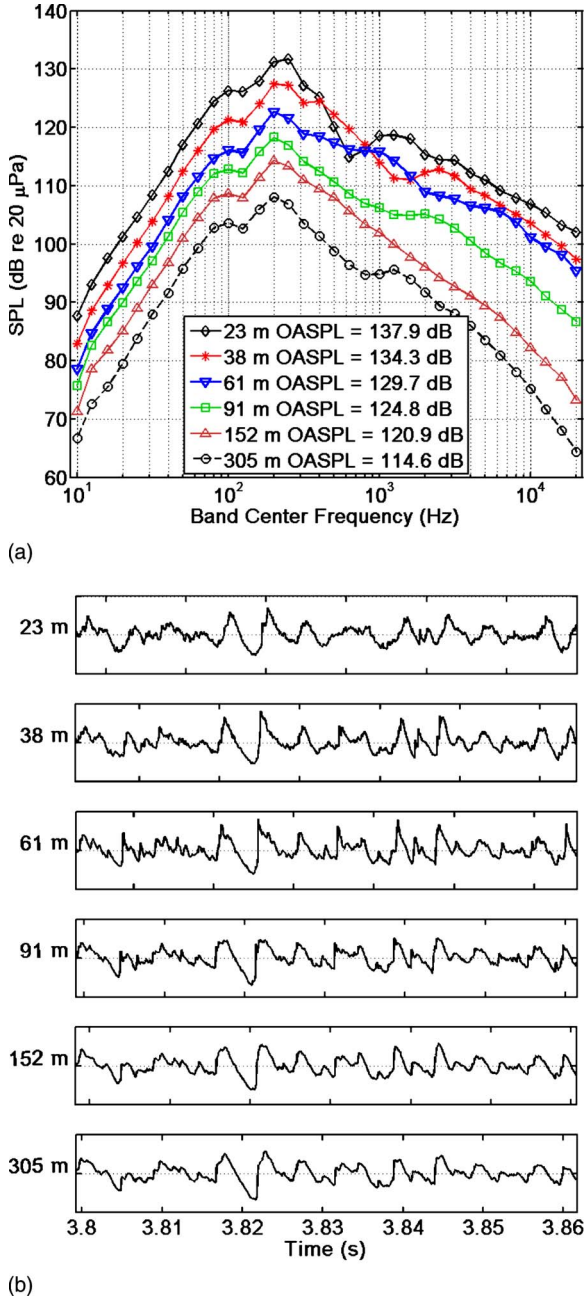


FIG. 4. (Color online) (a) Measured one-third octave levels along 125° for one engine at 90% rpm. (b) Amplitude-scaled and time-aligned waveform segments at 23–305 m. The ordinate limits for each waveform are ±800 Pa.

scale). This power-law roll-off will continue out to a frequency that is on the order of the characteristic inverse shock rise time, where it is replaced with an exponential roll-off. As the waveform travels farther, atmospheric absorption causes the shocks to begin to unsteepen and the frequency at which the power law gives way to the exponential roll-off to decrease. Examination of the high-frequency spectral trends in Fig. 4(a) shows that they appear to mimic this behavior. Between 23 and 91 m, little change occurs in the spectral slope out to 20 kHz, suggesting that significant shocks have formed. By 152 m, the power-law roll-off has given way to an exponential roll-off above 5 kHz and the frequency at which this transition occurs decreases by 305 m.

Additional evidence of this phenomenon can be seen by examining a small segment of the 90% rpm time waveform at each of the distances. Time-aligned portions of the waveforms that have been amplitude scaled by multiplying them by the ratio of the measurement distance and 23 m, which removes the effects of assumed spherical spreading, are displayed in Fig. 4(b). Visual analysis of the time waveform suggests a steepening trend that continues out to 61 or 91 m. Beyond 91 m, some of the shocks that have formed appear to thicken, which suggests the increasing relevance of atmospheric losses that manifest themselves as the exponential roll-off in the spectra in Fig. 4(a).

3. Afterburner

The measurement results for afterburner and 125°, displayed in Fig. 5, are similar to those of 90% rpm and so only significant differences are mentioned. First, Fig. 5(a) shows that, relative to 90% rpm, OASPL values have increased by 5–7 dB at each of the measurement locations. This suggests a greater importance of nonlinear effects, which is corroborated by the fact that the high-frequency spectral decay rate in Fig. 5(a), although somewhat complicated by the ground reflections, is clearly less than that for 90% rpm at 152 and 305 m. A qualitative comparison of the afterburner waveform segments (again amplitude scaled to remove spherical spreading) in Fig. 5(b) with the 90% rpm waveforms in Fig. 4(b) reveals that the shocks present in the afterburner waveform at 305 m appear to have shorter rise times than those in the 90% rpm waveform. The shock rise time comparison agrees with the observed high-frequency behavior in the spectra for the two engine conditions.

B. Function of angle

The microphone layout for the propagation experiments, displayed in Fig. 1, showed that four 6.35 mm microphones were located between 115° and 145° at 61 m. The measured one-third octave spectra for these microphones and the near engine at afterburner are displayed in Fig. 6(a). The spectra are similar at the four angles, with similar OASPL values, although the spectral peak frequency and the severity of the ground interference nulls both vary as a function of angle.

At 305 m, in addition to 115°–145°, a microphone was also located at 90°. The 305 m, afterburner spectra for these angles is shown in Fig. 6(b). Because the primary radiation direction of the jet mixing noise is to the rear of the aircraft, the peak-frequency region of the spectrum at 90° is somewhat ill defined and the OASPL is 10–12 dB less than the other angles. Given the lower OASPL at this angle, nonlinear effects would likely play a lesser role, so it is not surprising that the rate of the high-frequency roll-off is much greater at 90° than at the other angles. This is not to say that nonlinear effects do not play some role at 90°, however. Comparisons between measurement and predicted spectra along 90° and other angles are shown in Sec. V to determine the level of agreement between models and experiment as well as the relative significance of nonlinear propagation effects for a given case.

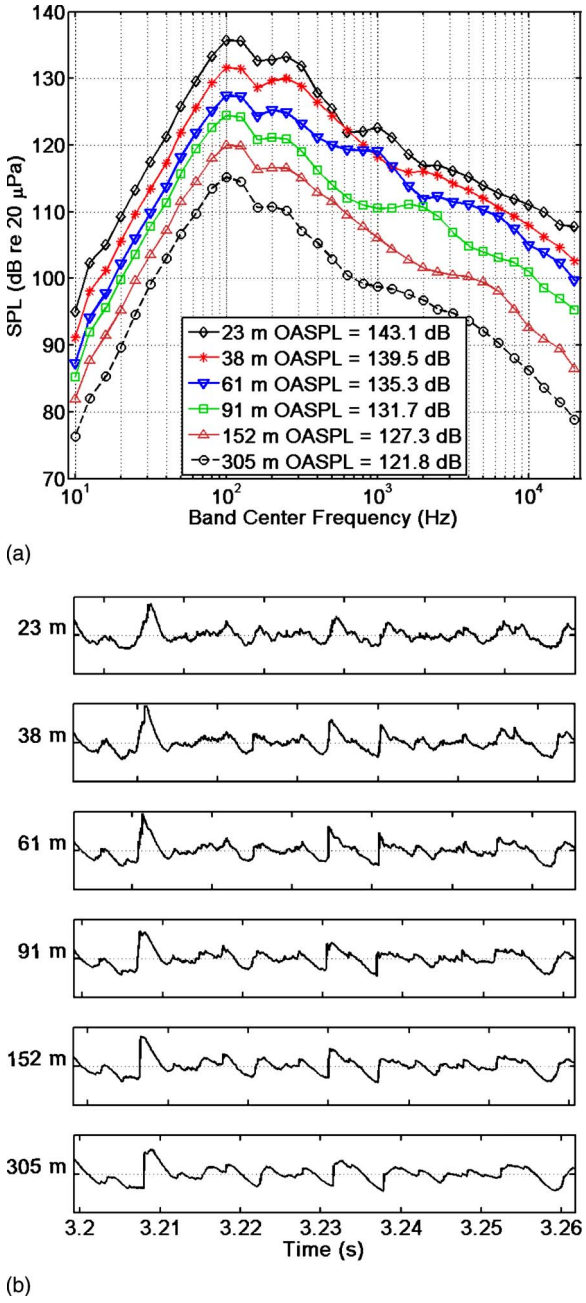
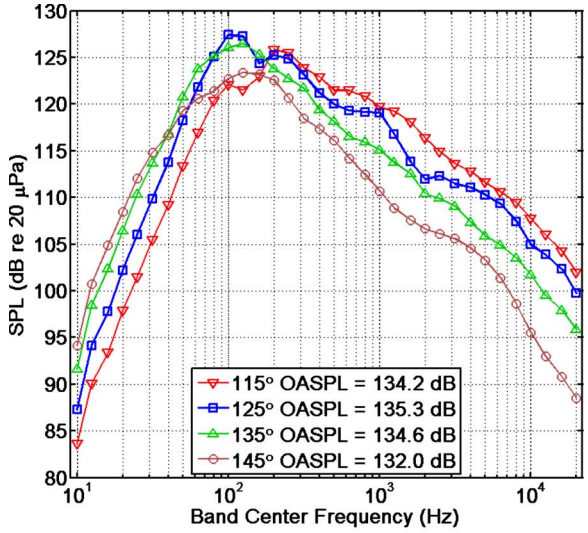


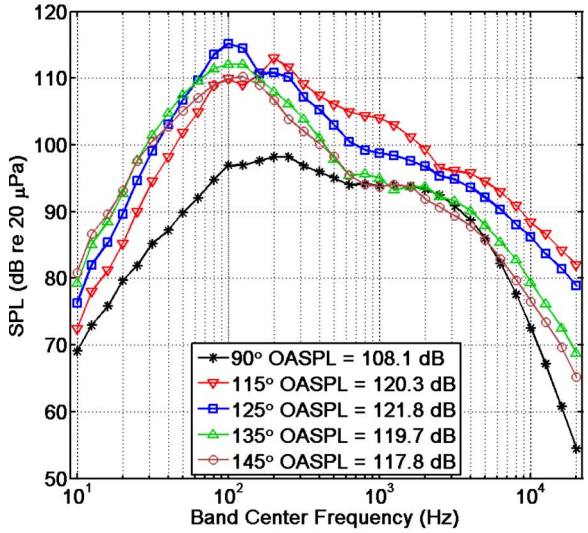
FIG. 5. (Color online) (a) Measured one-third octave levels along 125° for one engine at afterburner. (b) Amplitude-scaled and time-aligned waveform segments at 23–305 m. The ordinate limits for each waveform are ±1800 Pa.

V. COMPARISONS WITH MODELS

This section contains several comparisons between measured spectra and spectra predicted from free-field linear and nonlinear GBE-based propagation models. The comparisons are divided into three sections: (a) comparisons as a function of engine power (idle, 90% rpm, and afterburner) along 125°, (b) GBE model comparisons as a function of algorithm input distance, and (c) comparisons as a function of propagation angle for both 90% rpm and afterburner. It is noted at the outset that these comparisons are largely visual, with mention of the level of agreement at specific frequencies made. A prior study,[12] in which similar model-versus-measurement comparisons were made using a high-



(a)



(b)

FIG. 6. (Color online) Measured afterburner one-third octave levels at (a) 61 m and (b) 305 m.

amplitude controlled source, employed a quantity called the mean absolute error to quantify differences between measurement and predictions. However, the significance of this quantity is somewhat artificial in the sense that the greatest contributions to the mean absolute error for the case of significant nonlinearity and even moderate propagation ranges will come at the highest frequencies analyzed, where the difference between a linear prediction and the measurement will be the greatest. Because of the drastically different functional forms of the high-frequency spectral behavior for the linearly predicted and the measured spectra, use of the mean absolute error has been discarded in this article, leaving the comparisons between model and experiment mainly visual in nature.

As described in Sec. II, nonlinearly and linearly predicted waveforms were obtained by evaluating the nonlinear and linear[26] models out to the propagation comparison distance, which is 305 m. The input waveforms consisted of approximately 10.9 s of data (2^{20} samples), the same waveform length used to obtain the one-third octave spectra in Figs. 3, 4(a), and 5(a). To provide the most consistent comparison between models and experiment, the input waveforms used were time aligned via the retarded time to account for the propagation delay between input and receiver distances. Once the linearly and nonlinearly predicted waveforms were obtained, the predicted one-third octave spectra were calculated with the same one-third octave filtering procedure used to process the experimental data.

A. Function of engine power

The first set of comparisons carried out is for the three different engine powers with the propagation range and angle held constant. For ease of discussion, the low and high engine power results are presented first, followed by the results from the intermediate engine setting.

1. Idle

The measured and predicted spectra for both engines at idle and 125° are displayed in Fig. 7. The agreement between the 23–305 m predictions and the 305 m measurement is not extremely good; a free-field, homogeneous atmosphere model does not match the measurement with great success. Based on a study of all the data sets acquired between 6:30 and 8:00 a.m., we believe that the cause of the disagreement between the models and the measurement at low frequencies (<200 Hz) is meteorological in nature because it is not always present in the data sets (e.g., the 90% rpm test to be shown subsequently). In contrast, the disagreement between 500 Hz and 2 kHz is principally caused by ground reflections. The near equivalence of the two predicted spectra is significant and indicates that the GBE-based model predicts negligible nonlinear effects, at least out to the maximum comparison range of 4 kHz. Finally, of particular importance to subsequent comparisons between engine conditions is the fact that above 2 kHz, the measurement and models agree to within 1 dB.

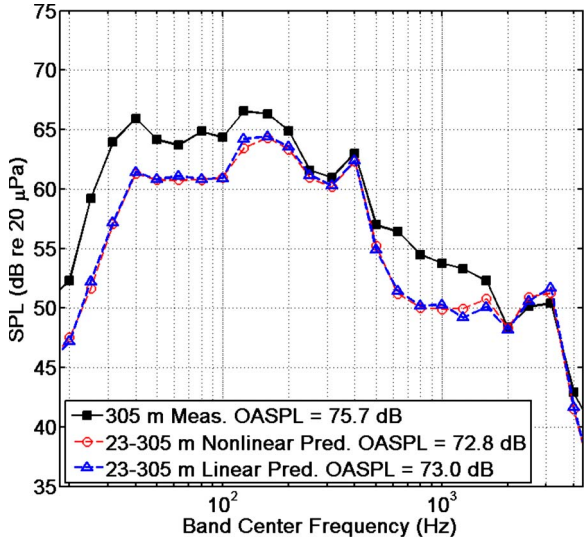
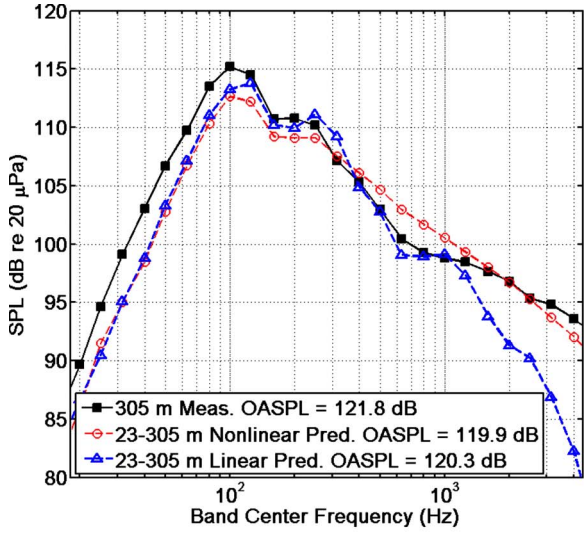
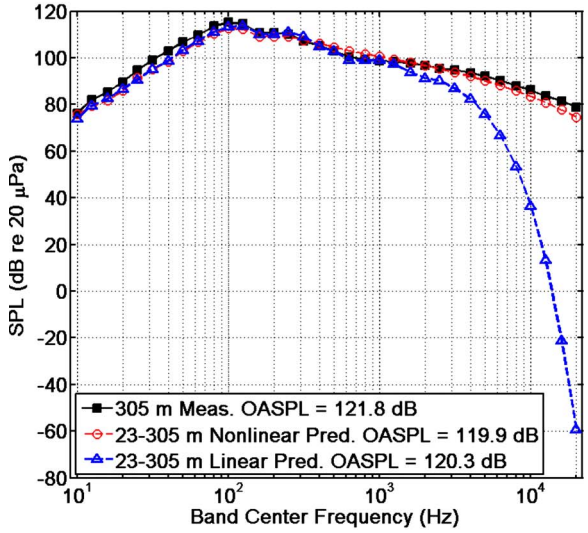


FIG. 7. (Color online) Comparison between predicted spectra and measurement for idle along 125°.



(a)



(b)

FIG. 8. (Color online) Comparison between predicted spectra and measurement for afterburner along 125°. In (a), the axes have been restricted to the same ranges used in the idle comparison (see Fig. 7). In (b), an expanded scale is shown, which shows the level of disagreement between the linear prediction and the measurement at high frequencies.

2. Afterburner

The afterburner run-up measurement whose results were discussed previously in Sec. IV A 3 was taken less than 1 min after the idle measurement. The results of the comparison between the predicted and measured spectra are displayed in Fig. 8(a), over the same limited frequency range as the idle comparison and also with a 40 dB vertical axis. A comparison between Figs. 8(a) and 7 reveals a similar low-frequency discrepancy between the model and measurement. Because the change in meteorological conditions during the course of the measurements was gradual, and the fact that this same discrepancy occurs in the low-amplitude idle measurement taken less than 1 min previously, it is believed that this low-frequency increase in spectral level relative to predictions is a linear, as opposed to a nonlinear, phenomenon.[29]

In contrast to the idle results from Fig. 7, the greatest disparity between the linear model and measurement for afterburner does not occur at low frequencies but rather above 2 kHz. In Fig. 8(a), the nonlinear prediction and measurement agree within 2 dB over this range, whereas there is a 12 dB difference between the linear prediction and the measurement. Extrapolating the spectral trends out to higher frequencies indicates that the difference between measurement and linear prediction should be much greater at the full analysis bandwidth of 20 kHz. This hypothesis is confirmed in Fig. 8(b), where the same afterburner prediction is shown on an expanded scale, from 10 Hz to 20 kHz. The difference between the measured spectrum and the nonlinear prediction is only about 4 dB for the 20 kHz one-third octave band, whereas the difference between the linear prediction and measured spectrum is 140 dB. Therefore, Fig. 8(b) represents strong evidence of how highly nonlinear the propagation of noise from a high-power jet can be.

The emphasis in these comparisons between predictions and measurement is on the one-third octave spectra because they conveniently demonstrate the average agreement between the predicted and measured noise waveforms and permit observation of behavior not readily seen in the time waveforms. However, because the models do, in fact, provide predicted waveforms and these waveforms are time aligned with the measured waveforms at 305 m, direct comparison of the predicted and measured waveforms is possible. The afterburner case has been chosen because the differences between the nonlinear and linear propagation predictions are easily seen. The nonlinearly and linearly predicted waveform segments at afterburner and for the same time segment as displayed in Fig. 5(b) are displayed in Fig. 9. The similarities between the waveform steepening in the nonlinear prediction from 23 m and the measured waveform are easily noted. On the other hand, the linear prediction at 305 m is only slightly smoother than the 23 m waveform in

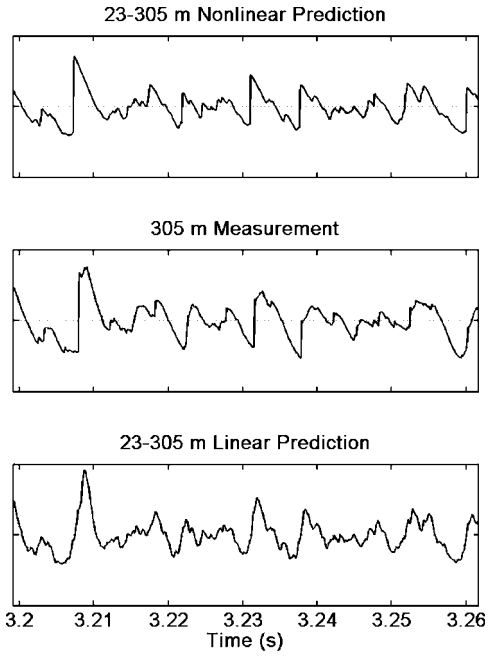


FIG. 9. Nonlinearly predicted, measured, and linearly predicted waveforms at 305 m. The ordinate limits for each waveform are ±140 Pa.

Fig. 5(b), which is due to the relatively low atmospheric absorption coefficient in the characteristic-frequency region and the inability to see the absorption-induced roll-off of the high-frequency content on this time scale.

3. 90% rpm

The previous results for afterburner showed that the propagation can be highly nonlinear, but what about other engine conditions where less thrust is provided? An intermediate engine condition is well represented by 90% rpm, which means that the turbomachinery in the engine is running at 90% of the rate of what it would be at military power (also commonly referred to as "Mil" power), which is 100% rpm. (Note that the afterburner yields both greater thrust and noise levels than military power.) Analysis of the measured one-third octave spectra, coupled with examination of the waveforms themselves, provided evidence that the propagation was, in fact, nonlinear. A comparison between the two models and the measured spectrum along 125° confirms this, first for the restricted range used in the comparison for idle and, second, for the expanded range, 10 Hz–20 kHz. These two comparisons are displayed in Figs. 10(a) and 10(b). The agreement between measurement and models at low frequencies is significantly better for this case than for the idle and afterburner comparisons, for which the data were taken about 10 min earlier. In Fig. 10(a), the disagreement between the measurement and linear model is only 7 dB at 4 kHz, less than it was for the afterburner case. However, with the expanded scale in Fig. 10(b), the disagreement between the measurement and linear prediction grows to more than 120 dB at 20 kHz, whereas there is only a 7 dB difference between the nonlinear model and the measurement. This again shows that the propagation along the peak radiation angle can be highly nonlinear, even for an intermediate-thrust engine setting. In other words, this result begins to establish the possible commonness of nonlinear propagation effects in noise radiated from a high-thrust military aircraft.

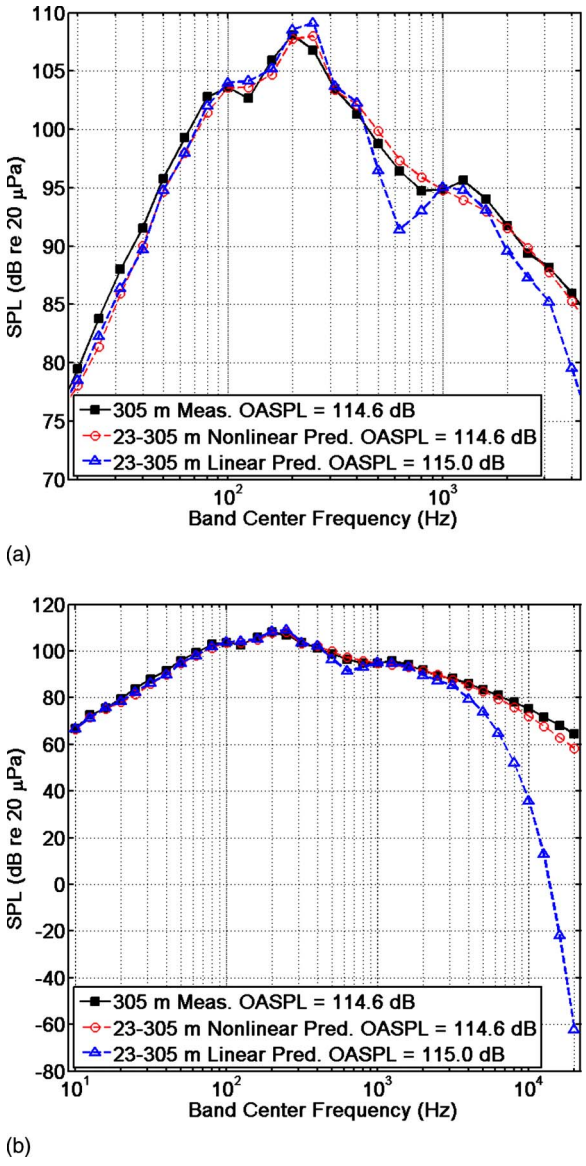


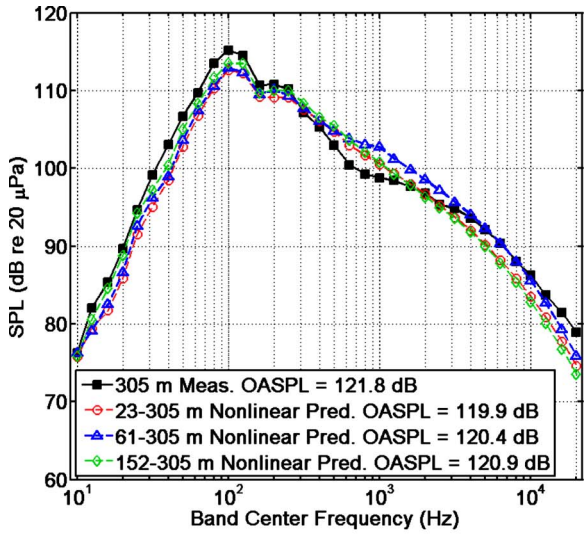
FIG. 10. (Color online) Comparison between predicted spectra and measurement for 90% rpm along 125°. In (a), the axes have been restricted to the same ranges used in the idle comparison (see Fig. 7). In (b), an expanded scale is shown.

B. Function of algorithm input distance

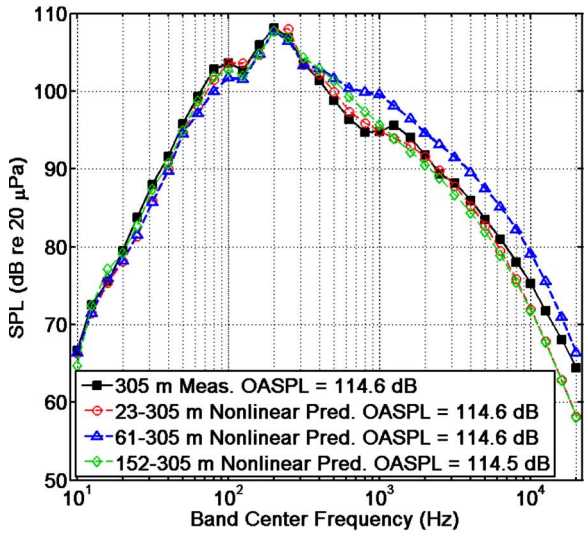
Before proceeding to a discussion of nonlinear propagation as a function of angle, it is first important to consider the issue of the algorithm input distance. The previously discussed comparisons for afterburner and 90% rpm showed good agreement between the nonlinear prediction and measurement, especially relative to the linear prediction. However, that agreement was only for one starting distance and it is possible that input waveforms acquired at different distances could yield different results, especially since it is likely that a microphone located at 23 m is not yet in the true geometric far field of the jet. To show the consistency of the propagation model results as a function of starting distance, the 125° afterburner and 90% rpm waveform data at 23, 61, and 152 m have been propagated with the GBE-based model out to 305 m. Their resulting one-third octave spectra are compared against the measured spectrum for the two engine settings in Fig. 11. The afterburner comparison in Fig. 11(a) reveals that the maximum difference between predicted spectra at any frequency is only about 1 dB. For the 90% rpm comparison in Fig. 11(b), the spread between nonlinearly predicted spectra is greater in that the 61–305 m predicted levels are too great above 400 Hz and the other prediction levels are too low above about 3 kHz. However, these results bound the measured spectrum at 305 m and represent relatively little error when compared to the spectra obtained with the linear model.

C. Function of angle

At this point, it has been established that nonlinearity affects the propagation along the peak radiation angle (125°) for intermediate- and maximum-thrust engine conditions. Furthermore, it has been verified that the results are relatively insensitive to the algorithm input distance. The remaining question to consider, therefore, is whether nonlinear propagation occurs only at the peak directivity angle or if it

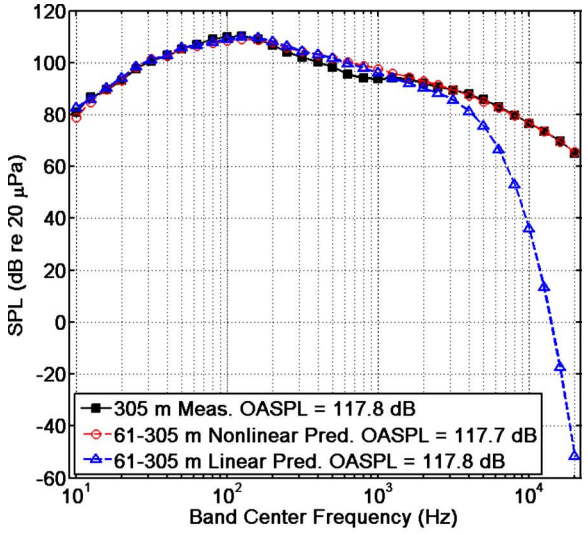


(a)

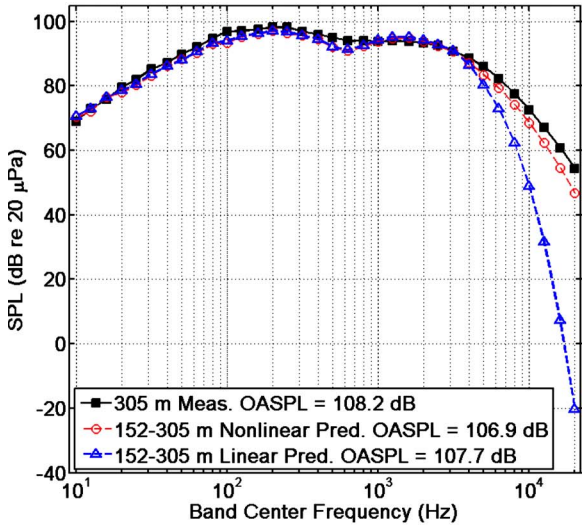


(b)

FIG. 11. (Color online) Comparison of various nonlinear predictions vs measurement along 125° as a function of algorithm starting distance for (a) afterburner and (b) 90% rpm.



(a)



(b)

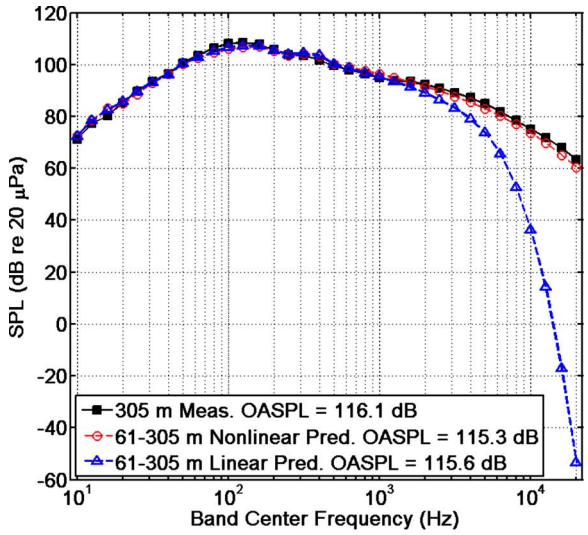
FIG. 12. (Color online) Comparisons between predicted spectra and measurement for afterburner along (a) 145° and (b) 90°.

is common at other angles as well. Because 145° and 90° represent the farthest angles from 125° over which predictions can be made, comparisons for these two angles are shown for both engine conditions.

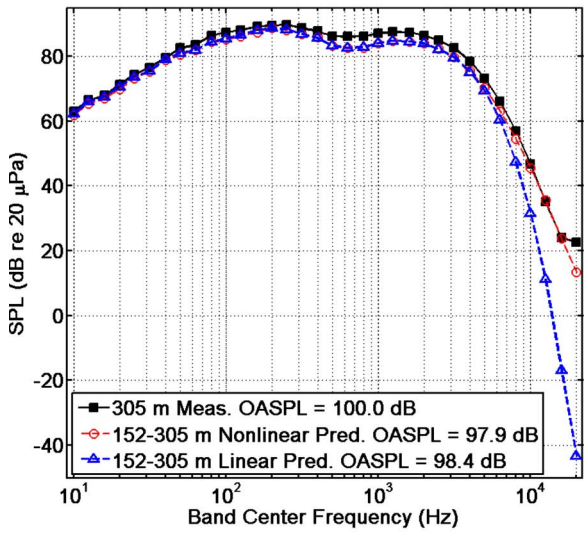
Comparisons along 145° and 90° are shown for afterburner in Fig. 12 and for 90% rpm in Fig. 13. Figure 12(a) confirms what might have been supposed from an examination of the measured spectra as a function of angle in Fig. 6, that the nonlinearity present in the propagation along 145° is similar to that experienced along 125°. This is also true for the 90% rpm comparison in Fig. 13(a). However, the results along 90° are especially significant in that, although the afterburner case exhibits greater nonlinear effects [see Fig. 12(b)], the noise propagation at 90% rpm and 90° is also nonlinear, as demonstrated by the agreement between the nonlinear model and the measurement at high frequencies in Fig. 13(b). Note that the leveling off in the measured spectrum in Fig. 13(b) at 20 kHz is caused by the measurement system noise floor being reached somewhere in that one-third octave band. The fact that the propagation appears to be appreciably nonlinear at 90° could be considered surprising in that a bispectral analysis of afterburner measurements made on the F/A-18E Super Hornet[3] revealed no evidence of nonlinearity at that angle despite the fact that the F/A-18E's OASPL at close range was several decibels greater than that of the F-22A.

VI. CONCLUDING DISCUSSION

The results of the numerous comparisons between linear and nonlinear propagation models and measurements of the F-22A Raptor have resulted in several findings. First, it has been shown that the propagation can be significantly nonlinear, even at an intermediate-thrust engine setting (90% rpm). Second, these nonlinear effects, including 90% rpm, are not limited to the peak radiation angle but extend at least over the measurement aperture of 90°–145°. Third, despite the fact that the generalized-Burgers-equation-based model does



(a)



(b)

FIG. 13. (Color online) Comparisons between predicted spectra and measurement for 90% rpm along (a) 145° and (b) 90°.

not contain all the phenomena encountered in the actual measurement (i.e., a variable atmosphere and multipath interference), favorable agreement between the measurement and the nonlinear model has been obtained for all cases considered, especially relative to the linear propagation model. These findings illustrate the likely prevalence of nonlinear propagation effects in the noise radiated by high-power jet aircraft, a question that has been debated for more than 30 years.

Although there is more to be done in terms of model refinement (e.g., the inclusion of multipath effects caused by the ground or turbulence), a potentially important direction for future research efforts is to investigate the impact of these nonlinear propagation effects on communities and individuals. From a community noise standpoint, one could study the results presented in this article and potentially conclude that, although significant nonlinear effects occur in the propagation, the impact of nonlinearity on the predicted OASPL is minimal, less than the experimental uncertainty associated with typical outdoor measurements. If this line of reasoning is coupled with recent findings[30] showing that nonlinear propagation of a waveform with a jet-noise-like spectrum may not significantly affect common single-number metrics, such as A-weighted OASPL, perceived noise level, Stevens Mark-VII loudness, and Zwicker loudness, it could result in a conclusion that nonlinear effects, though scientifically interesting, are unimportant to human perception.

However, we believe that, based on the conclusions of Ref. 30, the potential significance of nonlinear propagation effects related to human perception merits further investigation. Although Ref. 30 demonstrates that nonlinear propagation causes little change in calculated single-number metrics, it also contains multimedia content that encourages the reader to listen to the results of nonlinear versus linear propagation. Even though the nonlinearly and linearly propagated waveforms have nearly equal overall levels, as determined with the several common metrics discussed previously, the nonlinear waveform is perceived to be dramatically different than the linear waveform upon playback. Given the likely prevalence of nonlinear effects in high-power jet noise propagation, this significant difference in noise perception caused by nonlinearity is a topic that needs to be fully addressed in future research.

ACKNOWLEDGMENTS

K. L. Gee, V. W. Sparrow, J. M. Downing, M. M. James, and C. M. Hobbs were supported by the Strategic Environmental Research and Development Program. T. B. Gabrielson and A. A. Atchley were supported by the Office of Naval Research. The authors are grateful to Dr. Sally Anne McInerny of the University of Alabama-Birmingham for useful discussions during the course of this research. Robert McKinley, John Hall, and Frank Mobley of the Air Force Research Laboratory are also acknowledged for their roles in the measurement portion of this investigation.

[1]C. L. Morfey and G. P. Howell, "Nonlinear propagation of aircraft noise in the atmosphere," AIAA J. **19**, 986–992 (1981).
[2]K. L. Gee, T. B. Gabrielson, A. A. Atchley, and V. W. Sparrow, "Preliminary analysis of nonlinearity in military jet aircraft noise propagation," AIAA J. **43**, 1398–1401 (2005).
[3]K. L. Gee, A. A. Atchley, L. E. Falco, T. B. Gabrielson, and V. W. Sparrow, "Bispectral analysis of high-amplitude jet noise," AIAA (2005), Paper No. AIAA-2005-2937.
[4]J. A. Gallagher and D. K. McLaughlin, "Experiments on the non-linear characteristics of noise propagation from low and moderate Reynolds number supersonic jets," AIAA (1981), Paper No. AIAA-81-2041.
[5]B. P. Petitjean, D. K. McLaughlin, and K. Viswanathan, "Acoustic pressure waveforms measured in high speed jet noise experiencing nonlinear propagation," AIAA (2005), Paper No. AIAA-2005-0209.
[6]K. L. Gee, M. R. Shepherd, L. E. Falco, A. A. Atchley, L. S. Ukeiley, B. J. Jansen, and J. M. Seiner, "Identification of nonlinear and near-field effects in jet noise using nonlinearity indicators," AIAA (2007), Paper No. AIAA-2007-3653.
[7]D. F. Pernet and R. C. Payne, "Non-linear propagation of signals in air," J. Sound Vib. **17**, 383–396 (1971).
[8]F. M. Pestorius and D. T. Blackstock, "Propagation of finite-amplitude noise," in *Finite-Amplitude Wave Effects in Fluids*, edited by L. Bjorno (IPC Science and Technology, Guildford, 1973), pp. 24–29.
[9]A. D. Pierce, "Progressive wave equations and algorithms for sonic boom propagation," *Proceedings of Noise-Con 93* (Noise Control Foundation, Poughkeepsie, NY, 1993), pp. 157–162.
[10]D. T. Blackstock, "Nonlinear propagation of jet noise," in *Proceedings of*

the third Interagency Symposium on University Research in Transportation Noise (University of Utah, Salt Lake, UT, 1975), pp. 389–397.
[11]D. A. Webster and D. T. Blackstock, "Experimental investigation of outdoor propagation of finite-amplitude noise," NASA Contractor Report 2992, Applied Research Laboratories, The University of Texas at Austin, Austin, TX, 1978.
[12]K. L. Gee, V. W. Sparrow, M. M. James, J. M. Downing, and C. M. Hobbs, "Measurement and prediction of nonlinearity in outdoor propagation of periodic signals," J. Acoust. Soc. Am. **120**, 2491–2499 (2006).
[13]D. G. Crighton and S. Bashforth, "Nonlinear propagation of broadband jet noise," AIAA (1980), Paper No. AIAA-80-1039.
[14]J. F. Scott, "The nonlinear propagation of acoustic noise," Proc. R. Soc. London, Ser. A **383**, 55–70 (1982).
[15]J. Lighthill, "Some aspects of the aeroacoustics of high speed jets," NASA Contractor Report 191458, ICASE Report No. 93-20, 1993.
[16]J. N. Punekar, "Numerical simulation of nonlinear random noise," Ph.D. thesis, University of Southampton, Southampton, UK, 1996.
[17]P. Menounou and D. T. Blackstock, "A new method to predict the evolution of the power spectral density for a finite-amplitude sound wave," J. Acoust. Soc. Am. **115**, 567–580 (2004).
[18]K. L. Gee, "Prediction of nonlinear jet noise propagation," Ph.D. thesis, The Pennsylvania State University, University Park, PA, 2005.
[19]M. O. Anderson, "The propagation of a spherical N wave in an absorbing medium and its diffraction by a circular aperture," Technical Report No. ARL-TR-74-25, Applied Research Laboratories, The University of Texas at Austin, Austin, TX, 1974.
[20]K. L. Gee and V. W. Sparrow, "Quantifying nonlinearity in the propagation of noise from military jet aircraft," in *CD-ROM: Proceedings of Noise-Con 05*, edited by J. S. Bolton, P. Davies, and G. C. Maling, Jr. (Institute of Noise Control Engineering of the USA, Washington, DC, 2005), Paper No. nc05_194.
[21]K. L. Gee, V. W. Sparrow, M. M. James, J. M. Downing, C. M. Hobbs, T. B. Gabrielson, and A. A. Atchley, "Measurement and prediction of noise propagation from a high-power jet aircraft," AIAA J. **45**, 3003–3006 (2007).
[22]L. C. Sutherland and G. A. Daigle, "Atmospheric sound propagation," in *Encyclopedia of Acoustics*, edited by M. J. Crocker (Wiley, New York, 1997), Chap. 32, pp. 341–365.
[23]K. Attenborough, S. I. Hayek, and J. M. Lawther, "Propagation of sound above a porous half-space," J. Acoust. Soc. Am. **68**, 1493–1501 (1980).
[24]G. A. Daigle, "Effects of atmospheric turbulence on the interference of sound waves above a finite impedance boundary," J. Acoust. Soc. Am. **65**, 45–49 (1979).
[25]S. N. Gurbatov and O. V. Rudenko, "Statistical phenomena," in *Nonlinear Acoustics*, edited by M. F. Hamilton and D. T. Blackstock (Academic, San Diego, 1998), Chap. 13, pp. 377–398.
[26]These linear predictions differ slightly from those presented in Refs. 18, 20, and 21, which used a one-third octave band spectrum as input and simply applied the absorption coefficient for each of the band center frequencies. However, this can lead to erroneous predictions of atmospheric absorption losses in those bands where the energy is primarily distributed at one of the band edges. See Refs. 27 and 28 for relevant discussion.
[27]L. C. Sutherland andH. E. Bass, "Influence of atmospheric absorption on the propagation of bands of noise," J. Acoust. Soc. Am. **66**, 885–894 (1979).
[28]G. P. Howell and C. L. Morfey, "Finite bandwidth corrections applicable to noise spectra shaped by atmospheric attenuation," J. Acoust. Soc. Am. **72**, 1574–1582 (1982).
[29]The differences between the predictions and measurement for the idle case were used in Ref. 21 to empirically correct the predictions made for the afterburner case, which greatly improved the predictions over that frequency range.
[30]K. L. Gee, S. H. Swift, V. W. Sparrow, K. J. Plotkin, and J. M. Downing, "On the potential limitations of conventional sound metrics in quantifying perception of nonlinearly propagated noise," J. Acoust. Soc. Am. **121**, EL1–EL7 (2007).

Three-dimensional seismic array characterization study: Experiment and modeling[a)]

Arslan M. Tashmukhambetov, George E. Ioup, and Juliette W. Ioup
Department of Physics, University of New Orleans, New Orleans, Louisiana 70148

Natalia A. Sidorovskaia[b)]
Department of Physics, University of Louisiana at Lafayette, P.O. Box 44210, Lafayette, Louisiana 70504-4210

Joal J. Newcomb
Naval Research Laboratory-Stennis Space Center, Mississippi 39529

(Received 22 August 2007; revised 20 February 2008; accepted 5 March 2008)

In the summer of 2003, the Littoral Acoustic Demonstration Center conducted an acoustic characterization experiment for a 21-element marine seismic exploration airgun array of total volume of 0.0588 m^3 (3590 $in.^3$). Two Environmental Acoustic Recording System buoys, one with a desensitized hydrophone, were deployed at a depth of 758 m in a water depth of 990 m, near Green's Canyon in the Gulf of Mexico. Shots over a grid were recorded and calibrated to produce absolute broadband (up to 25 kHz) pressure-time dependencies for a wide range of offsets and arrival angles in the water column. Experimental data are analyzed to obtain maximum received zero-to-peak pressure levels, maximum received sound exposure levels, and pressure levels in 1/3-octave frequency bands for each shot. Experimental data are quantitatively modeled by using an upgraded version of an underwater acoustic propagation model and seismic source modeling packages for a variety of ranges and arrival angles. Experimental and modeled data show good agreement in absolute pressure amplitudes and frequency interference patterns for frequencies up to 1000 Hz. The analysis is important for investigating the potential impact on marine mammals and fish and predicting the exposure levels for newly planned seismic surveys in other geographic areas. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2902185]

PACS number(s): 43.30.Dr, 43.30.Zk, 43.80.Nd, 43.20.Mv [JAS] Pages: 4094–4108

I. INTRODUCTION

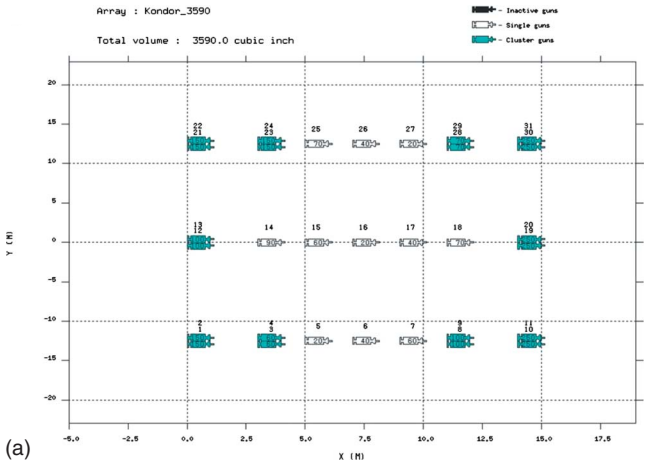
In the past decade, a considerable amount of effort has been focused on understanding how sound generated by human-made acoustic sources in the ocean may influence marine mammals. One of the important aspects of this effort is the measurement and prediction of the broadband acoustic energy distribution of such sources in complex variable ocean waveguides. Seismic exploration arrays are of interest for environmental impact assessment (Gordon *et al.*, 2004). These arrays comprise a collection of airguns distributed over an array geometry and towed behind a seismic vessel. They are designed to fire synchronously and produce powerful highly directional bottom-directed pulses to image acoustically the sub-bottom structure. The geophysical response is primarily analyzed in the low frequency band up to 300 Hz (Caldwell and Dragoset, 2000). Hence, the higher frequency component of acoustic radiation from a seismic array has been mostly overlooked until concerns were raised about the effect of this radiation on marine species, especially marine mammals, that rely on acoustics as a survival tool (for orientation, food foraging, communication, etc.). Recent studies of individual sperm whale communication codas strongly suggest that frequencies above 1000 Hz are of particular importance in sperm whale communication (Ioup *et al.*, 2005). This frequency range may overlap with the high frequency component of seismic array radiation. On-whale tag recordings during controlled exposure experiments conducted in the Northern Gulf of Mexico (GoM) in 2002 and 2003 showed that received peak pressures and sound exposure levels (SEL's) of tagged whales do not necessarily decrease as the range between the whale and the seismic array increases under certain circumstances, such as constructive interference of overlapping arrivals, the presence of a surface duct, etc. (Madsen *et al.*, 2006). Reported data show that absolute received pressure levels can be as high at 12 km as they are at 2 km. It strongly suggests that spherical and cylindrical spreading approaches should not be automatically used to determine impact zones and that animal SEL should be determined from existing waveguide propagation conditions and three-dimensional source array directional patterns.

There are discussions in the underwater acoustic community and oil industry about the results of quantitative studies of the effects of waveguide propagation including surface ducts, which are formed seasonally in the GoM, on acoustic

[a)]Portions of this work were presented in "Calibration and Analysis of Seismic Airgun Data from an EARS Buoy," Proceedings of 23rd Annual Gulf of Mexico Information Transfer Meeting, New Orleans, Louisiana, January 2005; "Modeling tools for 3-d airgun characterization studies," Proceedings of the Eighth ECUA, Carvoeiro, Portugal, June 2006; "3-D airgun source characterization and propagation modeling," SEG Technical Program Expanded Abstracts, New Orleans, Louisiana, October 2006.

[b)]Electronic mail: nas@louisiana.edu

 0001-4966/2008/123(6)/4094/15/$23.00



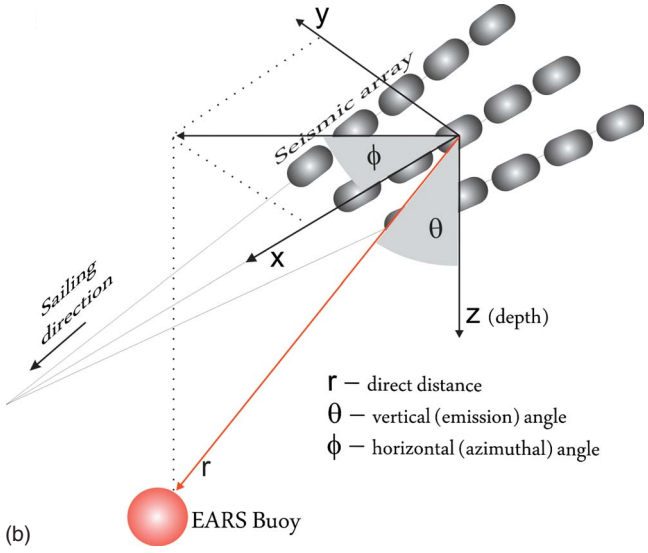


FIG. 1. (Color online) (a) The M/V Kondor seismic array configuration for the seismic characterization experiment. The numbers inside each airgun indicate the individual volume in in.^{3} of each airgun. (b) Reference coordinate system with the origin at the array center.

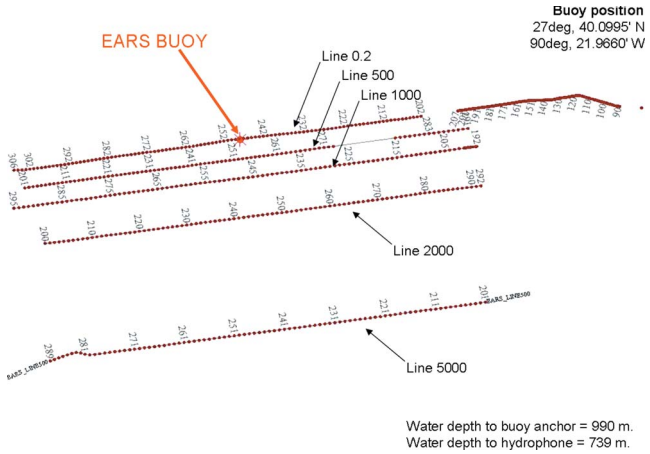


FIG. 2. (Color online) The M/V Kondor line/shot diagram in the horizontal plane. The nominal seismic array depth is 6.7 m below the surface.

energy distribution (MacGillivray, 2006; DeRuiter *et al.*, 2006; Tolstoy *et al.*, 2004). Surface ducts can form a series of energetically powerful precursor pulses (arriving before the main energy associated with the direct arrival) spread throughout the entire depth of the water column with a range decay rate slower than that of the direct arrival (Labianca, 1972; Monjo and DeFerrari, 1994; Sidorovskaia and Werby, 1995; Sidorovskaia, 2004). Therefore, an animal at any depth can be exposed to significant levels of acoustic energy that are not associated with the direct arrival. Hence, waveguide propagation modeling should become an indispensable part of the development of any mitigation metrics. Both calibrated measurements and quantitative modeling of a seismic array energy distribution for a full range of angles and emitted frequencies become the first steps in our ability to predict and mitigate any potentially negative effects.

In the first part of this paper, we present experimental calibrated measurements of the broadband absolute pressure output from an industrial seismic exploration airgun array, which has been collected by the Littoral Acoustic Demonstration Center (LADC) in June 2003 for three-dimensional seismic source characterization studies. LADC, which was founded in 2001, currently is a consortium of scientists from three universities (the University of New Orleans, the University of Southern Mississippi, and the University of Louisiana at Lafayette) and the Naval Research Laboratory at Stennis Space Center. Since 2001, LADC has conducted or participated in eight experiments in the Northern GoM and the Mediterranean Sea to study natural and anthropogenic noise in marine environments and the potential impact on marine mammals (Newcomb *et al.*, 2002a, 2002b; Newcomb *et al.*, 2005; Sidorovskaia *et al.*, 2006; Tashmukhambetov *et al.*, 2006). In the second part of the paper, we present the results of quantitative modeling of measured absolute pressures by using enhanced modeling techniques based on the standard underwater acoustic propagation model [the range dependent acoustic model (RAM)].

II. EXPERIMENT

A. Source/receiver configuration

LADC deployed Environmental Acoustic Recording System (EARS) buoys developed by the Naval Oceanographic Office. Two single channel EARS buoys (25 kHz bandwidth) were colocated on the same mooring near Green's Canyon in the Northern GoM (27° 40.0995′ N, 90°21.9660′ W) during June 2003 for a seismic characterization experiment. One buoy hydrophone recorded ambient noise and the other was desensitized (by 12.7 dBV) to record marine seismic array emissions without clipping the data. The hydrophone of each buoy was approximately 250 m from the bottom in a water depth of about 990 m. Only the data from the desensitized EARS hydrophone are discussed in this paper. The M/V Kondor towed a 21-element seismic airgun array of total volume of 3590 in.^{3} (0.0588 m^3) on five parallel linear tracks with horizontal closest approach points to the EARS buoy position of 63, 500, 1000, 2000, and 5000 m. The seismic array configuration is shown in Fig. 1(a). Figure 1(b) shows the reference coordinate system used in the paper to characterize the array directionality. The emission angle θ is the angle between the vertical and a line connecting the position of the array center and a receiving

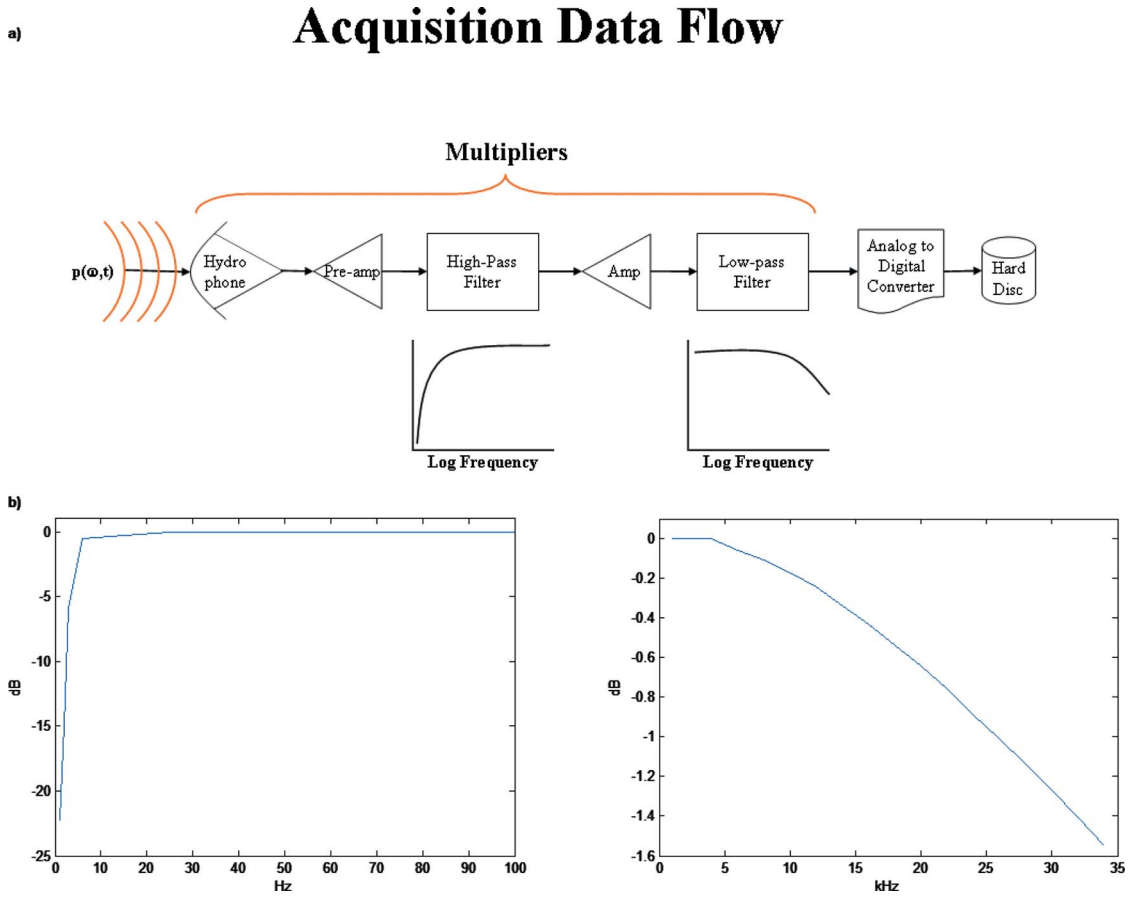


FIG. 3. (Color online) (a) Block diagram illustrating the data flow in a typical acoustic data acquisition system and the components that affect final data calibration. (b) Desensitized hydrophone frequency response curve used for calibration (low- and high-frequency extremes are shown separately for clarity).

hydrophone. The azimuthal angle ϕ is measured in the horizontal plane with 0° directly in front of the array, 180° directly behind, 90° to starboard, and 270° to port. The tracks provide a wide range of measured emission angles (6°–84°, with 0° corresponding to the vertical) and horizontal ranges up to 7 km from the array center to the EARS buoys. The Kondor tracks (labeled as line 0.2, line 500, line 1000, line 2000, and line 5000) are illustrated in Fig. 2. The total number of shots recorded was about 500.

B. Experimental data calibration

In order to obtain absolute measured sound pressure levels, it is important that the recording equipment calibrations be fully understood. Figure 3 is a block diagram of the data acquisition flow in a typical EARS buoy. Two calibration methods have been implemented for the EARS buoys. The first method, which is often called a frequency-domain method since the result is a direct function of frequency, involves injecting a single narrowband sine wave into the electronics downstream of the hydrophone. The input voltage magnitude and phase of the injected signal are compared to the output voltage. This is repeated for many different frequencies to obtain the transfer function of the equipment across a broad frequency band. In the other method, which is often called the time-domain method since the result is a direct function of time, a temporally very short signal (4.7 μs long) is injected into the electronics downstream of the hydrophone. The temporally short characteristic of this "impulse" results in a very wide band of frequencies. The output is recorded and is a direct measure of the impulse response of the equipment. Ideally, the impulse response of the equipment and the transfer function of the equipment are Fourier transform pairs and will lead to the same final results

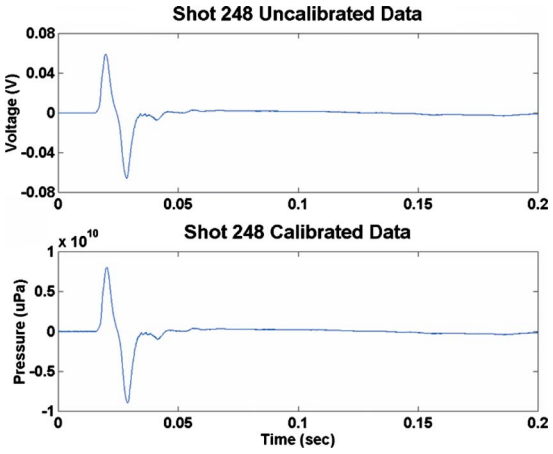


FIG. 4. (Color online) Calibration results for 200 ms of acoustic data corresponding to the direct arrivals from an airgun shot near the CPA of the array to the EARS buoy. The upper plot shows the raw data in V and the bottom plot shows the same data segment in μPa after all the calibrations have been applied. The EARS response (including the hydrophone) is nearly flat from 6 Hz to 25 kHz so that the two plots have very small differences except in units.

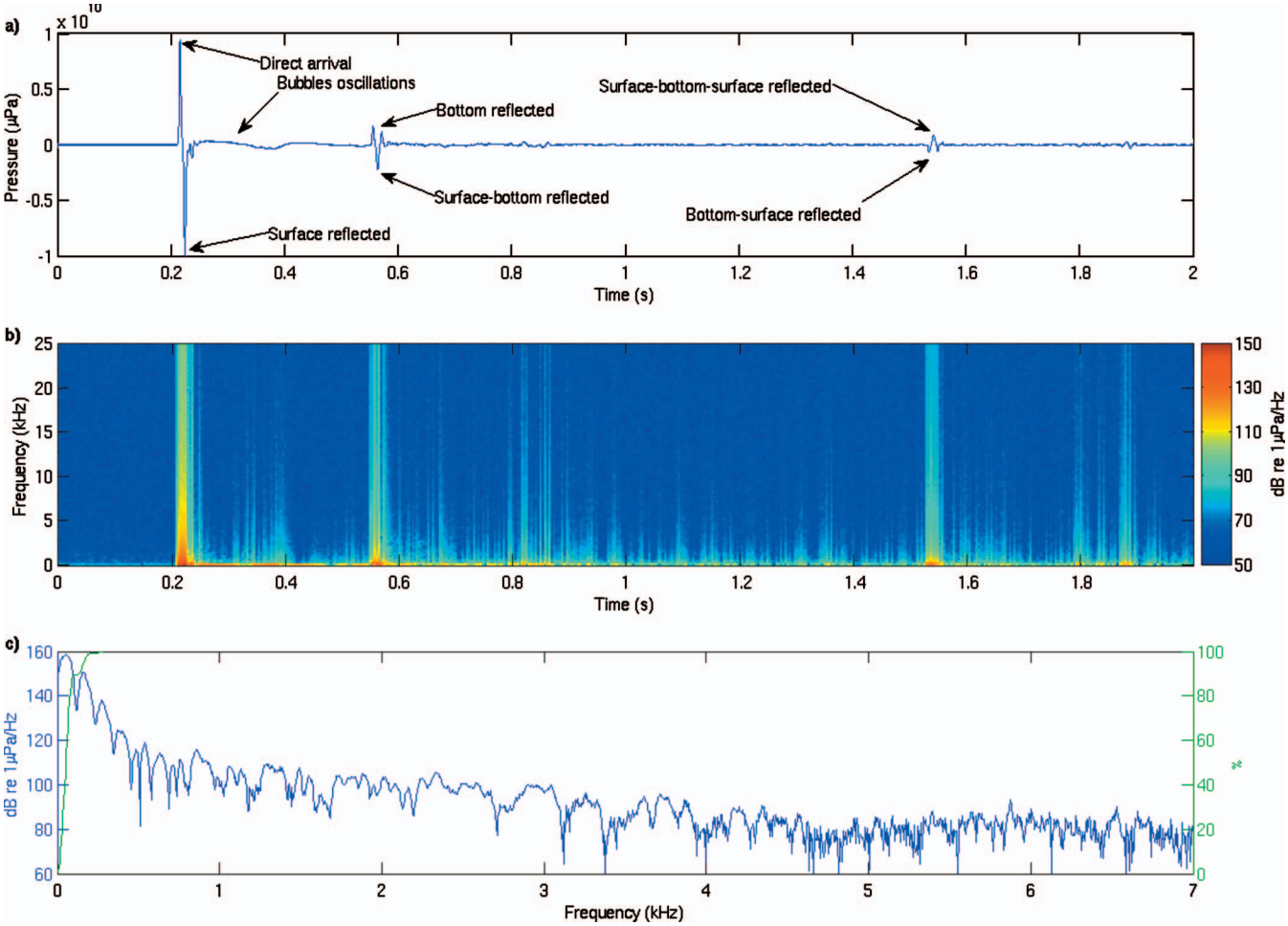


FIG. 5. (a) Measured absolute calibrated acoustic pressure for the CPA shot 249 on line 0.2 vs time. The horizontal range is 63 m, the direct distance to the hydrophone is 736 m, the emission angle is 5°, and the azimuthal angle is 202°. (b) Spectrogram of the signal in (a) using a 5 ms rectangular window with 20% overlap from 6 Hz to 25 kHz. (c) The calibrated amplitude spectrum over a 2 s rectangular window with a start time corresponding to the 0.2 s temporal mark of the spectrogram in (b) and cumulative energy flux in % vs frequency.

when appropriately applied to the raw data. For the LADC 2003 experiment, a comparison of the two methods for the desensitized EARS buoy yielded the same results between 6 Hz and 25 kHz. Since the time-domain method requires the use of more complicated deconvolution techniques to remove the impulse response from the recorded data, all final calibrations of the recorded data were performed using the frequency-domain method. It must be noted that neither of the above methods of calibration includes the response of the hydrophone itself. This must be included in the final calibration of the acoustic data to obtain absolute pressure levels. The hydrophone transfer functions have been determined by the manufacturer. Figure 4 illustrates 200 ms of acoustic data corresponding to the direct arrivals from an array emission near the closest point of approach (CPA) of the array to the EARS buoy. The upper plot is the raw data in volts and the bottom plot is the same data segment in micropascals after all the calibrations have been applied. The EARS response (including the hydrophone) is nearly flat from 6 Hz to 25 kHz so that the two plots have very little difference except in units. We will restrict our analyses of the data to this calibrated frequency band (6 Hz to 25 kHz).

III. EXPERIMENTAL DATA ANALYSIS: METHODS AND RESULTS

Seismic arrays are designed to be highly directional in order to focus the low-frequency sound energy in the vertical direction for the purpose of seismic exploration. The probability that a marine mammal will be exposed to the near vertical downward propagating direct pulse is fairly small. This is not so with off-axis acoustic emissions, so studies of off-axis acoustic signatures are of special interest. Hence, multipath propagation and leakage of high-frequency energy from the airgun array into the ocean waveguide are critical issues for studying the impact on marine mammals. Figures 5–7 show a series of absolute acoustic pressures versus time recorded during the experiment and the corresponding spectrograms for individual shots on different tracks with different horizontal ranges from the center of the array to the buoy location and different emission and azimuthal angles. The spectrograms $S(f_k,t_m)$ are calculated over a 5 ms window with 20% overlap,

$$S(f_k,t_m) = 20\log\{|\sqrt{2}F(k,m)|\}, \quad k = 1, \ldots, N/2 - 1,$$

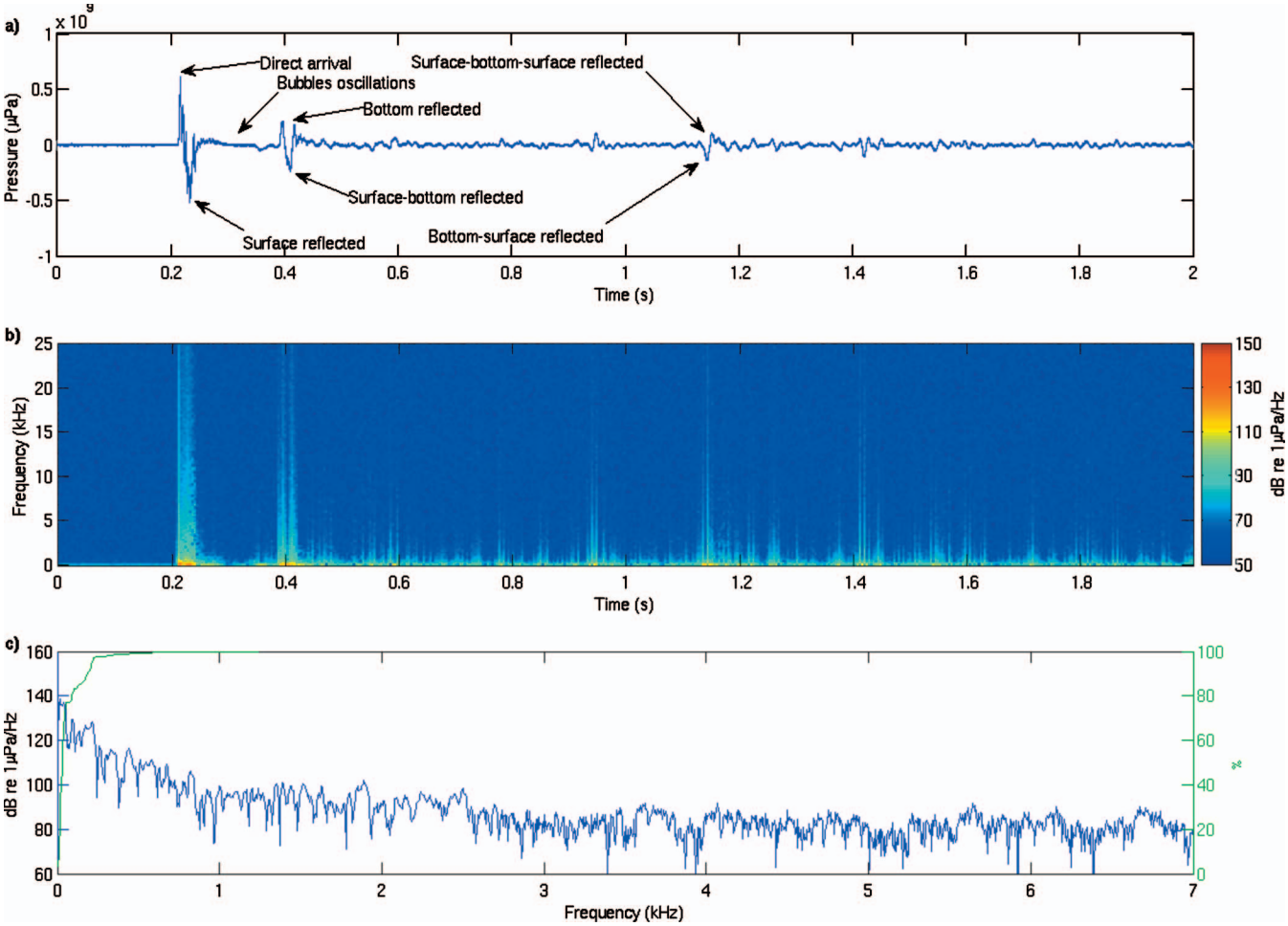


FIG. 6. (a) Measured absolute calibrated acoustic pressure for shot 235 on line 1000 vs time. The horizontal range is 1655 m, the direct distance to the hydrophone is 1810 m, the emission angle is 66°, and the azimuthal angle is 144°. (b) Spectrogram of the signal in (a) over a 5 ms rectangular window with 20% overlap from 6 Hz to 25 kHz. (c) The calibrated amplitude spectrum over a 2 s rectangular window with a start time corresponding to the 0.2 s temporal mark of the spectrogram in (b) and cumulative energy flux in % vs frequency.

$$F(k,m) = \Delta t \sum_{j=0}^{N-1} p[(j+N_s m)\Delta t] e^{-i2\pi jk/N},$$

$$k = 0,1,\ldots,N-1, \quad m = 0,\ldots,M, \tag{1}$$

where $F(k,m)$ are complex Fourier coefficients obtained from a standard fast Fourier transform program; $p(j\Delta t)$ are calibrated temporal pressure samples; $N=390$ is the number of pressure samples in a 5 ms analysis window; $\Delta t = 1.28 \times 10^{-5}$ s is the sampling interval for the collected data; $f_k = \Delta f k, \Delta f = (\Delta t N)^{-1}, k = 0,1,\ldots,N/2$; $N_s = N \cdot 0.8$ is the temporal index shift in terms of pressure sample number for 20% overlap; $M = N_0/N_s$ is the integer number of spectral windows in a 2 s spectrogram. The calculation of the Fourier coefficients in Eq. (1) reflects the transient nature of the measured seismic signatures that should be considered finite-energy signals, not power signals. (Fricke *et al.*, 1985; Johnston *et al.*, 1988). Instead of the power flux spectral density traditionally analyzed for infinitely long stationary signals, an energy flux spectral density $\varepsilon(k)$ is quantified in the calibration procedure for marine seismic source transient signals (Fricke *et al.*, 1985):

$$\varepsilon(k) = \frac{1}{\rho c}|F(k)|^2, \tag{2}$$

where $F(k)$ is the discrete Fourier transform coefficient, which is defined in Eq. (1) for a single m value, ρ is the water density at the receiver position, and c is the speed of sound at the measuring point. The energy flux spectral density curve has the same shape as the amplitude spectrum (absolute values of the Fourier coefficients) but different units (J/(m^2 Hz)). For a decibel scale, the amplitude spectrum level (referenced to 1 μPa) is 182 dB larger than the energy flux spectral density level (referenced to 1 (J/(m^2 Hz)) if the acoustic impedance of sea water is approximated by the constant value

$$Z = \rho c = 1026 \ \frac{\text{kg}}{\text{m}^3} \times 1500 \ \frac{\text{m}}{\text{s}} \approx 1.54 \times 10^6 \ \text{Pa s/m}.$$

Following SEG standards for specifying marine seismic energy sources (Johnston *et al.*, 1988), cumulative energy flux $u(k)$ and total energy flux $u(N/2)$ are calculated for the experimental data,

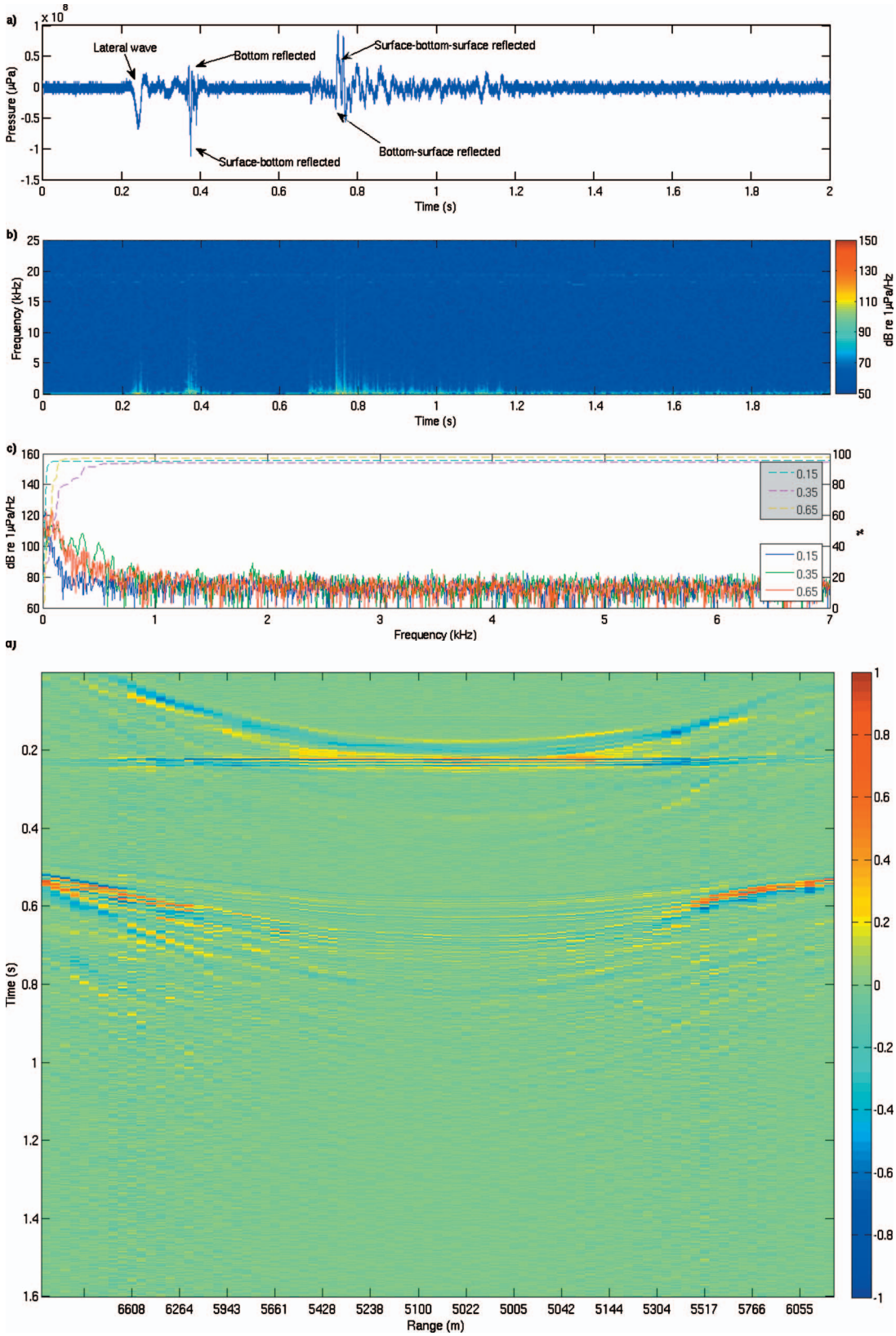


FIG. 7. (a) Measured absolute calibrated acoustic pressure for shot 211 on line 5000 vs time. The horizontal range is 6197 m, the direct distance to the hydrophone is 6240 m, the emission angle is 83°, and the azimuthal angle is 128°. (b) Spectrogram of the signal in (a) over a 5 ms rectangular window with 20% overlap from 6 Hz to 25 kHz. The lateral (head) wave precursor is the first arrival. High frequencies are attenuated as would be expected for a lateral wave. (c) The calibrated amplitude spectra over a 0.2 s rectangular window with start times corresponding to the 0.15, 0.35, and 0.65 s temporal marks of the spectrogram in (b) and cumulative energy fluxes (%) vs frequency. (d) Normalized signal moveout map for line 5000 shots. The time is synchronized on the first bottom reflection. Each shot pressure function is normalized by the absolute value of the maximum pressure in this shot. The separation between a precursor and a reference (strongest) arrival increases with range.

$$u(k) = \Delta f \varepsilon(0) + 2\Delta f \sum_{l=1}^{k} \varepsilon(l), \quad k = 1, \ldots, N/2 - 1,$$
$$u(0) = \Delta f \varepsilon(0), \quad u(N/2) = u(N/2 - 1) + \Delta f \varepsilon(N/2). \tag{3}$$

The cumulative energy flux corresponds to the amount of energy flux in a frequency band from 0 Hz to $k\Delta f$. The total energy flux is the cumulative energy flux in the full recorded frequency band. The cumulative energy flux is usually expressed as a percentage of the total energy flux,

$$\tilde{u}(k) = \frac{u(k)}{u(N/2)} \times 100\,\%, \quad k = 0, 1, \ldots, N/2. \tag{4}$$

Figure 5(a) shows the measured calibrated pressure in micropascals for the closest approach point on line 0.2, which corresponds to a horizontal range of 63 m, with a direct distance to the hydrophone of 736 m, emission angle of 5°, and azimuthal angle of 202°. The 2 s shot spectrogram is shown in Fig. 5(b). The amplitude spectrum level and cumulative energy flux for the 200 ms Fourier analysis window with a start time corresponding to the 0.2 s mark on the spectrogram plot are in Fig. 5(c). The direct arrival, surface reflected arrival, bottom reflected arrival, bubble oscillation cycle, and multiples can be clearly identified in Figs. 5(a) and 5(b). The separation between the direct and bottom reflected arrivals is 340 ms. The maximum amplitude spectrum power level is 159 dB re 1 μPa/Hz, with the level reaching 110 dB re 1 μPa/Hz at 1000 Hz for the direct arrival and 85 dB re 1 μPa/Hz at 5000 Hz for the direct arrival. The calculated total energy flux is 0.32 J/m^2. The sound propagation geometry to the EARS buoy for this shot is nearly vertical for the direct and bottom reflected pulses. The seismic arrays are tuned for optimal (near vertical) transmission of low frequencies for this geometry. The cumulative energy flux plot in Fig. 5(c) shows that most of the energy is under 300 Hz. This is consistent with the array design. The high frequencies are about 35 dB lower than the 300 Hz level. Semiquantitative comparison from Fig. 5(b) shows that the direct path signal energy flux spectral density level is about 20 dB greater at most high frequencies than the bottom reflected arrival and the multiples.

Figures 6(a)–6(c) show similar plots for shot 235 on line 1000. The horizontal range is 1655 m, the direct distance to the hydrophone is 1810 m, the emission angle is 66°, and the azimuthal angle is 144°. The calculated total energy flux is 0.0017 J/m^2. The arrival structure is still identifiable and labeled in Figs. 6(a) and 6(b). The separation between the direct and bottom reflected arrivals is decreased to 200 ms. This may potentially indicate an increased sound exposure level vs range to the shot for an animal having a 200 ms energy integration window (as discussed below). The maximum amplitude spectrum power level is 125 dB re 1 μPa/Hz, with the level reaching 100 dB re 1 μPa/Hz at 1000 Hz for the direct arrival and 80 dB re 1 μPa/Hz at 5000 Hz for the direct arrival, which is close to the background noise level. Figure 6(c) shows again that most of the energy is at a low frequency, under 500 Hz. At this range, the difference is about 25 dB between the high frequency and the 500 Hz levels.

Figures 7(a)–7(c) present the data for shot 211 on line 5000. The horizontal range is 6197 m, the direct distance to the hydrophone is 6240 m, the emission angle is 83°, and the azimuthal angle is 128°. The signal is more complicated and the interpretation of the arrival pattern is not as straightforward as for the shots shown in Figs. 5 and 6. The spectra reveal that most of the energy of the precursor is below 300 Hz. As one can see from the moveout of the signal with different shots on line 5000 in Fig. 7(d), the temporal separation between the precursor and the main energy arrival increases with range. An additional analysis of the signal move-out curves for other lines indicates that the precursor starts appearing at ranges larger than 4.5 km. These features of the precursor arrival strongly suggest that it is a lateral (head, interface) wave. Correlation of experimental time delays between arrivals with modeled ones is required to gain more confidence concerning the analysis of the precursor. The frequency partition of energy for the various components of this shot is similar to that shown in the previous two figures. The amplitude spectrum level and cumulative energy flux for the 200 ms Fourier analysis window for three identifiable arrivals with start times corresponding to 0.15, 0.35, and 0.65 s on the spectrogram plot are shown in Fig. 7(c). The calculated total energy fluxes are 0.028×10^{-3}, 0.023×10^{-3}, and 0.051×10^{-3} J/m^2. The high-frequency level reaches 90 dB re 1 μPa/Hz at 1000 Hz for the strongest arrival and this is close to the background noise level.

Figures 5–7 clearly demonstrate that there is a significant multipath energy in the sound field of the seismic array. The conclusion is that the acoustic energy in the multipath must be taken into account when calculating marine mammal exposure metrics, as suggested by Madsen *et al.* (2006). This can only be done accurately by using propagation models to calculate the full sound field for the waveguide environment.

Sequential 2 s amplitude spectra for all calibrated shots are collected in Fig. 8. The high-frequency part of the spectrum (16–21 kHz) is shown separately in Fig. 8(b), which allows better identification of the narrow spectral lines centered at 18 kHz. These represent the spectral content of the on-board echo-sounder signal. A Simrad EA500 echo sounder was part of the M/V Kondor equipment suite and emitted a 3 ms pulse every 12 s throughout the duration of the experiment. It is apparent from Fig. 8 that the high-frequency acoustic power levels from the seismic array as recorded by the EARS buoy at 739 m depth do not approach the levels of the echo sounder, at least for ranges below 7 km.

Various analysis attributes are generated to quantify and characterize the acoustic output of the seismic airgun array in the ocean in addition to the time and frequency analyses already given. The results shown here can easily be compared to the other studies presented in the literature (Blackwell *et al.*, 2004; Madsen *et al.*, 2006). The first characteristic widely accepted in the oil industry is the maximum received pressure level, zero to peak. (Some authors report a peak-to-peak value for far-field signatures, which will not be

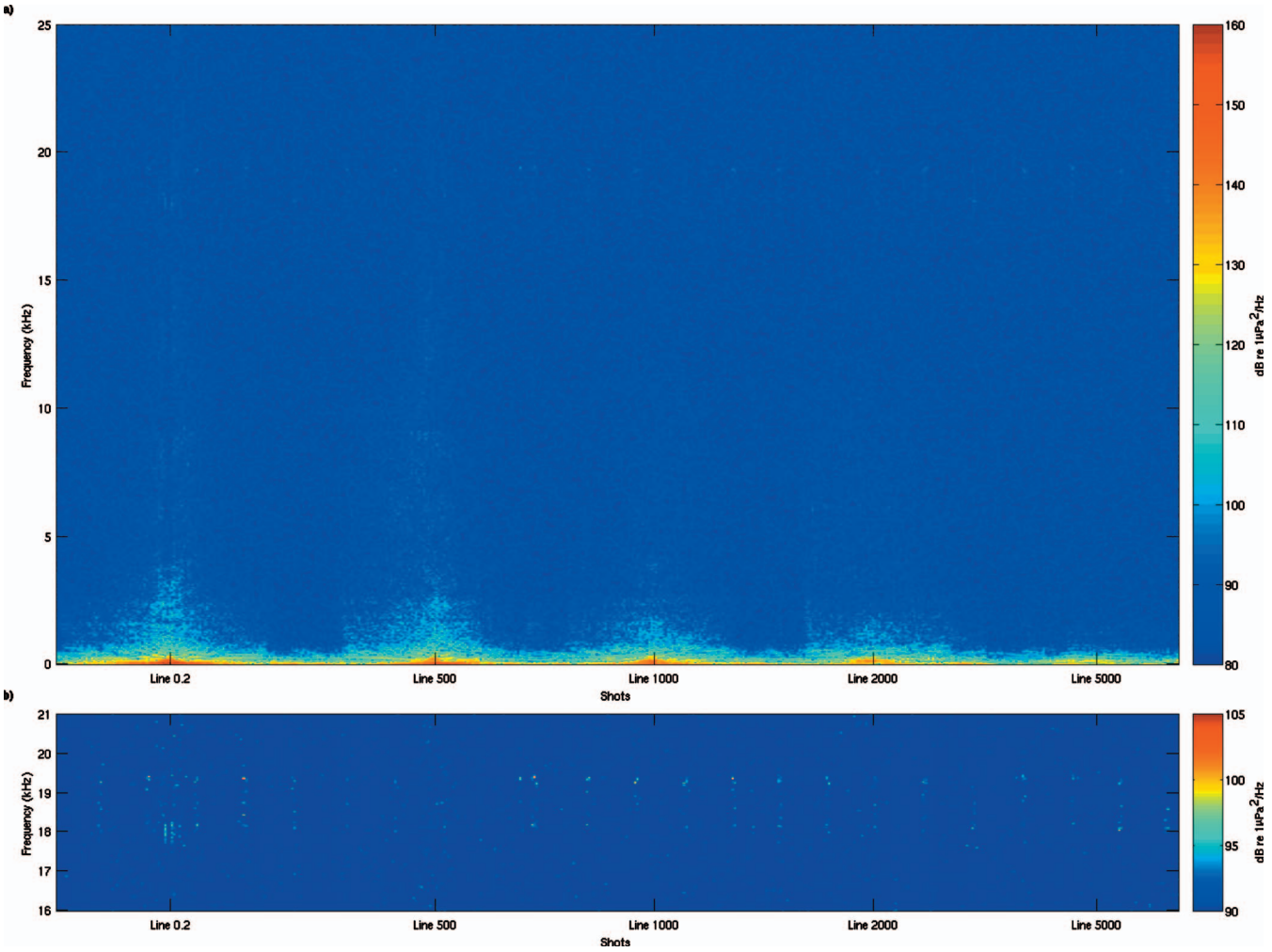


FIG. 8. (a) Sequential spectra of all calibrated shots collected over a 2 s rectangular window during the seismic characterization experiment from 6 Hz to 25 kHz. (b) High-frequency band (16–21 kHz) of the sequential spectra presented in (a). The short vertical lines centered at 18 kHz are spectra of the 3 ms pulses from an 18 kHz echo sounder on the M/V Kondor. It had a repetition rate of 12 s.

more than 3 dB greater than the zero-to-peak level.) Figure 9(a) shows the maximum received pressure level for each shot collected during the experiment. The maximum level for the closest shot almost directly overhead (horizontal range of 63 m, direct distance to the hydrophone of 736 m, and emission angle of 5°) is 200 dB re 1 μPa. Figure 9(b) shows the same data as a function of the horizontal range to the EARS buoy. The multivalued levels at a fixed range are due to array directionality and gun volume differences on the front versus the back of the array [see Fig. 1(a)]. The maximum received pressure levels do not gradually decrease with increasing range beyond 3 km for off-axis shots. They can be as high at the 5 km range as at the 3 km range due to waveguide propagation effects. These results are consistent with data recorded on sperm whales using acoustic tags during controlled exposure experiments (Madsen *et al.*, 2006). Solid and dashed curves represent the modeled maximum levels as a function of range in the vertical 0° plane aligned with the central line of the array obtained by using the parabolic equation model, RAM (Collins, 1993), and two notional source signature models: GUNDALF and NUCLEUS (Hatton, 2004; Nucleus) The details of the modeling are described in the next section. The modeled data do not reproduce all the features of measured

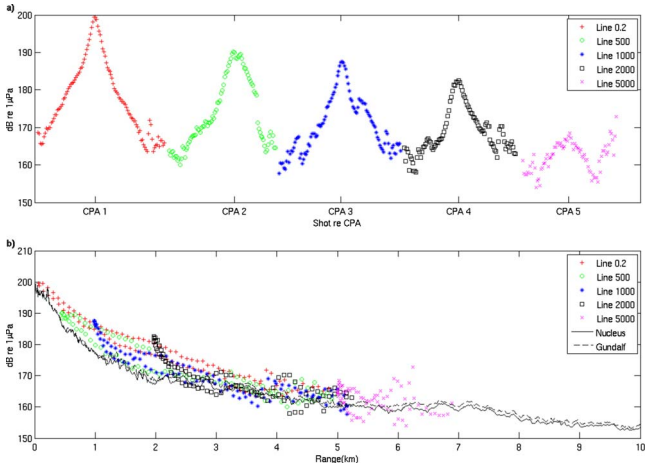


FIG. 9. (Color online) (a) Maximum received calibrated zero-to-peak sound pressure levels for each shot relative to the CPA indicated by the CPA marker on the horizontal axis for each line. (b) Maximum received zero-to-peak sound pressure levels for all collected shots as a function of range. Different symbols correspond to different shot lines. Note that the maximum levels monotonically decrease only for the first 3 km in range. They then start increasing again for ranges larger than 3 km, which indicates that the bottom reflected pulse dominates over the direct arrival. Solid and dashed lines are the modeled maximum received zero-to-peak sound pressure levels in the zero degree fixed vertical plane.

data because the array directionality in different vertical planes is not taken into account due to computational time limitations. The authors are moving to parallel cluster computers to implement full three-dimensional field modeling.

Maximum levels of direct and reflected arrivals are important measures of the seismic array signal directionality and attenuation in a waveguide and provide meaningful information for seismic interpretation characterizing the reflection strength of different sub-bottom reflectors, but they cannot be used as standalone parameters to account for acoustic sensation by a marine mammal because they do not take the duration of the transient seismic pulses into account. It is suggested that most biological receivers, including marine mammals, are best modeled as energy integrators, which integrate intensity over a frequency-dependent time window (Au *et al.*, 1997; Madsen, 2005). The integration time of 200 ms is chosen because it is believed to be used as an integration time by the auditory system of the endangered sperm whale. Therefore, a second attribute, SEL, is calculated over the time of each shot as

$$\begin{aligned} \mathrm{SEL}(i,t_j) &= 10\log\left\{\Delta t\sum_{m=0}^{N-1} p_i^2(t_j+m\Delta t)\right\} \\ &= 10\log\left\{\Delta f\sum_{k=0}^{N-1} |F_i(k,t_j)|^2\right\} \\ &= 10\log\{Zu_i(N/2,t_j)\}, \end{aligned} \tag{5}$$

where i is the shot number in the line, $\Delta t = 1.28\times10^{-5}$ s is the temporal sampling interval of the recorded data, $N = 15\,625$ corresponds to a 200 ms integration window, p_i is the sampled recorded calibrated pressure (in micropascals) for shot i, and t_j is the initial time for a 200 ms analysis window for every possible start time within each shot including 200 ms of ambient noise recording before the first seismic arrival for each shot. The maximum SEL calculated for each shot is selected to characterize that shot.

The maximum SEL for each shot in every line (sequentially) is displayed in Fig. 10(a). The maximum value for the above-mentioned closest shot is 177 dB re 1 μPa^2 s. In Fig. 10(b), the maximum SEL is shown as a function of the horizontal range between the center of the array and the receiving hydrophone. Solid and dashed lines are modeled sound exposure levels in the vertical 0° plane passing through the central line of the array. There are several factors that cause the maximum SEL to increase with range at ranges larger than 3 km. The first factor is that the temporal separation between the first (direct) and the second (bottom reflected) arrivals becomes less than the integration window. The second factor is that the SEL maxima are determined by energy in the multipaths for large range off-axis shots. To support this statement, Fig. 11 shows the SEL for the entire multipath shot as a function of time for the shots shown in Figs. 5–7.

The third attribute used for the recorded data is 1/3-octave band analysis (ANSI/ASA, 2004). 1/3-octave bandwidths are reported to represent the likely lower and upper limits of auditory filters in marine mammal auditory systems for which sparse laboratory bioacoustic data are available (Richardson *et al.*, 1995; Southall *et al.*, 2000; Southall *et al.*, 2003). The results of 1/3-octave band analysis for all collected shots are presented in Figs. 12(a) and 12(b). The 1/3-octave band received levels are calculated for the entire received signal (2 s temporal window) including all multipath arrivals received over 2 s. Figure 12(a) shows 1/3-octave band analysis of all shots sequentially plotted both within line number and by line number. Central frequencies of the bands are on the vertical axis. Band numbers 11–43 are included. Figure 12(b) shows 1/3-octave band analysis of shots within a line plotted as a function of range. The panels correspond to lines 0.2, 500, 1000, 2000, and 5000.

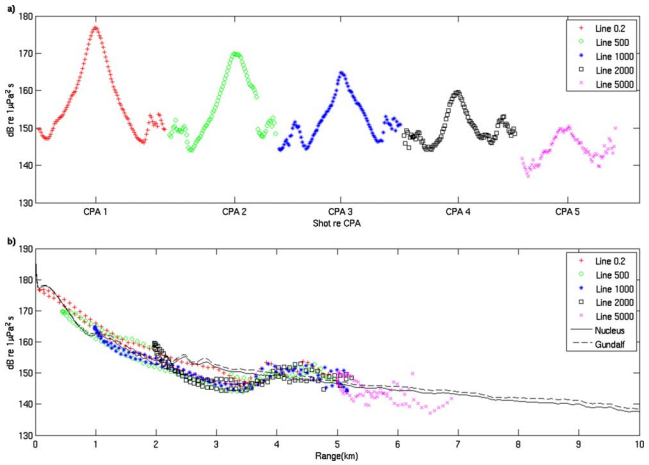


FIG. 10. (Color online) (a) Maximum sound exposure levels for a 200 ms sliding integration window for each shot plotted relative to the line CPA indicated by the CPA marker on the horizontal axis for each line. (b) Maximum sound exposure levels for a 200 ms sliding integration window for each shot shown for all shots as a function of range. Different symbols correspond to different shot lines. Solid and dashed lines are the modeled maximum sound exposure levels in the zero degree fixed vertical plane.

IV. ACOUSTIC MODELING: METHODS AND RESULTS

The seismic source acoustic energy distribution in the ocean depends not only on seismic source parameters but also on the propagation channel. Any meaningful mitigation

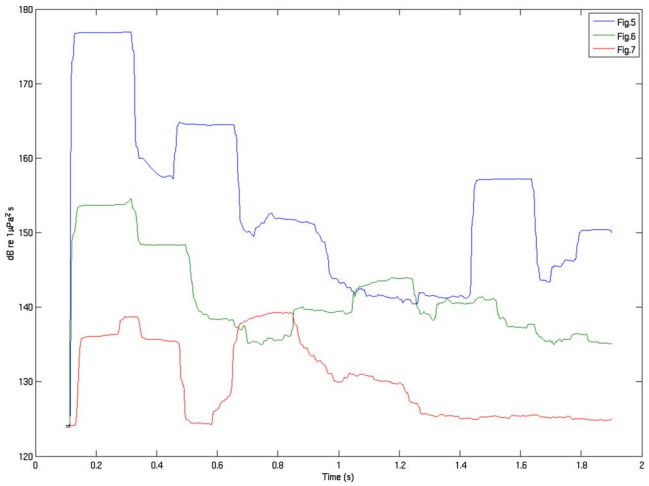


FIG. 11. (Color online) Sound exposure level vs the temporal position of the center of a 200 ms integration window for the entire shot (including multipath arrivals) for the three shots presented in Figs. 5–7.

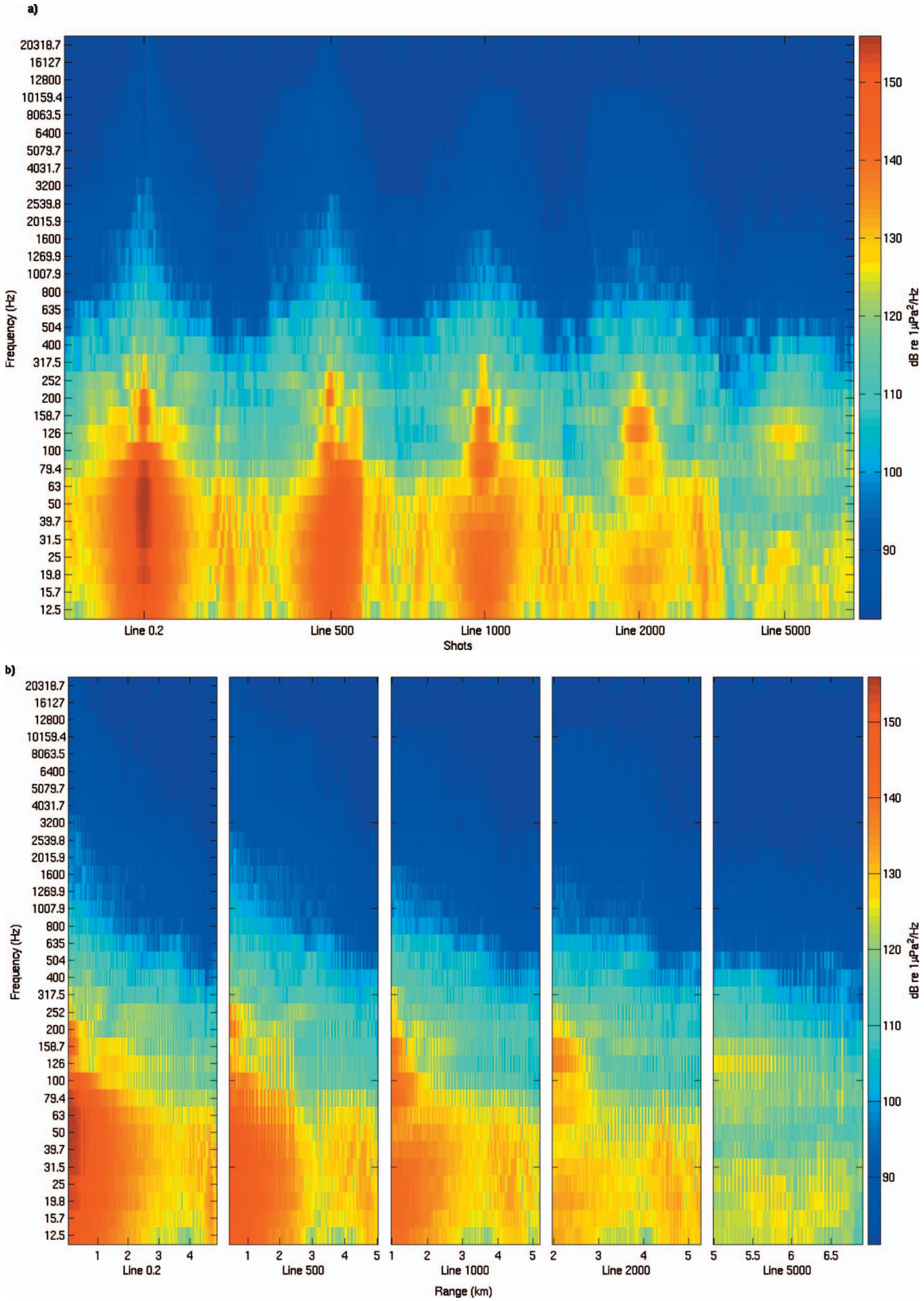


FIG. 12. (a) 1/3-octave band analysis of all shots plotted sequentially both within line number and by line number. Central frequencies of the bands are on the vertical axis and 1/3-octave bands are as defined in ANSI/ASA (2004). Band numbers 11–43 are included. (b) 1/3-octave band analysis of shots within a line plotted as a function of range. The panels correspond to lines 0.2, 500, 1000, 2000, and 5000.

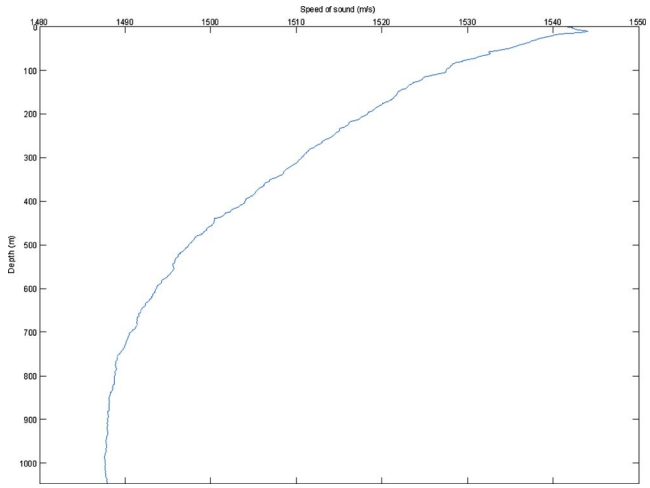


FIG. 13. (Color online) Sound speed profile in the water column during the experiment. Depth in m is plotted vs sound speed in m/s.

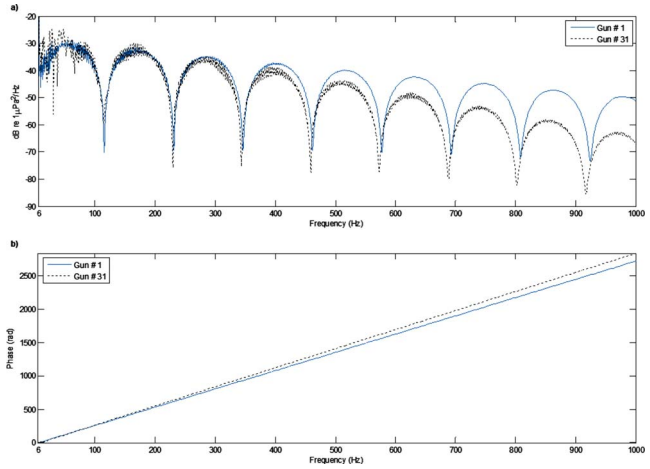


FIG. 14. (Color online) (a) Modeled waveguide transfer function levels (in dB re 1 μPa^2/Hz) for airguns 1 and 31 (airgun numbering shown in Fig. 1) for the closest approach shot 249 on line 0.2 [temporal received pressure signature is shown in Fig. 5(a)] vs frequency from 6 to 1000 Hz. (b) The arriving phase for airguns 1 and 31 for the same shot vs frequency. The waveguide transfer functions are generated by the underwater acoustic propagation model RAM adapted to model a broadband planar array of airguns.

efforts will be dependent on our ability to model quantitatively the acoustic energy distribution from a given seismic array in a particular ocean waveguide. There are several standard acoustic propagation models available to model sound propagation in range-dependent ocean waveguides: RAM, KRAKEN, UMPE, SWAMP, etc. (Collins, 1993; Porter, 1995; Smith and Tappert, 1993; Sidorovskaia, 2004). However, most of the standard models are inherently two dimensional and produce the acoustic pressure distribution of a point harmonic source in the vertical plane of a source and a receiver. There are several issues that have to be addressed when using these models for quantitative modeling of the acoustic pressure distribution from a seismic array: (1) the broadband nature of the seismic pulse produced by each airgun in the array, (2) the complex temporal/angular structure of notional signatures for each airgun in the array due to bubble interactions after firing (Ziolkowski, 1970; Ziolkowski *et al.*, 1982; Laws *et al.*, 1990; Hatton, 2004; Nucleus), and (3) the different ranges to the receiver position for different sources in the array. The last becomes especially important in accounting for the correct relative phases of the high-frequency components at the receiver location. The quality of the calculation will be sensitive to the completeness and accuracy of the parameters describing the propagation channel and the adequacy of notional airgun source signatures to reproduce the near field of the seismic array. The sound speed profile along the propagation path for modeling was derived from expendable bathythermographs and from conductivity-temperature-depth measurements taken during the experiment (see Fig. 13). A very thin surface duct about 10 m thick was present during the experiment. No bottom structure information was collected during the experiment, so the bottom model for the propagation code was based on a historic database (Hamilton, 1980) and previously collected data near the experimental site (Turgut *et al.*, 2002). The bottom model consists of three layers typically present in this area of the Gulf of Mexico: silty clay about 10 m deep, sand deposits up to 1 km deep from the bottom-water interface, and rock formations 1 km below the bottom-water interface.

The calibrated pressure data are modeled using the standard parabolic equation model RAM by Collins (1993), which is upgraded to generate waveguide transfer functions for a broadband multisource array. The measured individual frequency pressure components at the receiver location, $P(f,r_s,z_s)$, are modeled in the frequency domain as

$$P(f,r_s,z_s) = \sum_{i(\text{airguns})} C(f)G(f,r_s,z_s,r_i,z_i)S_i(f), \tag{6}$$

where $G(f,r_s,z_s,r_i,z_i)$ is the complex waveguide transfer function from an individual airgun to the receiver location generated by RAM, $C(f)$ is a highpass filter to cutoff RAM output below 6 Hz and $S_i(f)$ is the Fourier transform of the temporal notional signature of an individual airgun generated by two different airgun characterization models: GUNDALF and NUCLEUS. The waveguide transfer function is generated up to 1000 Hz with a frequency resolution of 0.5 Hz. This frequency resolution provides a sufficiently detailed fine structure for the transfer function to account for the arrival of reflected pulses. The upper frequency is limited by the computational time required by the Parabolic Equation (PE) model to model broadband high-frequency transfer functions for a planar source. Other modeling methods for higher frequencies or a parallel processing approach will be considered to expand the frequency range of modeling in future research. Figures 14 and 15 show the transfer function levels (TFL$=10\log\{|G|^2\}$) and arrival phases of two airguns for the closest approach shot on line 0.2 (Fig. 14) and shot 255 on line 500 (Fig. 15), which has a horizontal range of 448 m, a direct distance to the hydrophone of 859 m, an emission angle of 31°, and an azimuthal angle of 260°. The deep minima in Figs. 14(a) and 15(a) correspond to the interference structure due to a Lloyd's mirror effect. Analysis of this structure in the measured data allows us to correct the nominal average depth of the airguns in the array from 6 m recorded during the experiment to 6.7 m that is used in mod-

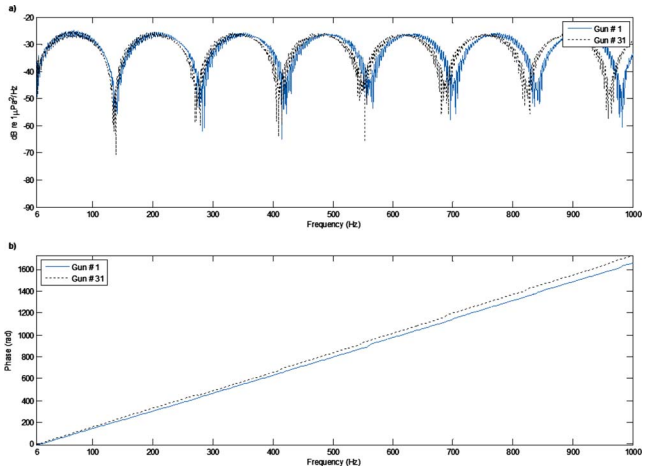


FIG. 15. (Color online) (a) Modeled waveguide transfer function levels (in dB re 1 μPa^2/Hz) for airguns 1 and 31 (airgun numbering shown in Fig. 1) for shot 255 on line 500 vs frequency from 6 to 1000 Hz. (b) The arriving phase for airguns 1 and 31 for the same shot vs frequency. The waveguide transfer functions are generated by the underwater acoustic propagation model RAM adapted to model a broadband planar array of airguns.

eling. The fine structure of the TFL carries information about reflections from bottom layers and multipaths. The TFL and arriving phase structure indicate that a point source model is not suitable for quantitative prediction of the seismic array energy distribution in the water column.

Figure 16 shows the temporal notional signatures of selected airguns, which are generated by the NUCLEUS and GUNDALF models, for the seismic array used in the experiment. The notional signature of each airgun in the array is transformed into the frequency domain using a standard fast Fourier transform program and multiplied by RAM-generated broadband transfer functions to model the frequency content of the calibrated shots [refer to Eq. (6)]. Figures 17(a) and 17(b) are a comparison between experimental and simulated data with the source notional signatures generated by GUNDALF and NUCLEUS for the closest approach shot on line 0.2 (a nearly on-axis shot) and for shot 255 on line 500 (an off-axis shot). The NUCLEUS model has a high-frequency cutoff filter above 800 Hz, so its modeling is only valid up to 800 Hz. GUNDALF is designed to include the high-frequency components up to 25 kHz. There are several factors contributing to the discrepancies between experimental and simulated data. The notches in the experimental data near 500 and 750 Hz are most probably due to the first bottom layer reflection that is inadequately specified based on the historical database. Errors in the bottom properties have an effect on the fine structure of the modeled signal. Both airgun modeling codes show better agreement with the experimental data for on-axis shots. The notional signatures used for this calculation were generated and calibrated for on-axis use and so are not the most appropriate for off-axis use (Hatton, 2002).

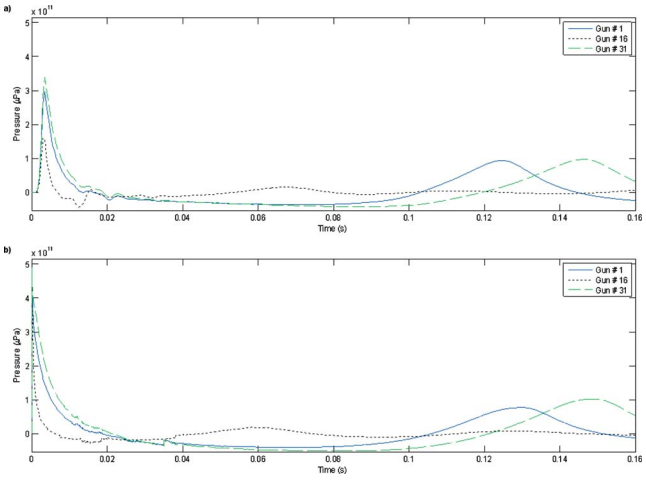


FIG. 16. (Color online) (a) Notional temporal pressure signatures (in μPa) for airguns 1, 16, and 31 (airgun numbering shown in Fig. 1) generated by NUCLEUS vs time in s. (b) Notional temporal pressure signatures (μPa) for the same airguns generated by GUNDALF vs time.

The Fourier synthesis technique for digitized signals is used to model the time-domain response that was measured in the experiment and used as a starting point for the calculation of the exposure levels in the time domain. We use frequencies above 6 Hz for comparison with experimental data both because of the rolloff in the receiving system frequency response and because the modeled frequency components at very low frequencies are not considered fully reliable. Figures 18(a) and 18(b) show the quantitative comparison between measured and modeled signatures in the time domain for shot 255 on line 500. We have also calculated SELs over a 200 ms window for the modeled received pulses in accordance with Eq. (5). Figure 18(c) shows the comparison of the modeled SEL with one calculated from the experimental data for shot 255 on line 500. The modeled and experimental sound exposure levels agree well for the direct and surface reflected arrivals and the bubble oscillation cycle. The discrepancies between modeled and experimental sound exposure levels for the times corresponding to later arrivals are due to inaccurate information about the bottom structure and for initial times are due to wraparound.

The good agreement between measured and calculated data allows us to model reliably the full three-dimensional acoustic energy distribution from the seismic array in the water column. Figure 19 shows the modeled received pressure level as a function of range, depth, and frequency for a point source placed at the center of the array. The power of the point source is equal to the total power of the array used in the experiment. Figures 20 and 21 show the modeled received pressure level as a function of range, depth, and frequency in two different vertical planes for the seismic array used in the experiment taking into account the full array geometry and spectral power components of individual array sources extracted from the notional signature frequency components. These characteristics are more meaningful to describe the broadband array radiation field in the waveguide, where array directionality is superimposed on waveguide energy channeling, than the traditional directional pattern of an array in free space. The array acoustic field structure in a waveguide is considerably different from a point source field structure in a waveguide and from the array free-space field. Generation of a series of such maps covering the full frequency band of interest and a set of vertical planes, which

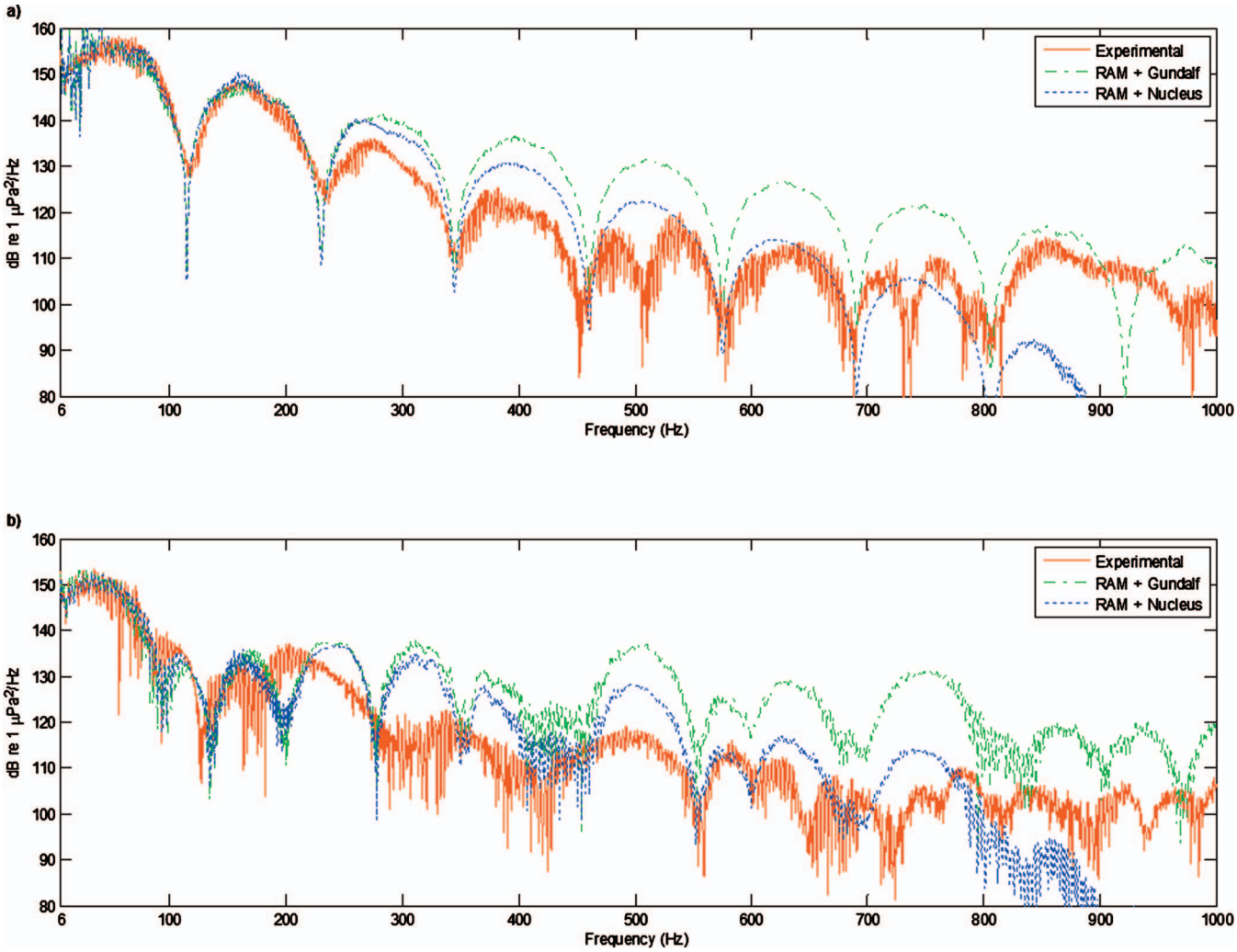


FIG. 17. Spectrum comparison (in dB re $1\mu Pa^2/Hz$) between experimental and modeled data with the source notional signatures generated by GUNDALF and NUCLEUS vs frequency from 6 to 1000 Hz: (a) for the closest approach shot 249 on line 0.2 (nearly on-axis shot) and (b) for shot 255 on line 500 (off-axis shot).

allows the generation of three-dimensional SEL maps for a particular array in a particular environment, is the subject of future research.

V. DISCUSSION AND CONCLUSIONS

The 2003 LADC calibrated data from a typical marine seismic exploration array is a significant contribution to three-dimensional broadband seismic source characterization studies. The data set measures the absolute calibrated pressures for a wide range of angles with frequencies up to 25 kHz. This data set also provides the opportunity to test available modeling tools by quantitative comparison of measured and modeled data. However, the angular/range density of the collected data does not allow detailed testing of the reconstruction of the array directivity for a fixed range or the energy distribution in an arbitrary fixed vertical plane solely based on measured data to compare with modeled results. Additional field data should be collected for validation of the model prediction for a full three-dimensional seismic array characterization. Special attention in future experimental designs should be paid to collecting data for the close-to-horizontal propagation direction that can be critical in study-

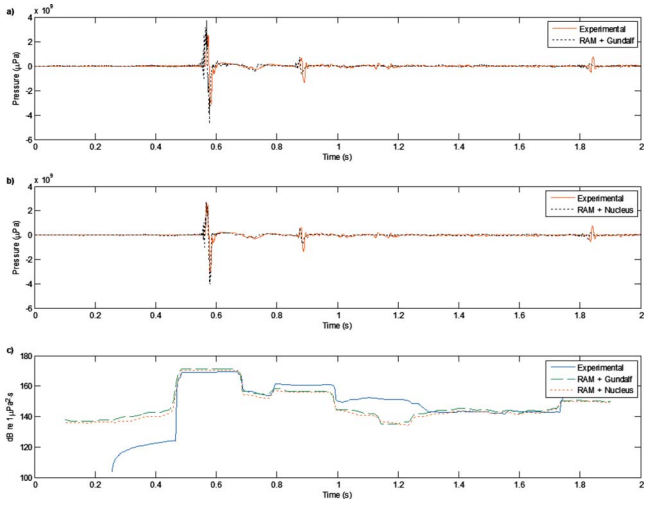


FIG. 18. (Color online) Comparison between experimental and modeled data (in μPa) vs time in s using frequency components from 6 to 1000 Hz: (a) for shot 255 on line 500 (off-axis shot) simulated with GUNDALF notional signatures, (b) for shot 255 on Line 500 (off-axis shot) simulated with NUCLEUS notional signatures, and (c) sound exposure levels calculated from experimental data and from modeled data for shot 255.

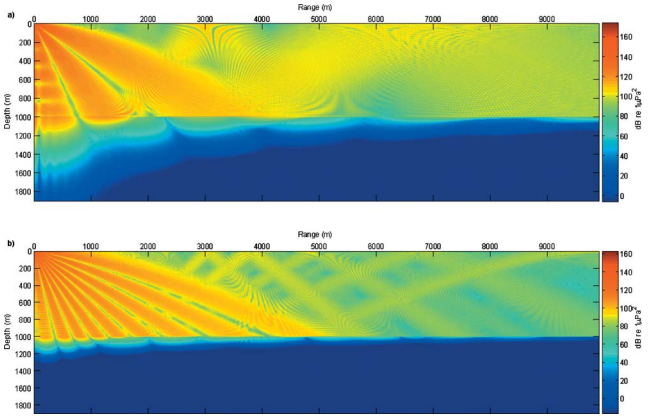


FIG. 19. (Color online) Modeled received pressure levels in dB re 1 μPa^2 as a function of range from 0.01 to 10 km and depth from 0 to 990 m for a point harmonic source at depth of 6.7 m at (a) 300 Hz and (b) 1000 Hz.

ing the issue of energy capture in a near-surface duct. Moreover, the arrival ranges should be extended to address waveguide acoustic propagation issues, such as the existence of convergence and shadow zones and surface duct effects. For example, the number of acoustic precursors formed by the surface duct will be dependent on the range (Sidorovskaia, 2004).

Propagation codes combined with notional signature models predict the broadband data reasonably well. All presented modeling results are *ab initio* calculations with no adjustable parameters. The accuracy of prediction is limited by uncertainties in environmental information and by the accuracy of the source models. Modeling is a useful tool in the prediction of the three-dimensional acoustic energy distribution in an ocean volume of interest. It can be used to determine three-dimensional acoustic energy distribution variations due to anticipated changes in the details of future surveys including changes in ocean environmental conditions and source configuration, without necessarily conducting field experiments. Modeling allows a fairly accurate prediction of sound exposure levels for marine mammals to aid in planning future seismic surveys.

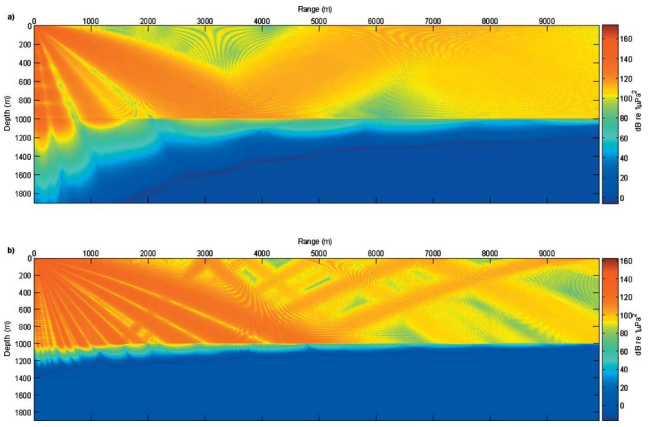


FIG. 20. (Color online) Modeled received pressure levels in dB re 1 μPa^2 as a function of range from 0.01 to 10 km and depth from 0 to 990 m for the seismic array in the 0°-azimuthal plane (a vertical plane through the central line of the array) at (a) 300 Hz and (b) 1000 Hz.

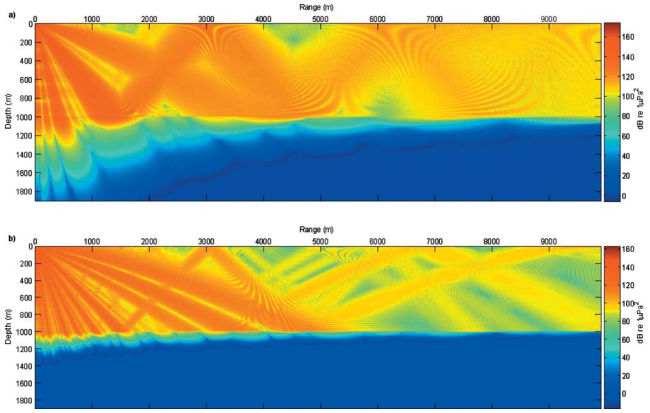


FIG. 21. (Color online) Modeled received pressure levels in dB re 1 μPa^2 as a function of range from 0.01 to 10 km and depth from 0 to 990 m for the seismic array in the 90°-azimuthal plane (a vertical plane through the array center perpendicular to the travel direction) at (a) 300 Hz and (b) 1000 Hz.

ACKNOWLEDGMENTS

This research has been funded by the Industry Research Funding Coalition through the International Association of Geophysical Contractors and the Joint Industry Project through the International Association of Oil and Gas Producers. The authors thank Phil Fontana of Veritas for supplying the source notional signatures from GUNDALF and NUCLEUS. GUNDALF is a product of Oakwood Computing, Limited, and NUCLEUS is by PGS. The authors are grateful to Phil Fontana, Les Hatton of Oakwood Computing, colleagues from the University of Southern Mississippi, particularly Grayson Rayborn and James Stephens and students Chris Walker and Ben Brack, and members of the Industry Research Funding Coalition for helpful discussions. We would like to thank the anonymous reviewers for the valuable comments that helped to improve our paper. In particular, we would like to express special thanks to one of them, whose insightful suggestions on the interpretation of experimental data, added scientific value to this publication.

ANSI/ASA (**2004**). "American National Standard Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters," ANSI Report No. S1.11-2004.

Au, W. W. L., Nachtigall, P. E., and Pawloski, J. L. (**1997**). "Acoustic effects of the ATOC signal (75 Hz, 195 dB) on dolphins and whales," J. Acoust. Soc. Am. **101**, 2973–2977.

Blackwell, S. B., Lawson, J. W., and Williams, J. T. (**2004**). "Tolerance by ringed seals (*Phoca hispida*) to impact pipe-driving and construction sounds at an oil production island," J. Acoust. Soc. Am. **115**, 2346–2357.

Caldwell, J., and Dragoset, W. (**2000**). "A brief overview of seismic air-gun arrays," The Leading Edge **19**, 898–902.

Collins, M. D. (**1993**). "A split-step Padé solution for the parabolic equation method," J. Acoust. Soc. Am. **93**, 1736–1742.

DeRuiter, S. L., Tyack, P. L., Lin, Y.-T., Newhall, A. E., Lynch, J. F., and Miller, P. J. O. (**2006**). "Modeling acoustic propagation of air gun array pulses recorded on tagged sperm whales (*Physeter macrocephalus*)," J. Acoust. Soc. Am. **120**, 4100–4114.

Fricke, J. R., Davis, J. M., and Reed, D. H. (**1985**). "A standard quantitative calibration procedure for marine seismic sources," Geophysics **50**, 1525–1532.

Gordon, J., Gillespie, D., Potter, J., Frantzis, A., Simmonds, M. P., Swift, R., and Thompson, D. (**2004**). "A review of the effects of seismic surveys on marine mammals," Mar. Technol. Soc. J. **37**, 16–34.

Hamilton, E. L. (**1980**). "Geoacoustic modeling of the sea floor," J. Acoust. Soc. Am. **68**, 1313–1340.

Hatton, L. (**2002**). "GUNDALF, a software package for predicting the acoustic

signature of high pressure airguns in exploration seismology," http://www.gundalf.com Last viewed on 8/01/07.
Hatton, L. (**2004**). "Incorporating marine mammal hearing sensitivity into a high-grade air gun modelling package (extended abstract)," PETEX 2004, November 2004.
Ioup, G. E., Ioup, J. W., Sidorovskaia, N. A., Walker, R. T., Kuczaj, S. A., Walker, C. D., Rayborn, G. H., Brack, B., Wright, A., Newcomb, J., and Fisher, R. (**2005**). "Analysis of Bottom-Moored Hydrophone Measurements of Gulf of Mexico Sperm Whale Phonations," Proceedings of 23rd Annual Gulf of Mexico Information Transfer Meeting, January 2005, pp. 109–136.
Johnston, R. C., Reed, D. H., and Desler, J. F. (**1988**). "Special Report of the SEG Technical Standards Committee SEG standards for specifying marine seismic energy sources," Geophysics **53**, 566–575.
Labianca, F. M. (**1972**). "Normal modes, virtual modes, and alternative representations in the theory of surface-duct sound propagation," J. Acoust. Soc. Am. **53**, 1137–1147.
Laws, R. M., Hatton, L., and Haartsen, M. W. (**1990**). "Computer modelling of clustered airguns," First Break **8**(9), 331–338.
MacGillivray, A. O. (**2006**). M.S. thesis, University of Victoria.
Madsen, P. T. (**2005**). "Marine mammals and noise: Problems with root mean square sound pressure levels for transients," J. Acoust. Soc. Am. **117**, 3952–3957.
Madsen, P. T., Johnson, M., Miller, P. J. O., Aguilar de Soto, N., Lynch, J., and Tyack, P. L. (**2006**). "Quantitative measures of air gun pulses recorded on sperm whales (*Physeter macrocephalus*) using acoustic tags during controlled exposure experiments," J. Acoust. Soc. Am. **120**, 2366–2379.
Monjo, C. L., and DeFerrari, H. A. (**1994**). "Analysis of pulse propagation in a bottom-limited sound channel with a surface duct," J. Acoust. Soc. Am. **95**, 3129–3148.
Newcomb, J., Fisher, R., Turgut, A., Field, R., Ioup, G. E., Ioup, J. W., Rayborn, G., Kuczaj, S., Caruthers, J., Goodman, R., and Sidorovskaia, N. (**2002a**). "Modeling and Measuring the Acoustic Environment of the Gulf of Mexico," Proceedings of the 21st Annual Gulf of Mexico Information Transfer Meeting, January 2002, pp. 509–521.
Newcomb, J., Fisher, R., Field, R., Rayborn, G., Kuczaj, S., Ioup, G. E., Ioup, J. W., and Turgut, A. (**2002b**). "Measurements of Ambient Noise and Sperm Whale Vocalizations in the Northern Gulf of Mexico Using Near Bottom Hydrophones," Marine Frontiers MTS/IEEE Proceedings of OCEANS'02, pp. 1365–1371.
Newcomb, J., Sanders, W., Stephens, J. M., Walker, C., Brack, B., Rayborn, G. H., Sidorovskaia, N. A., Tashmukhambetov, A. M., Ioup, G. E., Ioup, J. W., and Chapin, S. R. (**2005**). "Calibration and Analysis of Seismic Air gun Data from an EARS Buoy," Proceedings of 23rd Annual Gulf of Mexico Information Transfer Meeting, January 2005, pp. 83–100.
"NUCLEUS,-Advanced tools for Survey Planning, Seismic Modelling and Feasibility Studies," http://www.pgs.com/upload/Nucleus.pdf last viewed 5/6/2008.
Porter, M. B. (**1995**). *The KRAKEN Normal Mode Program* (SACLANT Undersea Research Center, La Spezia).
Richardson, W. J., Greene, Jr., C. R., Malme, C. I., and Thomson, D. H. (**1995**). *Marine Mammals and Noise* (Academic, San Diego, CA).
Sidorovskaia, N. A. (**2004**). "Systematic studies of pulse propagation in ducted oceanic waveguides in normal mode representation," Eur. Phys. J.: Appl. Phys. **25**, 113–131.
Sidorovskaia, N. A., and Werby, M. F. (**1995**). "Broad-band pulse signals and the characterization of shallow water ocean properties," Proceedings of the SPIE Conference, April 1995, pp. 97–108.
Sidorovskaia, N. A., Ioup, G. E., Ioup, J. W., Tashmukhambetov, A. M., Newcomb, J. J., Stephens, J. M., and Rayborn, G. H. (**2006**). "Modeling tools for 3-d air gun source characterization studies," Proceedings of the eighth European Conference on Underwater Acoustics, June 2006, edited by S. M. Jesus, and O. C. Rodriguez, pp. 95–100.
Smith, K. B., and Tappert, F. D. (**1993**). "UMPE: The University of Miami Parabolic Equation Model, Version 1.3," MPL Technical Memorandum No. 432.
Southall, B. L., Schusterman, R. J., and Kastak, D. (**2000**). "Masking in three pinnipeds: Underwater, low-frequency critical ratios," J. Acoust. Soc. Am. **103**, 1322–1326.
Southall, B. L., Schusterman, R. J., and Kastak, D. (**2003**). "Acoustic communication ranges for northern elephant seals (*Mirounga angustirostris*)," Aquat. Mamm. **29.2**, 202–213.
Tashmukhambetov, A. M., Sidorovskaia, N. A., Ioup, G. E., Ioup, J. W., Newcomb, J., Walker, C., Brack, B., and Rayborn, G. H. (**2006**). "3-D airgun source characterization and propagation modeling," SEG Technical Program Expanded Abstracts, Vol. **25**, pp. 26–30.
Tolstoy, M., Diebold, J. B., Webb, S. C., Bohnenstiehl, D. R., Chapp, E., Holmes, R. C., and Rawson, M. (**2004**). "Broadband calibration of R/V Ewing seismic sources," Geophys. Res. Lett. **31**, L14310.
Turgut, A., McCord, M., Newcomb, J., and Fisher, R. (**2002**). "Chirp sonar sediment characterization at the northern Gulf of Mexico Littoral Acoustic Demonstration Center experimental site," Proceedings of the Oceans '02 MTS/IEEE, Vol. **4**, pp. 2248–2252.
Ziolkowski, A. (**1970**). "A method for calculating the output pressure waveform from an air gun," Geophys. J. R. Astron. Soc. **21**, 137–161.
Ziolkowski, A., Parkes, G., Hatton, L., and Haugland, T. (**1982**). "The signature of an air gun array: Computation from near-field measurements including interactions," Geophysics **47**, 1413–1421.

Passive acoustic detection and localization of whales: Effects of shipping noise in Saguenay–St. Lawrence Marine Park[a)]

Yvan Simard[b)] and Nathalie Roy[c)]
Maurice Lamontagne Institute, Fisheries and Oceans Canada, 850 Route de la Mer, Mont-Joli, Québec G5H-3Z4, Canada. Marine Sciences Institute, University of Québec at Rimouski, 310 Allée des Ursulines, Rimouski, Québec G5L-3A1, Canada

Cédric Gervaise[d)]
E3I2, EA3876, ENSIETA, GIS Europole Mer, 2 Rue François Verny, 29200 Brest, France

(Received 19 September 2007; revised 10 March 2008; accepted 1 April 2008)

The performance of large-aperture hydrophone arrays to detect and localize blue and fin whales' 15–85 Hz signature vocalizations under ocean noise conditions was assessed through simulations from a normal mode propagation model combined to noise statistics from 15 960 h of recordings in Saguenay–St. Lawrence Marine Park. The probability density functions of 2482 summer noise level estimates in the call bands were used to attach a probability of detection/masking to the simulated call levels as a function of whale depth and range for typical environmental conditions. Results indicate that call detection was modulated by the calling depth relative to the sound channel axis and by modal constructive and destructive interferences with range. Masking of loud infrasounds could reach 40% at 30 km for a receiver at the optimal depth. The 30 dB weaker blue whale *D*-call were subject to severe masking. Mapping the percentages of detection and localization allowed assessing the performance of a six-hydrophone array under mean- and low-noise conditions. This approach is helpful for optimizing hydrophone configuration in implementing passive acoustic monitoring arrays and building their detection function for whale density assessment, as an alternative to or in combination with the traditional undersampling visual methods.
 [DOI: 10.1121/1.2912453]

PACS number(s): 43.30.Sf, 43.30.Nb, 43.50.Rq, 43.30.Bp [WWA] Pages: 4109–4117

I. INTRODUCTION

Passive acoustic monitoring (PAM) systems for continuously detecting and tracking whales from their specific calls over ocean basins have been explored since a few decades [cf., review by Mellinger *et al.* (2007)]. With the technological developments in electronics and computers, they became more accessible and are now rapidly spreading worldwide to monitor whales in their habitats over long periods. The capacity of these systems to achieve the targeted objectives in particular environments relies on the actual characteristics of the vocalizations, local ocean noise, and propagation conditions [Stafford *et al.* (2007)].

To assess this efficiency in monitored habitats, local noise characteristics must first be established, ideally over a long-term period representative of what is sought for the PAM application. Ocean noise actually experienced by marine mammal in their critical habitats can have several effects on the animals, notably on their communications [Richardson *et al.* (1995); NRC (2003); Southall *et al.* (2007)]. Low signal to noise ratio (SNR) hinders call detection and whale tracking from PAM hydrophone arrays as well as communication with conspecifics. SNR and propagation conditions determine the detection and localization functions of PAM systems that could be implemented to estimate whale local density time series [e.g., Clark and Fristrup (1997); Phillips *et al.* (2006); Simard *et al.* (2006b); Stafford *et al.* (2007)].

Given the propagation conditions in the monitored basin, which could be estimated with a ground-truthed numerical model, and the measured noise probability density function (PDF) in the vocalization band, the performance of a hydrophone array configuration in detecting and localizing whales can be assessed. This is the objective of the present paper for a special summer feeding habitat of North Atlantic baleen whales, located in an environment that is strongly affected by shipping noise.

The baleen whale feeding ground of the Saguenay–St. Lawrence Marine Park is located in a ~100-km-long segment of the Lower St. Lawrence Estuary at the head of the 300-m-deep Laurentian Channel (Fig. 1) [Simard and Lavoie (1999)]. This area is also a portion of the St. Lawrence Seaway, a major continental seaway of North America where cargo transits to and from the Atlantic and Great Lakes. About 6000 merchant ships annually transit through the area, with up to 5 ships/h on busy summer days. During summer, an important whale watching activity from a fleet of a few hundred–passenger ships and large zodiacs is taking place in the area [Michaud *et al.* (1997); Tecsult Environment Inc. (2000); Hoyt (2001)]. Recent observations indicate that the

[a)]Part of this work was presented in "Masking of blue and fin whales low-frequency vocalizations by shipping noise in the Saguenay–St. Lawrence Marine Park," Proceedings of International Conference on Effects of Noise on Aquatic Life, Nyborg, 13–17 August 2007.
[b)]Author to whom correspondence should be addressed. Electronic addresses: yvan.simard@dfo-mpo.gc.ca and yvan_simard@uqar.qc.ca.
[c)]Electronic mail: nathalie.roy@dfo-mpo.qc.ca
[d)]Electronic mail: Cedric.Gervaise@ensieta.fr.

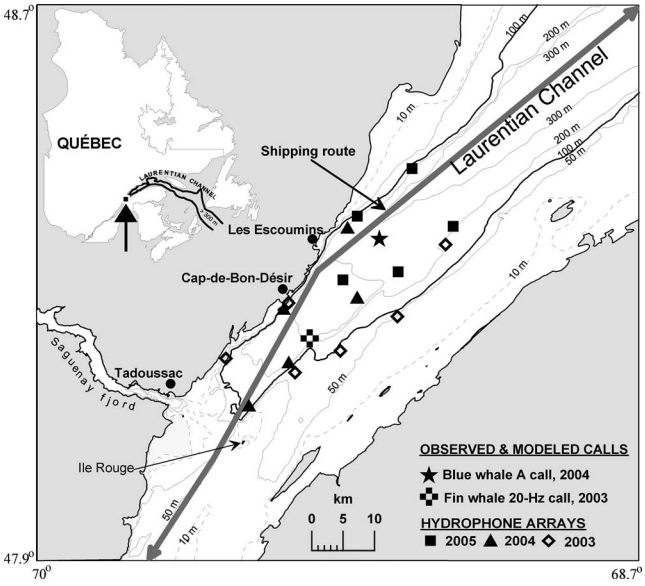


FIG. 1. Map of the study area with bathymetry, shipping route axis, positions of recording hydrophone arrays in 2003–2005 summers and of the calling blue and fin whales used for ORCA validation.

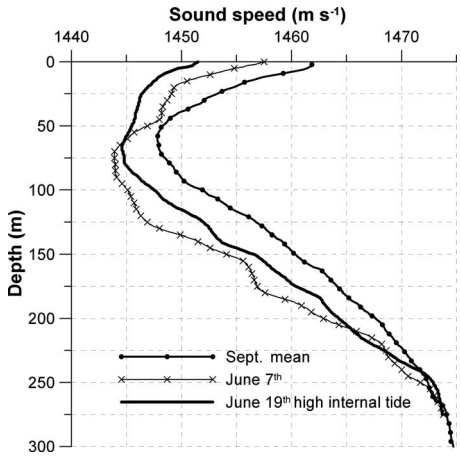


FIG. 2. Sound speed profiles in the area in summer of 2003 from Fisheries and Oceans Canada oceanographic data.

noise in the 10 Hz–1 kHz band often exceeds power spectrum density (PSD) levels for heavy traffic in ocean [Wenz (1962)], up to half of the time at some locations (unpublished data). This heavy shipping noise overlaps with the low-frequency vocalizations produced by baleen whales, notably blue [Berchok *et al.* (2006); Mellinger and Clark (2003)] and fin whales [Watkins *et al.* (1987)] feeding in the area. Call SNRs in the area mostly depend on distance between ships and whales, their respective source levels (SLs) at the corresponding frequencies, and propagation conditions [Simard *et al.* (2006a)].

In this paper, the importance of masking for detecting and localizing blue and fin whales signature calls is assessed from the summer noise level PDF in the call bands from a multiyear monitoring time series of the study area, and modeling of call propagation with a normal mode propagation model, validated with *in situ* observations. Call masking is first assessed for a single hydrophone under the seaway mean- and low-noise levels (representing quieter environments) as function of source depth and range, for the summer range of oceanographic conditions. The effects on spatial detection and two-dimensional (2D) localization from the autonomous hydrophone PAM array configuration used in 2003 are then estimated for a possible range of source depths for the different calls.

II. MATERIAL AND METHODS

Three hydrophone arrays were deployed in the study area during summers of 2003–2005 (Table I, Fig. 1). All hydrophones were HTI 95-min with a nominal receiving sensitivity in the low-frequency band (<2 kHz) of −164 dB re 1 V/μPa, which was confirmed by calibration at the Defense Research and Development Canada (Dartmouth, NS, Canada), acoustic calibration facility. The 16 bit acquisition systems were AURAL autonomous hydrophone systems (Multi-Electronique Inc, Rimouski, Qc, Canada) programmed to sample continuously over the 1 kHz band. They were deployed as oceanographic moorings with special care to minimize possible noise sources, with the hydrophones at intermediate depths in the water column close to the summer sound channel axis [Fig. 2, Table I, Simard and Roy (2008)]. In 2003, two hydrophones from a cabled coastal array deployed along a cape completed the seven-hydrophone array [Fig. 1, Cap-de-Bon-Désir, Table I, Simard and Roy (2008)]. The arrays were synchronized with a combination of means: starting and stopping the AURALs with a pulse per second impulse from a global positioning system (GPS) receiver, simultaneous recording of same acoustic signals, time drift cross-checks with the coastal array, and linear time interpolations assuming constant drift.

TABLE I. Hydrophone arrays' deployment characteristics and number of 5 min power spectral densities systematically sampled along the time series for the noise PDF.

Year	Start date	Duration (d)	Hydrophone mean depths (m)	Number of hydrophones	Array aperture (km)	Total recording (h)	Noise PSDs (No.)
2003	09/03	26	43–54	5	40	3120	
	08/19	43	93–144	2		2064	
2004	08/14	72	126–179	5	32.5	8640	1378
2005	05/03	72	113–170	4	21	6912	1043
	05/03	17	120	1		408	61

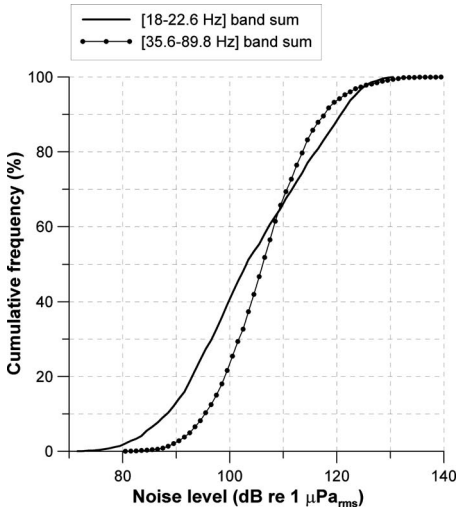


FIG. 3. Cumulative probability of summer noise levels in blue and fin whale signature calls' bands in the study area from 2482 systematic periods of 5 min recordings at ten stations. The low-noise conditions correspond to the first 25% below ~100 dB re 1 μPa$_{rms}$

The acoustic time series at the stations (Table I) were first examined for the presence of flow noise in the recordings, by monitoring the sound pressure level in an indicative low-frequency narrow band. Valid recording periods were those where this indicator level was below a threshold determined by cross-checks on the spectrograms of the signals for the absence of flow noise. The series of valid periods were then randomly subsampled to extract a 5 min recording for each 6 h consecutive periods to compute the noise level for ANSI third-octave bands that were summed over the call bands to get the noise levels. The summer noise level PDFs integrated over the infrasound (18–22.6 Hz) and audible *D*-call (35.6–89.8 Hz) bands were estimated for 2482 periods of 5 min extracted from the 15 960 h of recording during summers of 2004 and 2005 from the hydrophones deployed at ten stations in the study area (Figs. 1 and 3).

ORCA normal mode propagation model [Westwood *et al.* (1996)] was used to propagate representative blue whale A, B, and *D*-call and the 20 Hz pulse of fin whales [e.g., Berchok *et al.* (2006); Watkins *et al.* (1987)]. It was configured for the average seafloor conditions in the study area at the head of the 300-m-deep Laurentian channel (Table II) [Loring and Nota (1973); Massé (2001); Table 1.3 in Jensen *et al.* (2000)], and the September 2003 sound speed profile from Fisheries and Oceans Canada oceanographic data (http://www.osl.gc.ca/sgdo) (Fig. 2). Two additional sound speed profiles from June were used to assess the effect of the environmental variability in response to summer warming and the local internal tide. Four simplified profiles covering the expected envelope of summer variability and a 75 m internal tide were also used to explore the effect of variability in propagation medium.

The validity of the model was first examined by propagating actual blue and fin whales infrasounds, localized with the hydrophone arrays deployed in 2003 and 2004 by using hyperbolic and isodiachron algorithms [Spiesberger and Fristrup (1990); Spiesberger and Whalberg (2002); Spiesberger (2004)] (Fig. 1), and estimating propagation losses by ORCA. These latter were added to the measured received levels on the arrays (in decibels re 1 μPa$_{rms}$ for the call duration corresponding to 90% of total energy) to estimate the SLs. The resulting SLs (167–194 dB re1 μPa$_{rms}$ at 1 m) for possible source depths of 20–50 m generally agreed with published estimates [McDonald *et al.* (2001); Charif *et al.* (2002); Oleson *et al.* (2007); Širović *et al.* (2007)], except for low estimates at a few stations where sloping bathymetry differs from the model and sound speed profiles may be affected by tidal upwelling (cf. Discussion).

The SLs used for the simulations were 190 and 160 dB re 1 μPa$_{rms}$ at 1 m for infrasounds [McDonald *et al.* (2001); Charif *et al.* (2002); Oleson *et al.* (2007); Širović *et al.* (2007)] and audible *D*-call [Berchok *et al.* (2006)], respectively. The simulated calls were downsweep chirps, from 22.5 to 17.8 Hz lasting for 1 s for representing infrasound calls, especially the 20 Hz pulse of fin whales (but also blue whale infrasounds, which obey to same propagation conditions), and from 85 to 35 Hz lasting for 2 s for the blue whale *D*-call. Simulations were run for calling depths from 1 to 50 m, the possible range of depths where air-driven calls can likely be produced [Aroyan *et al.* (2000)] and were observed *in situ* [Oleson *et al.* (2007)]. The calling depths used for assessing the performance of the hydrophone array to detect and localize whales in the feeding ground were 25

TABLE II. Parameters used for Laurentian channel silt bottom description for ORCA normal mode propagation model.

Variable	Layer 1	Layer 2	Bed rock
Gradient	Linear	Linear	
Thickness (m)	1	200	
Compressional wave speed top of layer (m s^{-1})	1473	1575	1700
Compressional wave speed bottom of layer (m s^{-1})	1575	1575	
Shear wave speed top of layer (m s^{-1})	0	80	0
Shear wave speed bottom of layer (m s^{-1})	80	80	
Density top of layer (kg l^{-1})	1.0	1.7	1.8
Density bottom of layer (kg l^{-1})	1.7	1.7	
Compressional wave attenuation top of layer (dB λ^{-1})	−0.1	−1.0	−0.5
Compressional wave attenuation bottom of layer (dB λ^{-1})	−1.0	−1.0	
Shear wave attenuation top of layer (dB λ^{-1})	0	−1.5	0
Shear wave attenuation bottom of layer (dB λ^{-1})	−1.5	−1.5	

and 50 m, which cover the likely bounds of calling depth range.

Cumulative PDFs of noise levels in the calling bands were used to assess call masking by attributing a probability of detection/masking to the modeled call levels as function of range and depth. These functions were then used to map the detection and localization expectancy for the array deployed in 2003 (Fig. 1) for source depths of 25 and 50 m and receiver depth of 100 m. Masking probability was assessed for SNRs of 0 dB for mean summer noise level conditions and for low-noise conditions corresponding to the first quartile of the noise PDFs. Such masking corresponds to a detector (whale or signal processor) that would integrate the noise at frequencies in the vicinity of the specific call frequency band for the call duration. These conditions best match the 20 Hz pulse of fin whales or the A and B calls of blue whales, for which the narrow bandwidth and/or short duration prevents any detection gain through signal processing.

For the larger bandwidth blue whale *D*-call downsweep, which lasts a few seconds, it can be shown that a 10 dB processing gain can be obtained from an optimal time-frequency detector [e.g., Mellinger and Clark (2000)]. This possibility is therefore considered in assessing the expected detection and localization of *D*-call from the hydrophone array. The detection probability of the array at a given location is defined as the maximum probability level obtained for that point among all hydrophones of the array. 2D localization requires detection on a minimum of four hydrophones [cf. Spiesberger (2001)]. The localization probability expected at a given location is the fourth highest rank of the detection probability for that location from all hydrophones.

III. RESULTS

A. Single hydrophone configuration

In the upper water column where the whale is expected to call, the probability to detect a call under mean noise conditions increases as the source depth moves toward the ~60-m-deep sound channel axis (Fig. 2) in response to the downward refraction (Fig. 4). The probability of detecting a whale calling in the upper 20 m is less than 50% for infrasounds at distances larger than 60 km [Fig. 4(a)] and almost nil for *D*-calling blue whales at ranges larger than 5 km for a 100-m-deep hydrophone [Fig. 5(a)]. Constructive and destructive modal interferences generate up to $\sim \pm 5$ dB fluctuations around the decreasing trend in received levels as function of range in the first 20 km, which translates in $\sim \pm 5\%$ fluctuations in expected detections (Fig. 6). These fluctuations are minimal in the sound channel at depths of ~100–130 m, from the modeled total loss as function of depth and range for sources ranging from 20 to 40 m and the range of oceanographic conditions (not shown). The detectability of *D*-call is improved under low-noise conditions but remains less than 50% at ranges larger than 12 km [Fig. 5(b)]. A 10 dB processing gain significantly improves their detectability [Figs. 5(c) and 5(d)], but it still remains lower than that of infrasounds.

The different sound speed profiles at the beginning of the whale season in June, yet only superficially affected by summer warming (Fig. 2), produce a 20 Hz call detection pattern that is very similar to that of September ($\pm 7\%$), except for the positions of the zones of modal interference, which change in range by ± 1 km [Fig. 4(b)]. Similarly, the semidiurnal change of sound speed profile due to the internal tide (e.g., Fig. 2) also generates slight local changes ($\pm 8\%$) in call detection in response to similar range shifts in modal interference [Fig. 4(c)]. The June conditions appear to be slightly more favorable to *D*-call detection [Figs. 5(e) and 5(f)]. These conditions also slightly shift the band pattern of detection with range.

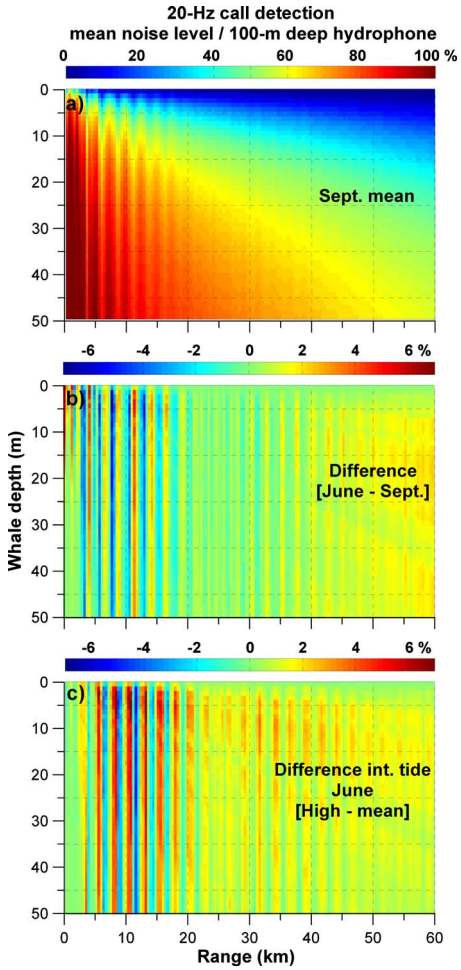


FIG. 4. (Color online) Percentage of calls expected to be detected by a 100-m-deep receiver as function of whale calling depth and range, for infrasounds for the September mean sound speed profile of Fig. 2. (a) The difference with the June 7 sound speed profile (b) and the variation due to the high internal tide in June (c).

B. Sparse array configuration

Infrasound spatial detection expectancy for the possible range of source depths under mean noise conditions is higher than 70% within the simulated array inner space [Figs. 7(a) and 7(c)]. The maps show annuli patterns around the hydrophones resulting from the above mentioned modal interferences. The localization expectancy within the array varies from 55% to 93% and presents a mosaic blueprint generated by the hydrophone detection patterns and the array configuration [Figs. 7(b) and 7(d)]. The detectability maps of the

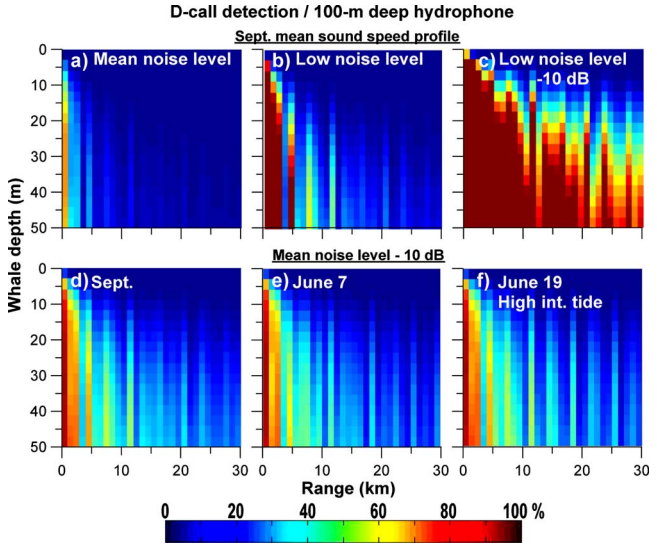


FIG. 5. (Color online) Percentage of calls expected to be detected by a 100-m-deep receiver as function of whale calling depth and range, for audible *D*-call for mean and low noise with the September mean sound speed profile [(a)–(c)] and with a 10 dB processing gain for the three sound speed profiles of Fig. 2: September mean (d), June (e), and June high internal tide (f).

higher-frequency *D*-call with a 10 dB gain processor are more variable and include areas where only 10%–20% of the calls are expected to be detected [Figs. 8(a) and 8(d)]. The maps of localization expectancy mirror this lower detectability with values varying from 5% to 55% for the mean noise conditions [Figs. 8(b) and 8(e)]. This is considerably improved under low-noise conditions, where values range from 20% to 100% [Figs. 8(c) and 8(f)].

C. Discussion

Noise level PDFs used in this study come from multiyear sampling effort distributed over the whole study area in summer, for depths corresponding to that used in the modeling. The 15 960 h of recordings this sampling effort represents make us confident that the PDF estimates are robust and representative of the average summer conditions at the head of the Laurentian channel in the St. Lawrence estuary. Flow noise at the hydrophone and strumming from the mooring are often a problem in estimating noise levels in low-frequency bands such as the whale call bands considered here. Much of these possible interferences were eliminated when selecting the time periods for estimating PSDs. For PAM implementation, however, these flow related interferences would hinder whale detection during a part of the tidal cycle, varying in length with the fortnight cycle. The three-dimensional pattern of tidal currents predicted by an operational circulation model [e.g., Lavoie *et al.*, (2000); Saucier and Chassé (2000)] could help optimizing hydrophone locations to minimize such interferences. For example, locating the hydrophones on the Laurentian channel slopes at a depth of ~50 m, as in 2003, may seem *a priori* favorable but stronger tidal currents in this area make this choice less advantageous than deeper depths in the basin just below the sound channel, as used in the modeled array.

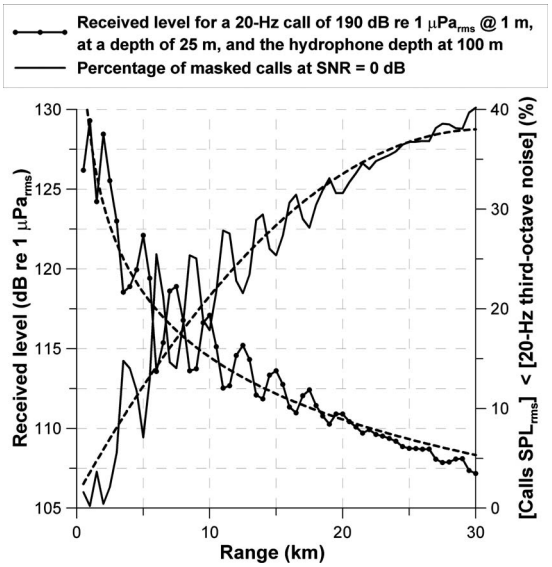


FIG. 6. 20 Hz call received level and percentage of masked calls under mean noise conditions as function of range for a 25-m-deep calling whale and a 100-m-deep receiver.

Sound channel characteristics should also be simultaneously considered in configuring the optimal array. The 100–130 m receiver depths used in 2004 and 2005 and for the simulations appear as the optimal layer to put the hydrophones to detect 20–40 m calling whales. In this layer where the sound is steadily channeled in summer, the influence of modal interference is largely reduced compared to upper and lower depths, and the seasonal and tidal changes in oceanographic characteristics have little effects. Therefore, the detection range is increased and the detection probability as function of range is more stable. The detection radius is thus larger and the amplitude of the annulus detection pattern is reduced compared to other possible receiver depths in the water column. A receiver is also expected to be less strongly imprinted by noise from transiting ships at a depth of 100 m than at a depth of 50 m. Therefore, the duration of the masking periods would be reduced.

The results clearly show the interest of taking into account the regional noise conditions and propagation characteristics in assessing the detectability and localization expectancy of large aperture hydrophone arrays to monitor whales over a basin from their calls, especially when these overlap with the main spectral band of local noise. Although baleen whales' powerful infrasounds can be detected over distances exceeding hundreds of kilometers in deep oceans [Stafford *et al.* (1998)], at their traditional feeding ground in the Saguenay—St. Lawrence Marine Park, it appears that, beyond ~60 km, the majority of these calls are most likely masked by the seaway shipping noise. Detectability of the audible blue whale *D*-call is less because of their 30 dB lower SL and higher noise levels in this band, which is closer to shipping noise spectral peak [Wales and Heitmeyer (2002); Simard *et al.* (2006a)]. However, their larger bandwidth and time-frequency structure allows processing gain to significantly improve their detectability. This allows maintaining this call in the list of valid prospects for whale monitoring over large distances from sparsely distributed hydrophones, despite its lower SL.

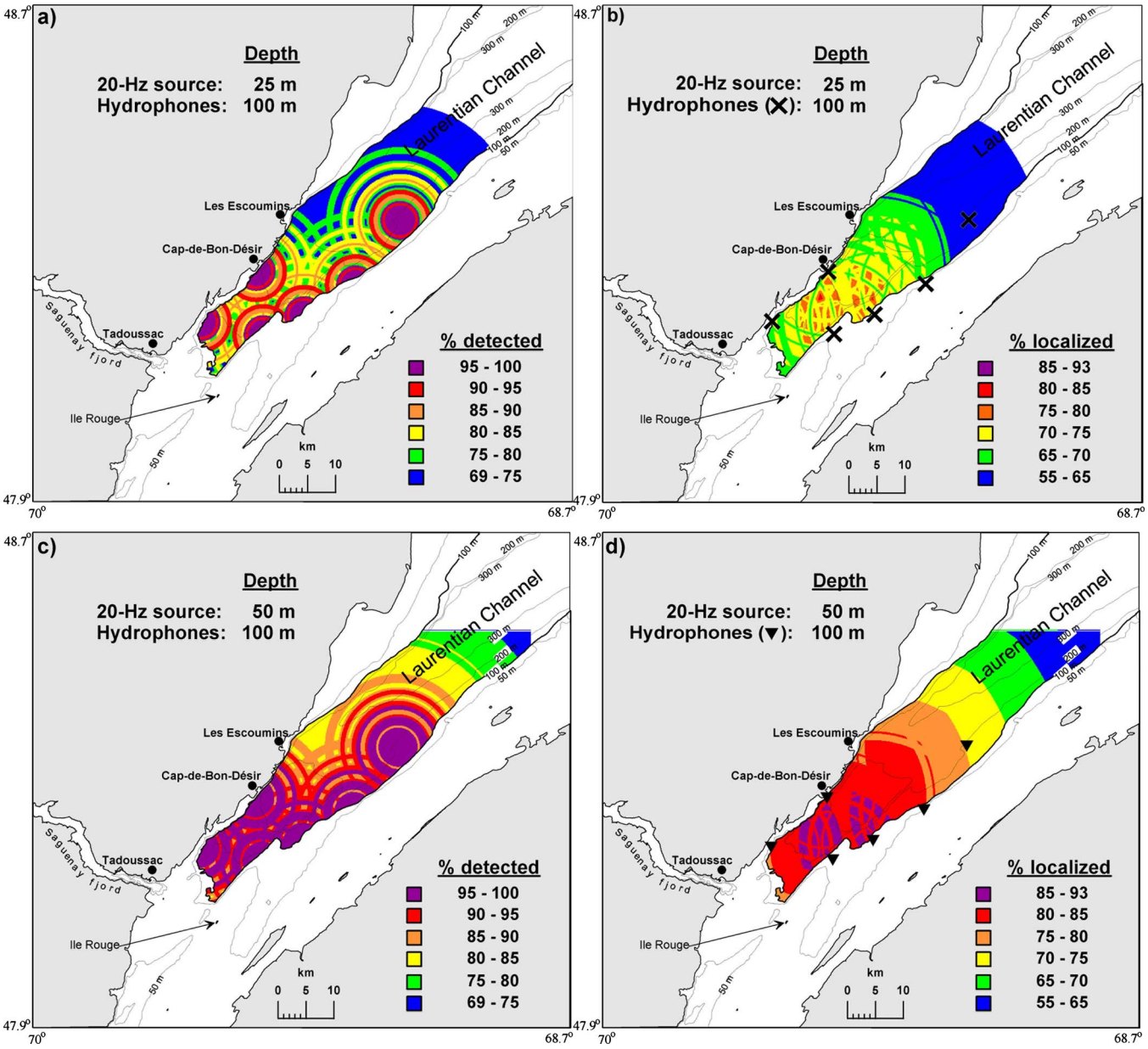


FIG. 7. (Color online) Maps of percentage of infrasound calls from 25 and 50 m calling whales expected to be detected [(a)–(c)] and localized [(b) and (d)] by a 100-m-deep hydrophone array at the 2003 array location under mean noise conditions on the seaway. Nonlinear palettes.

The presence of a well defined sound channel at intermediate depths in summer, due to the cold intermediate layer of North West Atlantic, provides notable gain to received levels as calling whales are approaching the channel axis. This latter may, however, be too deep to provide maximum gain based on present knowledge on calling depths of baleen whales [Aroyan *et al.* (2000); Oleson *et al.* (2007)]. Strong internal tides and higher-frequency internal waves at the head of Laurentian channel [e.g., Saucier and Chassé (2000)] generate semidiurnal vertical oscillations of the sound channel that are modulated by the fortnight tidal cycle. One can expect that call propagation and detectability would increase during flood and decrease during ebb, assuming that whales are calling between 25 and 50 m depths. However, the simulations showed that the gain may average zero but variations in detection probability by ±7% over ±1 km should be expected depending on receiver depth. Whales' shallow night dives [Michaud and Giard (1998)] during the nocturnal migration of their krill prey to feed in the ~20 m surface layer [Sourisseau *et al.* (2008)] could result in less propagating and less detectable shallow calls.

The kilometer-scale pattern in modeled received levels with range is a feature that would likely exist *in situ* because of the constructive interferences of the various propagating modes of the call, which are traveling at different speeds. However, the actual realizations of the interferences at a given location and time would likely be modulated by the time and range-dependent characteristics of the water masses, bottom gradients, and basin shape, which are not taken into account in the present normal mode propagation model. The simulations showed that seasonal and tidal variabilities can generate ±1 km shift in the positions of these annulus patterns. They also showed that these modal interference patterns can introduce ±15–20 dB local variations

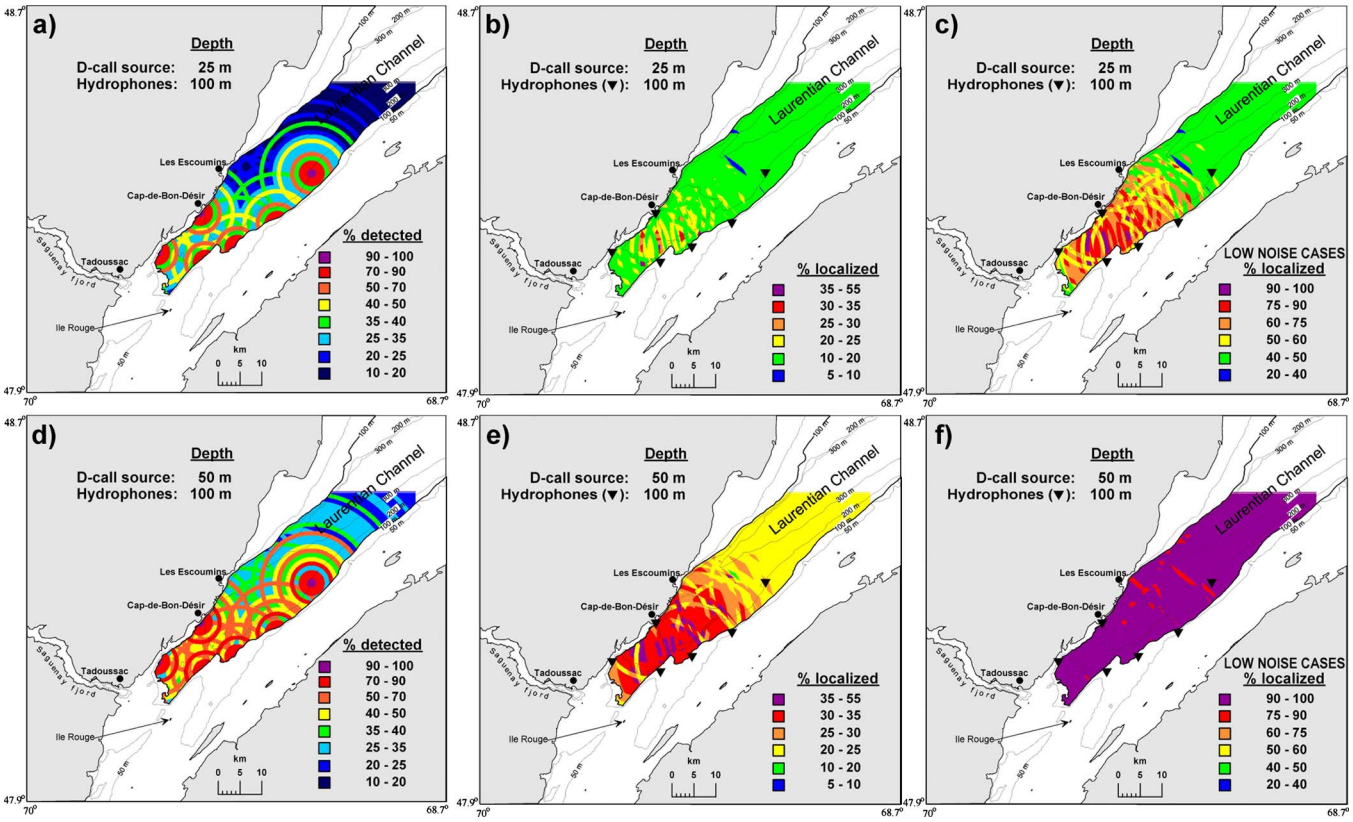


FIG. 8. (Color online) Maps of percentage of blue whale audible *D*-call from 25 and 50 m calling whales expected to be detected [(a) and (d)] and localized [(b), (c), (e), and (f)] with a 10 dB gain processor by a 100-m-deep hydrophone array at the 2003 array location under mean noise conditions [(a), (b), (d), and (e)] on the seaway and low-noise conditions occurring 25% of the time [(c) and (f)]. Nonlinear palettes.

in propagation loss and expected received levels from given sources, for both the seasonal and the tidal scales. The estimates of call SLs from the addition of propagation loss to measurements at a sparse array of distant hydrophones should show such a variability. The variability of our SLs estimates for observed fin whale 20 Hz pulses and blue whale A calls appears therefore realistic given the local characteristics of the internal tide and higher-frequency propagating internal wave [Saucier and Chassé (2000); Lavoie *et al.* (2000)]. The SL estimate range includes the published values for these calls, which is indicative of a reasonably unbiased adjustment of the propagation model to the regional conditions.

A 100-m deep PAM array of hydrophones distributed 10–15 km apart on either sides of the Laurentian channel allows at least ~70% and 55% infrasound detection and 2D localization respectively, for a whale calling at a depth of 25 m. The equivalent minima for blue whale *D*-call are about one-fourth to one-third of these percentages when a 10 dB processing gain is taken into account. To get spatial detectability and localization expectancies comparable to that of infrasounds, the detection and localization must be limited to lowest noise levels present in the area one quarter of the time. These modeling results indicate that an array of hydrophones, deployed at the optimal depth and regularly spaced about 10 km apart around the Laurentian channel, may provide a reasonably good configuration for monitoring blue and fin whales in the study area. Its detection expectancy would be $>\sim 90\%$ and 2D localization expectancy $>\sim 75\%$ for infrasounds in continue, and 25% of the time for blue whale *D*-call. A safe whale monitoring strategy with such a PAM would be to attach a confidence level to the whale distribution maps integrated over given time periods based on the noise levels measured at the hydrophones used for the localization. Likewise, when the hydrophones are individually used as whale detectors, call detection functions determining the detection radius of a hydrophone could be linked to the noise level in estimating the local call density for a given time period.

The sound field around a transiting ship is characterized by a three-dimensional anisotropic pattern [Arveson and Vendittis (2000); Wales and Heitmeyer (2002)] that extends over a few kilometers and affects the nearby hydrophone for a period of about 1 h at the average ship speed. Further work should explore the effect of taking this shipping noise time-space structure into account in modeling the detection and localization probability. This study was limited to the relatively flat and homogeneous Laurentian channel trench, for which the normal mode propagation model of ORCA was most appropriate and where the noise data were recorded. To extend the study to the shallower surrounding areas, the consideration of range-dependent environmental characteristics through a parabolic equation model and noise time series from these areas would be needed. Substantial efforts would then be required to properly take into account the effects of the complex bathymetry and bottom characteristics surrounding the basin, notably the steep Laurentian channel slopes where whales are often observed [Michaud *et al.*

(1997)] in response to the local aggregation of their food [Lavoie *et al.* (2000); Simard *et al.* (2002); Cotté and Simard (2005)]. The temporal variability in sound speed profile over the season and at higher frequencies should be simultaneously taken into account for accurate modeling.

Such modeling coupled with measured noise PDFs is required to determine the performance of different hydrophone array setups for detecting and localizing a proportion of whales' calls. Further work could try to evaluate the effect of the array configuration and the noise spatial pattern on the precision of the localization. The implementation of PAM to track whales in high-noise environments is truly challenging [e.g., Phillips *et al.* (2006); Buaka Muanke and Niezrecki (2007)]. The present study for the noisy Saguenay—St. Lawrence Marine Park whale feeding ground indicates that adequate protocols can be developed to optimize such PAM task even under difficult conditions.

ACKNOWLEDGMENTS

This work has been funded by Fisheries and Oceans Canada, NSERC Res. Grants to Y.S., ISMER-UQAR, FQRNT Québec-Ocean research network fund. We thank the scientific teams and the crews of the RV Coriolis II, FRV Calanus II, CCGS Isle Rouge and Cap d'Espoir for their generous contribution to the work at sea. We also thank Parks Canada Saguenay—St. Lawrence Marine Park for their steady collaboration and Department of National Defense DRDC for the calibration of the hydrophones.

Aroyan, J. L., McDonald, M. A., Webb, S. C., Hildebrand, J. A., Clark, D., Laitman, J. T., and Reidenberg, J. S. (**2000**). "Acoustic models of sound production and propagation," in *Hearing by Whales and Dolphins*, edited by W. W. L Au, A. N. Popper, and R. R. Fay (Springer-Verlag, New York), Chap. 10, pp. 409–469.

Arveson, P. T., and Vendittis, D. J. (**2000**). "Radiated noise characteristics of a modern cargo ship," J. Acoust. Soc. Am. **107**, 118–129.

Berchok, C. L., Bradley, D. L., and Gabrielson, T. B. (**2006**). "St. Lawrence blue whale vocalizations revisited: Characterization of calls detected from 1998 to 2001," J. Acoust. Soc. Am. **120**, 2340–2354.

Buaka Muanke, P., and Niezrecki, C. (**2007**). "Manatee position estimation by passive acoustic localization," J. Acoust. Soc. Am. **121**, 2049–2059.

Charif, R. A., Mellinger, D. K., Dunsmore, D. K., Fristrup, K. M., and Clark, C. W. (**2002**). "Estimated source levels of fin whale (*Balaenoptera physalus*) vocalizations: adjustments for surface interference," Marine Mammal Sci. **18**, 81–98.

Clark, C. W., and Fristrup, K. M. (**1997**). "Whales'95: A combined visual and acoustic survey of blue and fin whales off Southern California," Rep. Int. Whal. Comm. **47**, 583–600.

Cotté, C., and Simard, Y. (**2005**). "The formation of rich krill patches under tidal forcing at whale feeding ground hot spots in the St. Lawrence Estuary," Mar. Ecol.: Prog. Ser. **288**, 199–210.

Hoyt, E. (**2001**). "Whale watching 2001: Worldwide tourism numbers, expenditures and expanding socioeconomic benefits," IFAW, Yarmouth Port, MA, USA.

Jensen, F. B., Kuperman, W. A., Porter, M. B., and Schmidt, H. (**2000**). *Computational Ocean Acoustics* (Springer-Verlag, New York) p. 38.

Lavoie, D., Simard, Y., and Saucier, F. J. (**2000**). "Aggregation and dispersion of krill at channel heads and shelf edges: the dynamics in the Saguenay–St. Lawrence Marine Park," Can. J. Fish. Aquat. Sci. **57**, 1853–1869.

Loring, D. H., and Nota, D. J. (**1973**). "Morphology and sediments of the Gulf of St. Lawrence," Bulletin of Fishery Research Board of Canada No. 182.

Massé, M. (**2001**). "Évolution générale des dépôts quaternaires sous l'Estuaire du St-Laurent entre l'Ile aux Lièvres et Rimouski (General trend of quaternary deposits of the St. Lawrene Estuary between Ile aux Lièvres and Rimouski)," M.Sc. Thesis, Université du Québec à Rimouski, Rimouski, Qc, Canada.

McDonald, M. A., Calambokidis, J., Teranishi, A. M., and Hildebrand, J. A. (**2001**). "The acoustic calls of blue whales off California with gender data," J. Acoust. Soc. Am. **109**, 1728–1735.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529.

Mellinger, D., and Clark, C. W. (**2003**). "Blue whale (*Balaenoptera musculus*) sounds from the North Atlantic," J. Acoust. Soc. Am. **114**, 1108–1119.

Mellinger, D. K., Stafford, K. M., Moore, S. E., Dziak, R. P., and Matsumoto, H. (**2007**). "An overview of fixed passive acoustic observation methods for cetaceans," Oceanogr. **20**, 36–45.

Michaud, R., Bédard, C., Mingelbier, M., and Gilbert, M. C. (**1997**). "Les activités d'observation en mer des cétacés dans l'estuaire maritime du Saint-Laurent 1985–1996 (The 1985–1996 whale watching activity in Lower St. Lawrence Estuary)," GREMM, 108 de la Cale Sèche, Tadoussac, Québec G0T-2A0. Final report to Parks Canada Saguenay-St. Lawrence Marine Park, Can. Heritage Dept., Ottawa.

Michaud, R., and Giard, J. (**1998**). "Les rorquals communs et les activités d'observation en mer des cétacés dans l'estuaire maritime du Saint-Laurent en 1994–1996 (The fin whales and the whale watching activity in Lower St. Lawrence Estuary in 1994–1996)," GREMM, 108 de la Cale Sèche, Tadoussac, Québec G0T-2A0. Final report to Parks Canada Saguenay–St. Lawrence Marine Park, Can. Heritage Dept. Ottawa.

NRC (**2003**). *Ocean Noise and Marine Mammals* (The National Academies Press, Washington D.C.).

Oleson, E. M., Calambokidis Burgess, J. W. C., McDonald, M. A., LeDuc, C. A., and Hildebrand, J. A. (**2007**). "Behavioral context of call production by eastern North Pacific blue whales," Mar. Ecol.: Prog. Ser. **330**, 269–284.

Phillips, R., Niezrecki, C., and Beusse, D. O. (**2006**). "Theoretical detection ranges for acoustic based manatee avoidance technology," J. Acoust. Soc. Am. **120**, 153–163.

Richardson, W. J., Greene, C. R., Malme, C. I., and Thomson, D. H. (**1995**). *Marine Mammals and Noise*, Academic, New York.

Saucier, F. J., and Chassé, J. (**2000**). "Tidal circulation and buoyancy effects in the St. Lawrence estuary," Atmos.-Ocean. **38**, 505–556.

Simard, Y., and Lavoie, D. (**1999**). "The rich krill aggregation of the Saguenay–St. Lawrence Marine Park: hydroacoustic and geostatistical biomass estimates, structure, variability and significance for whales," Can. J. Fish. Aquat. Sci. **56**, 1182–1197.

Simard, Y., Lavoie, D., and Saucier, F. J. (**2002**). "Channel head dynamics: Capelin (Mallotus villosus) aggregation in the tidally-driven upwelling system of the Saguenay–St. Lawrence Marine Park's whale feeding ground," Can. J. Fish. Aquat. Sci. **59**, 197–210.

Simard, Y., and Roy, N. (**2008**). "Detection and localization of blue and fin whales from large-aperture autonomous hydrophone arrays: a case study from the St. Lawrence Estuary," Can. Acoust. **36**, 104–110.

Simard, Y., Roy, N., and Gervaise, C. (**2006a**). "Shipping noise and whales: World tallest ocean liner vs largest animal on earth," in OCEANS'06 MTS/IEEE—Boston, IEEE, Piscataway, NJ, USA. DOI: 10.1109/OCEANS.2006.307053, p. 1–6.

Simard, Y., Bahoura, M., Park, C. W., Rouat, J., Sirois, M., Mouy, X., Seebaruth, D., Roy, N., and Lepage, R. (**2006b**). "Development and experimentation of a satellite buoy network for real-time acoustic localization of whales in the St. Lawrence," in OCEANS'06 MTS/IEEE—Boston, IEEE Piscataway, NJ, USA. DOI: 10.1109/OCEANS.2006.307052, p. 1–6.

Širović, A., Hildebrand, J. A., and Wiggins, S. M. (**2007**). "Blue and fin whale call source levels and propagation range in the Southern Ocean," J. Acoust. Soc. Am. **122**, 1208–1215.

Sourisseau, M., Simard, Y., and Saucier, F. (**2008**). "Krill diel vertical migration fine dynamics, nocturnal overturns, and their roles for aggregation in stratified flows," Can. J. Fish. Aquat. Sci. **65**, 574–587.

Southall, B. L., Bowles, A. E., Ellison, W. T., Finneran, J. J., Gentry, R. L., Greene, C. R., Kastak, D., Ketten, D. R., Miller, J. H., Nachtigall, P. E., Richardson, W. J., Thomas, J. A., and Tyack, P. L. (**2007**). "Marine mammal noise exposure criteria: initial scientific recommendations," Aquat. Mamm. **33**, 411–522.

Spiesberger, J. L. (**2001**). "Hyperbolic location errors due to insufficient numbers of receivers," J. Acoust. Soc. Am. **109**, 3076–3079.

Spiesberger, J. L. (**2004**). "Geometry of locating sounds from differences in travel time: Isodiachrons," J. Acoust. Soc. Am. **116**, 3168–3177.

Spiesberger, J. L., and Fristrup, K. M. (**1990**). "Passive localization of calling animals and sensing of their acoustic environment using acoustic tomography," Am. Nat. **135**, 107–153.
Spiesberger, J. L., and Whalberg, M. (**2002**). "Probability density functions for hyperbolic and isodiachronic locations," J. Acoust. Soc. Am. **112**, 3046–3052.
Stafford, K. M., Fox, C. G., and Clark, D. S. (**1998**). "Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean," J. Acoust. Soc. Am. **104**, 3616–3625.
Stafford, K. M., Mellinger, D. K., Moore, S. E., and Fox, C. G. (**2007**). "Seasonal variability and detection range modeling of baleen whale calls in the Gulf of Alaska, 1999–2002," J. Acoust. Soc. Am. **122**, 3378–3390.
Tecsult Environment Inc. (**2000**). Étude socio-économique d'un secteur retenu pour l'identification d'une zone de protection marine pilote: Estuaire du Saint-Laurent (Socioeconomical study of an area considered for a pilot marine protected area: Lower St. Lawrence Estuary)," Tecsult Environment Inc. Final report to Fisheries and Oceans Canada, Maurice-Lamontagne Institute, Mont-Joli, Québec, Canada.
Wales, S. C., and Heitmeyer, R. M. (**2002**). "An ensemble source spectra model for merchant ship-radiated noise," J. Acoust. Soc. Am. **111**, 1211–1231.
Watkins, W., Tyack, P., Moore, K., and Bird, J. (**1987**). "The 20-Hz signals of finback whales (Balaenoptera physalus)," J. Acoust. Soc. Am. **82**, 1901–1912.
Wenz, G. M. (**1962**). "Acoustic ambient noise in the Ocean: Spectra and sources," J. Acoust. Soc. Am. **34**, 1936–1956.
Westwood, E. K., Tindle, C. T., and Chapman, N. R. (**1996**). "A normal mode model for acousto-elastic ocean environments," J. Acoust. Soc. Am. **81**, 912–924.

Predicting absorption and dispersion in acoustics by direct simulation Monte Carlo: Quantum and classical models for molecular relaxation

Amanda D. Hanford[a)]
Graduate Program in Acoustics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Patrick D. O'Connor[b)] and James B. Anderson[c)]
Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Lyle N. Long[d)]
Department of Aerospace Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

(Received 24 October 2007; revised 2 April 2008; accepted 2 April 2008)

In the current study, real gas effects in the propagation of sound waves are simulated using the direct simulation Monte Carlo method for a wide range of frequencies. This particle method allows for treatment of acoustic phenomena at high Knudsen numbers, corresponding to low densities and a high ratio of the molecular mean free path to wavelength. Different methods to model the internal degrees of freedom of diatomic molecules and the exchange of translational, rotational and vibrational energies in collisions are employed in the current simulations of a diatomic gas. One of these methods is the fully classical rigid-rotor/harmonic-oscillator model for rotation and vibration. A second method takes into account the discrete quantum energy levels for vibration with the closely spaced rotational levels classically treated. This method gives a more realistic representation of the internal structure of diatomic and polyatomic molecules. Applications of these methods are investigated in diatomic nitrogen gas in order to study the propagation of sound and its attenuation and dispersion along with their dependence on temperature. With the direct simulation method, significant deviations from continuum predictions are also observed for high Knudsen number flows. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912831]

PACS number(s): 43.35.Ae, 43.35.Fj, 43.20.Hq [RR] Pages: 4118–4126

I. INTRODUCTION

Fluid dynamics models for a gas can be categorized into two groups: continuum methods and particle methods. Continuum methods, which are popular for acoustic problems, model the fluid as a continuous medium. This model macroscopically describes the state of the fluid using quantities such as density, velocity, and temperature. The continuum approximation is valid when the characteristic length of the problem is much larger than the molecular mean free path (λ_m). This condition is satisfied for many engineering problems, which can be described using continuum equations such as the Navier–Stokes or Euler equations.

However, the continuum model has its limitations. The macroscopic model assumes that deviations from thermal equilibrium are small, and it is the failure of the closure of the Navier–Stokes equations that limits the applications of this approach. The Knudsen number (Kn) is defined as the mean free path divided by a characteristic length and is a measure of the nonequilibrium or viscous effects of the gas. The Knudsen number is also used to distinguish the regimes where different governing equations of fluid dynamics are applicable. The Navier–Stokes equations are normally valid for $\mathrm{Kn} < 0.05$ and reduce to the Euler equations as Kn approaches zero. The Boltzmann equation is the mathematical model for particle methods and is valid for all Kn. Therefore, particle methods are necessary for, but not limited to, problems where the Knudsen number is greater than about 0.05.

Particle methods are based on molecular models that describe the state of the gas at the microscopic level. Despite the fact that the Boltzmann equation has been derived using a microscopic approach, it has been shown that the Boltzmann equation will reduce to the continuum conservation equations (e.g., Navier–Stokes) for low Kn.[1,2]

Direct simulation Monte Carlo (DSMC) is a stochastic, particle-based method developed by Bird which is capable of simulating real gas effects for all values of Kn that traditional continuum models cannot offer.[1] The Knudsen number is large for sound propagation in dilute gases or at high frequencies. Prior work with DSMC has shown that sound absorption heavily depends on Kn for acoustic wave propagation in monatomic gases.[3,4] The successful application of DSMC to nonlinear acoustic waves has also been demonstrated for monatomic and diatomic gases.[5] The flexibility of

[a)]Electronic mail: ald227@psu.edu.
[b)]Electronic mail: pdo109@psu.edu.
[c)]Electronic mail: jba@psu.edu.
[d)]Electronic mail: lnl@psu.edu.

 0001-4966/2008/123(6)/4118/9/$23.00

the DSMC algorithm has also allowed for modeling of sound in specific gas mixtures including models for Earth, Mars, and Saturn's moon Titan.[5–7]

Further investigation has led to the study of absorption and dispersion with emphasis on inelastic collisions and the phenomenological models used for the exchange of energy, which are the subject of this study. Both discrete and continuous internal energy models at a variety of temperatures and Knudsen numbers were examined in order to gain a more fundamental understanding of the effects of vibrational and rotational modes of molecules in acoustics. This new approach for internal energy models allows us to use DSMC to explore the particle nature of the gas in a realistic way without making the assumptions required in treating the gas as a continuum to further enhance the understanding of internal energy exchange in acoustic wave propagation.

II. DIRECT SIMULATION MONTE CARLO

The DSMC method is a simulation tool that describes the dynamics of a gas through direct physical modeling of particle motions and collisions. DSMC is based on the kinetic theory of gas dynamics modeled on the Boltzmann equation, where representative particles are followed as they move and collide with other particles, and is valid beyond the continuum assumption. The movement of particles is determined by their velocities. While the collisions between particles are statistically determined, they are required to satisfy mass, momentum, and energy conservation. Excellent introductory[3] and detailed[1] descriptions of DSMC, as well as formal derivations,[8] can be found in literature. Due to the particle nature of the method, DSMC offers considerable flexibility with regard to the type of system available for modeling: rarefied gas dynamics,[1,9–11] hypersonic flows,[1,12,13] nonlinear acoustics,[3] and even extending beyond the Boltzmann equation by simulating chemical reactions[14] and detonations[15,16]

Despite the fact that DSMC is valid for all Kn, DSMC is most efficient for high Kn flows and has in fact become the *de facto* tool for high Kn situations. The Knudsen number is large for sound propagation in dilute gases (e.g., high altitude conditions) or at high frequencies, requiring a particle method solution. Experimental work[17] and DSMC simulations[3,4] have shown that sound absorption heavily depends on Kn, which significantly deviates from traditional continuum theory at high Kn. Traditionally, DSMC has been primarily used in regimes where continuum methods fail. However, DSMC has many advantages even for low Kn situations. Without modification, DSMC is capable of simulating all physical properties of interest at the molecular level for sound propagation: absorption, dispersion, nonlinearity, and molecular relaxation.

The current DSMC program that was used for simulations in this paper contains several types of energy models to treat molecules in gas mixtures with internal energy. The internal energy is represented by rotational and vibrational modes (electronic energy is ignored) and has been programed to simulate either the classical or quantum behavior. Each case uses a phenomenological approach developed by

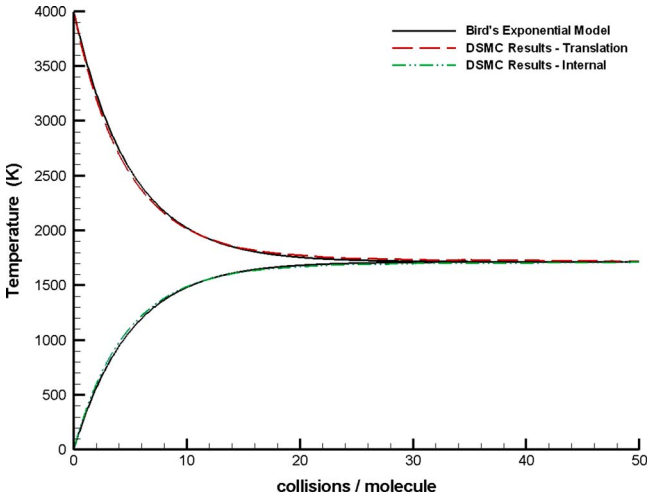


FIG. 1. (Color online) DSMC simulation of nitrogen molecules undergoing classical relaxation at 4000 K. Bird's exponential model (Ref. 1) are given by the solid lines and the DSMC results are in dashed lines.

Borgnakke and Larsen,[18] which treats only a fraction of intermolecular collisions as inelastic. This fraction is the reciprocal of the relaxation collision number Z, which is independently specified for rotational, Z_{rot}, and vibrational, Z_{vib}, degrees of freedom and is dependent on the molecular species of the colliding pair. If a collision is regarded as inelastic, the total energy of the particles involved in the collision is reassigned between the translational and internal modes by sampling from known equilibrium distributions.

In the case of molecules with vibrational internal degrees of freedom, postcollision energy is assigned through either a classical procedure that assigns a continuously distributed vibrational energy to each molecule or through a quantum approach that assigns a discrete vibrational level to each molecule. The vibrational energy exchange is independently treated from the energy exchange in rotational modes. With DSMC, each discrete vibrational energy level of a molecule can be modeled as a separate molecular species. The discrete exchange model used is primarily based on those developed by Anderson *et al.*[19] and Bergemann and Boyd[20] and uses characteristic temperatures (or energy levels) and collision energies to determine the populations of excited states for a given molecular species. This phenomenological model satisfies detailed balance and produces statistically accurate macroscopic behavior in addition to being very computationally efficient.[1]

Simulations of a nitrogenlike system with a fully classical rotation-vibration model and a quasiclassical rotation-vibration model were performed to test the molecular relaxation as a function of the number of collisions per molecule as shown in Figs. 1 and 2, respectively. Both simulations were completed under the same conditions in a single cell with a cell length of 1/2 the mean free path, an initial freestream temperature of 4000 K (translational energy only), and an average of one out of every five collisions treated as inelastic. Relaxation is the mechanism at which the system exchanges energy between translational and internal modes in order to reach a state of thermal equilibrium. The number of collisions required to bring the system to

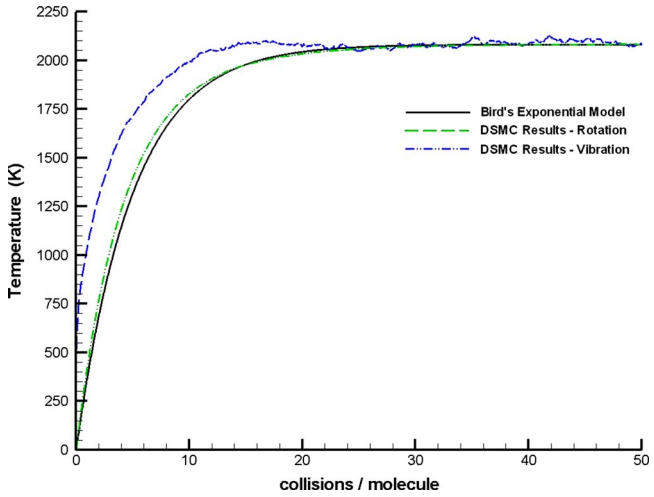


FIG. 2. (Color online) DSMC simulation of nitrogen molecules undergoing relaxation with a coupled discrete vibration/classical rotation model at 4000 K. Bird's exponential model (Ref. 1) are given by the solid lines and the DSMC results are in dashed lines.

approximately equilibrium, also known as the relaxation collision number (Z), was chosen to be 5 for this case. The collision number of nitrogen is approximately 5 at room temperatures.[21] While this choice of collision number does not reflect temperature dependence as seen in experimental results, it was our goal to investigate the relaxation of the internal energy models and their validity for the DSMC method. Therefore, this choice of collision numbers was based on the size of our test system, so that the relaxation effects could be monitored.

Both energy exchange models are compared to simple theoretical treatments (exponential model rates) derived by Bird.[1] Figure 1 shows excellent agreement between the exponential model and the classical rate of energy exchange. In Fig. 2, if no quantum behavior were present, the relaxation rate would be expected to match the model rate. The coupled classical rotation/quantum vibration model deviates from this idealized exponential rate and in fact gives a more realistic rate of exchange than that of the fully classical approach.

III. CONTINUUM THEORY

The physical properties that govern the absorption and dispersion of sound in a gas include classical losses associated with the transfer of acoustic energy into heat and relaxation losses associated with the redistribution of internal energy of molecules. Therefore, the internal energy losses are represented by the relaxation of the molecules' rotational and vibrational energies. A review of the continuum theory for the various losses in gaseous nitrogen will now be given.

A. Thermal-viscous losses

By substituting the harmonic plane wave expansion of the form $\exp[-i(\omega t+K_{cl}x)]$ into continuum thermodynamic equations, one can derive a dispersion relation for the classical thermal-viscous losses from the linearized continuity of momentum, mass, and energy equations. Written in terms of the complex-valued propagation constant $K_{cl}=\beta_{cl}+i\alpha_{cl}$, where $\beta_{cl}=\omega/c_{cl}$ and c_{cl} is the phase speed, and α_{cl} is the classical absorption coefficient, this dispersion relation is given by the equation[22,23]

$$\left(\frac{\omega}{c_0}\right)^2+\left[1+i\frac{\omega}{c_0^2}\left(\frac{4}{3}\frac{\mu}{\rho_0}+\frac{\kappa}{\rho_0 c_v}\right)\right]K_{cl}^2+\frac{\kappa}{\omega\rho_0 c_v}\left(\frac{i}{\gamma}-\frac{4}{3}\frac{\omega\mu}{c_0^2\rho_0}\right)K_{cl}^4=0. \tag{1}$$

Here, c_0 is the low frequency, low amplitude adiabatic speed of sound, ρ_0 is the ambient density, γ is the ratio of specific heats, μ is the coefficient of viscosity, κ is the coefficient of thermal conductivity, and c_v is the specific heat at constant volume. In the limit of low frequencies, Eq. (1) becomes the familiar expression

$$\alpha_{cl}=\frac{\omega^2}{2\rho_0 c_0^3}\left(\frac{4}{3}\mu+\frac{(\gamma-1)\kappa}{c_p}\right), \tag{2}$$

and the phase speed c_{cl} becomes the adiabatic sound speed given by

$$c_{cl}=\sqrt{\frac{\gamma RT_0}{M}}, \tag{3}$$

where R is the universal gas constant, M is the molecular weight, and T_0 is the equilibrium temperature.[24].

The classical thermal-viscous theory given by Eq. (1) can accurately describe the absorption and dispersion in gases with no internal energy at very small Kn. This theory predicts frequency squared dependence for low Kn (low frequency) as seen in Eq. (2), has a maximum value exhibiting a relaxation-type behavior, and is valid within the continuum assumption. When the continuum assumption breaks down, this theory can no longer correctly describe the transport phenomenon, and comparisons between experimental and computational results with predictions given by Eq. (1) for the absorption and dispersion of sound in monatomic gases show significant deviations at high Kn due to slow or incomplete translational relaxation.[3,4,25]

B. Rotational losses

Greenspan[26] presented theoretical expressions for the combined absorption and dispersion due to classical (translational) and rotational relaxation. It was found that if the gas is a Becker gas, with a Prandlt number $c_p\mu/\kappa$ of 3/4, then Eq. (1) is factorable and has an explicit solution.[26] By using this assumption, Greenspan showed that it is possible to write the combined absorption, α_{cr}, and dispersion, β_{cr}, due to translational and rotational relaxation as

$$\alpha_{cr}=\frac{\alpha_{cl}\beta_{\mathrm{rot}}}{\beta_0}+\frac{\alpha_{\mathrm{rot}}\beta_{cl}}{\beta_0} \tag{4}$$

and

$$\beta_{cr}=\frac{\beta_{cl}\beta_{\mathrm{rot}}}{\beta_0}-\frac{\alpha_{cl}\alpha_{\mathrm{rot}}}{\beta_0}, \tag{5}$$

where $\beta_0=\omega/c_0$ and α_{rot} is the absorption due to rotational relaxation. In addition, $\beta_{cl}=\omega/c_{cl}$ and $\beta_{\mathrm{rot}}=\omega/c_{\mathrm{rot}}$ are the scaled dispersion for translational and rotational relaxation,

respectively. β_{cl} and α_{cl} result from Eq. (1) and β_{rot} and α_{rot} are given by the equations

$$\frac{\alpha_{\text{rot}}}{\beta_{\text{rot}}} = \frac{\sigma - 1/\sigma}{2}\left(\frac{n}{r} + \frac{r}{n}\right)^{-1} \tag{6}$$

and

$$\frac{\beta_{\text{rot}}}{\beta_0} = \left(\frac{1}{\sigma}\left(\frac{n}{r} + \frac{r}{n}\right)\left(\frac{\sigma n}{r} + \frac{r}{\sigma n}\right)^{-1}\right)^{1/2}, \tag{7}$$

with $n=(4/5)[(c_p^\infty c_v^\infty)/(c_p^0 c_v^0)]^{1/2} Z_{\text{rot}}$, $\sigma=[(c_p^\infty c_v^\infty)/(c_p^0 c_v^0)]^{1/2}$, and $r=c_0^2\rho/\gamma\omega\mu$ as a nondimensional frequency, which is proportional to 1/Kn. The subscripts p and v represent the specific heat capacities at constant pressure and volume, respectively. The superscripts ∞ and 0 denote conditions that are well above or below the relaxation frequency for rotation. For the case of gaseous nitrogen at low temperatures where rotation is the only internal mode active, $c_p^\infty=5R/2$, $c_v^\infty=3R/2$, $c_p^0=7R/2$, and $c_v^0=5R/2$. This gives for gaseous nitrogen the values $n=(4/5)(3/7)^{1/2}Z_{\text{rot}}$ and $\sigma=5/\sqrt{21}$, where Z_{rot} is the rotational collision number.

At low frequencies, Eq. (4) reduces to $\alpha_{cr}=\alpha_{cl}+\alpha_{\text{rot}}$ and all absorption mechanisms are additive.

C. Vibrational losses

The total absorption coefficient α including vibrational relaxation losses is commonly written[24,27] as

$$\alpha = \alpha_{cr} + \alpha_{\text{vib}}, \tag{8}$$

where α_{cr} is given by Eq. (4) and α_{vib} is the absorption due to vibrational relaxation.

There have been many approaches for the theoretical development of absorption due to vibrational relaxation. The most common approach assumes that the absorption due to a single relaxation process takes the form

$$\alpha_{\text{relax}} = \frac{\pi s}{c_0}\frac{f^2/f_r}{1+(f/f_r)^2}, \tag{9}$$

where s is the relaxation strength and f_r is the relaxation frequency.[28] Therefore, the total absorption due to vibrational relaxation is the sum of the individual relaxation processes. While the development of Eq. (9) relies on microscopic information of the internal structure of the molecules, it is inherently a macroscopic relationship.

It has been shown that vibrational relaxation plays a important role in the absorption of sound at audible frequencies in Earth's atmosphere.[29] At high Kn where frequencies are well above the relaxation frequency, vibration does not contribute to the specific heat. However, at high Kn, the frequency ranges for rotational and translational relaxations overlap with each other and coincide with the breakdown on the continuum assumption. The theory presented in the previous section is an attempt to capture this phenomenon.

In addition, the relaxation frequency for vibration is a complicated function of temperature. Experimental work in shock tubes has given the relaxation time for nitrogen as a function of temperature. The results show that the relaxation of nitrogen is a relatively slow process[30,31] which occurs at relatively low frequencies at low temperatures. However, even for higher temperatures, the vibrational relaxation frequency of nitrogen is still well below the continuum limit.

The theoretical predictions given by Eqs. (1), (4), (5), and (9) compare well to experimental values for gases with internal energy at low frequencies. However, because of the breakdown of the continuum assumption, comparisons with experiment show poor agreement at high Kn. Several molecular-kinetics adjustments have been made to the theory to account for the discrepancy at high Kn with varying degrees of success. Sutherland and Bass[27] used an empirical adjustment to account for the high Kn behavior while Buckner and Ferziger[32] and Sirovich and Thurber[33] used approximations to the Boltzmann equation to describe deviation from the Navier–Stokes prediction for a monatomic gas. Nevertheless, large discrepancies between experiment and theory for gases with internal energy still exist. The DSMC approach and the internal energy models described here can help investigate these discrepancies for describing the acoustic phenomenon for gases with internal energy at high Kn.

IV. SIMULATION APPROACH

Due to the importance of relaxation effects on the absorption and dispersion of sound as presented above, our interest is in investigating the relative importance of internal energy models with DSMC as a function of Kn. Acoustic waves were simulated in a one dimensional simulation domain by creating a pistonlike boundary condition at one end of the domain. The piston was simulated as a rigid wall where particle collisions with the piston face would result in sinusoidally oscillating velocity components. In all cases, the macroscopic velocity amplitude of the piston source is 20 m/s. Specular wall reflections were implemented at the opposite side of the simulation domain. The domain length was 500 cells, with each cell initialized with 50 particles per cell on average. Results for varying Kn were simulated in a nitrogenlike gas with a fixed cross section (molecular weight $M=28.01$ g/mole and cross-sectional diameter $\sigma=3.78\times10^{-10}$ m). The variation in Kn was obtained by maintaining the number of cells per wavelength constant at 100 and varying the cell size from 1/2 of a mean free path to 1/200 of a mean free path.

The time step was taken to be at least an order of magnitude smaller than the mean collision time and is on the order of picoseconds for each case. Care was taken to ensure that the time step remained smaller than the acoustic period of oscillation.

In each case presented, the rotational collision number Z_{rot} was chosen to be 5 and the vibrational collision number Z_{vib} was chosen to be 200. While these choices for rotational and vibrational collision numbers do not reflect temperature dependence or even the correct order of magnitude for vibration as seen in experimental results for nitrogen, it was our goal to investigate the internal energy models and their validity for the DSMC method. The rotational collision number was chosen to be 5 based on experimental measurements of nitrogen at room temperature.[21] Under the worst conditions, the total number of collisions performed over the entire simulation was calculated to be only 60 000 given the small

scale of the system. The vibrational collision number given by experimental values should be close to 150 000.[30,31] This order of magnitude difference implies that no energy exchange would occur within the simulation. In order to simulate vibrational relaxation effects during the simulated time, the vibrational collision number was therefore chosen to be 200 given the goal of modeling the difference between internal energy models. Therefore, the choice of collision numbers was based on the size of our system, so that the relaxation effects could be more easily monitored given the frequency, temperature, and Kn ranges investigated.

Experimentation at high Kn is difficult[25] and results are sensitive to the distance between transmitting and receiving transducers,[4,34,35] especially at high Kn. Collisions of particles with the receiver introduce a wave number that is dependent on distance and also on receiver size.[35] Results taken closer than one mean free path from the transmitting receiver may be influenced by the free molecular flow.[33] In order to take these concerns into account, our results were computed based on the parameter $\omega x/c_m = 10^{34}$ as a nondimensional distance where c_m is the mean molecular velocity of the gas. For $\mathrm{Kn} \geq 0.2$, the computational results were fit for distances $\lambda_m > x > 10c_m/\omega$.

Calculations were performed at temperatures of 273, 2000, and 4000 K in order to monitor the temperature dependence of the excitation of the vibrational mode of nitrogen. For each case, the rotational degree of freedom for nitrogen was classically modeled, while the vibrational degree of freedom was modeled using either the classical model or the quantum model or modeled without vibration. The characteristic temperature of nitrogen was specified as 3371 K, which for low freestream temperatures allows for a single vibrational energy level to adequately model the vibrational temperature. It was shown by Bird that a single vibrational level is adequate to model higher temperature flows.[1] An example of this vibrational modeling at higher temperatures is shown in Fig. 2. For both approaches, nitrogen was modeled using two rotational degrees of freedom, with each degree of freedom represented by a single square term.

Each case was initialized in thermal equilibrium, but since molecular relaxation is a nonequilibrium process, deviation from equilibrium is expected. Section V A describes this in further detail.

Dissociation of diatomic nitrogen in the high temperature-low pressure systems was found to be negligible for the conditions at which the simulations were run. An independent simulation was completed using DSMC containing nitrogen molecules at a temperature of 4000 K. This simulation represented the system shown in Fig. 8 in Sec. V, which was predicted to have the greatest likelihood for dissociation. A reaction model was employed and the nitrogen molecules were given the opportunity to dissociate. Both the number of collision and dissociation events were recorded over the entire simulation time. It was determined that approximately 11 000 collisions were encountered, on average, between each dissociation. Based on the dissociation rate, the total simulation time, and the pressure, the simulations in Fig. 8 are predicted to have less than one molecule undergo dissociation over the entire simulation.

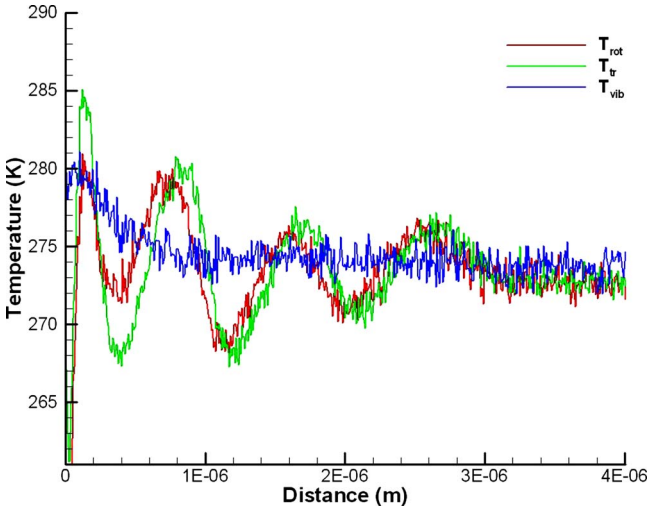


FIG. 3. (Color online) Nonequilibrium effects for Kn=0.02 at 273 K with a classical vibration model

A parallel, object-oriented DSMC solver was developed for this problem. The code was written in C++ and message passing interface for interprocessor communication and was run on massively parallel computers. The object-oriented approach allowed the DSMC algorithm to be divided into physical objects that are individually maintained. Cell and particle classes were created to govern the fundamental components of the algorithm. With this object-oriented technique, it was possible to develop a C++ code that was easy to read, maintain, and modify. Despite excellent parallel efficiency, CPU time and memory requirements were quite large, taking approximately 8 h on 32 processors for each run on NASA's most powerful machine, Columbia.[36]

V. RESULTS

A. Nonequilibrium

The physics of molecular relaxation imply that when a system that starts in equilibrium is subject to a sound source, there is a time delay between the exchange in energy between translational and internal modes. The details of this relaxation process for classical vibration are shown in Fig. 3, where the temperatures associated with the translational, rotational, and classical vibrational modes are computed for Kn=0.02 at 273 K. The relaxation time for the vibrational model is considerably longer than that for rotation and is evident by the minimal disturbance in the vibrational temperature. Similar nonequilibrium effects are seen at higher temperatures.

At high Kn when slow translational and rotational relaxation effects are more evident, the system is in a higher state of nonequilibrium. Relaxation is a very slow process at high Kn where the frequency of oscillation is well above the relaxation frequencies for rotation and vibration. The degree of nonequilibrium can be noted in Fig. 4 for Kn=1.0 at 273 K. Despite starting in equilibrium, the temperatures associated with the translational, rotational, and classical vibrational modes in this case are considerably different. Very little energy has relaxed from the translational mode into the internal modes where the internal modes are in their frozen state.

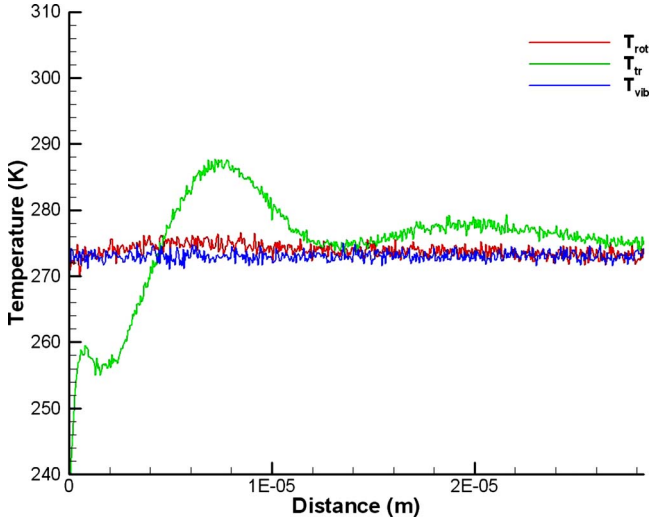


FIG. 4. (Color online) Nonequilibrium effects for Kn=1.0 at 273 K with a classical vibration model.

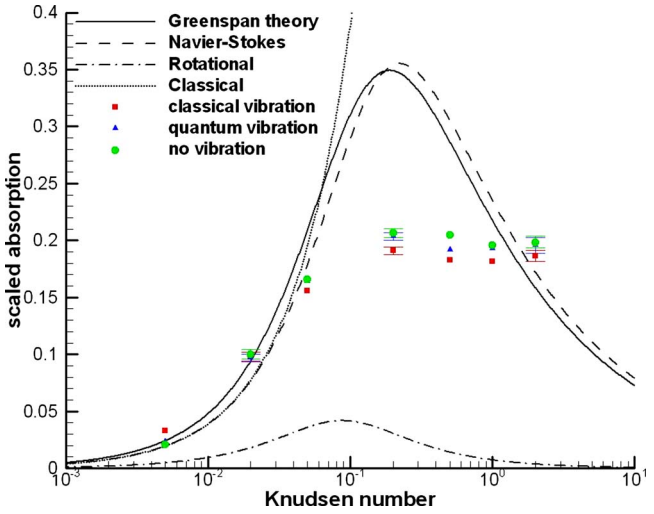


FIG. 6. (Color online) Scaled absorption in nitrogen for 273 K. DSMC results (symbols) are compared to continuum theory predictions (lines).

At low temperatures, the quantum vibrational model will exhibit almost no vibrational activity as a large majority of diatomic nitrogen molecules will remain in the ground state due to the molecules' high characteristic temperature. In contrast, the classical model allows for a continuous distribution of vibrational energy determined by the capacity of the molecule and energy available. Figure 5 shows the fraction of molecules in the excited state in each cell for the quantum vibration model at the temperatures of 273, 2000, and 4000 K for Kn=0.02. As the temperature increases, more molecules are in the excited state, as noted in the figure.

B. Absorption as a function of temperature

The scaled absorption α/β_0 as a function of Kn is shown for 273, 2000, and 4000 K in Figs. 6–8, respectively. DSMC results were computed by tracking the maximum pressure amplitude as a function of distance and computing an exponential line of best fit for extracting the absorption coefficient. DSMC results are also plotted against theoretical predictions given by Eqs. (1), (2), and (4). The results for the quantum, classical, and vibrationless models are shown for a range of Kn in each case. Large deviations from continuum theory are seen for high Kn, as expected due to the breakdown of the continuum. The simulation results approach the free molecular flow limit of 0.2 in each case.[34] In each temperature case, differences between the models are small for low Kn.

Due to the fundamental differences between the models, simulations by using continuous and discrete vibrational models at low temperatures are expected to produce different results. The classical harmonic-oscillator model assumes a fully active vibrational mode, while the quantum harmonic oscillator does not. The quantum harmonic-oscillator model is a more realistic model at low temperatures. For evenly spaced energy levels, the quantum and classical models become identical in the limit of high temperatures.

In the 273 K case, the quantum and vibrationless configurations yield similar results for all Kn. This is because at low temperatures, the fraction of molecules in excited states

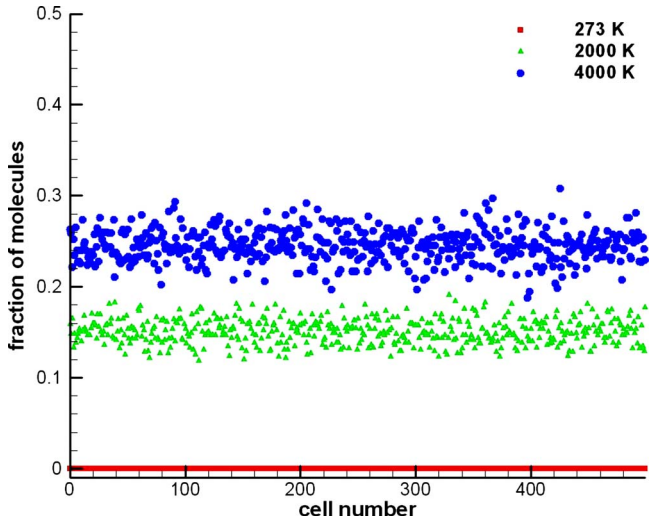


FIG. 5. (Color online) Fraction of molecules in the excited state for Kn =0.02 at 273 K (square), 2000 K (triangle), and 4000 K (circle).

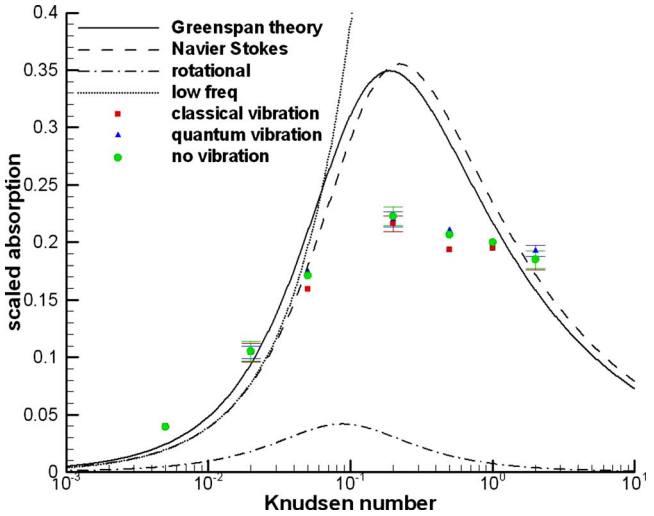


FIG. 7. (Color online) Scaled absorption in nitrogen for 2000 K. DSMC results (symbols) are compared to continuum theory predictions (lines).

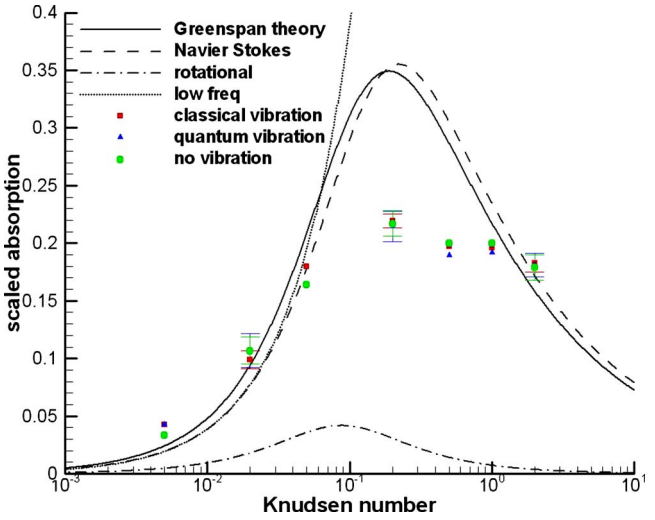


FIG. 8. (Color online) Scaled absorption in nitrogen for 4000 K. DSMC results (symbols) are compared to continuum theory predictions (lines).

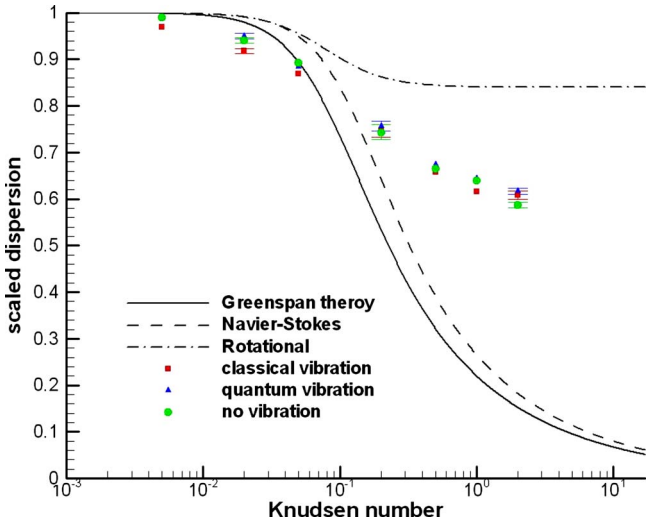


FIG. 9. (Color online) Scaled dispersion in nitrogen for 273 K. DSMC results (symbols) are compared to continuum theory predictions (lines).

is nearly negligible and the vibrational energy is very close to zero, as in the vibrationless case. The classical vibration results are lower than the quantum and vibrationless cases for high Kn and higher than the quantum and vibrationless cases for low Kn. In this case, the differences arise because of different nonequilibrium effects in the classical vibration case, which are a function of the choice of the vibration relaxation collision number. In addition, the theory plotted in the figures given by Eqs. (1)–(3) do not take into account the effects of vibrational relaxation and are only valid for one relaxation process.

As the temperature increases, the excited state for quantum vibration becomes more populated and behaves more like the classic harmonic oscillator. Therefore, the differences between continuous and discrete vibrational models become less, as is shown in Figs. 7 and 8. By 4000 K, all three internal energy configurations yield similar results within the error where the quantum vibration is highly populated.

Due to the stochastic nature of DSMC, there is an intrinsic degree of scatter in the simulation results. Scatter is reduced by averaging many independent ensembles and random error is estimated from variance in the repeated runs. As the temperature increases, the amount of scatter increases due to the increase in molecular speed. The amount of random error was found to have no significant effect on the results shown for 273 K. However, for 2000 and 4000 K, the error is more significant. Error bars indicating one standard deviation are given in the figures.

C. Dispersion as a function of temperature

The scaled dispersion β/β_0 as a function of Kn is shown for 273, 2000, and 4000 K in Figs. 9–11, respectively. DSMC results were computed by tracking zero crossings of the acoustic pressure as a function of distance and computing a line of best fit for extracting the phase speed. DSMC results are also plotted against theoretical predictions given by Eqs. (1), (5), and (7). The results for the quantum, classical, and vibrationless models are shown for a range of Kn in each case. The scaled dispersion data were again computed using the nondimensional distance $\omega x/c_m=10$. Large deviations from continuum theory are seen for high Kn, as expected.

In contrast to the absorption cases, the dispersion seems to be less dependent on internal energy configurations for all of the temperatures. Similarly to absorption, the classical vibration configuration consistently produces results that are lower than the vibrationless and quantum vibration cases. In addition, the intrinsic amount of scatter in the results does not significantly increase as the temperature increases.

VI. CONCLUSIONS

By using the DSMC method, it is possible to simulate the details of molecular relaxation. Investigations with three different methods for treating internal energies were performed as a function of Kn for multiple temperatures. Large deviations from continuum theory were seen for high Kn; therefore, the continuum theory presented in Sec. III fails for sufficiently large Kn. Differences between internal energy

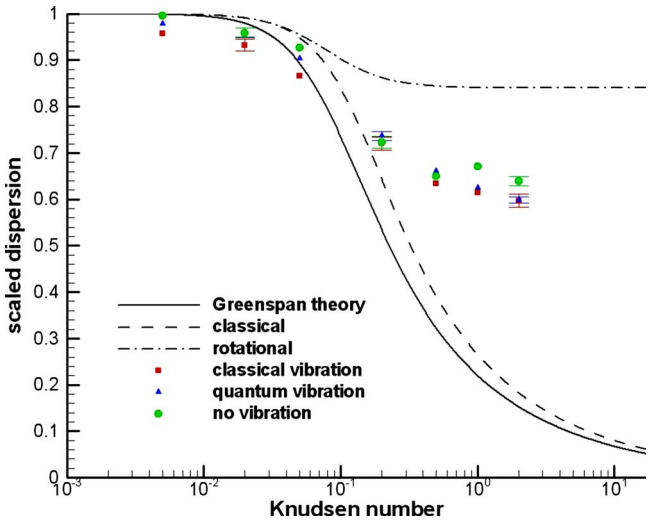


FIG. 10. (Color online) Scaled dispersion in nitrogen for 2000 K. DSMC results (symbols) are compared to continuum theory predictions (lines).

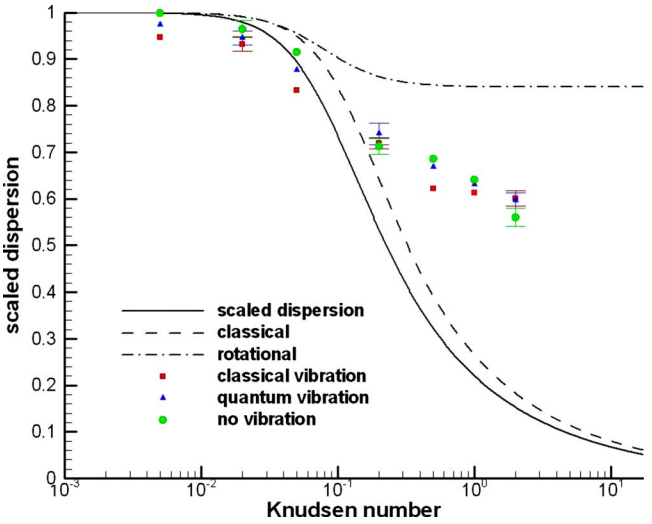


FIG. 11. (Color online) Scaled dispersion in nitrogen for 4000 K. DSMC results (symbols) are compared to continuum theory predictions (lines).

modes were small for low Kn at low temperatures and for all Kn at high temperatures. At high temperatures, the differences between the classical and quantum vibration models are small, implying that the classical model can be a good approximation for vibrational energy exchange. At low temperatures, the quantum model for vibrational energy is the most realistic and should be the model of choice. Nonequilibrium effects were evident in the simulations even at low Kn due to the infrequency of inelastic collisions and inherent differences in the internal energy models.

DSMC is a robust algorithm capable of simulating many systems. The intrinsic scatter of the DSMC algorithm is a universal drawback of DSMC and makes it difficult in certain cases to produce an adequate resolution especially for low acoustic amplitudes and is a source of limitation for this work. Despite this and the large computation time and memory requirements, DSMC is a powerful computational tool that can be used to study the effects of internal energy exchange on acoustic attenuation and dispersion for a large range of Knudsen numbers. The validity of the algorithm for high Kn makes it a powerful tool to study the acoustics in a regime where continuum theory is not valid. The flexibility of the DSMC algorithm has also allowed for modeling of sound in specific gas mixtures. DSMC, along with the internal energy models described here, should be the method of choice for describing the acoustic phenomenon for gases with internal energy at high Kn.

ACKNOWLEDGMENTS

Support by the NASA GSRP and NASA PSGC Fellowship Programs is gratefully acknowledged.

[1]G. A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows* (Clarendon, Oxford, 1994).
[2]S. Chapman and T. G. Cowling, *The Mathematical Theory or Non-Uniform Gases*, 3rd ed. (Cambridge University Press, Cambridge, 1970).
[3]A. L. Danforth and L. N. Long, "Nonlinear acoustic simulations using direct simulation Monte Carlo," J. Acoust. Soc. Am. **116**, 1948–1955 (2004).
[4]N. G. Hadjiconstantinou and A. L. Garcia, "Molecular simulations of sound wave propagation in simple gases," Phys. Fluids **13**, 1040–1046 (2001).
[5]A. Danforth-Hanford, P. D. O'Connor, L. N. Long, and J. B. Anderson, "Molecular relaxation simulations in nonlinear acoustics using direct simulation Monte Carlo," in *Innovations in nonlinear acoustics: ISNA 17, 17th International Symposium on Nonlinear Acoustics including the International Sonic Boom Forum*, edited by A. A. Atchley, V. W. Sparrow, R. M. Keolian (AIP, 2006), pp. 556–559.
[6]A. D. Hanford, L. N. Long, and V. W. Sparrow, "The propagation of sound on Titan using the direct simulation Monte Carlo," J. Acoust. Soc. Am. **121**, 3117–3117 (2007).
[7]A. D. Hanford and L. N. Long, "The absorption of sound on Mars using the direct simulation Monte Carlo," J. Acoust. Soc. Am. **119**, 3264–3264 (2006).
[8]W. Wagner, "A convergence proof for Bird's direct simulation Monte Carlo method for the Boltzmann equation," J. Stat. Phys. **66**, 1011–1044 (1992).
[9]E. P. Muntz, "Rarefied gas dynamics," Annu. Rev. Fluid Mech. **21**, 387–422 (1989).
[10]G. A. Bird, "Monte Carlo simulation of gas flows," Annu. Rev. Fluid Mech. **10**, 11–31 (1978).
[11]E. Oran, C. Oh, and B. Cybyk, "Direct simulation Monte Carlo: Recent advances and applications," Annu. Rev. Fluid Mech. **30**, 403–441 (1998).
[12]G. A. Bird, "Direct simulation and the Boltzmann equation," Phys. Fluids **13**, 2676–2681 (1970).
[13]L. N. Long, "Navier-Stokes and Monte Carlo results for hypersonic flows," AIAA J. **29**, 200–207 (1991).
[14]P. D. O'Connor, L. N. Long, and J. B. Anderson, "Accurate rate expressions for simulations of gas-phase chemical reactions," J. Comput. Phys. (to be published).
[15]P. D. O'Connor, L. N. Long, and J. B. Anderson, "The direct simulation of detonations," in *Proceedings from the AIAA/ASME/SAE/ASEE Joint Propulsion Conference* (AIAA, 2006), Paper No. 2006–4411.
[16]J. B. Anderson and L. N. Long, "Direct Monte Carlo simulation of chemical reaction systems: Prediction of ultrafast detonations," J. Chem. Phys. **118**, 3102–3110 (2003).
[17]M. Greenspan, "Propagation of sound in monatomic gases," J. Acoust. Soc. Am. **29**, 180–180 (1957).
[18]C. Borgnakke and P. S. Larsen, "Statistical collision model for Monte Carlo simulation of polyatomic gas mixture," J. Comput. Phys. **18**, 405–420 (1975).
[19]J. B. Anderson, J. D. Foch, M. J. Shaw, R. C. Stern, and B. J. Wu, "Statistical theory of electronic energy relaxation," in *Proceedings from the 15th International Symposium on Rarefied Gas Dynamics*, edited by V. Boffi and C. Cercignani (B. G. Teubner Stuttgart, 1986), Vol. **2**, pp. 413–421.
[20]F. Bergemann and I. D. Boyd, "New discrete vibrational energy model for the direct simulation Monte Carlo method," in *Proceedings from the 18th International Symposium on Rarefied Gas Dynamics*, edited by B. D. Shizgal and D. P. Weaver (AIAA, 1994), Vol. **158**, pp. 174–183.
[21]H. E. Bass and L. C. Sutherland, "On the rotational collison number for air at elevated temperatures," J. Acoust. Soc. Am. **59**, 1317–1318 (1976).
[22]L. Rayleigh, *Theory of Sound*, 2nd ed. (Dover, New York, 1945).
[23]M. Greenspan, "Propagation of sound in rarefied helium," J. Acoust. Soc. Am. **22**, 684–684 (1950).
[24]A. D. Pierce, *Acoustics—An Introduction to its Physical Principles and Applications* (The Acoustical Society of America, Woodbury, New York, 1989).
[25]M. Greenspan, "Propagation of sound in five monatomic gases," J. Acoust. Soc. Am. **28**, 644–648 (1956).
[26]M. Greenspan, "Combined translational and relaxational dispersion of sound in gases," J. Acoust. Soc. Am. **26**, 70–73 (1954).
[27]L. C. Sutherland and H. E. Bass, "Atmospheric absorption in the atmosphere up to 160 km," J. Acoust. Soc. Am. **115**, 1012–1032 (2004).
[28]K. F. Herzfeld and T. A. Litovitz, *Absorption and Dispersion of Ultrasonic Waves* (Academic, New York, 1959).
[29]H. E. Bass, L. C. Sutherland, A. J. Zuckerwar, D. T. Blackstock, and D. M. Hester, "Atmospheric absorption of sound: Further developments," J. Acoust. Soc. Am. **97**, 680–683 (1995).
[30]R. C. Millikan and D. R. White, "Systematics of vibrational relaxation," J. Chem. Phys. **39**, 3209–3213 (1963).
[31]R. C. Millikan and D. R. White, "Vibrational energy exchange between N_2 and CO: The vibrational relaxation of nitrogen," J. Chem. Phys. **39**, 98–101 (1963).

[32]J. K. Buckner and J. H. Ferziger, "Linearized boundary value problem for a gas and sound propagation," Phys. Fluids **9**, 2315–2322 (1966).
[33]L. Sirovich and J. K. Thurber, "Propagation of forced sound waves in rarefied gasdynamics," J. Acoust. Soc. Am. **37**, 329–339 (1965).
[34]R. Schotter, "Rarefied gas acoustics in the noble gases," Phys. Fluids **17**, 1163–1168 (1974).
[35]F. Sharipov, J. W. Marques, and G. M. Kremer, "Free molecular sound propagation," J. Acoust. Soc. Am. **112**, 395–401 (2002).
[36]Hardware description and documentation available at http://www.nas.nasa.gov/Resources/Systems/columbia.html. Last visited April, 2008.

On the influence of spatial correlations on sound propagation in concentrated solutions of rigid particles

Michael Baudoin[a)]
Université Pierre et Marie Curie-Paris 6, Institut Jean Le Rond D'Alembert (IJLRDA), UMR CNRS 7190 and Institut des NanoSciences de Paris (INSP), UMR CNRS 7588, 4 place Jussieu, 75252 Paris Cedex 05, France

Jean-Louis Thomas
INSP, CNRS, and Université Pierre et Marie Curie-Paris 6, 4 place Jussieu, Paris 75005, France

François Coulouvrat
IJLRDA, CNRS, and Université Pierre et Marie Curie-Paris 6, 4 place Jussieu, Paris 75005, France

(Received 29 May 2007; revised 26 March 2008; accepted 31 March 2008)

In a previous paper [J. Acoust. Soc. Am. **121**, 3386–3387 (2007)], a self-consistent effective medium theory has been used to account for hydrodynamic interactions between neighboring rigid particles, which considerably affect the sound propagation in concentrated solutions. However, spatial correlations were completely left out in this model. They correspond to the fact that the presence of one particle at a given position locally affects the location of the other ones. In the present work, the importance of such correlations is demonstrated within a certain frequency range and particle concentration. For that purpose, spatial correlations are integrated in our two-phase formulation by using a closure scheme similar to the one introduced by Spelt *et al.* ["Attenuation of sound in concentrated suspensions theory and experiments," J. Fluid Mech. **430**, 51–86 (2001)]. Then, the effect is shown through a careful comparison of the results obtained with this model, the ones obtained with different self-consistent approximations and the experiments performed by Hipp *et al.* ["Acoustical characterization of concentrated suspensions and emulsions. 2. Experimental validation," Langmuir, **18**, 391–404 (2002)]. With the present formulation, an excellent agreement is reached for all frequencies (within the limit of the long wavelength regime) and for concentrations up to 30% without any adjustable parameter.
 [DOI: 10.1121/1.2912445]

PACS number(s): 43.35.Bf [AJS] Pages: 4127–4139

I. INTRODUCTION

A precise prediction of the attenuation and dispersion of acoustical waves induced by the presence of particles in suspensions of different natures would be of great interest for acoustic spectroscopy.[1,2] Although the propagation in dilute suspensions is now well described (see Ref. 3 for bubbles, the ECAH theory,[4,5] and the coupled phase theory[6] for emulsions and the models of Gubaidullin and Nigmatulin,[7] Gumerov *et al.*,[8] and Duraiswami and Prosperetti[9] for aerosols), there remain some difficulties in concentrated suspensions as the interactions between neighboring particles must be taken into account. For that purpose, different methods have been used: first, numerical methods which are generally based on the so called "multipole expansion" (see Refs. 10 and 11 for the Helmholtz equation, and Refs. 12–15 for the Stokes and Brinkman equations). Although this numerical treatment of the problem is required for the study of particular configurations (when the particles are not homogeneously distributed), it does not take advantage of the average homogeneous distribution of the particles for randomly distributed spheres. That is why, a statistical treatment of the equations is interesting in this case. First, a hierarchy of mutually dependent averaged equations can be derived (see Ref. 16 for the multiple scattering theory developed for the Helmholtz equation and Ref. 17 for the two-phase Navier–Stokes equations). Then arises the problem of the efficient closure of this hierarchy. In many papers, this hierarchy is truncated at a certain order (generally at first or second order[18–20]). At first order, mutual interactions between neighboring particles are completely left out. At second order, only mutual interactions between two particles are taken into account. This truncation of the hierarchy cannot be used for concentrated suspensions because in this case, mutual interactions between N particles cannot be neglected compared to mutual interactions between $N+1$ particles. In order to avoid this truncation, self-consistent effective medium theories have been widely used in many branches of physics (see Ref. 21 for acoustical waves in bubbly liquids, Ref. 22 for elastic waves in composites, and Refs. 23–25 for two-phase flow, etc.). These methods consist of calculating the constitutive equations by considering a test particle surrounded by an effective medium whose properties are determined in a consistent way. To take into account spatial correlations (that is to say the modification of the particles location due to the presence of the test sphere), different approximations have been introduced (see Refs. 25–27 for a comparison of the different

[a)]Electronic mail: baudoin@lmm.jussieu.fr.

ones). Some take into account the continuous variation of the conditional volume fraction with the distance from the test particle and others approximate this variation by a step function and therefore reduce to "core shell models." A core-shell approximation (originally introduced by Dodd *et al.*[28]) has been successfully used by Spelt *et al.*[29] to compute the propagation of acoustical waves in suspensions of different natures. In their work, the suspension is considered as a whole and the plane acoustical wave impinging a test particle is decomposed according to the ECAH theory[4,5] into compressional, thermal, and viscous modes. Therefore, this model can be used for a large range of frequencies and different kinds of suspensions.

In the following, the same closure scheme (self-consistent approximation with core-shell approximation) is introduced to compute the exchange terms in our two-phase formulation. However, simpler equations are obtained by considering only the scattering mechanisms and pole orders necessary for the study of rigid particle in the long wavelength regime (LWR). Moreover, vectorial expressions of the closure terms (force and stresslet) are obtained and the acceleration of the particles is taken into account. These expressions can therefore be used for other purposes than plane acoustical waves. We must note that for a plane acoustical wave propagating in suspensions of rigid particles in the LWR, the two models should correspond.

The present model is used to study the influence of spatial correlations on the sound propagation in solutions of rigid particles. In the first section, we recall the linearized two-phase equations obtained in a previous paper[30] to describe the sound propagation in solutions of rigid particles. Then, we derive the equations for the conditionally averaged fields which should be solved to take into account spatial correlations and, in particular, the continuous variation of the conditional volume fraction with the distance from the test particle. To perform the explicit calculation of the dispersion equation, a simplification of this problem is used: the conditional volume fraction is approximated by a step function. In this way, a "core-shell" model is obtained but with a "core radius" related to the particle volume fraction and radius by a complex function calculated from Percus–Yevick (PY) theory for hard spheres.[31,32] The results are finally compared to the experiments of Hipp *et al.*[33] performed in solutions of silica particles for different frequencies, particle sizes, and concentration and an excellent agreement is reached.

II. COUPLED PHASE THEORY

A. Linearized ensemble averaged equations and the hierarchy of balance equations

In a precedent paper,[30] ensemble averaged equations have been derived to describe the propagation of acoustical waves in homogeneous solutions of monodisperse rigid particles. They were obtained from local-instant balance equations in each phase by the introduction of the phasic function:

$$\chi_k(\mathbf{x},t) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is in phase } k \text{ at time } t \\ 0 & \text{otherwise,} \end{cases}$$

which allows to broaden the validity of local balance equations to every position and time, and by the use of a statistical average (noted $\langle\ \rangle$ hereafter) such as

$$\langle G'(\mathbf{x},t)\rangle = \int G'(\mathbf{x},t|C_N)p(t,C_N)dC_N,$$

where $p(t,C_N)dC_N$ is the probability of finding the N particles in the vicinity of $C_N=(\mathbf{x}_1,\ldots,\mathbf{x_N})$, regardless of their order and $G'=\Sigma_k\chi_k G'_k$ is a local function generalized to every position and time by the use of the phasic function.

Once linearized, the balance equations stand under the following form.

Mass conservation,

$$\rho_{co}\left(\frac{\partial\alpha_c}{\partial t} + \alpha_{co}\,\mathrm{div}(\mathbf{v_c})\right) + \alpha_{co}\frac{\partial\rho_c}{\partial t} = 0, \tag{1}$$

$$\rho_{do}\left(\frac{\partial\alpha_d}{\partial t} + \alpha_{do}\,\mathrm{div}(\mathbf{v_d})\right) + \alpha_{do}\frac{\partial\rho_d}{\partial t} = 0, \tag{2}$$

$$\alpha_d = 1 - \alpha_c. \tag{3}$$

Momentum conservation,

$$\alpha_{co}\rho_{co}\frac{\partial\mathbf{v_c}}{\partial t} = -\nabla(\alpha_c p_c) + \mu_c\Delta\mathbf{v} + (\lambda_c+\mu_c)\nabla\ \mathrm{div}(\mathbf{v}) + \mathrm{div}\,\mathbf{S} - \mathbf{F}, \tag{4}$$

$$\alpha_{do}\rho_{do}\frac{\partial\mathbf{v_d}}{\partial t} = \mathbf{F}. \tag{5}$$

Equations of state,

$$\rho_c = p_c/c_{co}^2, \tag{6}$$

$$\rho_d = p_c/c_{do}^2. \tag{7}$$

In these equations, the subscripts o, c, and d denote, respectively, the equilibrium state, and the continuous and the dispersed phases, $\alpha_k=\langle\chi_k\rangle$ is the volume fraction of phase k while $\rho_k=\langle\chi_k\rho'\rangle$ is its mean density and $\mathbf{v_k}=\langle\chi_k\rho'\mathbf{v}'\rangle/\langle\chi_k\rho'\rangle$ its mean velocity. Finally, p_c, μ_c, and $\zeta_c=\lambda_c+2\mu_c/3$ are, respectively, the pressure, the shear, and the bulk viscosities of the continuous phase, c_{ko} the sound speed in phase k, $\mathbf{v}=\alpha_{co}\mathbf{v_c}+\alpha_{do}\mathbf{v_d}$ the average velocity of the suspension, $\mathbf{F}=\langle\chi_d\,\mathrm{div}(\mathbf{\Pi}')\rangle$ the interphase force, and $\mathbf{S}=\langle\chi_d\mathbf{\Pi}'\rangle$ the average of the local generalized stress tensor $\mathbf{\Pi}'=\Sigma_{k=c,d}\chi_k\Pi'_k$. We can note that in Eq. (2) (mass conservation for the dispersed phase), the compressibility of the particles has been taken into account in order to give a correct prediction of both attenuation and dispersion for concentrated suspensions of silica nanoparticle. Even if these particles are 35 times less compressible than water (ratio 2, 2 for density, 4 for the sound speed, and therefore 35 for the compressibility $\xi=1/\rho c^2$), their departure from a perfectly rigid behavior might have an influence on the effective sound speed in concentrated suspensions. However, this modifica-

tion does not affect the attenuation curves and they will only induce a global shift of the effective sound speed as we are far from the particle resonances in the frequency range considered here.

To achieve closure, the interphase force **F** and the stresslet **S** must be expressed in terms of the averaged fields. The first step is to establish the link between these expressions and the so-called test particle problem. This particular issue was addressed by Buyevich and Shchelchkova:[24]

$$\mathbf{F} = \langle \chi_d \, \mathrm{div}(\mathbf{\Pi}') \rangle \approx \frac{3\alpha_d}{4\pi a^3} \oint \langle \mathbf{\Pi}' \rangle_{\mathbf{x}} \, . \, \mathbf{n} dS, \tag{8}$$

$$\mathbf{S} = \langle \chi_d \mathbf{\Pi}' \rangle \approx \frac{3\alpha_d}{4\pi a^3} \oint^{(s)} \mathbf{a} \otimes (\mathbf{n} \, . \, \langle \mathbf{\Pi}' \rangle_{\mathbf{x}}) dS, \tag{9}$$

where a is the radius of the particle, **n** the normal vector, and $\langle \, \rangle_{\mathbf{x}}$ the statistical average conditioned by the knowledge of the position of one particle in **x**:

$$\langle G' \rangle_{\mathbf{x}}(\mathbf{x}', t) = \int G'(\mathbf{x}', t | C_N) p(t, C_{N-1} | \mathbf{x}) dC_{N-1}.$$

With these expressions, the interphase force and stresslet are related to the conditionally averaged fields. One could therefore decide to derive the balance equations for the conditionally averaged fields. The same equations would be obtained, but now the conditional force $\mathbf{F}_{\mathbf{x}}$ and stresslet $\mathbf{S}_{\mathbf{x}}$ would depend on the averaged fields with the positions of two particles being known and so on. In this way, an infinite hierarchy of interdependent equations, similar to the cluster expansion which appears in statistical physics, would be disclosed. It was rigorously established by Hinch[17] in 1977.

Now arises the problem of the efficient closure of this hierarchy. The first idea consists of truncating this hierarchy at a certain order. At first order, the constitutive Eqs. (8) and (9) are calculated for a sphere embedded in the pure continuous phase. In this case, interactions between particles are completely left out. This is the approximation classically used in two-phase models.[6,34] At second order, the conditional force $\mathbf{F}_{\mathbf{x}}$ and stresslet $\mathbf{S}_{\mathbf{x}}$ are calculated for a cluster of two spheres lying in the pure continuous phase. In this case, binary interactions of pairs of sphere are taken into account while ternary or higher order interactions are completely left out. One could of course calculate these expressions for higher order clusters but this method is limited. First, because the calculation of the constitutive equations becomes more and more difficult when the size of the cluster increases. Second, because interactions between $N+1$ particles are no more negligible compared to interactions between N particles in very concentrated solutions, and thus the whole hierarchy must be considered.

It is the merit of the pioneering work of Lundgren[23] and Buyevich *et al.*[24,35] to have proposed an alternative *self-consistent effective medium theory* that takes into account interactions at all order within a certain approximation. These self-consistent schemes have then been extended to even more concentrated medium, when spatial correlations must be considered. For this purpose, different approximations have been introduced and they are compared in a paper by Sangani and Yao.[26,27] In our precedent paper,[30] a self-consistent effective medium theory had been used to take into account the influence of hydrodynamic interactions between particles on the propagation of acoustical waves in suspensions of rigid particles. However, spatial correlations were not considered and the theory will now be modified to include these effects.

B. The long wavelength regime

Before delving into this crucial problem, the balance equations will be simplified in the neighborhood of a test particle, lying at position **x** to calculate the surface integrals (8) and (9). For that purpose, a mesoscopic scale l such that $a \ll l \ll \lambda$ can be introduced whenever the wavelength λ is much larger than the radius a of the particle, that is to say in the LWR. Within a cell of characteristic length l around the test particle, all terms linked to the compressibility of the continuous phase can be neglected and thus Eqs. (1)–(7) reduce to the following form after Fourier transform:

$$\mathrm{div}(\mathbf{v_c}) = \mathrm{div}(\mathbf{v_d}) = 0, \tag{10}$$

$$-\alpha_{co}\rho_{co}(i\omega)\mathbf{v_c} = -\nabla(\alpha_c p_c) + \mu_c \Delta \mathbf{v} + \mathrm{div}\, \mathbf{S} - \mathbf{F}, \tag{11}$$

$$-\alpha_{do}\rho_{do}(i\omega)\mathbf{v_d} = \mathbf{F}. \tag{12}$$

If we rewrite them in the convective frame of reference related to the velocity of the test particle, we simply obtain

$$\mathrm{div}(\mathbf{V_c}) = \mathrm{div}(\mathbf{V_d}) = 0, \tag{13}$$

$$-\alpha_{co}\rho_{co}(i\omega)\mathbf{V_c} = -\nabla(\alpha_c p_c) + \mu_c \Delta \mathbf{V} + \mathrm{div}\, \mathbf{S} - \mathbf{F} - \alpha_{co}\rho_{co} \nabla \Psi, \tag{14}$$

$$-\alpha_{do}\rho_{do}(i\omega)\mathbf{V_d} = \mathbf{F} - \alpha_{do}\rho_{do} \nabla \Psi, \tag{15}$$

where $\mathbf{V_k} = \mathbf{v_k} - \mathbf{v_d}|_{r=0}$ is the average velocity of phase k in the new frame of reference, $\Psi = -i\omega \mathbf{r}$. $\mathbf{v_d}|_{r=0}$ is a function introduced to take into account the acceleration of the test particle, $\mathbf{r} = \mathbf{x}' - \mathbf{x}$ is the distance from the test particle, and ω the frequency of the propagating wave.

Similar balance equations can also be derived for the conditionally averaged fields but this time, the conditional volume fraction $\alpha_{ko,\mathbf{x}}$ of phase k stands instead of the unconditional one:

$$\mathrm{div}(\alpha_{co,\mathbf{x}}\mathbf{V_{c,x}}) = \mathrm{div}(\alpha_{do,\mathbf{x}}\mathbf{V_{d,x}}) = 0, \tag{16}$$

$$-\alpha_{co,\mathbf{x}}\rho_{co}(i\omega)\mathbf{V_{c,x}} = -\nabla(\alpha_{c,\mathbf{x}} p_{c,\mathbf{x}}) + \mu_c \Delta \mathbf{V_x} + \mathrm{div}\, \mathbf{S_x} - \mathbf{F_x} - \alpha_{co,\mathbf{x}}\rho_{co} \nabla \Psi, \tag{17}$$

$$-\alpha_{do,\mathbf{x}}\rho_{do}(i\omega)\mathbf{V_{d,x}} = \mathbf{F_x} - \alpha_{do,\mathbf{x}}\rho_{do} \nabla \Psi, \tag{18}$$

where $\mathbf{V_{k,x}} = \mathbf{v_{k,x}} - \mathbf{v_d}|_{r=0}$.

On the test sphere ($r=a$), the conditional velocity of the continuous phase is null and far from it ($r \to \infty$), the influence of the test particle vanishes. We therefore obtain the following boundary conditions:

$$\mathbf{V_{c,x}} = 0 \quad \text{in } r = a, \tag{19}$$

$$\{\mathbf{V}_{\mathbf{c},\mathbf{x}}, \mathbf{V}_{\mathbf{d},\mathbf{x}}, p_{c,\mathbf{x}}\} \to \{\mathbf{V}_{\mathbf{c}}, \mathbf{V}_{\mathbf{d}}, p_c\} \quad \text{when } r \to \infty. \tag{20}$$

III. SPATIAL CORRELATIONS

A. The conditional volume fraction

For pointlike particle, there would be no difference between the conditional volume fraction $\alpha_{do,\mathbf{x}}$ and the unconditional one α_{do}. However, the nonoverlapping property of hard spheres modifies the distribution of particles in the neighborhood of the test particle. We will now see how the conditional volume fraction can be estimated for hard spheres.

First, the so-called distribution function $p(t,\mathbf{x}|\mathbf{x}')$ (which is nothing but the probability of finding one of the sphere center in $\mathbf{x}'$ when another particle is lying in $\mathbf{x}$) can be calculated with models inherited from statistical physics such as the PY (Refs. 31 and 32 or hypernetted chain[36,37] (HNC) models. Numerical methods (for example, Monte Carlo simulations) could also be used but PY theory provides accurate and easily computable estimate of this function.

Then the conditional volume fraction $\alpha_{do,\mathbf{x}}(\mathbf{x}')$ can be deduced from the distribution function $p(t,\mathbf{x}|\mathbf{x}')$ with the following formula[38] for hard spheres:

$$\alpha_{do,\mathbf{x}}(\mathbf{x}') = \int_{|\mathbf{x}''-\mathbf{x}'|\leqslant a} p(t,\mathbf{x}''|\mathbf{x}) d\mathbf{x}''. \tag{21}$$

For isotropically distributed spheres, the distribution function depends only on the distance $r'=|\mathbf{x}''-\mathbf{x}|$ and thus this formula reduces to

$$\alpha_{do,\mathbf{x}}(r) = \int_{\max(2a,r-a)}^{r+a} p(t,r') \frac{\pi r'}{r} [2rr' - r'^2 - r^2 + a^2] dr', \tag{22}$$

with $r=|\mathbf{x}'-\mathbf{x}|$. Figure 1 illustrates the evolution of the distribution function and the conditional volume fraction with the distance r from the test particle center, calculated from PY theory for hard spheres. We can note that the nonoverlapping condition does not mean that some parts of the particles cannot lie in the region $a \leqslant r \leqslant 2a$ but just that the centers of the particles are excluded from it. That is why the volume fraction progressively increases in this region contrarily to the distribution function which is null.

B. The self-consistent effective medium closure scheme

We will now apply the self-consistent condition (called method A in papers from Chang *et al.*[25,39] and Yao and Sangani[26,27]) in order to close the infinite hierarchy previously described. We can note that contrarily to the scheme proposed by Buyevich[38,40] (called method B in the papers of Chang *et al.*), closure will be obtained for the conditionaly averaged field and not for the perturbation (as defined by Buyevich in his paper).

To achieve closure, the force $\mathbf{F}$ and the stresslet $\mathbf{S}$ must necessarily be expressed in terms of the average fields and their space and time derivatives of appropriate tensor dimensionality.

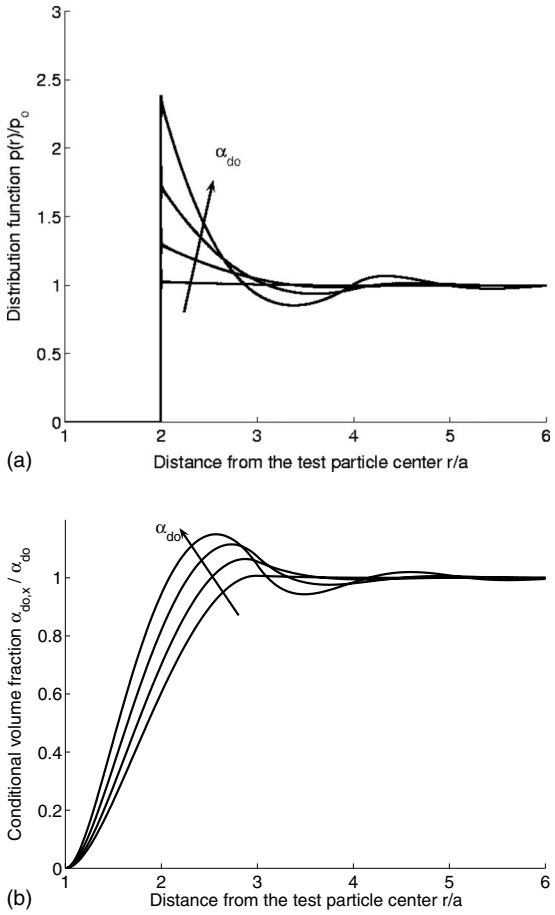


FIG. 1. Evolution of the distribution function and the conditional volume fraction with the distance r/a from the test particle center for volume fraction α_{do} of, respectively, 1%, 10%, 20%, and 30%. These curves are calculated with PY theory.

$$\mathbf{F} = f\langle \mathbf{V}_{\mathbf{c}}, \mathbf{V}_{\mathbf{d}}, \nabla p_c, \nabla \psi, \Delta \mathbf{V}_{\mathbf{c}}, \Delta \mathbf{V}_{\mathbf{d}}, \dots \rangle, \tag{23}$$

$$\mathrm{div}(\mathbf{S}) = s\langle \mathbf{V}_{\mathbf{c}}, \mathbf{V}_{\mathbf{d}}, \nabla p_c, \nabla \psi, \Delta \mathbf{V}_{\mathbf{c}}, \Delta \mathbf{V}_{\mathbf{d}}, \dots \rangle. \tag{24}$$

Since the two-phase equations considered here are linear, f and s must be linear functions of their arguments. Moreover, these functions must depend on the frequency ω to take into account the time derivatives in the Fourier space. For a moving sphere embedded in a pure ambient fluid, the expressions of f and s are well known:[41,42]

$$\mathbf{F} = \alpha_{do}[n_1(\mathbf{V}_{\mathbf{c}} - \mathbf{V}_{\mathbf{d}}) + n_2 \Delta \mathbf{V}_{\mathbf{c}} + n_3 \nabla \Psi], \tag{25}$$

$$\mathrm{div}(\mathbf{S}) = \nabla(\alpha_d p_c) + \alpha_{do} n_o \Delta \mathbf{V}_{\mathbf{c}}, \tag{26}$$

where n_o, n_1, n_2, and n_3 are the coefficients that depend on the pure fluid properties ($\mu_c \rho_{co}$) and on the frequency ω. In these expressions, $n_1(\mathbf{V}_{\mathbf{c}}-\mathbf{V}_{\mathbf{d}})$ corresponds to the sum of the Stokes drag, the Basset hereditary, and the total inertial forces, $n_2 \Delta \mathbf{V}_{\mathbf{c}}$ is the Faxen correction due to the nonuniformity of the ambient fluid velocity, and $n_3 \nabla \Psi$ is due to the acceleration of the test particle.

If we now take into account the influence of the other distributed spheres, the particle is no more embedded in the pure fluid but in an effective medium whose properties are unknown at this stage of the derivation. In this case, the force $\mathbf{F}$ and the stresslet $\mathbf{S}$ will be related to the averaged fields in

the same way but with new coefficients $\tilde{n}_k$ which depend on the effective properties of the surrounding fluid $(\mu_{\text{eff}}, \rho_{\text{eff1}}, \rho_{\text{eff2}})$:

$$\mathbf{F} = \alpha_{do}[\tilde{n}_1(\mathbf{V_c} - \mathbf{V_d}) + \tilde{n}_2 \Delta \mathbf{V_c} + \tilde{n}_3 \nabla \Psi], \tag{27}$$

$$\text{div}(\mathbf{S}) = \nabla(\alpha_d p_c) + \alpha_{do} \tilde{n}_0 \Delta \mathbf{V_c}, \tag{28}$$

where ρ_{eff1} and ρ_{eff2} are some effective densities, respectively, linked to the inertial phenomena and the change of frame of reference, and μ_{eff} is the effective viscosity of the solution. We must underline how important is the hypothesis of homogeneity of the suspension at this stage of the derivation because an inhomogeneous distribution of the particles might induce extra forces related to the gradient of the volume fraction.

The self-consistent condition consists of keeping the *same* coefficients $\tilde{n}_k$ to express $\mathbf{F_x}$ and $\mathbf{S_x}$ in terms of the conditionally averaged fields:

$$\mathbf{F_x} = \alpha_{do,\mathbf{x}}[\tilde{n}_1(\mathbf{V_{c,x}} - \mathbf{V_{d,x}}) + \tilde{n}_2 \Delta \mathbf{V_{c,x}} + \tilde{n}_3 \nabla \Psi], \tag{29}$$

$$\text{div}(\mathbf{S_x}) = \nabla(\alpha_{d,\mathbf{x}} p_{c,\mathbf{x}}) + \alpha_{do,\mathbf{x}} \tilde{n}_0 \Delta \mathbf{V_{c,x}}. \tag{30}$$

Of course, the conditional volume fraction replaces the unconditional one because the distribution of the particles is modified by the presence of the test sphere.

Now, the effective properties of the surrounding fluid must be determined in a consistent way. For that purpose, Eqs. (13)–(15) [with the expressions of **F** and **S** given by Eqs. (27) and (28)] must be properly combined to obtain a final set of equations in the effective medium similar to the equations which would stand in a pure fluid:

$$\text{div}(\mathbf{V_c}) = 0, \tag{31}$$

$$-\rho_{\text{eff1}}(i\omega)\mathbf{V_c} = -\nabla p_c + \mu_{\text{eff}} \Delta \mathbf{V_c} - \rho_{\text{eff2}} \nabla \Psi. \tag{32}$$

As mentioned earlier by the authors,[30] Eqs. (31), (32), and (13)–(15) form a closed system and thus the effective properties can be expressed in terms of the properties of the continuous and dispersed phases, and the coefficients $\tilde{n}_k$ (see Ref. 41 for more details about this calculation):

$$\rho_{\text{eff1}} = \alpha_{co}\rho_{co} + \frac{\alpha_{do}\rho_{do}\tilde{n}_1}{\tilde{n}_1 - i\omega\rho_{do}}, \tag{33}$$

$$\mu_{\text{eff}} = \alpha_{co}\mu_c + \alpha_{do}\tilde{n}_0 + \frac{\alpha_{do}\rho_{do}i\omega\tilde{n}_2 + \alpha_{do}\mu_c(\tilde{n}_1 - i\omega\tilde{n}_2\rho_{\text{eff1}}/\mu_{\text{eff}})}{\tilde{n}_1 - i\omega\rho_{do}}, \tag{34}$$

$$\rho_{\text{eff2}} = \alpha_{co}\rho_{co} + \alpha_{do}\rho_{do}\frac{\tilde{n}_1 - \tilde{n}_3 i\omega}{\tilde{n}_1 - i\omega\rho_{do}}. \tag{35}$$

Our expression of μ_{eff} slightly differs from the expression obtained by Buyevich because of a different choice in the definition of $\tilde{n}_0$, which is more appropriate for our study.

The final step consists of calculating integrals (8) and (9) to express the coefficients $\tilde{n}_k$ in terms of the effective properties of the surrounding fluid. For that purpose, Eqs. (16)–(18) with $\mathbf{F_x}$ and $\mathbf{S_x}$ given by Eqs. (29) and (30) have to be solved.

The solution of Eqs. (31) and (32) for the averaged fields (equivalent to the so-called Brinkman equations) is well known: it was independently solved by Howells[43] for porous media and Buyevich and Markov[35] for the calculation of the force applied on a moving sphere embedded in an unsteady nonuniform velocity field. However, the resolution of the equations for the conditionally averaged fields would be a challenging task because the variation of the conditional volume fraction with the distance r introduces new terms in the equation.

C. Approximation of the conditional volume fraction by a step function

To simplify this calculation, the evolution of the volume fraction obtained with PY theory in Sec. III A will be approximated by a step function:

$$\alpha_{co,\mathbf{x}} = \chi_p + \alpha_{co}\chi_e, \tag{36}$$

$$\alpha_{do,\mathbf{x}} = \alpha_{do}\chi_e, \tag{37}$$

where

$$\chi_p = \begin{cases} 1 & \text{if } a < r < R_c \\ 0 & \text{if } r > R_c \end{cases}, \quad \chi_e = \begin{cases} 0 & \text{if } a < r < R_c \\ 1 & \text{if } r > R_c, \end{cases}$$

and R_c is given by the following formula:

$$\int_{r=a}^{2a} \alpha_{do,\mathbf{x}}(r)d\mathbf{r} + \int_{r=2a}^{\infty} (\alpha_{do,\mathbf{x}}(r) - \alpha_{do})d\mathbf{r} = \int_{R_c<r<2a} \alpha_{do} d\mathbf{r}. \tag{38}$$

With this definition of R_c, the volume occupied by the particles is conserved and also the asymptotic behavior when $r \to \infty$. In this way, a "core-shell" model is obtained (see Fig. 2). The particle is surrounded by a layer of pure fluid which is itself embedded in a homogeneous effective medium. The condition (38) introduced here to calculate the evolution of the "core radius" R_c with the volume fraction (as illustrated by Fig. 3) is equivalent to the one introduced by Spelt *et al.*[29] The only difference is that these authors expressed R_c in terms of the number density, but it does not affect its estimation.

We can now split the conditionally averaged fields into their value in the pure fluid layer and their value in the homogeneous effective medium:

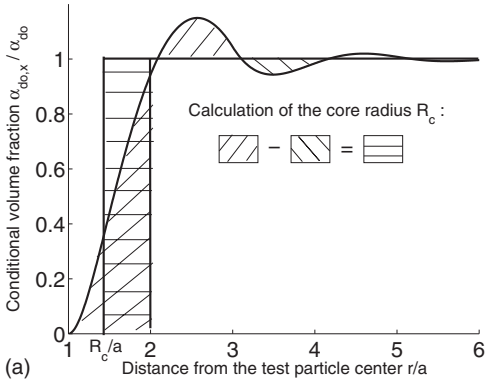
$$\alpha_{co,\mathbf{x}}\mathbf{V_{c,x}} = \chi_p \mathbf{V^p_{c,x}} + \chi_e \alpha_{co} \mathbf{V^e_{c,x}},$$

$$\alpha_{do,\mathbf{x}}\mathbf{V_{d,x}} = \chi_e \alpha_{do} \mathbf{V^e_{d,x}},$$

$$p^e_{c,x} = \chi_p p^p_{c,x} + \chi_e p^e_{c,x},$$

and thus deduce the balance equations in both parts. In the pure fluid layer $(a<r<R)$,

$$\text{div}(\mathbf{V^p_{c,x}}) = 0, \tag{39}$$



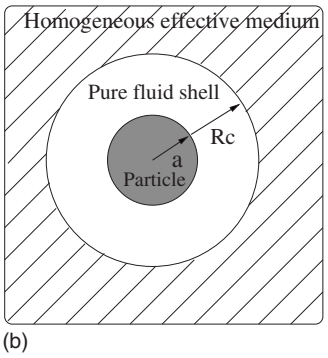


FIG. 2. Approximation of the conditional volume fraction by a step function: core shell model.

$$-\rho_{co}(i\omega)\mathbf{V}^{\mathbf{p}}_{\mathbf{c,x}} = -\nabla p^{p}_{c,x} + \mu_c \Delta \mathbf{V}^{\mathbf{p}}_{\mathbf{c,x}} - \rho_{co}\nabla\Psi. \tag{40}$$

In the homogeneous effective medium $(r>R)$,

$$\mathrm{div}(\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}}) = 0, \tag{41}$$

$$-\rho_{\mathrm{eff1}}(i\omega)\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}} = -\nabla p^{e}_{c,x} + \mu_{\mathrm{eff}}\Delta \mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}} - \rho_{\mathrm{eff2}}\nabla\Psi. \tag{42}$$

The problem is therefore reduced to the study of a particle embedded in a pure fluid shell (with a viscosity μ_c and a density ρ_{co}), which is itself surrounded by a homogeneous effective medium (with effective properties μ_{eff}, ρ_{eff1}, and ρ_{eff2}).

Now, the boundary conditions must be expressed for the different fields. On the particle surface $(r=a)$, the conditionally averaged velocity of the continuous phase is equal to zero that is to say

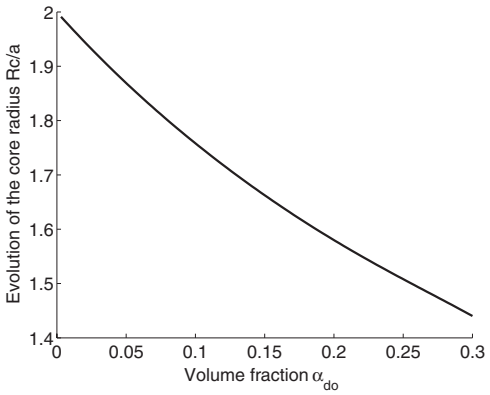


FIG. 3. Evolution of the core radius with the volume fraction.

$$\alpha_{co,\mathbf{x}}\mathbf{V}_{\mathbf{c,x}} = \mathbf{V}^{\mathbf{p}}_{\mathbf{c,x}} = 0.$$

Then, the mass conservation equation for the mean velocity $[\mathrm{div}(\mathbf{V_x})=0]$ imposes the following condition at the core-shell surface $(r=R_c)$:

$$\mathbf{V}^{\mathbf{p}}_{\mathbf{c,x}} = \alpha_{co}\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}} + \alpha_{do}\mathbf{V}^{\mathbf{e}}_{\mathbf{d,x}}, \tag{43}$$

where $\mathbf{V}^{\mathbf{e}}_{\mathbf{d,x}}$ can be expressed in terms of $\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}}$ and $\nabla\Psi$ through Eq. (18).

The momentum conservation gives the following condition in $r=R_c$:

$$\mathbf{\Pi}^{\mathbf{p}}_{\mathbf{x}} \cdot \mathbf{n} = \mathbf{\Pi}^{\mathbf{e}}_{\mathbf{x}} \cdot \mathbf{n}, \tag{44}$$

where $\mathbf{n}$ is the normal vector to the surface of the core and $\mathbf{\Pi}^{\mathbf{p}}_{\mathbf{x}}$ and $\mathbf{\Pi}^{\mathbf{e}}_{\mathbf{x}}$ are, respectively, the strain tensors in the pure layer and in the homogeneous effective medium:

$$\mathbf{\Pi}^{\mathbf{p}}_{\mathbf{x}} = -p^{p}_{c,x} + 2\mu_c\mathbf{D}^{\mathbf{p}}_{\mathbf{c,x}} \quad \text{with } \mathbf{D}^{\mathbf{p}}_{\mathbf{c,x}} = 1/2(\nabla\mathbf{V}^{\mathbf{p}}_{\mathbf{c,x}} + \nabla^t\mathbf{V}^{\mathbf{p}}_{\mathbf{c,x}}),$$

$$\mathbf{\Pi}^{\mathbf{e}}_{\mathbf{x}} = -p^{e}_{c,x} + 2\mu_{\mathrm{eff}}\mathbf{D}^{\mathbf{e}}_{\mathbf{c,x}} \quad \text{with } \mathbf{D}^{\mathbf{e}}_{\mathbf{c,x}} = 1/2(\nabla\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}} + \nabla^t\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}}).$$

Finally, far from the test particle $(r\to\infty)$, the perturbation induced by its presence vanishes so that

$$\{\mathbf{V}^{\mathbf{e}}_{\mathbf{c,x}}, p^{e}_{c,x}\} \to \{\mathbf{V_c}, p_c\}. \tag{45}$$

With Eqs. (31), (32), and (39)–(45), we have derived all the equations and boundary conditions necessary to compute integrals (8) and (9), which reduce to

$$\mathbf{F} = \frac{3\alpha_d}{4\pi a^3}\oint \mathbf{\Pi}^{\mathbf{p}}_{\mathbf{x}} \cdot \mathbf{n} dS, \tag{46}$$

$$\mathbf{S} = \frac{3\alpha_d}{4\pi a^3}\oint^{(s)} \mathbf{a}\otimes(\mathbf{n}\cdot\mathbf{\Pi}^{\mathbf{p}}_{\mathbf{x}})dS. \tag{47}$$

D. Calculation of the closure terms

To solve these equations, the velocity and pressure fields must be expressed in terms of spherical functions according to the method developed by Buyevich and Markov.[42] The details of this calculation can be found in Appendixes A and B. In this section, we only give the final expressions of the coefficient $\tilde{n_k}$ which result from the identification of the expression of $\mathbf{F}$ and $\mathbf{S}$ calculated in the Appendix and formulas (27) and (28).

1. Expression of $\tilde{n}_0$

$$\tilde{n}_0 = 3\frac{\mu_c\beta_c}{\beta_{\mathrm{eff}}}\left[\left(\dot{S}_2(\gamma) + \left(\frac{4}{\gamma}+\frac{\gamma}{2}\right)S_2(\gamma)\right)V_{C_1} + \left(\dot{Q}_2(\gamma) + \left(\frac{4}{\gamma}+\frac{\gamma}{2}\right)Q_2(\gamma)\right)V_{C_2} - \left(\frac{10}{\gamma}+\gamma\right)V_{C_3} - \gamma V_{C_4}\right],$$

where

$$S_2(X) = \frac{(X^2-3X+3)e^X - (X^2+3X+3)e^{-X}}{2X^4},$$

$$Q_2(X) = \frac{(X^2+3X+3)e^{-X}}{X^4},$$

and

$$\epsilon = \beta_{\text{eff}} a, \quad \gamma = \beta_c a,$$

$$\eta = \beta_{\text{eff}} R_c, \quad \delta = \beta_c R_c,$$

$$\beta_c^2 = -(i\omega)\rho_{co}/\mu_c, \quad \beta_{\text{eff}}^2 = -(i\omega)\rho_{\text{eff }1}/\mu_{\text{eff}},$$

$$\kappa = \beta_c \mu_c / \beta_{\text{eff}} \mu_{\text{eff}}.$$

Finally, $V_C = [V_{C_1} V_{C_2} V_{C_3} V_{C_4} V_{C_5} V_{C_6}]^t$ is a column vector whose expression is given by

$$V_C = M_2^{-1} V_{A2}, \tag{48}$$

where the expressions of M_2 and V_{A2} are given in Appendix C.

2. Expression of $\tilde{n}_1$, $\tilde{n}_2$, and $\tilde{n}_3$

$$\tilde{n}_1 = -\mu_c \beta_c^2 [S_1(\gamma) V_{E_1} + Q_1(\gamma) V_{E_2} - V_{E_3} - V_{E_4}], \tag{49}$$

$$\tilde{n}_2 = \frac{\mu_c \beta_c^2}{\beta_{\text{eff}}^2} [S_1(\gamma)(3V_{D_1} + V_{E_1}) + Q_1(\gamma)(3V_{D_2} + V_{E_2}) - (3V_{D_3} + V_{E_3}) - (3V_{D_4} + V_{E_4})], \tag{50}$$

$$\tilde{n}_3 = \rho_{co} + \mu_c \beta_c^2 [S_1(\gamma) V_{F_1} + Q_1(\gamma) V_{F_2} - V_{F_3} - V_{F_4}], \tag{51}$$

where

$$S_1(X) = \frac{(X-1)e^X + (X+1)e^{-X}}{2X^3},$$

$$Q_1(X) = \frac{(X+1)e^{-X}}{X^3},$$

and V_D, V_E, and V_F are some column vectors whose expressions are given by

$$V_D = M_1^{-1} V_{A1}, \tag{52}$$

$$V_E = M_1^{-1} V_{M1}, \tag{53}$$

$$V_F = M_1^{-1} V_{P1}, \tag{54}$$

where the expression of M_1, V_{A1}, V_{M1}, and V_{P1} are given in Appendix C.

With these expressions, the coefficients $\tilde{n_k}$ are related to the effective parameters μ_{eff}, $\rho_{\text{eff}1}$, and $\rho_{\text{eff}2}$, which are themselves related to the coefficients $\tilde{n_k}$ through Eqs. (33)–(35). Thus, the system is closed and the coefficients $\tilde{n_k}$ can be calculated either with a simple iterative procedure or with more elaborate numerical schemes such as the "globally convergent Newton's method."

IV. RESULTS AND COMPARISON WITH EXPERIMENTS

A. Final dispersion equation

Then, the expression of the force (27) and the stresslet (28) can be substituted in the linearized system (1)–(7), and the dispersion equation can be derived for a plane wave by considering fields of the form: $G = G_o + \bar{G} e^{i(k_* x - \omega t)}$, where $\bar{G}$ is the amplitude of the wave, G_o the equilibrium state, and k_* the complex effective wave number. As calculated in a previous paper,[30] the effective wave number k_* is the solution of the following quadratic equation:

$$Ak_*^4 + Bk_*^2 + C = 0,$$

$$A = d_r h_c \left[\frac{(\lambda_c + 2\mu_c)}{\rho_{do} i\omega} + \frac{r c_{co}^2}{\alpha_{co}\omega^2} \right],$$

$$B = -d_r \left[h_c + \frac{N_0^* + (\lambda_c + 2\mu_c)(\alpha_{co} + \alpha_{do} h_v)/\alpha_{do}\rho_{do}}{i\omega} \right] - \frac{c_{co}^2}{\alpha_{co}\omega^2} \frac{[1 + d_r r h_v]}{[1 + \alpha_{do}\xi_d/\alpha_{co}\xi_c]}, \tag{55}$$

$$C = 1 + d_r h_v,$$

where

$$h_v = \frac{N_1^*}{N_1^* + i\omega(N_3^* - 1)}, \quad h_c = \frac{N_2^*}{N_1^* + i\omega(N_3^* - 1)}$$

and

$$N_k^* = \frac{n_k^*}{\rho_{do}}, \quad d_r = \frac{\alpha_{do}\rho_{do}}{\alpha_{co}\rho_{co}}, \quad r = \frac{\rho_{co}}{\rho_{do}}.$$

We can note that the coefficient $\alpha_{do}\rho_{do}$ was missing in the expression of B in our previous paper (just in the manuscript, the good expression had been considered for the computation). We can also note that a new coefficient $[1 + \alpha_{do}\xi_d/\alpha_{co}\xi_c]$ (with ξ_k the compressibility of phase k) appears, as we have taken into account the compressibility of the dispersed phase in Eq. (2).

B. Comparison of the different theories with experiments

Now, the results of this corrected effective medium theory (that take into account spatial correlations) can be compared to previous results[30] obtained with the same theory but without spatial correlations (that is to say when a homogeneous effective medium is considered around the test particle) and also with the classical coupled phase theory (when the test particle is supposed to be surrounded by the pure continuous phase). For that purpose, we will consider the experiments performed by Hipp *et al.*[33] who measured the attenuation of acoustical waves in solutions of silica particle in water for different concentrations, frequencies, and particle sizes. Before analyzing these curves, let us recall some elements which will be useful to understand the influence of spatial correlations. When an acoustical wave propagates through a solution of rigid particles, the particles do not

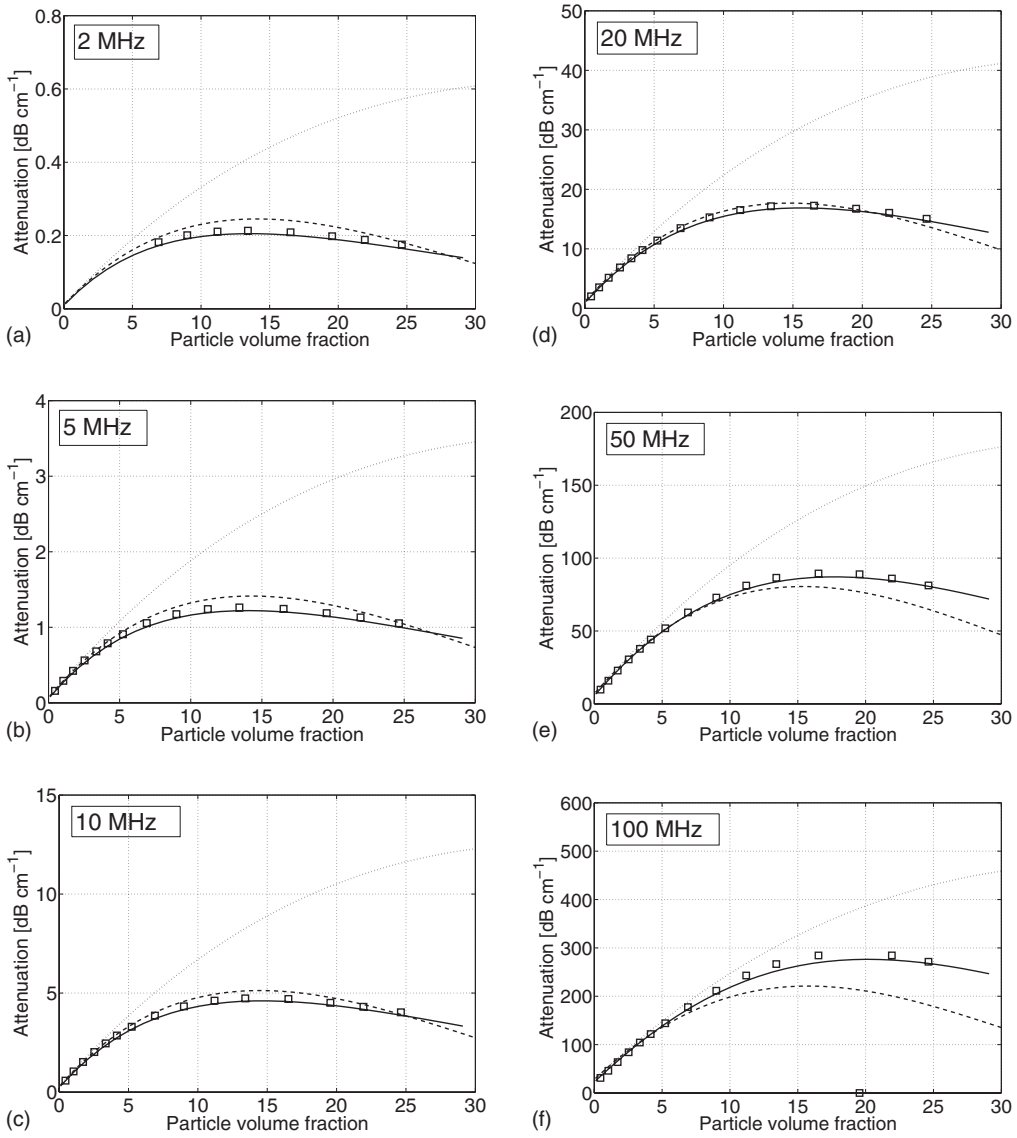


FIG. 4. Attenuation as a function of the volume fraction at various frequencies for silica particles of 56 nm radius in water. The solid lines (-) correspond to the new effective theory, the broken lines (–) to the effective theory without spatial correlationss, the dotted line (..) to the classical coupled phase theory, and the symbols to the experimental data.

move with the same velocity as the surrounding fluid because of the difference of density between them and the surrounding fluid. As a consequence, a dipolar wave is scattered and a part of the energy of the impinging wave is therefore redirected, inducing a loss of spatial coherence. Some energy is also dissipated because of the viscosity of the surrounding fluid which slows down the movement of the particle and therefore converts energy from the compressional propagation mode to a viscous lossy mode. The combination between these two scattering mechanisms is referred to as the "viscoinertial" phenomena. The characteristic length for the decrease of the viscous lossy wave is the size of the boundary layer δ_v which is related to the frequency ω according to the following formula:

$$\delta_v = \sqrt{\frac{2\mu_c}{\omega\rho_c}}.$$

In concentrated suspensions, viscous interactions between neighboring particles may appear according to their concentration and the frequency of the propagating wave. As long as $\delta_v \ll R_c - a$, the properties of the fluid in the boundary layer are very close to the properties of the surrounding pure fluid (because $\alpha_{do,\mathbf{x}} \approx 0$ in this area) and thus the force and the stresselet can be estimated by considering a particle embedded in the pure liquid (as it is done in the classical coupled phase theory). However, when $\delta_v \gg R_c - a$, the variation of the effective properties due to the spatial correlations only concerns a thin part of the boundary layer and thus the approximation of the surrounding fluid by a homogeneous effective medium for the calculation of the closure terms (as it was done in a previous paper[30]) should give good results. As $R_c - a \approx a$, the transition between these two limiting case should happen when $\delta_v \approx a$. For the two suspensions considered here, the corresponding characteristic frequencies are, respectively, of $f_c = 101$ MHz for Fig. 4 and $f_c = 11$ MHz for Fig. 5. Finally, as δ_v is inversely proportional to the frequency, the condition $\delta_v \ll a$ correspond to high frequencies $f \gg f_c$ and the condition $\delta_v \gg a$ to low frequencies $f \ll f_c$.

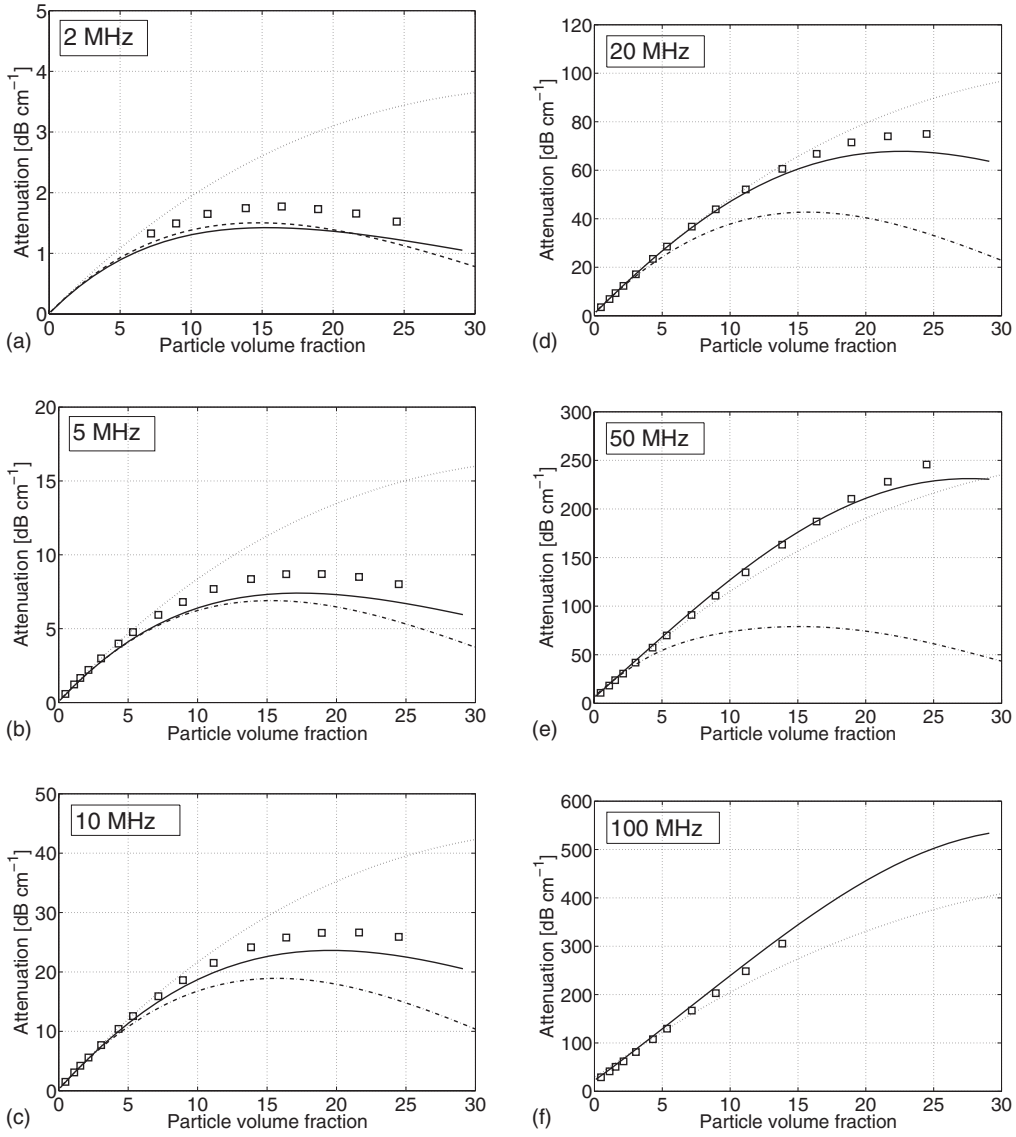


FIG. 5. Attenuation as a function of the volume fraction at various frequencies for silica particles of 164.5 nm radius in water. The solid lines (-) correspond to the new effective theory, the broken lines (–) to the effective theory without spatial correlationss, the dotted line (..) to the classical coupled phase theory, and the symbols to the experimental data.

Effectively, we can see on the different figures that for frequencies $f \ll f_c$, the experimental data are close to the results obtained with the homogeneous effective medium[30] (dash dotted lines), whereas for frequencies $f \gg f_c$, the results are closer to the curves obtained with the classical coupled phase theory (dotted line). As a consequence, none of these theories can properly describe the sound propagation in concentrated suspensions for a wide range of frequencies and particle sizes. For high frequencies and particle concentration, the homogeneous effective theory even gives unphysical effective parameters; that is why we have not plotted the corresponding curves (see Fig. 5 at 100 MHz). With the introduction of spatial correlations, we obtain a model (solid lines) which gives good results for both limiting cases. As expected, the results are less accurate for the transition ($f \approx f_c$) because the progressive evolution of the conditional volume fraction has been approximated by a step function. We can note at this point that equivalent results should be obtained with the model of Spelt *et al.*[29] based on the ECAH[4,5] decomposition. Of course, for low volume fraction ($\alpha_{do} < 5\%$), all these theories give the same results because in this case, the effective properties of the surrounding fluid are close to the properties of the pure fluid. Concerning the remaining discrepancies between our theory and the experiments, different phenomena might explain them such as the polydispersity of the solution or collisions between neighboring particles. Another possible effect might be the increase of the importance of thermal effects in concentrated suspension. It is well known that in dilute suspensions of silica particles in water, viscoinertial effects are more important than thermal ones[44,45] because the first ones are proportional to $(1-r)=O(1)$ in these suspensions and thermal effects to $(\gamma-1) \ll 1$ (where γ is the specific heat ratio). However, viscous interactions make the attenuation induced by viscoinertial effects decrease. In the same way, there will also be some thermal interactions due to the overlapping of viscous boundary layer, which will make the attenuation induced by this scattering mechanism decrease. However, as the viscous boundary layer δ_v is usually larger than the thermal boundary

layer δ_t for aqueous solutions (see, for example, Hipp *et al.*[46]), viscoinertial interactions will be more important than thermal interactions and therefore thermal attenuation might become more significant in concentrated suspensions. It would be interesting to investigate this question in a future work.

To conclude this discussion, we can notice that unlike interactions of compressional waves, viscous coupling tends to diminish the attenuation induced by the suspension when the concentration increases, which is quite unusual.

V. CONCLUSION

An effective medium coupled phase theory has been derived to properly describe the sound propagation in concentrated suspensions of rigid particles. An excellent agreement is obtained between this theory and the experimental data of Hipp *et al.* who measured the attenuation induced by the presence of silica particles in water for different particle sizes, concentration up to 30%, and frequencies between 2 and 100 MHz. Moreover, the influence of spatial correlations on the propagation in solutions of rigid particles has been clearly identified. This theory could be improved by considering the exact evolution of the conditional volume fraction with the distance from the test particle center instead of the core-shell aproximation. It could be also extended to polydisperse suspensions but it would require the knowledge of the distribution function in polydisperse suspensions which is not an easy matter.[47] Finally, we can underline that the expressions obtained here for the force and the stresslet could also be used for hydrodynamic studies of concentrated suspensions of particles, as long as the characteristic macroscopic length of the flow is much larger than the size of the particles, and as long as the flow does not modify the distribution of the particles.

ACKNOWLEDGMENTS

The authors would like to thank A. K. Hipp, G. Storti, and M. Morbidelli (Department of Chemical Engineering, ETH Zürich) for kindly providing us with the results of their experiments and D. Lhuillier (Institut Jean le Rond D'Alembert, CNRS) for numerous and fruitful discussions.

APPENDIX A: RESOLUTION OF THE SYSTEM

First, we can subtract Eqs. (31) and (32) for the averaged fields from Eqs. (41) and (42) for the conditionally averaged fields in the region $r > R_c$ to obtain equations for the perturbation due to the presence of the test sphere:

$$\mathrm{div}(\mathbf{V_c^{e*}}) = 0, \tag{A1}$$

$$-\rho_{\mathrm{eff}\,1}(i\omega)\mathbf{V_c^{e*}} = -\nabla p_c^{e*} + \mu_{\mathrm{eff}}\Delta\mathbf{V_c^{e*}}, \tag{A2}$$

where

$$\mathbf{V_c^{e*}} = \mathbf{V_{c,x}^{e}} - \mathbf{V_c},$$

$$p_c^{e*} = p_{c,x}^{e} - p_c.$$

Then, the boundary conditions can easily be rewritten in terms of the perturbation fields:

$$\mathbf{V_{c,x}^{p}} = 0 \quad \text{in } r = a \tag{A3}$$

$$\mathbf{V_{c,x}^{p}} - \alpha_{co}\mathbf{V_c^{e*}} - \alpha_{do}\mathbf{V_d^{e*}} = \alpha_{co}\mathbf{V_c} + \alpha_{do}\mathbf{V_d} \quad \text{in } r = R_c, \tag{A4}$$

$$\mathbf{\Pi_x^p} \cdot \mathbf{n} - \mathbf{\Pi^{e*}} \cdot \mathbf{n} = \mathbf{\Pi_c} \cdot \mathbf{n} \quad \text{in } r = R_c, \tag{A5}$$

$$\{\mathbf{V_c^{e*}}, p_c^{e*}\} \to 0 \quad \text{when } r \to \infty. \tag{A6}$$

Equations (31), (32), (39), (40), (A1), and (A2) can all be written under the form

$$\mathrm{div}(\mathbf{U}) = 0, \tag{A7}$$

$$(\Delta - \beta^2)\mathbf{U} = \nabla R, \tag{A8}$$

where

$$\{U = \mathbf{V_c}, R = 1/\mu_{\mathrm{eff}}(p_c + \rho_{\mathrm{eff}\,2}\psi), \beta = \beta_{\mathrm{eff}}\} \quad \text{for the first set of equations,}$$

$$\{U = \mathbf{V_{c,x}^{p}}, R = 1/\mu_c(p_{c,x}^{p} + \rho_{co}\psi), \beta = \beta_c\} \quad \text{for the second,}$$

$$\{U = \mathbf{V_c^{e*}}, R = 1/\mu_{\mathrm{eff}}(p_c^{e*}), \beta = \beta_{\mathrm{eff}}\} \text{ for the third one.}$$

Then the velocity and pressure fields can be expressed in terms of spherical functions:

$$U(\mathbf{r}) = \sum_{k=0}^{\infty}\left[F_k(r)s_k(\theta,\phi)\frac{\mathbf{r}}{r} + G_k(r)r\nabla s_k(\theta,\phi) + H_k(r)\mathbf{r} \times \nabla s_k(\theta,\phi)\right], \tag{A9}$$

$$R(\mathbf{r}) = \sum_{k=0}^{\infty} L_k(r)s_k(\theta,\phi), \tag{A10}$$

where r, θ, and ϕ are the spherical coordinates. $F_k s_k$ denotes the summation

$$F_k s_k = F_k^0(r)P_k(\cos(\theta)) + \sum_{k'=1}^{k}[F_{k+}^{k'}(r)P_k^{k'}(\cos(\theta))\cos(k'\phi) \tag{A11}$$

$$+ F_{k-}^{k'}(r)P_k^{k'}(\cos(\theta))\sin(k'\phi)], \tag{A12}$$

and P_k and $P_k^{k'}$ are, respectively, the principal and associated Legendre functions. Of course, $G_k s_k$ and $H_k s_k$ denote similar summations.

As the inertial terms due to the rotation of the particle are neglected $H_k=0$ and because of the symmetry of the problem (which is invariant by any rotation of ϕ angle), only the principal Legendre functions are required:

$$F_k s_k = F_k^0(r) P_k(\cos(\theta)), \tag{A13}$$

$$G_k s_k = G_k^0(r) P_k(\cos(\theta)). \tag{A14}$$

Then, by replacing the preceding expressions of $\mathbf{U}$ and R in the conservation Eqs. (A7) and (A8), Buyevich and Markov[42] obtain the following expressions for F_k^0 and G_k^0:

$$F_k^0 = A_k S_k(\xi) + B_k Q_k(\xi) - k M_k \left(\frac{\xi}{\beta}\right)^{k-1} + (k+1) N_k \left(\frac{\xi}{\beta}\right)^{-k-2}, \tag{A15}$$

$$G_k^0 = \frac{2}{k(k+1)} A_k \left(S_k(\xi) + \frac{\xi}{2}\dot{S}_k(\xi)\right) + \frac{2}{k(k+1)} B_k \left(Q_k(\xi) + \frac{\xi}{2}\dot{Q}_k(\xi)\right) - M_k \left(\frac{\xi}{\beta}\right)^{k-1} - N_k \left(\frac{\xi}{\beta}\right)^{-k-2}, \tag{A16}$$

where the coefficients A_k, B_k, M_k, and N_k are some constants which must be determined from the boundary conditions for each order k; $\dot{S}_k$ and $\dot{Q}_k$ are the derivatives of S_k and Q_k with respect to ξ; and the expressions of S_k and Q_k are given by the following expressions:

$$S_k = 2^k \xi^{k-1} \frac{d^k}{d(\xi^2)^k} \frac{\sinh \xi}{\xi}, \tag{A17}$$

$$Q_k = (-2)^k \xi^{k-1} \frac{d^k}{d(\xi^2)^k} \frac{\exp(-\xi)}{\xi}, \tag{A18}$$

with $\xi \equiv \beta r$.

We can now introduce the following notations: the constants $\{A_k, B_k, M_k, N_k\}$ are, respectively, equal to the following:

- $\{a_k, b_k, m_k, n_k\}$ for the first set of equations (for the average fields in the original frame of reference),
- $\{a_k^p, b_k^p, m_k^p, n_k^p\}$ for the second set of equations (for the conditionally averaged field in the pure fluid shell), and
- $\{a_k^e, b_k^e, m_k^e, n_k^e\}$ for the third set of equations (for the perturbation in the effective homogeneous medium).

We can easily deduce from the boundary conditions that: (a) b_k and n_k are null because the averaged fields are bounded when $r \to 0$, and (b) a_k^e and m_k^e are null because the perturbation vanishes when $r \to \infty$.

Now, there only remains to express the relations between the remaining coefficients, deduced from the boundary conditions (A3)–(A6). At first order ($k=1$), we obtain

$$M_1 V_1 = a_1 V_{A1} + m_1 V_{M1} - (i\omega)\, v_d|_{r=0} V_{P1} \tag{A19}$$

and at second order ($k=2$)

$$M_2 V_2 = a_2 V_{A2} + m_2 V_{M2}, \tag{A20}$$

where the expression of the column vectors V_{A1}, V_{M1}, V_{A2}, and V_{M2} and the matrices M_1 and M_2 are given in Appendix C and the expressions of V_1 and V_2 are given by

$$V_1 = \begin{bmatrix} a_1^p \\ b_1^p \\ m_1^p \\ n_1^p/a^3 \\ b_1^e \\ n_1^e/R_c^3 \end{bmatrix}, \quad V_2 = \begin{bmatrix} a_2^p \\ b_2^p \\ am_2^p \\ n_2^p/a^4 \\ b_2^e \\ n_2^e/R_c^4 \end{bmatrix}.$$

APPENDIX B: CALCULATION OF THE FORCE AND THE STRESSLET

The next step consists of expressing the force $\mathbf{F}$ and the stresslet $\mathbf{S}$ in terms of the coefficients a_k^p, b_k^p, m_k^p, and n_k^p. It is important to note that only the coefficients of first order ($k=1$) are required for the calculation of the force and the coefficients of second order ($k=2$) for the calculation of the stresslet because the contribution of the other terms vanishes when the integration over the surface of the sphere is performed. The following expressions are obtained:

$$\mathbf{F} = \alpha_{do}\mu_c\beta_c^2 \left[S_1(\gamma) a_1^p + Q_1(\gamma) b_1^p - m_1^p - \frac{n_1^p}{a^3} \right] \mathbf{e_z} + \alpha_{do}\rho_{co} \nabla \Psi,$$

$$\mathbf{S} = \frac{\alpha_{do}\mu_c\beta_c}{5} (3\mathbf{e_z} \otimes \mathbf{e_z} - \mathbf{I}) \times \left[\left(\dot{S}_2(\gamma) + \left(\frac{4}{\gamma} + \frac{\gamma}{2} \right) S_2(\gamma) \right) a_2^p + \left(\dot{Q}_2(\gamma) + \left(\frac{4}{\gamma} + \frac{\gamma}{2} \right) Q_2(\gamma) \right) b_2^p - \left(\frac{10}{\gamma} + \gamma \right) am_2^p - \gamma \frac{n_2^p}{a^4} \right].$$

The final step consists of expressing the coefficients a_1, m_1, a_2, and m_2 in terms of the averaged fields in $r=0$:

$$a_1 \mathbf{e_z} = 3/\beta_{\text{eff}}^2 \, \Delta \mathbf{V_c}|_{r=0},$$

$$m_1 \mathbf{e_z} = 1/\beta_{\text{eff}}^2 \, \Delta \mathbf{V_c}|_{r=0} - \mathbf{V_c}|_{r=0},$$

$$a_2 (3\mathbf{e_z} \otimes \mathbf{e_z} - \mathbf{I}) = 30/\beta_{\text{eff}}^2 \, \nabla^s \Delta \mathbf{V_c}|_{r=0},$$

$$m_2 (3\mathbf{e_z} \otimes \mathbf{e_z} - \mathbf{I}) = - \nabla^s \mathbf{V_c}|_{r=0} + 1/\beta_{\text{eff}}^2 \, \nabla^s \Delta \mathbf{V_c}|_{r=0}.$$

Finally, with the relation $\Delta^2 \mathbf{V_c} = \beta_{\text{eff}}^2 \Delta \mathbf{V_c}$ [which can be easily deduced from Eqs. (31) and (32)] and by comparing the above expressions with Eqs. (27) and (28), we obtain the expression of the coefficients $\tilde{n}_k$.

APPENDIX C: EXPRESSION OF THE MATRICES M_1, M_2 AND THE VECTORS V_{A1}, V_{M1}, V_{P1}, V_{A2}, V_{M2}

$$M_1 = \begin{bmatrix} S_1(\gamma) & Q_1(\gamma) & -1 & 2 & 0 & 0 \\ S_1(\gamma)+\gamma/2\dot{S}_1(\gamma) & Q_1(\gamma)+\gamma/2\dot{Q}_1(\gamma) & -1 & -1 & 0 & 0 \\ S_1(\delta) & Q_1(\delta) & -1 & 2(a/R_c)^3 & -\alpha_A Q_1(\eta) & -2\alpha_B \\ S_1(\delta)+(\delta/2)\dot{S}_1(\delta) & Q_1(\delta)+(\delta/2)\dot{Q}_1(\delta) & -1 & -(a/R_c)^3 & -\alpha_A(Q_1(\eta)+\eta/2\dot{Q}_1(\eta)) & \alpha_B \\ 2\kappa\dot{S}_1(\delta) & 2\kappa\dot{Q}_1(\delta) & -\kappa\delta & -\kappa(a/R_c)^3(\delta+12/\delta) & -2\dot{Q}_1(\eta) & \eta+12/\eta \\ \kappa(\delta/2\ddot{S}_1(\delta)+\dot{S}_1(\delta)) & \kappa(\delta/2\ddot{Q}_1(\delta)+\dot{Q}_1(\delta)) & 0 & 6\kappa a^3/\delta R_c^3 & -(\eta/2\ddot{Q}_1(\eta)+\dot{Q}_1(\eta)) & -6/\eta \end{bmatrix},$$

$$M_2 = \begin{bmatrix} S_2(\gamma) & Q_2(\gamma) & -2 & 3 & 0 & 0 \\ 1/3(S_2(\gamma)+\gamma/2\dot{S}_2(\gamma)) & 1/3(Q_2(\gamma)+(\gamma/2)\dot{Q}_2(\gamma)) & -1 & -1 & 0 & 0 \\ S_2(\delta) & Q_2(\delta) & -2R_c/a & 3(a/R_c)^4 & -\alpha_A Q_2(\eta) & -3\alpha_B \\ 1/3(S_2(\delta)+(\delta/2)\dot{S}_2(\delta)) & 1/3(Q_2(\delta)+(\delta/2)\dot{Q}_2(\delta)) & -R_c/a & -(a/R_c)^4 & -\alpha_A/3(Q_2(\eta)+(\eta/2)\dot{Q}_2(\eta)) & \alpha_B \\ 2\kappa\dot{S}_2(\delta) & 2\kappa\dot{Q}_2(\delta) & -\kappa R_c/a(\delta+4/\delta) & -\kappa a^4/R_c^4(\delta+24/\delta) & -2\dot{Q}_2(\eta) & (\eta+24/\eta) \\ \kappa((4/\delta+\delta/2)S_2(\delta)-\dot{S}_2(\delta)) & \kappa((4/\delta+\delta/2)Q_2(\delta)-\dot{Q}_2(\delta)) & -6\kappa R_c/a\delta & 24\kappa a^4/\delta R_c^4 & \dot{Q}_2(\eta)-(4/\eta+\eta/2)Q_2(\eta) & -24/\eta \end{bmatrix},$$

$$V_{A1} = \begin{bmatrix} 0 \\ 0 \\ \alpha_A S_1(\eta) \\ \alpha_A(S_1(\eta)+\eta/2\dot{S}_1(\eta)) \\ 2\dot{S}_1(\eta) \\ \dot{S}_1(\eta)+\eta/2\ddot{S}_1(\eta) \end{bmatrix}, \quad V_{M1} = \begin{bmatrix} 0 \\ 0 \\ -\alpha_B \\ -\alpha_B \\ -\eta \\ 0 \end{bmatrix}, \quad V_{P1} = \begin{bmatrix} 0 \\ 0 \\ \alpha_C \\ \alpha_C \\ (\rho_{\text{eff}2}-\rho_{co})R_c/\mu_{\text{eff}}\beta_{\text{eff}} \\ 0 \end{bmatrix},$$

$$V_{A2} = \begin{bmatrix} 0 \\ 0 \\ \alpha_A S_2(\eta) \\ \alpha_A/3(S_2(\eta)+\eta/2\dot{S}_2(\eta)) \\ 2\dot{S}_2(\eta) \\ (4/\eta+\eta/2)S_2(\eta)-\dot{S}_2(\eta) \end{bmatrix}, \quad V_{M2} = \begin{bmatrix} 0 \\ 0 \\ -2\alpha_B R_c \\ -2\alpha_B R_c \\ -(4/\eta+\eta/2)R_c \\ -\dfrac{6}{\eta}R_c \end{bmatrix},$$

with

$$\alpha_A = \alpha_{co} + \alpha_{do}(c_1+c_2), \quad \alpha_B = \alpha_{co} + \alpha_{do}c_1, \quad \alpha_C = \alpha_{do}\frac{\tilde{n}_3 - \rho_{do}}{\tilde{n}_1 - i\omega\rho_{do}},$$

$$c_1 = \frac{\tilde{n}_1}{\tilde{n}_1 - i\omega\rho_{do}}, \quad c_2 = \frac{\tilde{n}_2\beta_{\text{eff}}^2}{\tilde{n}_1 - i\omega\rho_{do}}.$$

[1] R. E. Challis, M. J. W. Povey, M. L. Mather, and A. K. Holmes, "Ultrasound techniques for characterizing colloidal dispersions," Rep. Prog. Phys. **68**, 1541–1637 (2005).

[2] A. S. Dukhin and P. J. Goetz, "Acoustic and electroacoustic spectroscopy for characterizing concentrated dispersions and emulsions," Adv. Colloid Interface Sci. **92**, 73–132 (2001).

[3] E. L. Carstensen and L. L. Foldy, "Propagation of sound through a liquid containing bubbles," J. Acoust. Soc. Am. **19**, 481–501 (1947).

[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] P. S. Epstein and R. R. Carhart, "The absorption of sound in suspensions and emulsions. I. Waterfog in air," J. Acoust. Soc. Am. **25**, 553–565 (1953).

[6] J. M. Evans and K. Attenborough, "Coupled phase theory for sound propagation in emulsions," J. Acoust. Soc. Am. **102**, 278–282 (1997).

[7] D. A. Gubaidullin, and R. I. Nigmatulin, "On the theory of acoustic waves in polydispersed gaz-vapor-droplet suspensions," Int. J. Multiphase Flow **26**, 207–228 (2000).

[8] N. A. Gumerov, A. I. Ivandaev, and R. I. Nigmatulin, "Sound waves in monodisperse gas-particle or vapour-droplet mixtures," J. Fluid Mech. **193**, 53–74 (1988).

[9] R. Duraiswami and A. Prosperetti, "Linear pressure wave in fogs," J. Fluid Mech. **299**, 187–215 (1995).

[10] N. A. Gumerov and R. Duraiswami, "Computation of scattering from N spheres using multipole reexpansion," J. Acoust. Soc. Am. **112**(6), 2688–2701 (2002).

[11] N. A. Gumerov and R. Duraiswami, "Computation of scattering from clusters of spheres using fast multipole method," J. Acoust. Soc. Am. **117**,

1744–1761 (2005).
[12]A. J. C. Ladd, "Hydrodynamic interactions in a suspension of spherical particles," J. Chem. Phys. **88**(8), 5051–5063 (1988).
[13]A. J. C. Ladd, "Hydrodynamic interactions and the viscosity of suspensions of freely moving spheres," J. Chem. Phys. **90**, 1149–1157 (1989).
[14]A. J. C. Ladd, "Hydrodynamic transport coefficients of random dispersions of hard spheres," J. Chem. Phys. **93**, 3484–3494 (1990).
[15]G. Mo and A. S. Sangani, "A method for computing Stokes flow interactions among spherical objects and its application to suspensions of drops and porous particles," Phys. Fluids **6**, 1637–1652 (1994).
[16]L. L. Foldy, "The multiple scattering of waves. I. General theory of isotropic scattering by randomly distributed scatterers," Phys. Rev. **67**, 107–119 (1945).
[17]E. J. Hinch, "An averaged-equation approach to particle interactions in a fluid suspension," J. Fluid Mech. **83**, 695–720 (1977).
[18]F. S. Henyey, "Corrections to Foldy's effective medium theory for propagation in bubble clouds and other collections of very small scatterers," J. Acoust. Soc. Am. **105**(4), 2149–2153 (1999).
[19]A. S. Sangani, "A pairwise interaction theory for determining the linear acoustic properties of dilute bubbly liquids," J. Fluid Mech. **232**, 221–284 (1991).
[20]Z. Ye and L. Ding, "Acoustic dispersion and attenuation relations in bubbly mixture," J. Acoust. Soc. Am. **98**, 1629–1636 (1995).
[21]S. G. Kargl, "Effective medium approach to linear acoustics in bubbly liquids," J. Acoust. Soc. Am. **111**, 168–173 (2002).
[22]S. K. Kanaun, V. M. Levin, and F. J. Sabina, "Propagation of elastic waves in composites with random set of spherical inclusions," (effective medium approach), Wave Motion **40**, 69–88 (2004).
[23]T. S. Lundgren, "Slow flow through stationary random beds and suspensions of spheres," J. Fluid Mech. **51**, 273–299 (1972).
[24]Yu. A. Buyevich and I. N. Shchelchkova, "Flow of dense suspensions," Prog. Aerosp. Sci. **18**, 121–151 (1978).
[25]E. Chang, B. S. Yendler, and A. Acrivos, *Advances in Multiphase Media*, edited by G. Papanicolaou (SIAM, Philadelphia, PA, 1986).
[26]A. S. Sangani and C. Yao, "Transport processes in random arrays of cylinders. i thermal conduction," Phys. Fluids **31**, 2426–2444 (1988).
[27]A. S. Sangani and C. Yao, "Transport processes in random arrays of cylinders. II viscous flow," Phys. Fluids **31**, 2435–2444 (1998).
[28]T. L. Dodd, D. A. Hammer, A. S. Sangani, and D. L. Koch, "Numerical simulations of the effect of hydrodynamic interactions on diffusivities of integral membrane proteins," J. Fluid Mech. **293**, 147–180 (1995).
[29]P. D. M. Spelt, M. A. Norato, A. S. Sangani, M. S. Greenwood, and L. L. Tavlarides, "Attenuation of sound in concentrated suspensions, theory and experiments," J. Fluid Mech. **430**, 51–86 (2001).
[30]M. Baudoin, J. L. Thomas, F. Coulouvrat, and D. Lhuillier, "An extended coupled phase theory for the sound propagation in polydisperse suspensions of rigid particles," J. Acoust. Soc. Am. **121**, 3386–3397 (2007).
[31]J. K. Percus and G. J. Yevick, "Analysis of classical statistical mechanics by means of collective coordinates," Phys. Rev. **110**, 1–13 (1957).
[32]M. S. Wertheim, "Exact solution of the Percus-Yevick integral equation for hard spheres," Phys. Rev. Lett. **10**, 312–323 (1963).
[33]A. K. Hipp, G. Storti, and M. Morbidelli, "Acoustic characterization of concentrated suspensions and emulsions. 2. Experimental validation," Langmuir **18**, 391–404 (2002).
[34]R. I. Nigmatulin and J. C. Friedly, *Dynamics of Multiphase Media* (Hemisphere, Washington, 1991), Vols. **1** and **2**.
[35]Yu. A. Buyevich and V. G. Markov, "Continual mechanics of monodisperse suspensions, rheological equations of state for suspensions of moderate concentration," Prikl. Mat. Mekh. **37**, 1059–1077 (1973).
[36]T. Morita and K. Hiroike, "A new approach to the theory of classical fluids," Prog. Theor. Phys. **23**, 1003–1027 (1960).
[37]J. M. J. Van Leeuwen, J. Groeneveld, and J. De Boer, "New method for the calculation of the pair correlation function," Physica (Amsterdam) **25**, 792–808 (1959).
[38]Yu. A. Buyevich, "Heat and mass transfer in disperse media. II. Constitutive equations," Int. J. Heat Mass Transfer **35**(10), 2453–2463 (1992).
[39]E. Chang and A. Acrivos, "Rate of heat conduction from a heated sphere to a matrix containing passive spheres of a different conductivity," J. Appl. Phys. **59**(10), 3375–3382 (1986).
[40]Yu. A. Buyevich, "Heat and mass transfer in disperse media. I. Averaged field equations," Int. J. Heat Mass Transfer **35**(10), 2445–2452 (1992).
[41]Yu. A. Buyevich, "Interphase interaction in fine suspension flow," Chem. Eng. Sci. **50**, 641–650 (1995).
[42]Yu. A. Buyevich and V. G. Markov, "Rheology of concentrated mixtures of fluids with small particles," Prikl. Mat. Mekh. **36**, 480–493 (1972).
[43]I. D. Howells, "Drag due to the motion of a Newtonian fluid through a sparse random array of small fixed rigid objects," J. Fluid Mech. **64**, 449–475 (1974).
[44]A. K. Hipp, G. Storti, and M. Morbidelli, "Acoustic characterization of concentrated suspensions and emulsions. 1. Model analysis," Langmuir **18**, 391–404 (2002).
[45]S. Temkin, "Attenuation and dispersion of sound in dilute suspensions of spherical particles," J. Acoust. Soc. Am. **108**, 126–145 (2000).
[46]A. K. Hipp, G. Storti, and M. Morbidelli, "On multiple-particle effects in the acoustic characterization of colloidal dispersions," J. Phys. D **32**, 568–576 (1999).
[47]H. Hansen-Goos and R. Roth, "A new generalization of the Carnahan–Starling equation of state to additive mixture of hard spheres," J. Chem. Phys. **124**, 1–8 (2006).

Caustic and anticaustic points in the phonon focusing patterns of cubic crystals

Litian Wang[a)]
Østfold University College, 1757 Halden, Østfold 1757 Norway

(Received 14 November 2007; revised 14 February 2008; accepted 10 March 2008)

Phonon focusing patterns are dependent on the existence of concave/saddle regions and acoustic axes in the slowness surface. The main feature of the focusing patterns in cubic crystals can be characterized by the caustic and anticaustic points in the symmetry planes. By applying the Stroh formalism, the caustic and anticaustic points in the symmetry planes are investigated in relation to degeneracies in the Stroh eigenvalue equation. A set of analytical expressions for the locations of the caustic and anticaustic points is derived for cubic crystals. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2903874]

PACS number(s): 43.35.Gk, 43.20.Bi [RLW] Pages: 4140–4146

I. INTRODUCTION

Acoustic wave propagation in anisotropic media is governed by the Christoffel equation.[1,2] The equation yields three-sheeted slowness surface with many fascinating features. The three slowness sheets can intersect with each other along the so-called acoustic axes, and the outer and medial sheets can be concave locally while the inner sheet is globally convex. By employing the concept of Gaussian curvature, the two outer slowness sheets can be divided into concave/saddle and convex regions separated by the so-called parabolic line with zero Gaussian curvature.[3]

The group velocity, associated with a point on a slowness sheet, was shown parallel to the surface normal to the slowness sheet at the point. The group velocities can in turn be described by the so-called wave surface.[1,2] Geometric features of the wave surface can also be described by the so-called Gauss map which is defined by the normals of a closed surface.[4] The most important features of the wave surface are the caustics and anticaustics. The caustics (or cusps) on the wave surface result from the folds associated with the local concavities on the slowness surface, while the anticaustics result from the conical acoustic axis and they form an anticaustic cone.[4–7]

Because of the presence of concave regions on the slowness sheets, the energy flow associated with a wave package will exhibit a focusing effect. The boundaries of the focusing pattern will be determined by the parabolic lines and the anticaustic cone.[3] In the long wavelength limit ($k \rightarrow 0$), the focusing pattern is identical to the polar projection of the Gauss map of the slowness surface. The phonon focusing patterns have been well documented in the experimental and theoretical studies.[6–11]

Theoretically, a set of critical conditions for the presence of various types of caustics in cubic crystals was elegantly formulated by calculating group velocities in the vicinity of symmetry directions.[6,7] The Gaussian curvature in the vicinity of symmetry axes in the crystals of various symmetries was also studied in order to formulate critical conditions for some caustics.[12–14] But the analysis of Gaussian curvature failed to produce analytical results for the caustic points because of geometric complexity of the slowness surface. Numerically, extensive simulations have been done to illustrate the focusing patterns and to determine the size of focusing patterns in cubic crystals using the Monte Carlo method.[11] Although the simulations gave a good phenomenological description of the focusing patterns, there remains a minor discrepancy compared with theoretical analysis[6] concerning the focusing patterns in the vicinity of the {100} directions.

In elastodynamics,[15–17] the existence of the caustic points can be understood as a degeneracy problem in the two-dimensional elastodynamics. In the symmetry planes, the three slowness branches are defined in such manner that one of them (S_b) is pure elliptic and associated with the transverse polarized bulk waves, while the two other branches (S_a, S_c) are nonelliptic and associated with quasi-transverse and quasi-longitudinal waves. The surface normal associated with each branch can be determined analytically by solving the Stroh eigenvalue equation, which defines the wave surface because the normal for a point in the slowness branch is then identical to the surface normal due to the symmetry.[18,19] The caustic point (the cuspidal point in the wave surface) is then a direct result of a triplication point in the slowness branch where the tangent to the point is associated with three coalescing points in its neighborhood (see Fig. 1). The anticaustic cone, however, can be partially described by the two curve normals associated with the two outer slowness branches at the acoustic axis.[19,20]

In this paper, we will apply the Stroh formalism for two-dimensional elastodynamics to the symmetry planes in cubic crystals. Both the caustic and anticaustic points are investigated, and their locations in the phonon focusing patterns are given as roots of simple and soluble equations.

[a)]Electronic mail: litian.wang@hiof.no

 0001-4966/2008/123(6)/4140/7/$23.00

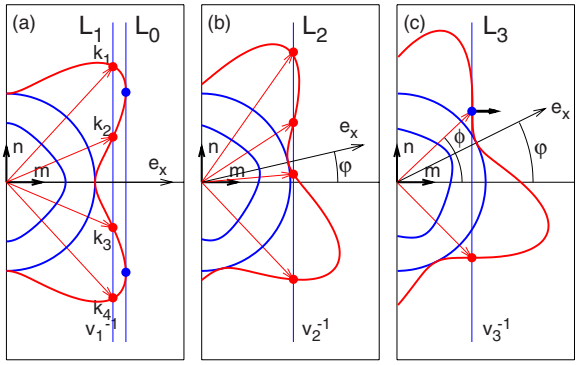


FIG. 1. (Color online) Interpretation of the triple degeneracy in the Stroh formalism and the triplication point. The inclination angle ϕ is the angle between the wavevector $\mathbf{k}$ and the vector $\mathbf{m}$.

II. STROH FORMALISM AND DEGENERACIES

A. Stroh formalism

Consider a two-dimensional infinite homogeneous elastic medium. The equation of motion for steady states is given by

$$C_{ijkl}\frac{\partial^2 u_k}{\partial x_j \partial x_l} = \rho \frac{\partial^2 u_i}{\partial t^2}. \tag{1}$$

For a steady state solution defined by $\mathbf{u}(x)=\mathbf{a}\exp(ikz)$, where $z=(\mathbf{m}+p\mathbf{n})\cdot\mathbf{x}-vt$, with the wave vector $\mathbf{k}=k(\mathbf{m}+p\mathbf{n})$, the polarization vector $\mathbf{a}$ obeys

$$\Gamma\mathbf{a} \equiv \{(mm) - \rho v^2 I + p[(mn)+(nm)] + p^2(nn)\}\mathbf{a} = 0, \tag{2}$$

where the term (mn) etc. is given by $(mn)_{ik}=m_j c_{ijkl} n_l$, and the repeated subscripts denote summation. The plane spanned by $(\mathbf{m},\mathbf{n})$ is referred as the reference plane. The traction $\mathbf{t}(\mathbf{x})$ exerted on the planes parallel to $\mathbf{n}\cdot\mathbf{x}=0$ will be given by $t_i(\mathbf{x})=\sigma_{ij}n_j=c_{ijkl}\partial u_k/\partial x_l n_j$ and $\mathbf{t}(\mathbf{x})=\mathbf{b}\exp(ikz)$, and $\mathbf{b}$ can then be expressed in terms of $\mathbf{a}$

$$\mathbf{b} = -[(nm)+p(nn)]\mathbf{a}. \tag{3}$$

The Stroh formalism[15] is based on the construction of a six-dimensional vector $\xi=[\mathbf{a},\mathbf{b}]^T$ so that Eqs. (2) and (3) can be merged into a six-dimensional eigenvalue equation

$$N\xi_\alpha = p_\alpha \xi_\alpha, \quad \alpha = 1\ldots 6, \tag{4}$$

where

$$N = -\begin{pmatrix} (nn)^{-1}(nm) & (nn)^{-1} \\ (mn)(nn)^{-1}(nm) - (mm) + \rho v^2 I & (mn)(nn)^{-1} \end{pmatrix}$$

and I is a 3 by 3 identity matrix. When the velocity is sufficiently low, the eigenvalues appear as three pairs of complex conjugated numbers. The Stroh eigenvalue p_α can be determined by the characteristic equation for Eqs. (2) or (4), namely, a sextic equation in p:

$$\sum_{n=0}^{6} a_n(v)p^n = 0. \tag{5}$$

The Stroh eigenvalue p, as a function of velocity v and orientations of the basis vectors $\mathbf{m}$ and $\mathbf{n}$, has a very simple geometrical interpretation.[16,17] Consider a situation where the reference plane $(\mathbf{m},\mathbf{n})$ is defined in such a way that the basis vector $\mathbf{m}$ is tilted from the crystal axis $\mathbf{e}_x$ with an angle φ as shown in Fig. 1. The solutions for the eigenvalue Eqs. (2), $p(v,\varphi)$, can be interpreted by drawing a vertical line L normal to $\mathbf{m}$ at the velocity v (or slowness v^{-1}). We start with $\varphi=0$. Letting $v=v_1$ [Fig. 1(a)], the vertical line L_1 can be drawn and it intersects the outer slowness curve at four points. The four points represent four bulk-wave solutions with wave vectors $\mathbf{k}_i \| (\mathbf{m}+p_i\mathbf{n})$ and phase velocities $v/\cos\phi_i$ $(i=1,\ldots,4)$, where ϕ_i are the angles between the wavevectors $\mathbf{k}_i$ and the basis vector $\mathbf{m}$. The Stroh eigenvalue is just the tangent of the inclination angle: $p_i(v,\varphi)=\tan\phi_i$. In this case there are four real eigenvalues and two complex ones. If we reduce the velocity v so that the vertical line moves to L_0 and it tangentially touches the slowness curve, there will be two simple duplex degeneracy: $p_1=p_2$ and $p_3=p_4$, and such a situation is usually referred as a Type 4 transonic state,[16] at which four Stroh eigenvalues coalesce pairwise. By increasing the tilting angle φ [Fig. 1(b)], the upper three intersection points approach toward the inflection point. Eventually, with proper tilting angle φ and velocity v, the vertical line will tangentially intersect the inflection point [see line L_3 in Fig. 1(c)]. Under such a circumstance, there will be a triple degeneracy among three real Stroh eigenvalues, $p_1=p_2=p_3=\tan\phi$. The velocity v_3 represents then the group velocity associated with the inflection point and the basis vector $\mathbf{m}$ defines the surface normal at the triplication point. Such an inflection point can be found in the {100} and {110} planes referred as points 3 and 2 in Fig. 8 in Ref. 6.

Analytically, when the reference plane $(\mathbf{m},\mathbf{n})$ is defined within a symmetry plane, both the Γ matrix (2) and its characteristic equation can be decomposed into two parts, i.e.,

$$\begin{pmatrix} \Gamma_{11} & \Gamma_{12} & 0 \\ \Gamma_{12} & \Gamma_{22} & 0 \\ 0 & 0 & \Gamma_{33} \end{pmatrix}, \tag{6}$$

$$(a_4p^4 + a_3p^3 + a_2p^2 + a_1p + a_0)(b_2p^2 + b_1p + b_0) = 0 \tag{7}$$

and the Stroh eigenvalues can then be partitioned into (p_1,p_2,p_3,p_4) and (p_5,p_6) correspondingly. The two-dimensional part represents in-plane polarized waves, while the one-dimensional part represents ex-plane polarized (transverse) waves. Note that in the symmetry plane, only the slowness branch related to the in-plane polarized wave can be concave. This leads to the condition for the existence of the triplication point:

$$p_1 = p_2 = p_3, \tag{8}$$

which in turn determines the caustic point.

The anticaustics in the focusing pattern of cubic crystals result from the conical acoustic axis along the [111] direction (see Fig. 2). The surface normals in the vicinity of the [111] direction constitute an anticaustic cone with its axis along the acoustic axis.[1] In the Gauss map and the phonon focusing pattern, the cone is projected into a circle centered at the [111] direction. In the $(0\bar{1}1)$ plane, we will have two anticaustic points located symmetrically with respect to the [111]

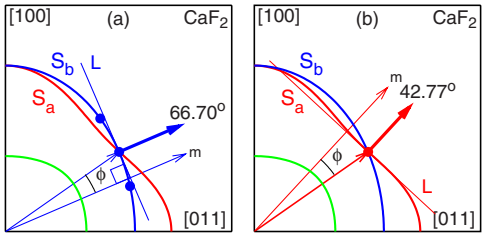


FIG. 2. (Color online) Formation of the anticaustic circle about [111] direction in CaF_2 with respect to the $(0\bar{1}1)$ plane. The basis vector **m** defines the direction of group velocity (the thick arrow) at the acoustic axis with respect to (a) the elliptic sheet S_b and (b) the nonelliptic sheet S_a.

direction. Figure 2 shows how these two anticaustic points are originated from the surface normal consideration in CaF_2. For example, the surface normal with the angle 66.70° from $\mathbf{e}_x$ is associated with a transonic state,[16] where the line L [Fig. 2(a)] tangentially touches the elliptic sheet S_b. This implies a degeneracy of two Stroh eigenvalues $p_5=p_6$ where p_5 and p_6 are depicted by the two intersection points before the vertical line approaches L in Fig. 2(a). Since the line also intersects the nonelliptic sheet S_a leaving a real eigenvalue p_1, we have then a new triple degeneracy $p_1=p_5=p_6$. The similar condition can also be reached for the surface normal pertaining to S_a with the angle 42.78° and it is given by $p_1=p_2=p_5$ [Fig. 2(b)]. Because these two anticaustic points are symmetrical with respect to the [111] direction, we choose

$$p_1 = p_5 = p_6 \tag{9}$$

as a condition for the existence of the anticaustic point.

There are many types of degeneracies among six eigenvalues in the Stroh eigenvalue equation. Analysis of these degeneracies has mainly been done in the study of surface waves in anisotropic media,[17] but the significance of the triple degeneracies has not been fully recognized. In this work, we will show that investigation of triple degeneracies enables us to resolve the caustic and anticaustic points in symmetry planes in simple fashion.

B. Degeneracy conditions

Since the condition for the existence of the caustic point refers to $p_1=p_2=p_3$, which are repeated roots of the in-plane part of the characteristic Eq. (7), the condition can then be derived by evaluating the coefficients of the quartic equation. Let us consider a monic quartic equation Q_4. A triple degeneracy implies a common root among Q_4 and its first and second derivatives:

$$\begin{aligned} Q_4 &= x^4 + \alpha x^3 + \beta x^2 + \gamma x + \delta = 0, \\ Q_4' &= 4x^3 + 3\alpha x^2 + 2\beta x + \gamma = 0, \\ Q_4'' &= 12x^2 + 6\alpha x + 2\beta = 0. \end{aligned} \tag{10}$$

The common root among these equations requires that the resultants between them must vanish.[21] Moreover, for Q_4 and Q_4', the common root is of the second order, the first subresultant for Q_4 and Q_4' must then vanish as well. The vanishing resultants above lead to two independent conditions as follows:

$$\begin{aligned} R_1 &= \beta^2 - 3\alpha\gamma + 12\delta = 0 \\ R_2 &= 9\alpha^2\delta - \alpha\beta\gamma + 9\gamma^2 - 32\beta\delta = 0. \end{aligned} \tag{11}$$

The condition for the anticaustic point, on the other hand, refers to $p_5=p_6=p_1$, where p_5 and p_6 are roots of the ex-plane part of the characteristic Eq. (7). Therefore, its existence condition involves both the quadratic equation (which defines p_5, p_6) and the quartic equation (which yields p_1). Specifically, the triple degeneracy $p_5=p_6=p_1$ requires existence of a common root among

$$\begin{aligned} Q_2 &= x^2 + \kappa x + \sigma = 0, \\ Q_2' &= 2x + \kappa = 0, \\ Q_4 &= x^4 + \alpha x^3 + \beta x^2 + \gamma x + \delta = 0, \end{aligned} \tag{12}$$

which leads also to two independent conditions:

$$\begin{aligned} R_5 &= 4\sigma - \kappa^2 = 0, \\ R_6 &= \kappa^4 - 2\alpha\kappa^3 + 4\beta\kappa^2 - 8\gamma\kappa + 16\delta = 0. \end{aligned} \tag{13}$$

In the following section, we will explicitly define the reference plane $(\mathbf{m},\mathbf{n})$ within the symmetry planes in cubic crystals so that the Γ matrix (2) can be substantiated. The characteristic equations and their coefficients can then be derived before the conditions (11) and (13) are applied.

III. CAUSTIC POINTS AND ANTICAUSTIC POINTS IN CUBIC CRYSTALS

In the cubic crystals, there are two sets of principal symmetry planes: {100} and {110}. The caustic points can be located in both sets of planes while the anticaustic point only in the {110} planes in the vicinity of the ⟨111⟩ directions. We will first examine the caustic point in the (001) plane and then caustic/anticaustic points in the $(0\bar{1}1)$ plane. Since the slowness surface and phonon focusing patterns are uniquely determined by the relative elastic constants $a=c_{11}/c_{44}$ and $b=c_{12}/c_{44}$ and the anisotropy factor $\Delta=a-b-2$, we will carry out analysis in terms of a and b.

A. Caustic point in (001) plane

In the (001) plane, we define the reference plane $(\mathbf{m},\mathbf{n})$ with $\mathbf{m}=\cos\varphi\mathbf{e}_x+\sin\varphi\mathbf{e}_y$ and $\mathbf{n}=-\sin\varphi\mathbf{e}_x+\cos\varphi\mathbf{e}_y$. The Γ matrix (2) is then given by the following elements:

$$\begin{aligned} \Gamma_{11} &= p^2(\cos^2\varphi + a\sin^2\varphi) + p(1-a)\sin 2\varphi + a\cos^2\varphi \\ &\quad + \sin^2\varphi - \rho v^2, \\ \Gamma_{12} &= -\frac{1}{2}(b+1)(p^2\sin 2\varphi - 2p\cos 2\varphi - \sin 2\varphi), \\ \Gamma_{22} &= p^2(a\cos^2\varphi + \sin^2\varphi) - p(1-a)\sin 2\varphi + \cos^2\varphi \\ &\quad + a\sin^2\varphi - \rho v^2, \end{aligned}$$

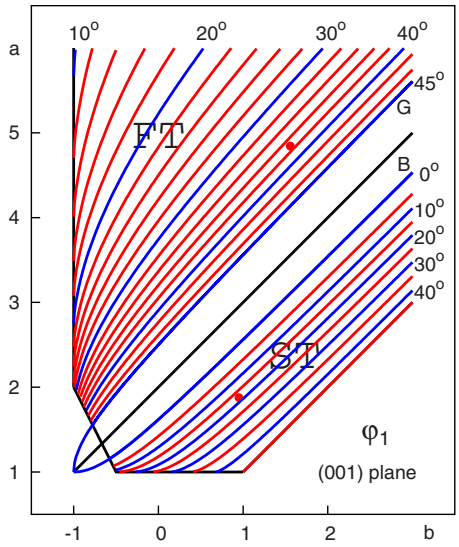


FIG. 3. (Color online) Variation of the caustic points $\hat{\varphi}_1$ in the (001) plane. The two filled circles mark two cubic crystals: CaF_2 (upper) and GaAs (lower).

$$\Gamma_{33} = p^2 - \rho v^2 + 1. \tag{14}$$

The characteristic equation can then be decomposed into an in-plane part ($\Gamma_{11}\Gamma_{22}-\Gamma_{12}^2=0$) and an ex-plane part ($\Gamma_{33}=0$). The in-plane part yields a quartic equation

$$a_4 p^4 + a_3 p^3 + a_2 p^2 + a_1 p + a_0 = 0, \tag{15}$$

which the triple degeneracy refers to.

By converting Eq. (15) into a monic equation and substituting its coefficients into the conditions for the caustic point (11), one obtains two simultaneous equations:

$$R_1(a,b,\varphi,\rho v^2) = 0, \quad R_2(a,b,\varphi,\rho v^2) = 0. \tag{16}$$

The resultant for R_1 and R_2 with respect to ρv^2 will leave a cubic trigonometric equation

$$f(a,b,\hat{\varphi}_1) = \sum_{n=0}^{3} c_n(a,b)\cos^n 4\hat{\varphi}_1 = 0, \tag{17}$$

where $c_n(a,b)$ are polynomials of a and b. The root $\hat{\varphi}_1$ defines then the caustic point within $[0,\pi/4]$ in the (001) plane, or the point $3'$ shown in Fig. 8 in Ref. 6. Another caustic point with $\hat{\varphi}=\pi/2-\hat{\varphi}_1$ can also be identified because of the symmetry. For the sake of conciseness, the pseudocode for deriving $f(a,b,\hat{\varphi}_1)$ is given in the Appendix .

Figure 3 illustrates the variation of the caustic point $\hat{\varphi}_1$ as a function of a and b. One can recognize that the caustic point lies in the vicinity of the [100] direction for crystals with $\Delta<0$ and in the vicinity of the [110] direction for crystals with $\Delta>0$, respectively.

The condition $f(a,b,\hat{\varphi}_1)=0$ verifies also the critical existence conditions for the caustic point in the phonon focusing patterns. By setting $\hat{\varphi}_1=0$ and $\hat{\varphi}_1=\pi/4$, we will have

$$f|_{\hat{\varphi}_1=0} \sim \{a[(a-1)a-(b+1)^2](a+b^2+2b)\}^3$$

$$f|_{\hat{\varphi}_1=\pi/4} \sim [(a-b)(a+b+2)(2a^2+ab-3a-b^2-b+2)(2a^2-ab-5a-b^2-3b)]^3, \tag{18}$$

respectively, which implies $a(a-1)-(b+1)^2=0$ and $\Delta(2a+b-1)-2(b+1)=0$, and they are represented by the curves B and G in the $a-b$ plot, respectively, consistent with Ref. 6.

B. Caustic points in $(0\bar{1}1)$ plane

Concerning the caustic/anticaustic points in the $(0\bar{1}1)$ plane, we can rotate the coordinate system by $\pi/4$ about $\mathbf{e}_x$ axis so that a parallel discussion can be carried out in terms of the transformed elastic constants in the new coordinate system with $c_{22}=c_{33}=(a+b+2)/2$, $c_{23}=(a+b-2)/2$ and $c_{44}=(a-b)/2$ while the rest remains unchanged.

The Γ-matrix (2) for the $(0\bar{1}1)$ plane is then given by

$$\begin{aligned}
\Gamma_{11} &= p^2(\cos^2\varphi + a\sin^2\varphi) + (1-a)p\sin 2\varphi - (\rho v^2 - a\cos^2\varphi + \sin^2\varphi)\\
\Gamma_{12} &= -\tfrac{1}{2}(b+1)(p^2\sin 2\varphi - 2p\cos 2\varphi - \sin 2\varphi)\\
\Gamma_{22} &= \tfrac{1}{4}\{p^2[2(a+b)\cos^2\varphi + 4] + 2(a+b)p\sin 2\varphi + [2(a+b)\sin^2\varphi - 4(\rho v^2 - 1)]\}\\
\Gamma_{33} &= \tfrac{1}{2}\{p^2[\Delta\cos^2\varphi + 2] + p\Delta\sin 2\varphi + [\Delta\sin^2\varphi - 2(\rho v^2 - 1)]\}
\end{aligned} \tag{19}$$

and the characteristic equation can again be expressed by an in-plane (quartic) part and an ex-plane (quadratic) part. By converting the quartic part into a monic equation and substituting its coefficients into the conditions for the caustic point (11), we get again two simultaneous equations $R_3(a,b,\varphi,\rho v^2)=0$ and $R_4(a,b,\varphi,\rho v^2)=0$. The resultant of R_3 and R_4 removes ρv^2 and leaves again another trigonometric equation (see Appendix):

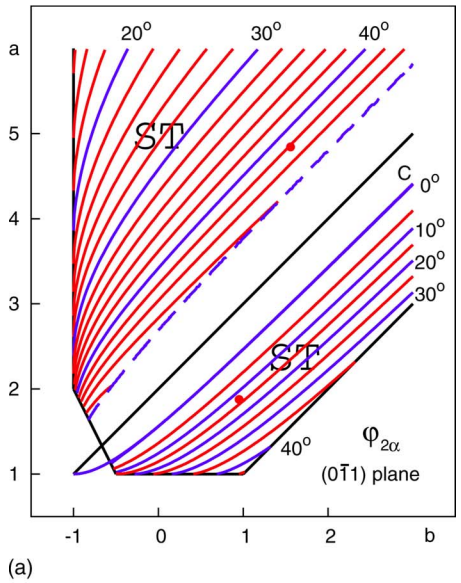
$$g(a,b,\hat{\varphi}_2) = \sum_{n=0}^{6} d_n(a,b)\cos^n 2\hat{\varphi}_2 = 0, \tag{20}$$

where $d_n(a,b)$ are polynomials of a and b. The solution $\hat{\varphi}_2$ defines the caustic point $\hat{\varphi}_2$ within $[0,\pi/2]$ in the $(0\bar{1}1)$ plane, or the point $2'$ shown in Fig. 8 in Ref. 6.

Figure 4 illustrates the variation of the caustic points $\hat{\varphi}_2$ as function of a and b. The solution for $\hat{\varphi}_2$ always appears in pair: One ($\hat{\varphi}_{2\alpha}$) is in the neighborhood of the [001] direction, and the other ($\hat{\varphi}_{2\beta}$) is near the [011] direction.

The condition $g(a,b,\hat{\varphi}_2)=0$ verifies again the critical existence conditions for the caustics in the phonon focusing patterns. By setting $\hat{\varphi}_2=0$ and $\hat{\varphi}_2=\pi/2$, we obtain

$$g|_{\hat{\varphi}_2=0} \sim [(a+b+2)(a+a^2-5b+ab-2b^2-4)\times(a+2b+b^2)]^3$$



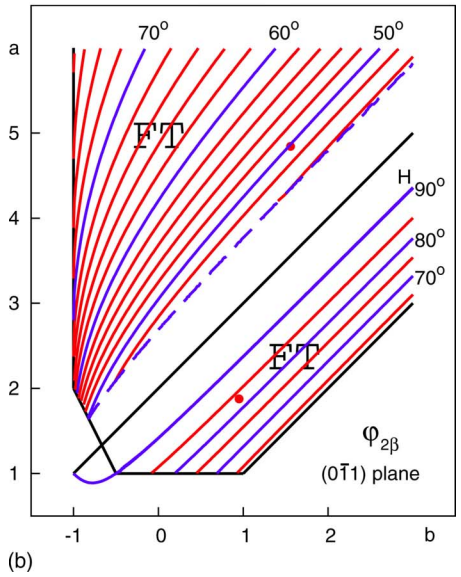


FIG. 4. (Color online) Variation of the caustic points (a) $\hat{\varphi}_{2\alpha}$ and (b) $\hat{\varphi}_{2\beta}$ in the $(0\bar{1}1)$ plane. The dash curve marks the limit along which $\hat{\varphi}_{2\alpha}=\hat{\varphi}_{2\beta}$.

$$g|_{\hat{\varphi}_2=\pi/2} \sim \{a[a(a+b)-2(b+1)^2](a+5b+2b^2+2)\}^3, \tag{21}$$

respectively, which means $(a-1)(a+b+2)-2(b+1)^2=0$ and $a(a+b)-2(b+1)^2=0$, and they are represented by the curves C and H in the $a-b$ plot, respectively, also consistent with Ref. 6.

C. Anticaustic point in $(0\bar{1}1)$ plane

The anticaustic points are originated from the acoustic axis between the outer nonelliptic sheet and the elliptic sheet and they always appear in pairs. Since the condition involves coefficients of both the quartic equation $(\Gamma_{11}\Gamma_{22}-\Gamma_{12}^2=0)$ and the quadratic equation $(\Gamma_{33}=0)$, we will convert them both into monic equations and set their coefficients into the conditions for the anticaustic point (13), i.e., $R_5(a,b,\varphi,\rho v^2)=0$ and $R_6(a,b,\varphi,\rho v^2)=0$. By calculating the resultant of R_5 and R_6, we can remove ρv^2 and get a simple trigonometric equation (see Appendix):

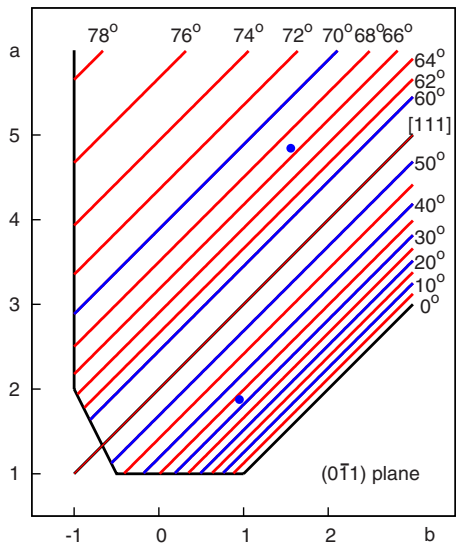


FIG. 5. (Color online) Variation of the anticaustic point $\hat{\varphi}_x$ in the $(0\bar{1}1)$ plane.

$$h(a,b,\varphi_x) = -(b+1)\Delta[(a-b)\cos^2\varphi_x + 2\sin^2\varphi_x]^2 \times [(a-b)^2\cos^2\varphi_x - 2\sin^2\varphi_x]\sin^2\varphi_x = 0. \tag{22}$$

The nontrivial solution for Eq. (22) will define the anticaustic point $\hat{\varphi}_x$ as follows:

$$\hat{\varphi}_x = \tan^{-1}\frac{a-b}{\sqrt{2}}. \tag{23}$$

The dependence of $\hat{\varphi}_x$ upon a and b is shown in Fig. 5. The symmetric anticaustic point with respect to the [111] direction can be determined from $2\varphi_{[111]}-\hat{\varphi}_x$, where $\varphi_{[111]}=\cos^{-1}(1/\sqrt{3})$, and the radius of the anticaustic circle is simply given by $\Delta_\varphi=|\varphi_{[111]}-\hat{\varphi}_x|$, or

$$\Delta_\varphi = \tan^{-1}\frac{|\Delta|}{\sqrt{2}(\Delta+3)}, \tag{24}$$

consistent with the earlier study.[1]

It should be noted that the analysis above produces simple and explicit results (17, 20, 23) without solving either the Christoffel equation or the Stroh eigenvalue equation. The critical existence conditions for the caustic points agree with the earlier study, and the scheme can readily be applied to crystals with lower symmetries.

IV. EXAMPLES AND DISCUSSIONS

It is well known that the cubic crystals can be grouped into two classes according to their signs of the anisotropic parameter Δ. Such a classification plays a minor role in the present scheme.

In the present discussion we take CaF_2 (with $\Delta>0$) and GaAs (with $\Delta<0$) as demonstration examples. Table I and Figs. 6 and 7 summarize the calculation results and the pho-

TABLE I. Calculation results for CaF_2 and GaAs. Elastic constants (c_{11}, c_{12}, and c_{44} in GPa) used in calculations are: CaF_2 (1.74, 0.56, 0.3593) and GaAs (1.126, 0.571, 0.6).

Planes	{001}	{011}		
	$\hat{\varphi}_1$	$\hat{\varphi}_{2\alpha}$	$\hat{\varphi}_{2\beta}$	$\hat{\varphi}_z$
CaF_2 ($\Delta>0$)	37.20	42.22	49.78	66.70
GaAs ($\Delta<0$)	18.07	13.11	83.06	33.19

non focusing patterns and they agree with each other. The patterns are simulated by calculating Gauss map using the Monte Carlo method with 10^5 points. In order to enhance the anticaustic circle, the random values $\theta \in [0,\pi]$ and $\varphi \in [0,2\pi]$ in the spherical coordinate system $(1,\theta,\varphi)$ are uniformly distributed.

An extensive numerical simulation is carried out and it agrees with the analytical results. Compared to earlier works,[6,11] it is noted that there is a major difference between $\hat{\varphi}_1$ (Fig. 3) and β_{1-} by Hurley and Wolfe [see Fig. 2(c) in

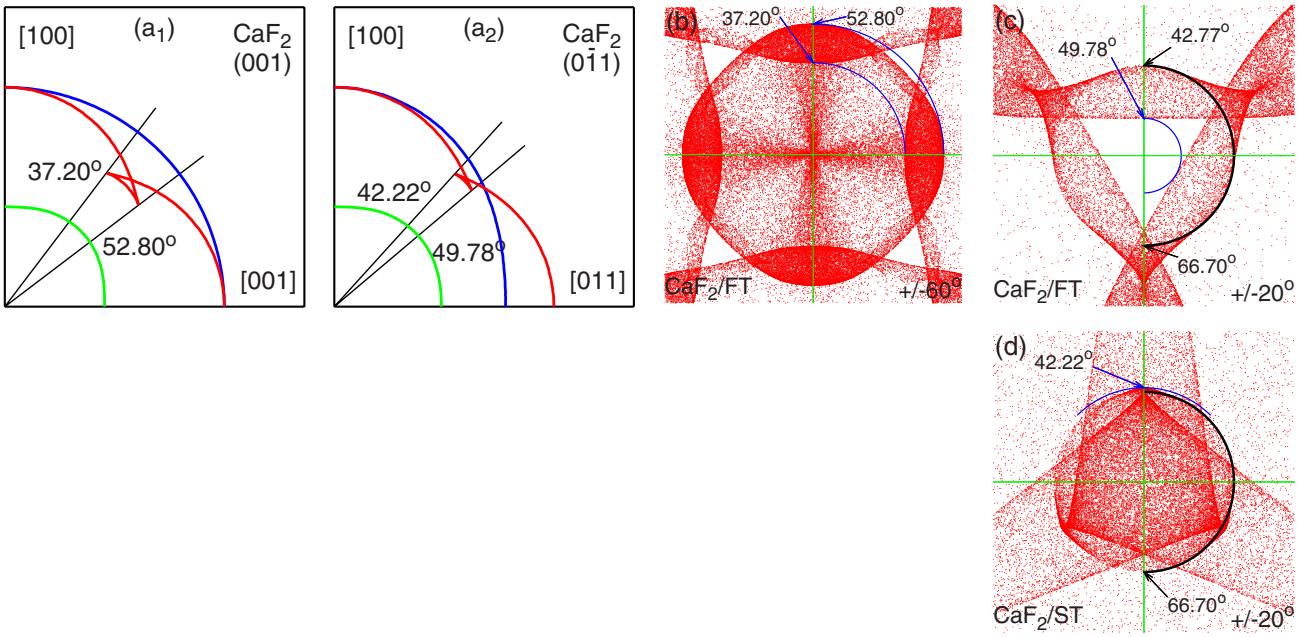


FIG. 6. (Color online) (a) Wave surface sections (slowness plot) in the (001) and (0$\bar{1}$1) planes for CaF_2. The phonon focusing patterns: (b) FT mode along [100] direction, (c) FT mode along [111] direction, and (d) ST mode along [111] direction. The thick half circles partially mark the anticaustic circles.

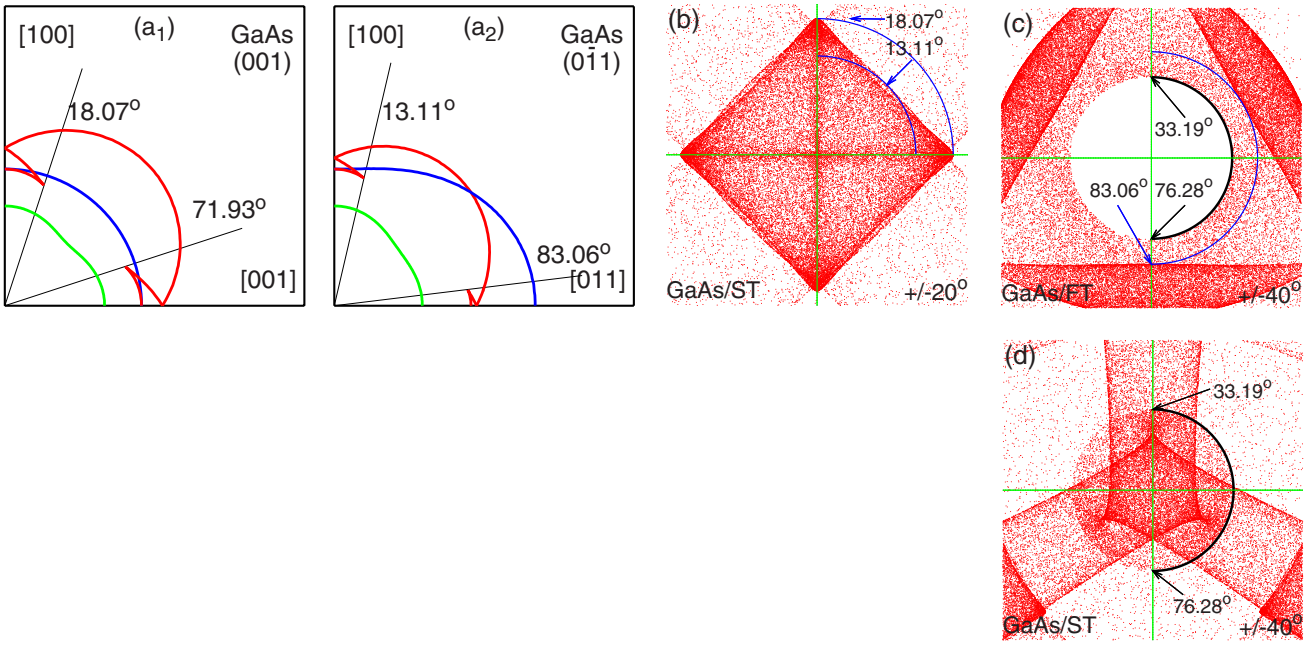


FIG. 7. (Color online) (a) Wave surface sections (slowness plot) in the (001) and (0$\bar{1}$1) planes for GaAs. The phonon focusing patterns: (b) ST mode along [100] direction, (c) FT mode along [111] direction, and (d) ST mode along [111] direction. The thick half circles partially mark the anticaustic circles.

Ref. 11] in the region with $a \approx 1.0$. Our simulation shows that, in the region with $a \gg 1.0$, β_{1-} refers to the same caustic point as $\hat{\varphi}_1$, or the point $3'$ shown in Fig. 8(b) in Ref. 6. In the region $a \approx 1.0$, β_{1-} describes, in fact, another caustic point associated with the point $4'$, rather than the point $3'$. The point $4'$ is actually originated from a parabolic point, not the inflection point, on the outer slowness sheet.

The present investigation provides not only an analytical scheme to rigorously determine the positions of the caustic/anticaustic points in cubic crystals but also offers a possibility to recover the relative elastic constants a and b from a given phonon focusing pattern. Since the functions $f(a,b,\varphi)$ and $g(a,b,\varphi)$ are analytically well defined by Eqs. (17) and (20), respectively, we can recover a and b directly from a pair of caustic points. This can be done by solving two equations ($f=0$ and $g=0$) simultaneously (see Appendix). Compared to the earlier work based upon solely numerical simulations,[11] our scheme provides a precise and effective way to determine the location/size of the caustic patterns.

V. CONCLUSION

The study of the phonon focusing patterns based on the Gaussian curvature analysis failed to produce analytical results for the caustic points because of the geometric complexity of the slowness surface. In the present work, by confining to the symmetry sections of the slowness/wave surface, the Stroh formalism is applied and it makes explicit deduction of the caustic/anticaustic points possible. By recognizing the connection between the caustic/anticaustic points and the triple degeneracies in the Stroh eigenvalue problem, we are able to obtain existence conditions for these points in terms of simple trigonometric equations. The analytical formulations of the caustic and anticaustic points also provide us with a direct and simple scheme for explicitly recovering the relative elastic constants from the characteristic points of phonon focusing patterns in cubic crystals.

ACKNOWLEDGMENT

The author would like to thank Kent G. Ryne at Østfold University College for many discussions.

APPENDIX: PSEUDOCODES FOR MATHEMATICA

Equations (17), (20), and (23) can be deduced by using MATHEMATICA. Given the Γ-matrix (2), the characteristic equation can be decomposed as a product of a quartic equation and a quadratic equation as Eq. (7). The conditions derived from the resultants $R_i, i=1,\ldots,6$ and the functions f, g, and h can be carried out as follows:

$$R_1 = \beta^2 - 3\alpha\gamma + 12\delta/.\{\alpha \to a_3/a_4, \ldots, \delta \to a_0/a_4\}$$

$$R_2 = 9\alpha^2\delta - \alpha\beta\gamma + 9\gamma^2 - 32\beta\delta/.\{\alpha \to a_3/a_4, \ldots\}$$

$$f = \text{Resultant}[R_1, R_2, \rho v^2]$$

$$R_5 = 4\sigma - \kappa^2/.\{\kappa \to b_1/b_2, \sigma \to b_0/b_2\}$$

$$R_6 = \kappa^4 - 2\alpha\kappa^3 + 4\beta\kappa^2 - 8\gamma\kappa + 16\delta/.\{\alpha \to a_3/a_4, \ldots\}$$

$$h = \text{Resultant}[R_5, R_6, \rho v^2].$$

Note that the function g [Eq. (20)] can be deduced in the same manner as f.

The caustic points for given a and b can be calculated by Solve$[f=0, \hat{\varphi}_1]$ and Solve$[g=0, \hat{\varphi}_2]$ directly.

For a given pair of caustic points, e.g., $\hat{\varphi}_1 = 26.0°$ and $\hat{\varphi}_{2\alpha} = 19.0°$, recovery of the relative elastic constants a and b can be done by using FindRoot

$$\text{FindRoot}[f(\hat{\varphi}_1) = 0, g(\hat{\varphi}_{2\alpha}) = 0, \{a, 2.5\}, \{b, 2.0\}]$$

with the initial guess $a=2.5$, $b=2.0$.

[1] M. J. P. Musgrave, *Crystal Acoustics* (Holden-Day, San Francisco, 1970).
[2] B. A. Auld, *Acoustic Waves and Fields in Solids* (Wiley, New York, 1973), Vol. **1**.
[3] J. P. Wolfe, *Imaging Phonons* (Cambridge University Press, Cambridge, 1998).
[4] J. A. Thorpe, *Elementary Topics in Differential Geometry* (Springer-Verlag, New York, 1979).
[5] V. Vavryčuk, "Parabolic lines and caustics in homogeneous weakly anisotropic solids," Geophys. J. Int. **152**, 318–334 (2003).
[6] A. G. Every, "Ballistic phonons and the shape of the ray surface in cubic crystals," Phys. Rev. B **24**, 3456–3467 (1981).
[7] A. G. Every, "Formation of phonon-focusing caustics in crystals," Phys. Rev. B **34**, 2852–2862 (1986).
[8] A. G. Every, "Acoustic symmetry in phonon imaging," J. Phys. C **20**, 2973–2982 (1987).
[9] G. A. Northrop and J. P. Wolfe, "Ballistic phonon imaging in germanium," Phys. Rev. B **22**, 6196–6212 (1980).
[10] R. L. Weaver, M. R. Hauser, and J. P. Wolfe, "Acoustic flux imaging in anisotropic media," Z. Phys. B: Condens. Matter **90**, 27–46 (1993).
[11] D. C. Hurley and J. P. Wolfe, "Phonon focusing in cubic crystals," Phys. Rev. B **32**, 2568–2587 (1985).
[12] A. L. Shuvalov and A. G. Every, "Curvature of acoustic slowness surface of anisotropic solids near symmetry axes," Phys. Rev. B **53**, 14906–14916 (1996).
[13] A. L. Shuvalov and A. G. Every, "Shape of the acoustic slowness surface of anisotropic solids near points of conical degeneracy," J. Acoust. Soc. Am. **101**, 2381–2383 (1997).
[14] A. L. Shuvalov and A. G. Every, "Transverse curvature of the acoustic slowness surface in crystal symmetry planes and associated phonon focusing cusps," J. Acoust. Soc. Am. **108**, 2107–2113 (2000).
[15] A. N. Stroh, "Steady state problems in anisotropic elasticity," J. Math. Phys. (Cambridge, Mass.) **41**, 77–103 (1962).
[16] P. Chadwick and G. D. Smith, "Foundations of the theory of surface waves in anisotropic elastic materials," *Advances in Applied Mechanics* (Academic, New York, 1977), Vol. **17**, pp. 303–376.
[17] T. C. T. Ting, *Anisotropic Elasticity: Theory and Applications* (Oxford University Press, Oxford, 1996).
[18] L. Wang, "Determination of the ray surface and recovery of elastic constants of anisotropic media," J. Phys.: Condens. Matter **7**, 3863–3880 (1995).
[19] V. Vavryčuk, "Calculation of the slowness vector from the ray vector in anisotropic media," Proc. R. Soc. London, Ser. A **462**, 883–896 (2006).
[20] V. Vavryčuk, "Generation of triplications in transversely isotropic media," Phys. Rev. B **68**, 054107 (2003).
[21] M. Bôcher, *Introduction to Higher Algebra* (MacMillan, New York, 1936).

Dispersion of circumferential waves in cylindrically anisotropic layered pipes in plane strain

R. Y. Vasudeva[a)]
Applied Mathematics Department, Andhra University, Visakhapatnam, 530 003 India

G. Sudheer
Mathematics Department, G.V.P. College of Engineering, Visakhapatnam, 530 003 India

Anu Radha Vema
Applied Mathematics Department, Andhra University, Visakhapatnam, 530 003 India

(Received 24 August 2007; revised 12 February 2008; accepted 4 March 2008)

Dispersion spectra of circumferential waves along the periphery of circular pipes made of layered anisotropic materials do not seem to be available in literature. This note attempts to partially fill this gap by providing the dispersion spectra in two and three layered cylindrically anisotropic pipes in plane strain motion. The spectra for pipes executing time harmonic vibrations in plane strain condition are obtained as roots of a numerical characteristic equation derived extending a weighted residual method of solution of the governing equations for a single layer pipe [Towfighi *et al.*, J. Appl. Mech. **69**, 283–291 (2002)] to a general N layered pipe. The anisotropic elastic coefficients are considered to be independent of position coordinates and the bond condition at interfaces of the layers is assumed to be perfect. Numerical illustrations are presented for two and three layered pipes with anisotropy directions differing in adjacent layers. Increase in curvature of the pipe and inclination of the fiber orientation in the outermost layers to propagation direction are factors that seem to influence the mode number and pattern within the limited examples worked out.
 [DOI: 10.1121/1.2902180]

PACS number(s): 43.35.Zc, 43.20.Mv, 43.40.Le [RLW] Pages: 4147–4151

I. INTRODUCTION

The particle displacement components appear coupled in the system of partial differential equations that characterize wave propagation in elastic media. In the simplest case of elastic isotropy of the medium of propagation there are well known techniques such as the Helmholtz decomposition that can be used to decouple the displacements for a solution of the equations of motion. In the case of material anisotropy no general technique of separating the variables in the governing Navier-type equations exists. In the case of cylindrical anisotropy in which the elastic moduli do not depend on the position coordinates, Towfighi *et al.*[1] developed a numerical weighted residual method—without resorting to decoupling—to obtain the dispersion spectra of time harmonic circumferential waves propagating along the periphery of a single layered hollow pipe. We refer to Ref. 1 for a survey of literature on studies in circumferential waves in pipes.

In this note we modify and extent Towfighi *et al.*[1] to a general N layered cylindrical pipe of cylindrical anisotropy and obtain dispersion of time harmonic circumferential waves in the annulus vibrating in plane strain condition. The method can be extended to the three-dimensional case. As an instance of recovery of results of Ref. 1 for a single layered pipe, it is shown in the section on numerical results that the modification improves their results in tracing the modes.

II. THEORY

A. Problem formulation

We consider time harmonic circumferential wave motion in an annulus of uniform thickness and the structural element is termed a pipe (P).

The symbolic equation

$$L_{p\alpha}^{(k)}[\bar{u}^{(k)}(r,\theta,t)]=0, \quad \alpha=1,2 \tag{1}$$

denotes the two partial differential (p) Navier-type equations of motion for the typical kth layer of the N layered cylindrically anisotropic pipe vibrating freely in plane strain condition. Layers are assumed to be concentric; $\rho^{(k)}$ the density and $C_{ij}^{(k)}$ the five elastic coefficients are taken as constants for the material of the kth layer. The axis of P is along the z axis of (r,θ,t) frame. The inner and the outermost bounding surfaces are, respectively, at r_o and r_N. The kth interface of the k and $(k+1)$ layers is at r_k. We consider perfect bond conditions at all interfaces. For free vibration, the two stress components, denoted by σ_{rr}, $\sigma_{r\theta}$, must vanish at r_o and r_N. The perfect bonds at all interface layers impose additional conditions of continuity on the displacement components and the stress components at the interface of the kth, $(k+1)$th layers. Thus we have in addition to the coupled partial differential Eqs. (1) four boundary and 4N-4 interface conditions totaling 4N conditions.

For circumferential waves, we take

a)Author to whom correspondence should be addressed. Electronic mail: r.y.vasudeva@gmail.com

$$\bar{u}^{(k)}(r,\theta,t) = [u_r^{(k)}(r), u_t^{(k)}(r)]e^{i(p\theta-\omega t)}, \tag{2}$$

where p is a real nonintegral wave number, ω is the frequency. The velocity of the propagating phase is $v_{\text{ph}}(r) = c_b r/(b)$, where c_b is assumed to be the phase velocity at the outer surface with radius b. The stress components in the kth layer in terms of the displacement components are

$$\bar{\sigma}^{(k)} = \bar{C}^{(k)}\bar{e}^{(k)}, \tag{3}$$

where

$$\bar{\sigma}^{(k)} = (\sigma_{\theta\theta}^{(k)}, \sigma_{zz}^{(k)}, \sigma_{rr}^{(k)}, \sigma_{r\theta}^{(k)})^t,$$

$$\bar{C}^{(k)} = \begin{pmatrix} C_{11}^{(k)} & C_{12}^{(k)} & C_{13}^{(k)} & C_{14}^{(k)} \\ C_{12}^{(k)} & C_{22}^{(k)} & C_{23}^{(k)} & C_{24}^{(k)} \\ C_{13}^{(k)} & C_{23}^{(k)} & C_{33}^{(k)} & C_{34}^{(k)} \\ C_{14}^{(k)} & C_{24}^{(k)} & C_{34}^{(k)} & C_{44}^{(k)} \end{pmatrix}$$

$$\bar{e}^{(k)} = (e_{\theta\theta}^{(k)}, 0, e_{rr}^{(k)}, 2e_{r\theta}^{(k)})^t,$$

$$e_{r\theta}^{(k)} = \frac{1}{2}\left(\frac{\partial u_r^{(k)}}{r\,\partial\,\theta} + \frac{\partial u_\theta^{(k)}}{\partial r} - \frac{u_\theta^{(k)}}{r}\right),$$

$$e_{rr}^{(k)} = \frac{\partial u_r^{(k)}}{\partial r}$$

and $C_{ij}^{(k)}$ are the elastic coefficients. Equation (2) put in Eq. (1) makes it a system of two ordinary differential equations

$$(\rho^{(k)}\omega^2 r^2 - p^2 C_{44}^{(k)} - C_{11}^{(k)})u_r^{(k)} - ip(C_{11}^{(k)} - C_{44}^{(k)})u_\theta^{(k)} + rC_{33}^{(k)}u_r'^{(k)}C_r^{(k)} + ipr(C_{13}^{(k)} - C_{44}^{(k)})u_r^{(k)} + r^2 C_{33}^{(k)}u_r''^{(k)} = 0 \tag{4a}$$

and

$$ip(C_{11}^{(k)} + C_{44}^{(k)})u_r^{(k)} + (\rho^{(k)}\omega^2 r^2 - p^2 C_{11}^{(k)} - C_{44}^{(k)})u_\theta^{(k)} + ipr(C_{13}^{(k)} + C_{44}^{(k)})u_r'^{(k)} + rC_{44}^{(k)}u_\theta'^{(k)} + r^2 C_{44}^{(k)}u_\theta''^{(k)} = 0. \tag{4b}$$

We have with Eqs. (4a) and (4b), the traction free boundary conditions

$$\sigma_{rr}^{(1)} = \sigma_{r\theta}^{(1)} = 0 \quad \text{at } r = r_0; \quad \sigma_{rr}^{(N)} = \sigma_{r\theta}^{(N)} = 0 \quad \text{at } r = r_N \tag{5}$$

together with the interface conditions, insisting on continuity of displacements and stresses

$$u_r^{(k)} = u_r^{(k+1)}, \quad u_\theta^{(k)} = u_\theta^{(k+1)} \quad \text{at } r = r_k,\ k = 1,\ldots,N-1, \tag{6}$$

$$\sigma_{rr}^{(k)} = \sigma_{rr}^{(k+1)}, \quad \sigma_{r\theta}^{(k)} = \sigma_{r\theta}^{(k+1)} \quad \text{at } r = r_k,\ k = 1,\ldots,N-1. \tag{7}$$

Thus we have in Eq. (5) four boundary conditions and in Eqs. (6) and (7) 4N-4 interface conditions, leading to a total of 4N conditions to be satisfied.

B. A Weighted residual method (w.r.m) of solution

We assume finite Fourier series expansions with $(m+1)$ terms for $u_r^{(k)}(r)$ and $u_\theta^{(k)}(r)$ with m being the same for each of the N layers, i.e.,

$$u_r^{(k)}(r) = x_{or}^{(k)} + \sum_{n=1}^{m} x_{nr}^{(k)} \cos\left(\frac{n\pi}{L_k}\right)r + y_{nr}^{(k)} \sin\left(\frac{n\pi}{L_k}\right)r, \tag{8a}$$

$$u_\theta^{(k)}(r) = x_{o\theta}^{(k)} + \sum_{n=1}^{m} x_{n\theta}^{(k)} \cos\left(\frac{n\pi}{L_k}\right)r + y_{n\theta}^{(k)} \sin\left(\frac{n\pi}{L_k}\right)r \tag{8b}$$

and denote the $4m+2$ vector of coefficient in Eq. (8) by $X^{(k)}$

$$X^{(k)} = [x_{or}^{(k)}, x_{o\vartheta}^{(k)}, x_{1r}^{(k)}, y_{1r}^{(k)} \cdots y_{m\theta}^{(k)}]^t. \tag{9}$$

We substitute the finite Fourier series in the left hand side of Eq. (4). It results in two residuals and we denote them by $f_\alpha(r, X^{(k)})$. The same sine and cosine functions in the finite Fourier series in Eq. (8) are used as weights, unlike in Ref. 1, where linear functions in r are used as weights. From the residuals $f_\alpha(r, X^{(k)})$ we construct the weighted residuals $R_\alpha^{(k)} \times(r)$ keeping in view the $4m+2$ unknowns.

$$R_\alpha^{(k)}(r) = \int_{r_{k-1}}^{r_k} w_\beta^{(k)} f_\alpha(r, X^{(k)})dr \quad \alpha = 1,2. \tag{10}$$

We equate each $R_\alpha^{(k)}(r)$ to zero and solve for the coefficient vector $X^{(k)}$ from the resulting system of equations. The weights $w_\beta^{(k)}$ that appear in the integrand of Eq. (10) are taken as

$$w_1^{(k)} = 1, \quad w_{2q}^{(k)} = \cos\left(\frac{q\pi r}{L_k}\right) \quad \text{and} \quad w_{2q+1}^{(k)} = \sin\left(\frac{q\pi r}{L_k}\right), \quad q = 1,2,\ldots,m. \tag{11}$$

Now $R_\alpha^{(k)}(r)$ when written in matrix form is a $4m+2$ system given by

$$A^k X^{(k)} = 0^{(k)}. \tag{12}$$

We get four independent solution vectors

$$\bar{\boldsymbol{u}}^{\eta(k)} = [\boldsymbol{u}_r^{\eta(k)}, \boldsymbol{u}_\theta^{\eta(k)}]^t \quad \text{for } \eta = 1,2,3,4 \tag{13}$$

for the kth layer by determining the $4m-2$ elements of $X^{(k)}$ in terms of the four elements $x_{mr}^{(k)}$, $y_{mr}^{(k)}$, $x_{m\theta}^{(k)}$, $y_{m\theta}^{(k)}$. We choose these elements as $(1,0,0,0)$ for the first solution and as $(0,1,0,0)$, $(0,0,1,0)$ and $(0,0,0,1)$, respectively, for the second, third, and fourth solutions. The general solution $\Sigma_{\eta=1}^{4} C^{\eta(k)}\bar{u}^{\eta(k)}$ when substituted in the boundary and interface conditions (5)–(7) gives the dispersion relation. One can then solve the dispersion relation numerically. In the present work, we obtained a numerical solution using the bisection algorithm. We employed MATHEMATICA 5.0 for the computation.

III. NUMERICAL RESULTS

The theory developed in the previous section holds for composite cylindrical pipes made of any N number of layers. The anisotropy is restricted to cylindrical anisotropy in which the elasticity tensor is independent of the position co-

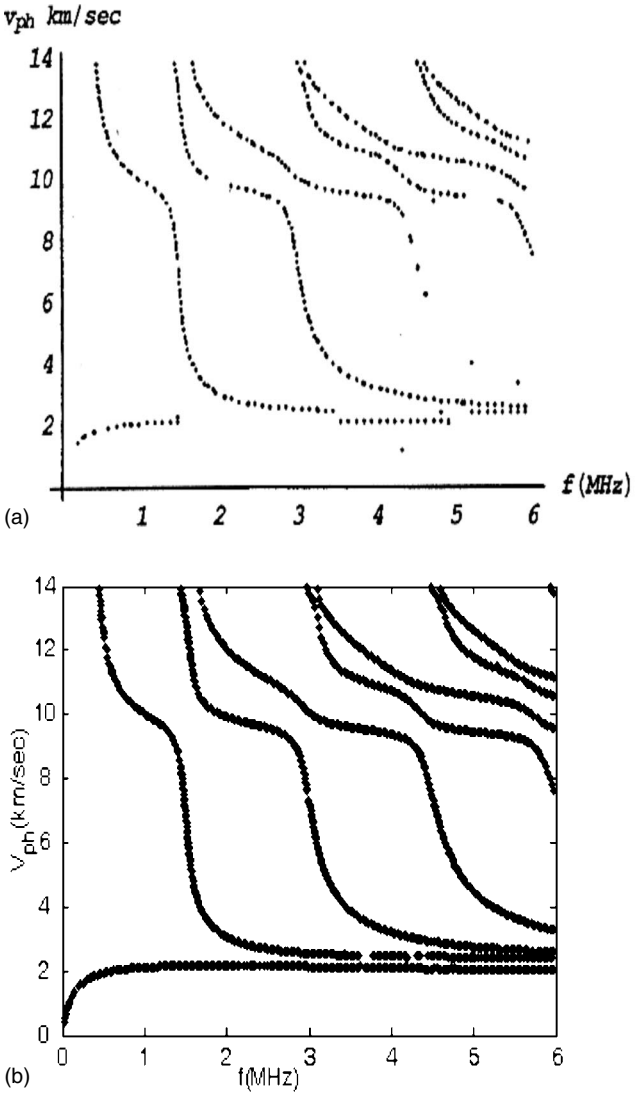


FIG. 1. Dispersion curves in a single layered anisotropic of 5 mm a) from [tow (1)] and (b) present scheme.

ordinates and the vibrations are in plane strain condition. Anisotropic directions in adjacent layers as well as the curvature of the pipe influence the dispersion of these circumferential waves. This can be seen in the dispersion spectra—Figs. 1 and 2—obtained using the above theory for two and three layered pipe structures. The numerical values of the elastic coefficients C_{ij} used in these illustrations are calculated from the five propagating velocities in the material medium given by Markham's[2] measurements for carbon fiber-epoxy resin composites. Earlier the values of velocities provided in Ref. 2 have been employed in numerical illustrations in Lamb wave propagation in transversely anisotropic plates made of fiber reinforced materials by Baylis,[3] and Baylis and Green.[4] More recently, these values have been employed by Kaplunov *et al.*[5] to illustrate the mode dispersion in transversely isotropic plates near their cutoff frequencies and also by Vasudeva and Anu Radha Vema[6] in validating the present w.r.m scheme against two and three layered transversely isotropic sandwich plate modes provided in Baylis[3] and Baylis and Green.[4] In Ref. 6, Vasudeva and Anu Radha Vema have also shown that the present w.r.m scheme fares better than Towfighi *et al.*[1]

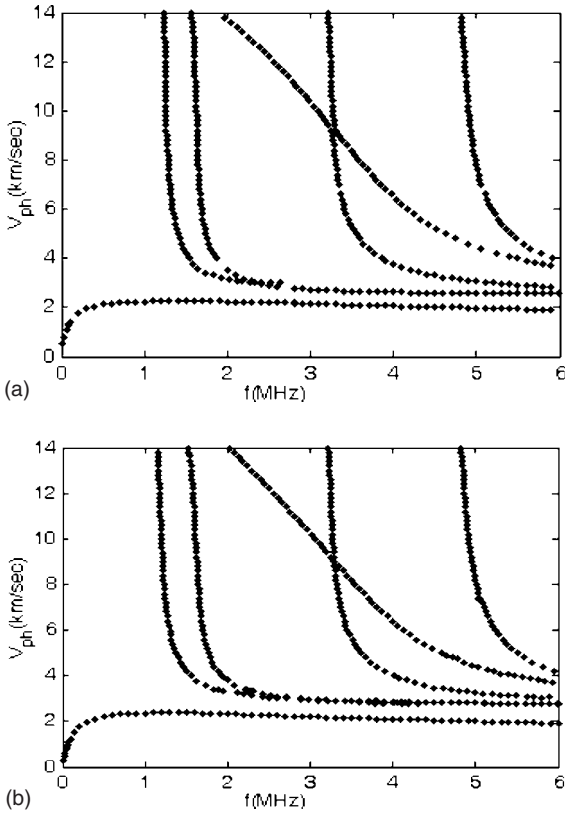


FIG. 2. Dispersion curves in a two layered anisotropic pipe when fibers in the inner layer are oriented in the circumferential direction; outer radius of pipe is (a) 10 mm, (b) 5 mm.

A. Plane strain circumferential waves: Mode dispersion spectra in anisotropic layered small diameter pipes

The dispersion of circumferential waves in layered pipes is relatively a recent attraction to workers in nondestructive evaluation. In layered pipes made of anisotropic materials, to the best of the authors' knowledge there are no numerical illustrations of the spectra in literature. There are, however, the dispersion spectra available in pipes of all isotropic layers. Recently Luo *et al.*[7] have obtained theoretical dispersion of shear harmonic circumferential waves in an N-layered pipe made of isotropic visco-elastic materials making use of correspondence principle. They presented numerical results in a—isotropic-elastic and isotropic visco-elastic—two layer pipe. Mention must be made here of a paper by Adamou and Craster.[8] Adamou and Craster's[8] work is more general in its scope. It describes how to trace dispersion curves and displacement profiles in a large class of elastic waveguides—curved, layered, damped, inhomogeneous and anisotropic—using a numerical method called the spectral method. Even so, they have not shown any numerical results in anisotropic layered pipes. The only one dispersion diagram, Fig. 8 of Ref. 8, they presented in the case of anisotropic structures is the dispersion of in plane circumferential waves in a single layered large diameter pipe. Thus their illustration is essentially the dispersion diagram of Lamb waves in an anisotropic single layered plate. They show it by way of comparison of their scheme with the analytically obtained dispersion available in Rose[9] for such a plate. Towfighi *et al.*[1] also used

the same illustration from Ref. 9 as part of their validation. In the following we present dispersion of circumferential waves graphically in relatively very small diameter pipes of two and three layered anisotropic media. In all the figures, the anisotropic directions differ in adjacent layers.

1. Validation of the present scheme

The present scheme is validated against the results of Towfighi *et al.*[1] for single layered anisotropic pipes of 5 mm diameter. The elastic constants of the material of the pipes are as given in Towfighi *et al.*[1], Eq. (12). Figure 1(b) shows the improvement over Towfighi *et al.*[1] reproduced here in Fig. 1(a). It can be seen that the present scheme not only recovers the results of Towfighi *et al.*[1] but also computes the dispersion completely without any gaps even in the lower frequencies improving over Ref. 1.

2. Two layered anisotropic pipes

Figures 2(a) and 2(b) show the circumferential wave dispersion in two layered anisotropic pipes of different outer radii. All the figures are drawn for structures made of the same materials. The fiber directions in the two layers are at right angles to each other with the fiber direction in the inner layer along the wave propagation direction. The matrix $\bar{C}^{(k)}$ that appears in Eq. (3) in units of Giga Pascal (GPa) for the materials of the two layers calculated from the velocity data from Ref. 3 are

$$\bar{C}^{(1)} = \begin{pmatrix} 241.71 & 4.37 & 4.37 & 0 \\ 4.37 & 10.57 & 5.65 & 0 \\ 4.37 & 5.65 & 10.57 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \tag{14}$$

$$\bar{C}^{(2)} = \begin{pmatrix} 10.57 & 5.65 & 5.63 & 0 \\ 5.65 & 10.57 & 5.65 & 0 \\ 5.63 & 5.65 & 10.57 & 0 \\ 0 & 0 & 0 & 2.46 \end{pmatrix}.$$

The curves are drawn for 30 terms of the Fourier series.

3. Dispersion spectra in three layered anisotropic pipes

The circumferential wave dispersion in three layered anisotropic symmetric sandwich pipes of different outer radii is exhibited in Figs. 3(a) and 3(b). The fiber direction in outer layers ($k=1,3$) is at right angles to that in the core layer ($k=2$) and is in the direction of wave propagation. The constitutive matrices of the material in the outer layers obtained from the velocity data of Ref. 2 in units of GPa (i.e., $\bar{C}^{(1)}=\bar{C}^{(3)}$) are the same as those of $\bar{C}^{(1)}$ of Eq. (14) while that of the material in the core is given by

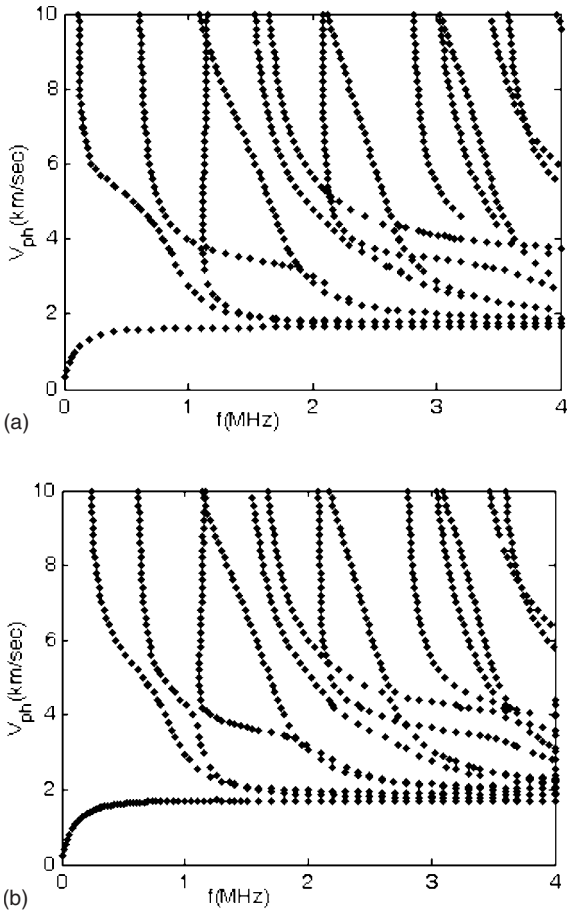


FIG. 3. Dispersion curves in a three layered anisotropic pipe when fibers in the outer layers are oriented in the circumferential direction; outer radius of pipe is (a) 10 mm, (b) 5 mm.

$$\bar{C}^{(2)} = \begin{pmatrix} 10.57 & 4.37 & 5.65 & 0 \\ 4.37 & 241.71 & 4.57 & 0 \\ 5.65 & 4.57 & 4.57 & 0 \\ 0 & 0 & 0 & 2.46 \end{pmatrix}.$$

The curves are drawn for 30 terms of the series. For consistency the dispersion curves for two and three layered pipes are drawn for $m=30$. The curves could be traced continuously all through the range without any gaps. The outer diameters of the pipe are chosen to be 10 and 5 mm. Unlike in Towfighi *et al.*[1] the thicknesses of the pipes are also varied. The frequency is in MHz along the horizontal and phase velocity in km/s along the vertical, the curves are plotted from 0.01 MHz and 0.01 km/s, slightly away from the origin. Usually more layers give rise to more number of modes but what we observe even in the limited comparison that is carried out in this note is that the number of modes in a two layered anisotropic pipe as seen in Figs. 2(a) and 2(b) is less than those in a single layered pipe. This could be attributed to orientation of fibers in the outerlayer being normal to direction of propagation of waves. In the three layered case when the fiber orientation in the outerlayer coincides with the direction of propagation we find once again, the number of modes increasing. Clearly the increase in curvature is a measure of departure from plate geometry and we see some

of the lower modes tend to decouple as the curvature of the pipe increases.

IV. CONCLUSIONS

For circumferential wave propagation the dispersion mode spectra is available in literature for single and multilayered isotropic pipe guides. For anisotropic pipes, since exact solutions are hard to get, Towfighi *et al.*[1] obtained the dispersion using a well-tested numerical method for single layered pipes. In this note we present a modified extension of Towfighi *et al.*[1] for cylindrically anisotropic layered pipes vibrating in plane strain condition. In the present work, as in Ref. 1, the elastic moduli are assumed to be independent of coordinate functions. The numerical scheme developed here is valid for any general N layered pipe. The final matrix that gives the mode dispersion on employing a root finding algorithm is constituted from N copies of the algebra worked out for a single layer. This feature is common also with the spectral method.[9] The spectral method has its own implementation difficulties[9] and in fact any numerical technique is fraught with difficulties peculiar to it. It is welcome to have different methods for counter checking the respective outputs. The numerical illustrations of w.r.m can be used for such a comparison. The dispersion is demonstrated in the case of asymmetric two layered pipes with the upper layer 10 times thinner than the lower. Further the anisotropy directions due to fiber orientations are mutually orthogonal in the two layers. In the case of symmetric sandwich pipes, the upper and lower facings are 20 times thinner than that of the core. The fiber orientation in the facings is at right angles to that in the core, however all the layers are made of the same material. An effort is made to see the effect of curvature on dispersion by plotting the propagation modes in pipes of outer radii 10 and 5 mm. Since many variations in layering and radii are possible, what we presented by way of graphical illustrations constitutes only a limited study from which no generalization seems to be possible.

[1] S. Towfighi, T. Kundu, and M. Ehsani, "Elastic wave propagation in circumferential direction in anisotropic cylindrical curved plates," J. Appl. Mech. **69**, 283–291 (2002).

[2] M. F. Markham, "Measurement of the elastic constants of fiber composites by ultrasonics," Composites **1**, 145–149 (1970).

[3] E. R. Baylis, "Wave propagation in an asymmetric fiber—reinforced laminated plates," Acta Mech. **64**, 187–206 (1986).

[4] E. R. Baylis and W. A. Green, "Flexural waves in fiber–reinforced laminated plates," J. Sound Vib. **110**, 1–26 (1986).

[5] J. D. Kaplunov, Z. L. Yu, Y. Kossovich, and G. A. Rogerson, "Direct asymptotic integration of the equations of transversely isotropic elasticity for a plate near cut-off frequencies," Q. J. Mech. Appl. Math. **53**, 323–341 (2000).

[6] R. Y. Vasudeva and Anu Radha Vema, "In-plane circumferential vibrations of layered pipes," *15th US National Congress on Theoretical and Applied Mechanics*, University of Colorado at Boulder, June 25–30 (2006).

[7] W. Luo, J. L. Rose, J. K. Van Velsor, M. Avioli, and J. Spanner, "Circumferential guided waves for defect detection in coated pipe," *Review of Progress in Quantitative Nondestructive Evaluation***25**, 165–172 (2006).

[8] A. T. I Adamou and R. V. Craster, "Spectral methods for modeling guided waves in elastic media," J. Acoust. Soc. Am. **116**, 1524–1535 (2004).

[9] J. L. Rose, *Ultrasonic Waves in Solid Media* (Cambridge University Press, Cambridge, U.K.) pp. 264–271 (1999).

The trapped fluid transducer: Modeling and optimization

Lei Cheng[a)]
Department of Mechanical Engineering, University of Michigan, Ann Arbor, Michigan 48105-2125, USA

Karl Grosh[b)]
Department of Mechanical and Biomedical Engineering, University of Michigan, Ann Arbor, Michigan 48105-2125, USA

(Received 24 April 2007; revised 17 March 2008; accepted 21 March 2008)

Exact and approximate formulas for calculating the sensitivity and bandwidth of an electroacoustic transducer with an enclosed or trapped fluid volume are developed. The transducer is composed of a fluid-filled rectangular duct with a tapered-width plate on one wall emulating the biological basilar membrane in the cochlea. A three-dimensional coupled fluid-structure model is developed to calculate the transducer sensitivity by using a boundary integral method. The model is used as the basis of an optimization methodology seeking to enhance the transducer performance. Simplified formulas are derived from the model to estimate the transducer sensitivity and the fundamental resonant frequency with good accuracy and much less computational cost. By using the simplified formulas, one can easily design the geometry of the transducer to achieve the optimal performance. As an example design, the transducer achieves a sensitivity of around −200 dB (1 V/μPa) at 10 kHz frequency range with piezoelectric sensing. In analogy to the cochlea, a tapered-width plate design is considered and shown to have a more uniform frequency response than a similar plate with no taper. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2908301]

PACS number(s): 43.38.Ar, 43.40.At, 43.40.Dx [AJZ] Pages: 4152–4164

I. BACKGROUND

The continuous development of electroacoustic transducers for sensing and transmitting is essential to many navy and commercial applications such as sound navitation ranging SONAR and medical imaging and therapy. The transducer introduced in this paper, unlike conventional piezoelectric and condenser microphones or hydrophones, includes a fluid-filled back chamber and two separate membranes as input and output ports. The design of the fluid-filled chamber preserves the bandwidth of the transducer for underwater operations, which enables the transducer to operate underwater as well as in air. Generally, the bandwidth of a condenser microphone will be greatly reduced if it could be submerged due to the extra mass loading from the fluid environment.[1] However, in our design, the fluid inside the chamber increases the dynamic mass of the system, so that the external fluid loading has little influence on the transducer bandwidth.[1] The separation of the input and output ports is also beneficial to the packaging of the system.

In literature, the first reported fluid-filled hydrophone was developed by Bernstein.[2] It was designed as a condenser microphone with silicone oil or triacetin filled in the gap. The hydrophone achieved a sensitivity of around −206 dB (1 V/μPa) within 2 kHz frequency range. The notion of utilizing the fluid-structure interaction has also been implemented on other acoustic devices such as passive noise control systems. In recent years, two groups of researchers independently proposed a new type of passive noise control device—a structure acoustic silencer.[3–7] In these devices, a flexible panel forms part of a duct wall, which vibrates in response to incident waves. Noise control is achieved by energy absorption through the losses in the structure in addition to energy reflection by the impedance mismatch. The modeling techniques for this type of fluid-structure interaction device include the finite element method[3,4] and boundary integral method.[5,7] A new analytical mode-matching approach, developed by Lawrie and Kirby,[8] offers a different way to model the coupled system (restricted to two dimensional) without finding roots from the characteristic equation. The formulations and approximations developed in this paper will also find use in such engineered devices.

Although silencer and transducer applications are obviously different, they share a common structure—a fluid-filled duct and a flexible layer inserted in an otherwise rigid duct wall. Both of the designs are partly inspired by the structure of the mammalian cochlea.[9] The cochlea is a spiral-shaped organ in the inner ear, where sound vibrations are converted into nerve impulses. Inside the cochlea is the basilar membrane. The basilar membrane vibrates in response to incoming sounds and, due to the varying impedance along its length, a traveling structural acoustic wave peaks at different locations for each frequency and followed by a sharp spatial decay. Such broad-band filtering is the desired behavior sought by the silencer design. The cochlea serves as a structural acoustic filter, which divides the frequency spectrum into bands of frequencies. The inner hair cells arrayed along the length of the basilar membrane are mechanoelectrical transducers, which translate the mechanical vibrations into the neural signals sent to the brain.[9] This work is motivated by devising an electroacoustic transducer and adapting the passive cochlear structure into the engineered acoustic device. The objective of this paper is not to generate the sharp

[a)]Electronic mail: lchengz@umich.edu.
[b)]Electronic mail: grosh@umich.edu.

 0001-4966/2008/123(6)/4152/13/$23.00

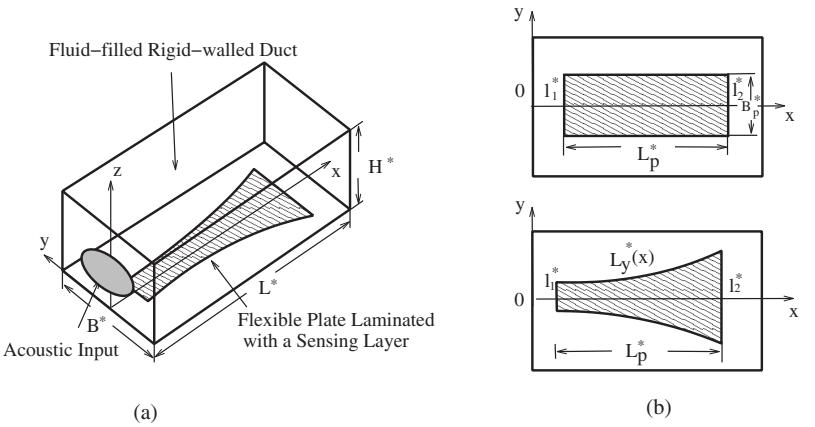


FIG. 1. Transducer model. (a) Straight and single rectangular duct model; (b) dimension of the constant-width and tapered-width plates.

frequency filtering as in the cochlea but rather to explore the benefits of the cochlea structure in the transducer design and to develop accurate and efficient design strategies.

A number of researchers have constructed electroacoustic devices by mimicking the cochlear structure, although most of them are used to demonstrating cochlear characteristics and answering questions arising from cochlear mechanics. Zhou *et al.*[10] pointed out that the life-size model of the cochlea that they developed and tested could also be an acoustical-optical transducer as an application. Lechner's[11] hydromechanical model included a nonlinear feedback which sharpened the cochlear filtering. The micromachined single channel hydrophone developed by White *et al.*[1] adapted the cochlear structure by using a fluid-filled chamber as the back chamber and two sets of membranes as input and output ports. The fluid-filled chamber allows for hydrophone designs that have the same sensitivity and bandwidth for underwater operations as for in air. The microengineered hydromechanical cochlear model of White and Grosh,[12] which was motivated by the possibility of building cochlearlike filters, acoustically demonstrated the traveling fluid-structure wave and measured the frequency-location mapping function. In the cochlear mechanics, the Wentzel–Kramer–Brillouin method[13] is well known for modeling the cochlea and cochlear-based fluid-structure coupled systems.

In the current study, we aim at providing an efficient prediction and optimization tool for the design of this type of the electroacoustic transducers. To this end, a coupled fluid structure model is developed for the transducer by using a boundary integral method. The model is used as the basis of an optimization methodology seeking to enhance the transducer performance. For a transducer made of a constant-width plate, simplified formulas are derived to predict the transducer sensitivity and the fundamental resonant frequency. By using the simplified expressions, the dependence of the transducer sensitivity and bandwidth on the geometry is discussed in detail. These relations are used to restrict the design space for the optimal performance of the transducer.

II. THEORETICAL MODELING

Figure 1(a) shows the geometry under consideration (which can be thought of as an unwrapped rectangular idealized cochlea). Inside the ear, the sound wave is transmitted to the cochlea by the movement of the stapes, a bone in the middle ear. The piston action of the stapes moves the fluid inside the cochlea, which initiates a traveling wave pattern propagated down the basilar membrane. The transducer model consists of a straight fluid-filled single duct with a rectangular cross section and a flexible plate on the bottom wall, which emulates the biological basilar membrane. The mechanical to electrical transduction is realized by a thin piezoelectric layer bonded to the surface of the plate. Figure 1(b) shows two plate models (constant-width and tapered-width) used in the analysis with the definition of the geometrical variables.

In the model, a harmonic pressure fluctuation is applied to the side wall, Γ_s, of the duct as an excitation. The flexible plate laminated with the sensing layer occupies part of the bottom wall, Γ_p. The other walls are considered to be rigid. The plate is assumed to be simply supported on the boundaries. In the current setting, the x axis longitudinally extends along the duct length direction, the y axis is oriented along its width, and the z axis is along the height of the duct.

To solve this fluid-structure coupled system, a Green's function is used in the fluid domain to express the fluid pressure as boundary integrals, and the modal expansion method is used in the structure to describe the plate displacement.[14] In the following analysis, all the dimensional variables are normalized by three basic quantities: The fluid density ρ_f^*, the speed of sound c^*, and the duct length L^*:

$$x = \frac{x^*}{L^*}, \quad y = \frac{y^*}{L^*}, \quad z = \frac{z^*}{L^*},$$

$$P = \frac{P^*}{\rho_f^* c^{*2}}, \quad W_p = \frac{W_p^*}{L^*}, \quad \omega = \frac{\omega^* L^*}{c^*}, \tag{1}$$

where the asterisks denote the dimensional variables and the dimensionless ones are without asterisks. P, W_p, and ω are the dimensionless fluid pressure, plate displacement, and excitation frequency, respectively.

By assuming the fluid inside the duct to be inviscid, compressible, and irrotational, the fluid pressure P satisfies the Helmholtz equation:

$$\nabla^2 P + \omega^2 P = 0. \tag{2}$$

Here, we consider the response to be time harmonic solutions with assumed $e^{j\omega t}$ dependence.

At the wall Γ_p, the pressure is related to the plate displacement by Euler's equation:

$$\left.\frac{\partial P}{\partial n}\right|_{\Gamma_p} = -\omega^2 W_p, \tag{3}$$

and at Γ_s, the magnitude of the incident sound pressure is known,

$$P|_{\Gamma_s} = \frac{P_s^*}{\rho_f^* c^{*2}}, \tag{4}$$

where n is the unit outward normal vector and P_s^* is the magnitude of the uniform input pressure.

Before solving this fluid-structure coupled problem, we first suppose that we have solved a simpler problem: Finding the response of this transducer model to a point source excitation located on the plate. Let us define a three-dimensional Green's function G to satisfy the following equation:

$$\nabla^2 G + \omega^2 G = \delta(\boldsymbol{x} - \boldsymbol{x}_0), \tag{5}$$

with the Neumann boundary conditions on all boundaries except on input boundary Γ_s, where $G=0$. The point source is located as $\boldsymbol{x}_0$. Here, all the vectors are indicated by bold letters, for example, $\boldsymbol{x}=\{x,y,z\}^T$.

The solution to Green's function is

$$G(\boldsymbol{x},\boldsymbol{x}_0) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \gamma_{lmn}\Phi_{lmn}(\boldsymbol{x})\Phi_{lmn}(\boldsymbol{x}_0) \quad \text{for } \boldsymbol{x}_0 \in \Omega, \tag{6}$$

where

$$\Phi_{lmn}(\boldsymbol{x}) = \sin\frac{(2l+1)\pi x}{2}\cos\frac{m\pi(y+\epsilon_B/2)}{\epsilon_B}\cos\frac{n\pi z}{\epsilon_H},$$

$$\gamma_{lmn} = \frac{2\alpha_m\alpha_n}{\lambda_{lmn}^2\epsilon_B\epsilon_H},$$

$$\lambda_{lmn}^2 = \omega^2 - \frac{(2l+1)^2\pi^2}{4} - \left(\frac{m\pi}{\epsilon_B}\right)^2 - \left(\frac{n\pi}{\epsilon_H}\right)^2,$$

$$\alpha_i = \begin{cases} 1 & \text{for } i=0 \\ 2 & \text{for } i\neq 0, \end{cases} \tag{7}$$

where Ω is the total volume of the domain. $\epsilon_B=B^*/L^*$ and $\epsilon_H=H^*/L^*$, where B^* and H^* are the width and the height of the duct, respectively.

If we multiply Eq. (2) by G and Eq. (5) by P, integrate $G\nabla^2 P - P\nabla^2 G$ over the total volume Ω, and convert the volume integral to a surface integral by Green's theorem, the acoustic pressure can be written in terms of Green's function G:

$$P(\boldsymbol{x}) = -\int_{\Gamma_{s0}} \frac{\partial G(\boldsymbol{x},\boldsymbol{x}_0)}{\partial x_0} P_s d\Gamma_{s0} + \int_{\Gamma_{p0}} G(\boldsymbol{x},\boldsymbol{x}_0)\omega^2 W_{p0}(\boldsymbol{x}_0) d\Gamma_{p0} \quad \text{for } \boldsymbol{x}\in\Omega. \tag{8}$$

The choice of Green's function G and the formulation of the pressure P in the above equations are based on the assumption that the source is located inside the domain. If $\boldsymbol{x}_0$ is on the boundary, for example, $\boldsymbol{x}_0=(0.5,0.5\epsilon_B,0)$, the integration of the dirac delta function is

$$\int_\Omega A(\boldsymbol{x})\delta(\boldsymbol{x}-\boldsymbol{x}_0)dV = \frac{1}{2}A(\boldsymbol{x}_0) \tag{9}$$

for any continuous function $A(\boldsymbol{x})$ in the domain. Therefore, for any point $\boldsymbol{x}$ on the boundary, $P(\boldsymbol{x})$ becomes $P(\boldsymbol{x})/2$ and $G(\boldsymbol{x},\boldsymbol{x}_0)$ becomes $G(\boldsymbol{x},\boldsymbol{x}_0)/2$ in Eq. 8. Canceling the factor $1/2$ on both sides recovers the same equation for a point source on the boundary as in the domain for our particular choice of Green's function G.

For an isotropic Kirchhoff plate with residual stress T, the plate displacement W_p satisfies

$$D\nabla^4 W_p - T\nabla^2 W_p - m_p\omega^2 W_p = -P(x,y,0), \tag{10}$$

where T, D and m_p are the dimensionless residual stress, plate bending stiffness, and plate mass, respectively. The relations to their dimensional analogs are

$$T = \frac{T^*}{\rho_f^* c^{*2} L^*}, \quad D = \frac{D^*}{\rho_f^* c^{*2} L^{*3}}, \quad m_p = \frac{\rho_p^* t_p^*}{\rho_f^* L^*}, \tag{11}$$

and $D^*=E^* t_p^{*3}/12(1-\nu^2)$, where E^*, ν, and t_p^* are Young's modulus, Poisson's ratio, and the plate thickness, respectively.

The fluid pressure P in Eq. (10) can be replaced by the boundary integrals in Eq. (8). Equation (10) is thus left with only one unknown: The plate displacement. If we assume a cross mode shape (as discussed in Ref. 4) in the y direction and use a modal expansion in the x direction, the plate displacement is given by

$$W_p = \sum_{k=1}^{N} a_k \Psi_k(x,y), \tag{12}$$

where N is the number of modes used in the x direction, a_k is the kth modal coefficient, and Ψ_k is the kth mode shape for a simply supported plate,

$$\Psi_k = \sin\frac{k\pi(x-\epsilon_1)}{\epsilon_2-\epsilon_1}\cos\frac{\pi y}{L_y(x)}, \tag{13}$$

where $\epsilon_1=l_1^*/L^*$ and $\epsilon_2=l_2^*/L^*$. l_1^* and l_2^* are the x coordinates of the plate [see Fig. 1(b)]. $L_y(x)$ is the normalized width function of the tapered plate.

In the above assumed modes, the orthogonality in the x direction no longer exists for a tapered plate since the y mode is also a function of x. Of course, the fluid-structure coupling also destroys the structural mode orthogonality. However, by using a Galerkin scheme, we can still solve for the modal coefficients of the plate displacement. By multiplying Eq. (10) by each mode of the plate and integrating over the plate surface, a matrix equation is obtained to solve for the modal expansion coefficients,

$$\boldsymbol{Ha} = \boldsymbol{f}, \tag{14}$$

where $\boldsymbol{H}$ is a square matrix of size $N \times N$, and $\boldsymbol{f}$ is an $N \times 1$ forcing vector. Their components are expressed as

$$H_{rs} = \omega^2 \sum_{lmn} \gamma_{lmn} \int_{\Gamma_{p0}} \Psi_s(x_0, y_0) \Phi_{lmn}(x_0, y_0) d \times \Gamma_{p0} \int_{\Gamma_p} \Psi_r(x,y) \Phi_{lmn}(x,y) d\Gamma_p + \beta_{rs},$$

$$\beta_{rs} = \int_{\Gamma_p} (D\nabla^4 \Psi_s - T\nabla^2 \Psi_s - m_p \omega^2 \Psi_s) \Psi_r d\Gamma_p,$$

$$f_r = \sum_{lmn} \gamma_{lmn} \int_{\Gamma_p} \Psi_r(x,y) \Phi_{lmn}(x,y) d\Gamma_p \times \int_{\Gamma_{s0}} P_s \frac{\partial \Phi_{lmn}(x_0, z_0)}{\partial x_0} d\Gamma_{s0}. \tag{15}$$

The matrix $\boldsymbol{\beta}$ is a full matrix due to lack of orthogonality in the modes for the tapered plate. The plate displacement can be obtained by solving the above algebraic matrix equation. Note that the mechanics of the piezoelectric material are neglected in this derivation. Hence we restrict ourselves to piezoelectric patches that are much thinner than the plate. The voltage output from the sensing material can be solved from its displacement by using the constitutive law of the piezoelectric material,[15]

$$V_{\text{out}}^* = \frac{t_e^* e_{31}^*}{2\epsilon_{33}^*} \frac{t_p}{A_e} \int_{A_e} \left(\frac{\partial^2 W_p}{\partial x^2} + \epsilon_e \frac{\partial^2 W_p}{\partial y^2} \right) dA_e, \tag{16}$$

where V_{out}^* is the voltage output from the piezoelectric layer, t_e^* is the thickness of the piezoelectric layer, and A_e is the electroded area. e_{31}^* and e_{32}^* are the piezoelectric coupling coefficients for the stress-charge form in the x and y directions, respectively ($\epsilon_e = e_{32}^*/e_{31}^*$).

III. APPROXIMATION OF SENSITIVITY AND RESONANT FREQUENCY

The sensitivity and bandwidth of a transducer can be directly read out from the frequency response curve. The modeling in the previous section provides a numerical approach to calculate the frequency response of the transducer. However, using such a frequency sweep analysis to determine the transducer sensitivity and bandwidth is usually time consuming and all the calculations would need to be repeated if any of the transducer geometrical dimensions changes. Clearly, it is desirable to find an alternative to this formulation for the transducer design because of the considerable computational cost. Figure 2 shows a cartoon of typical frequency response of this type of transducer devices after performing a frequency sweep analysis. Two key features are evident: (1) The sensitivities at low frequencies are asymptotic to a certain value. (2) The first resonant frequency ends the region of the flat response. The cartoon indicates that the sensitivity of the transducer can be approximated by the low frequency sensitivity asymptote and the bandwidth is determined by the first resonant frequency. For a constant-width plate with no residual stress, simplified expressions can be derived from the full formulation of Green's function method to approximate the low frequency sensitivity and the first resonant frequency. Next, we will show several approximation techniques for the transducer behavior followed by an evaluation of the accuracy of these approximations.

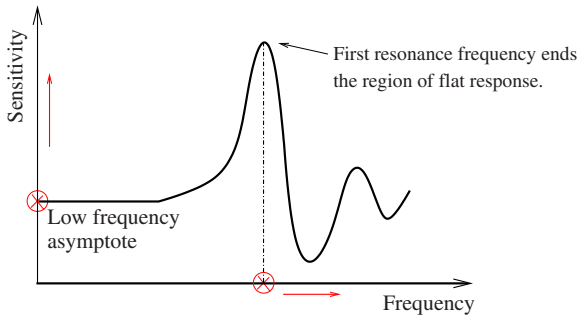


FIG. 2. (Color online) Cartoon of a typical transducer sensitivity output.

A. Low frequency sensitivity approximation

At very low frequencies ($\omega \to 0$), the weak coupling between the fluid and structure in Euler's equation [Eq. (3)] simplifies the fluid pressure in Eq. (8) to

$$P(\boldsymbol{x}) = -\int_{\Gamma_{s0}} \frac{\partial G(\boldsymbol{x}, \boldsymbol{x}_0)}{\partial x_0} P_s d\Gamma_{s0}. \tag{17}$$

At the limit $\omega \to 0$, the pressure P is reduced to P_s, the uniform pressure input. This result is also obvious if we directly solve the pressure from the Laplace equation $\nabla^2 P = 0$ with boundary conditions $P = P_s$ at $x = 0$ and $\partial P / \partial n = 0$ at all the other walls.

By substituting the pressure into the plate equation [Eq. (10)], the plate displacement can be solved for, and therefore, voltage sensitivity is obtained from Eq. (16),

$$V_{\text{out}}^* = \frac{t_e^* e_{31}^*}{\epsilon_{33}^*} \frac{t_p}{D} \sum_{k=1,3,5,\ldots}^{\infty} \frac{32 L_p^2 B_p^2 (k^2 B_p^2 + \epsilon_e L_p^2)}{k^2 \pi^6 (k^2 B_p^2 + L_p^2)^2}, \tag{18}$$

where $L_p (= \epsilon_2 - \epsilon_1)$ and B_p are the normalized length and width of the constant-width plate, respectively. Since at low frequencies the fluid pressure loading on the plate is equal to the input pressure, the dimensions of the duct do not affect the voltage sensitivity. Typically, the summation in Eq. (18) is dominated by the first three longitudinal modes.

B. First resonant frequency approximation

For a coupled system, there is no closed form solution to solve for the resonant frequency. The approximations discussed next make use of the structure of the matrix $\boldsymbol{H}$ in Eq. (14) under different fluid compressibility conditions to derive the simplified solutions for the resonant frequency.

1. Incompressible fluid

For an incompressible fluid ($c \to \infty$), γ_{lmn} in Eq. (15) no longer depends on frequency ω. The matrix Equation (14) can be rearranged to

$$(\omega^2 \boldsymbol{U}^I - \boldsymbol{V}^I)\boldsymbol{a} = \boldsymbol{f}, \tag{19}$$

where the $\boldsymbol{U}^I$ and $\boldsymbol{V}^I$ matrices are both derived from $\boldsymbol{H}$:

$$U^I_{rs} = \sum_{lmn} \gamma_{lmn} \int_{\Gamma_{p0}} \Psi_s \Phi_{lmn}(\boldsymbol{x}_0) d\Gamma_{p0} \int_{\Gamma_p} \Psi_r \Phi_{lmn}(\boldsymbol{x}) d\Gamma_p - \int_{\Gamma_p} m_p \Psi_s \Psi_r d\Gamma_p,$$

$$V^I_{rs} = -\int_{\Gamma_p} D\nabla^4 \Psi_s \Psi_r d\Gamma_p. \tag{20}$$

Equation (19) is a standard eigenvalue problem therefore, the resonant frequencies can be numerically solved using the $\boldsymbol{U}^I$ and $\boldsymbol{V}^I$ matrices from the following equation:

$$\det\left[\boldsymbol{U}^I(\boldsymbol{V}^I)^{-1} - \frac{1}{\omega^2}\boldsymbol{I}\right] = 0. \tag{21}$$

However, a further simplification can be made by noting that the matrix $\boldsymbol{V}^I$ is diagonal and the diagonal terms of the full matrix $\boldsymbol{U}^I$ are dominant over the off-diagonal ones. We introduce two approximate resonant frequencies by neglecting the intermodal couplings in the matrix $\boldsymbol{U}^I$. By using only the first diagonal terms U^I_{11} and V^I_{11}, a closed form approximation of the first resonant frequency can be obtained,

$$\omega^I_A = \sqrt{V^I_{11}/U^I_{11}}, \tag{22}$$

where

$$U^I_{11} = \frac{2(\int_{\Gamma_p} \Psi_1 \Phi_{000} d\Gamma_p)^2}{\lambda^2_{000}\epsilon_B\epsilon_H} + \frac{4(\int_{\Gamma_p} \Psi_1 \Phi_{010} d\Gamma_p)^2}{\lambda^2_{010}\epsilon_B\epsilon_H} + \cdots - m_p \int_{\Gamma_p} \Psi_1^2 d\Gamma_p,$$

$$V^I_{11} = -\int_{\Gamma_p} D\nabla^4 \Psi_1 \Psi_1 d\Gamma_p. \tag{23}$$

This approximation amounts to assuming a mode shape of the structure. Since U^I_{11} is obtained by summing up all the fluid modes, the approximation is designated as an all-mode incompressible approximation for convenience. As a further simplification, if we only consider the first fluid mode in Green's function (i.e., $l=m=n=0$), the approximate resonant frequency is simplified to

$$\omega^I_1 = \sqrt{\tilde{V}^I_{11}/\tilde{U}^I_{11}}, \tag{24}$$

where

$$\tilde{U}^I_{11} = -\frac{512 B_p^2 L_p^2}{\pi^6 \epsilon_B \epsilon_H (4 - L_p^2)^2}\left[\sin\left(\frac{\pi\epsilon_1}{2}\right) + \sin\left(\frac{\pi\epsilon_2}{2}\right)\right]^2 - \frac{1}{4} m_p L_p B_p,$$

$$\tilde{V}^I_{11} = -\frac{D\pi^4 (B_p^2 + L_p^2)^2}{4 B_p^3 L_p^3}. \tag{25}$$

This approximation is designated as a one-mode incompressible approximation. It is less accurate but much simpler than the all-mode incompressible approximation since it only accounts for the first fluid mode.

2. Compressible fluid

The effect of fluid compressibility becomes more significant at high frequencies when the dimension of the duct approaches a quarter of a wavelength (when ω approaches $\pi/2$). At those frequencies, the incompressible approximation does not hold and fluid compressibility should be included in the calculation.

For a compressible fluid, the matrix $\boldsymbol{U}$ in Eq. (19) is a function of frequency ω, which poses difficulty in solving the system resonant frequency since Eq. (19) is no longer a standard eigenvalue problem. However, intermodal coupling is still relatively weak and the first diagonal term in the $\boldsymbol{U}$ matrix is dominant. Similarly to ω^C_I from Eq. (22), the first resonant frequency can be approximated from

$$U^C_{11}(\omega^C_A) - V^C_{11}/(\omega^C_A)^2 = 0, \tag{26}$$

which, when expanded, has the form

$$\frac{2(\int \Psi_1 \Phi_{000} d\Gamma_p)^2}{(\omega^C_A)^2 - (\pi/2)^2} + \frac{4(\int \Psi_1 \Phi_{010} d\Gamma_p)^2}{(\omega^C_A)^2 - (\pi/2)^2 - (\pi/\epsilon_B)^2} + \frac{4(\int \Psi_1 \Phi_{001} d\Gamma_p)^2}{(\omega^C_A)^2 - (\pi/2)^2 - (\pi/\epsilon_H)^2} + \cdots + \frac{\epsilon_B\epsilon_H \int D\nabla^4 \Psi_1 \Psi_1 d\Gamma_p}{(\omega^C_A)^2} - \epsilon_B\epsilon_H m_p \int \Psi_1^2 d\Gamma_p = 0, \tag{27}$$

where all the terms are known except ω^C_A. This approximation is designated as an all-mode compressible approximation. The special structure of the left-hand side in Eq. (27) clearly indicates that it has poles at $\omega^C_A = 0, \pi/2, \sqrt{(\pi/2)^2 + (\pi/\epsilon_B)^2}, \ldots$ for different combinations of fluid modes l, m, and n. The first resonant frequency of the transducer ω^C_A, i.e., the smallest root of Eq. (27), lies between the smallest two poles 0 and $\pi/2$, i.e., $0 < \omega^C_A < \pi/2$, where $\pi/2$ is also the nondimensional resonant frequency for a rigid-walled duct. It is important to note that the nondimensional first resonant frequency of the fluid-structure system is always bounded on the upper side by $\pi/2$ and fluid compressibility is negligible if the first resonant frequency of the transducer is much smaller than $\pi/2$. For a high-order polynomial equation like Eq. (27), there are no efficient analytical root-finding schemes available and the Newton–Raphson method is used here to locate the roots. To initiate the

Newton–Raphson scheme, an initial guess has to be provided and in this problem, since we are only interested in the smallest root, it should be chosen in the predicted range $(0 < \omega_A^C < \pi/2)$ to ensure the convergence of the procedure.

If we only consider the first fluid modes in the x, y, and z directions, the approximate frequency will be

$$\tilde{U}_{11}^C(\omega_1^C) - \tilde{V}_{11}^C/(\omega_1^C)^2 = 0, \tag{28}$$

where

$$\tilde{U}_{11}^C = \frac{512 B_p^2 L_p^2}{\pi^4 \epsilon_B \epsilon_H [4(\omega_1^C)^2 - \pi^2](4 - L_p^2)^2} \times \left[\sin\left(\frac{\pi\epsilon_1}{2}\right) + \sin\left(\frac{\pi\epsilon_2}{2}\right) \right]^2 - \frac{1}{4} m_p L_p B_p,$$

$$\tilde{V}_{11}^C = -\frac{D\pi^4 (B_p^2 + L_p^2)^2}{4 B_p^3 L_p^3}. \tag{29}$$

This approximation is designated as a one-mode compressible approximation. All these expressions are readily evaluated for a variety of geometries as embodied by L_p, B_p, and any of the other parameters.

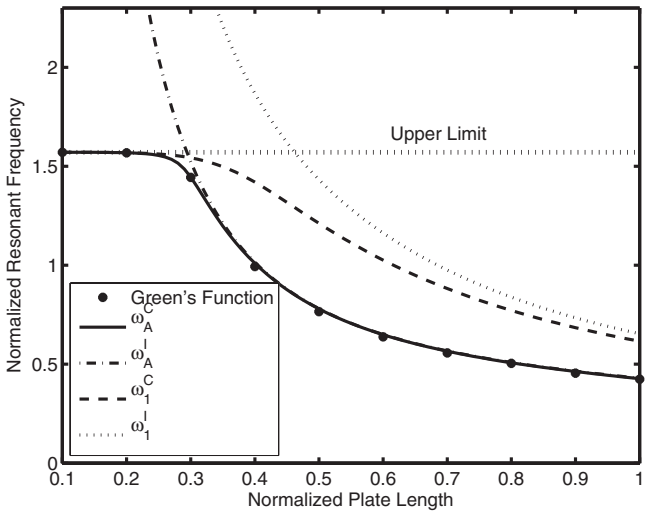


FIG. 3. The comparison of four first resonant frequency approximations with the exact results calculated in Green function's formulation when varying the plate length. The other dimensions of the system are $L^*=0.032$ m, $\epsilon_B=0.8$, $\epsilon_H=1.2$ [defined beneath Eq. (7)], $B_p=0.4$, and $t_p=0.016$. 50 x modes, 50 y modes, and 100 z modes are used in Green's function formulation. The upper limit of the first resonant frequency for the transducer system is $\pi/2$. Note that the dots are the discrete points from Green's function calculation and the solid line is a separate approximation, not an interpolation of the exact results.

IV. RESULTS AND DISCUSSION

The simplified formulas provided in the previous section to approximate the low frequency sensitivity and the first resonant frequency are evaluated here by comparing to the exact solutions from full Green's function formulation. By using these approximations, the dependence of the low frequency sensitivity and the fundamental resonant frequency on the transducer geometry is also investigated. With this knowledge, the transducer dimensions can be optimized to achieve the desired performance.

A. Accuracy of the approximations

1. Accuracy of the resonant frequency

In Fig. 3, four approximations to the natural frequency of the coupled system are compared to the exact results from Green's function formulation. As an example, Fig. 3 shows the dependence of the system natural frequency on the plate normalized length L_p. The transducer is modeled as a fluid-filled rectangular duct coupled with a constant-width aluminum plate on the lower wall. The material properties are fixed in the calculations: $\rho_f^*=1000$ kg/m^3, $c^*=1500$ m/s, $E^*=71$ GPa, $\nu=0.3$, and $\rho_p^*=2700$ kg/m^3. In order to obtain ω_1^I, we have a closed form expression [Eq. (24)]. For ω_1^C, we solve a polynomial equation [Eq. (28)], a slightly higher computational burden. Solving for ω_A^I requires first summing over all of the fluid modes, but once that is completed a closed form solution [Eq. (22)] is used for the final computation. Finding ω_A^C is somewhat more involved as the matrix elements depend on the frequency. The algorithm for finding ω_A^C is as described above [Eq. (26)]. The exact results are obtained by performing a frequency sweep in Green's function formulation for a particular selection of the plate length and then selecting the first peak frequency of the displacement response. This procedure is then repeated for each plate length.

In Fig. 3, it can be seen that both incompressible approximations fail to predict the correct trend in the resonant frequency below a normalized plate length of 0.3. The incompressible approximations fail at these plate lengths because the fluid compressibility becomes more important than the plate compliance. As a result, the first resonant frequency approaches its limit $(\pi/2)$ at these plate lengths. For longer plate lengths, ω_1^I gives the correct trend while ω_A^I accurately predicts the resonant frequency. The all-mode compressible approximation (ω_A^C) predicts the correct resonant frequency over the entire nondimensional length range within 3% error, while ω_1^C is only quantitative for normalized plate lengths below 0.3. All of the approximations still neglect intermodal coupling.

In the sense of accuracy, the approximation ω_A^C gives the best overall prediction on the first resonant frequency of the system, however, solving the resonant frequency in this case involves finding roots from a high-order polynomial equation, which makes the computational cost comparatively high and does not provide a closed form design formula. The most favorable approximation should have both good accuracy and high efficiency. Equation (22) (ω_A^I) is closed form and accurate for frequencies below $\pi/2$ (for $L_p > 0.3$). Hence, this approximation can be used in this regime. For plate lengths at which ω_A^I rises above $\pi/2$, we realize that from our analysis of poles and zeros of Eq. (26), the first resonant frequency must lie on or near $\pi/2$. The results of these simulations point to an efficient calculation method to estimate the first resonant frequency:

TABLE I. Fundamental resonant frequency of the coupled system.

Geometry	ω_A^I (rad/s)	Eigenvalue (rad/s)	ω_A^C (rad/s)	Green's function (rad/s)	Error (%)
L^*, B^*	5 761.7	5 730.3	5 711.4	5 724.0	0.22
$10L^*, B^*$	546.6	490.1	540.4	490.1	10.26
$L^*, B^*/10$	362 967.0	338 801.9	72 991.7	72 803.3	0.26

$$\omega = \begin{cases} \omega_A^I & \text{when } \omega_A^I < \pi/2 \\ \pi/2 & \text{for other smaller non-dimensional plate lengths.} \end{cases} \tag{d30}$$

Next, we discuss whether or not these results will be affected by varying the geometrical dimensions or material properties. We found that the approximation was correct when the coupling between the fluid and the structure was weak or the aspect ratio of the plate (i.e. length to width) was close to 1. Under the incompressible assumption, the approximation ω_A^I in Sec. III is directly obtained by neglecting the cross terms in the matrix $\boldsymbol{U}^I(\boldsymbol{V}^I)^{-1}$ in Eq. (21) (note that $\boldsymbol{U}^I$ is a full matrix but $\boldsymbol{V}^I$ is diagonal). We can achieve sufficient accuracy from this approximation if the fluid-structure coupling is weak, so that the matrix $\boldsymbol{U}^I$ has dominant diagonal terms (and the first diagonal term is the largest element). In that case, one would expect that the mode shape of the structure remains nearly sinusoidal in the coupled system. Further, we found that the structure modal stiffness matrix $\boldsymbol{V}^I$ could also help build the strength of the first diagonal term in the matrix $\boldsymbol{U}^I(\boldsymbol{V}^I)^{-1}$ if it has unique diagonal terms. For example, if the first diagonal term in the matrix $\boldsymbol{V}^I$ is much smaller than the other diagonal terms, multiplying $\boldsymbol{U}^I$ by $(\boldsymbol{V}^I)^{-1}$ will make the first column of the resulting matrix $\boldsymbol{U}^I(\boldsymbol{V}^I)^{-1}$ stronger than that of the matrix $\boldsymbol{U}^I$. As a result, the dominance of the first diagonal term over all the other non-first-column terms is increased.

The diagonal terms of the matrix $\boldsymbol{V}^I$ are determined by the stiffness in the structure longitudinal and transverse directions. The kth diagonal term represents the stiffness of the kth longitudinal mode and the first cross mode. For a constant-width plate, the diagonal terms can be written as

$$V_{kk}^I = \frac{D\pi^4}{4}\frac{(L_p^2 + k^2 B_p^2)^2}{B_p^3 L_p^3}. \tag{31}$$

We can see that the values of the diagonal terms increase as the longitudinal modal number k increases, however, the difference between the different diagonal terms is dependent on the relative stiffness in the transverse direction and longitudinal direction. For a long and narrow plate ($B_p \ll L_p$), the structure is very stiff in the transverse direction. Varying k in the above equation does not significantly change the values of the diagonal terms, meaning that the difference between the first several diagonal terms is small before kB_p reaches the same value as L_p and the values of the first elements of the matrix $\boldsymbol{V}^I$ are similar—hence, no term is dominant. An example of this type of structure is given next.

From the above discussion we know that one scenario that this approximation could lead to a bigger error is when the fluid-structure coupling is very strong and the structure is very stiff in the transverse direction, as we will see from the following example. When the coupling is strong, the fundamental structural mode shape for the coupled system will also shift from the *in vacuo* plate mode. Table I shows the first resonant frequency of the coupled system with the different geometries computed in four different ways: the approximation ω_A^I, the approximation ω_A^C, directly solving the eigenvalues from Eq. (19), and the full Green function formulation. The last column is the error between ω_A^C and the exact result from the full formulation. For convenience, the dimensional values are used to describe the geometry of the system. The duct size in the first analysis is $L^*=0.032$ m, $B^*=0.0254$ m, and $H^*=0.0381$ m (corresponding to $\epsilon_B=0.4$, $\epsilon_H=1.2$). The flexible plate covers the total area of the duct wall at $z=0$, so that $L_p^*=L^*$ and $B_p^*=B^*$. While the matrix $\boldsymbol{U}^I$ has some moderate cross-coupling terms (by no means dominated by the diagonal terms), after multiplication by $(\boldsymbol{V}^I)^{-1}$, the first diagonal term of the resulting matrix is dominant. As a result, the approximations match well with the exact resonant frequency from full Green's function formulation. For the next two examples, the ratios of the duct width to length and the plate width to length are both reduced. The results are shown in the second and third rows of Table I. By using these two geometries, the fluid-structure coupling (as embodied in $\boldsymbol{U}^I$) is similar to the first example, but the transverse stiffness of the plate is higher. From the results, we can see larger errors between ω_A^I and the true eigenvalues, as is expected. Note that in the third example, the eigenvalue calculated from Eq. (21) varies by 365% from the first resonant frequency of the system (see the result for Green's function in the fifth column) because the incompressible assumption is no longer valid in this case. The first *in vacuo* resonant frequency of the plate is much larger than that of the duct. From our previous analysis, we know that the normalized resonant frequency of the coupled system must lie on or near $\pi/2$ (corresponding to 73 631.1 rad/s for a 0.032 m long duct). For a system such as this, ω_A^C still gives the excellent prediction of the system resonant frequency with only 0.26% error. These two cases are presented here to show conditions where the approximations produce larger errors in predicting the resonant frequency of the system. For a more realistic transducer design, those geometries are not likely to be chosen since these choices do not optimize the bandwidth and sensitivity of the transducer because of the low resonant frequency or low sensitivity. Figures 4–6 show the comparison of the approximation ω_A^C with the exact

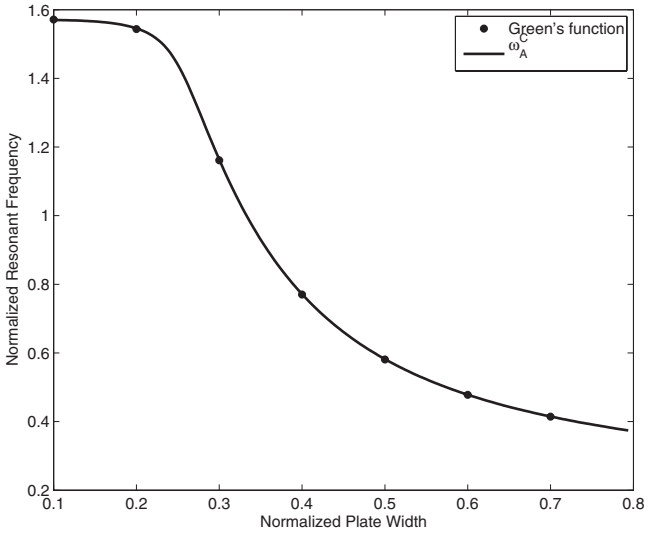


FIG. 4. The comparison of the approximation ω_A^C with the exact results calculated in Green's function formulation when varying the plate width. The other dimensions of the system are $L^*=0.032$ m, $\epsilon_B=0.8$, $\epsilon_H=1.2$, $L_p=0.5$, and $t_p=0.016$.

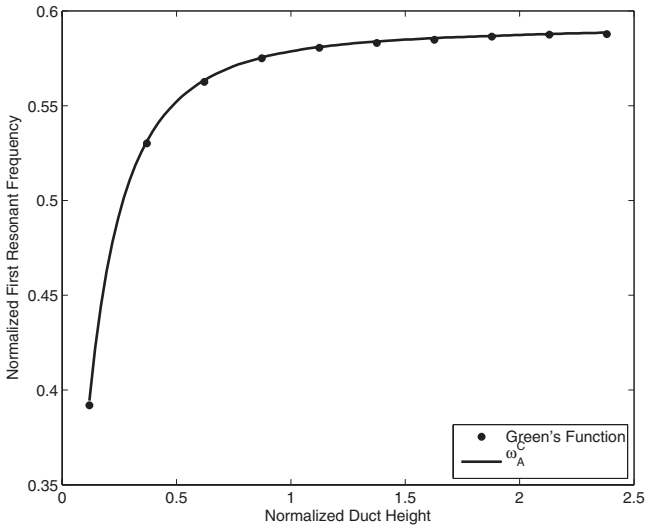


FIG. 6. The comparison of the approximation ω_A^C with the exact results calculated in Green's function formulation when varying the duct height. The other dimensions of the system are $L^*=0.032$ m, $\epsilon_B=0.8$, $L_p=0.5$, $B_p=0.5$, and $t_p=0.016$.

results in three other cases: varying the plate width, duct width, and duct height. All the results indicate very good agreement between the approximation ω_A^C and the exact solutions when varying the geometry of the system. The fluid and structural material properties do not appear in the off-diagonal terms of the matrix $\boldsymbol{U}^I$; therefore, the variation of the material properties have less influence on the accuracy of the approximation.

The computing times for the approximations and the exact solution are discussed next. For a sweep of 100 nondimensional plate lengths as shown in Fig. 3, the recommended method [Eq. (30)] requires 4 s to generate the full curve. Using the approximation ω_A is more time consuming, taking 40 s, while the exact solution takes 372 s for one plate length, so that 100 plate lengths need 37 200 s for Green's function formulation to complete the computation. From this comparison, we can see that the recommended approach is nearly 10 000 times faster than the exact solution. The enormous time saving by using these approximations in calculating the resonant frequency is beneficial for the transducer design. Note that the computing time for each approach does not change with the dimension or material properties of the system.

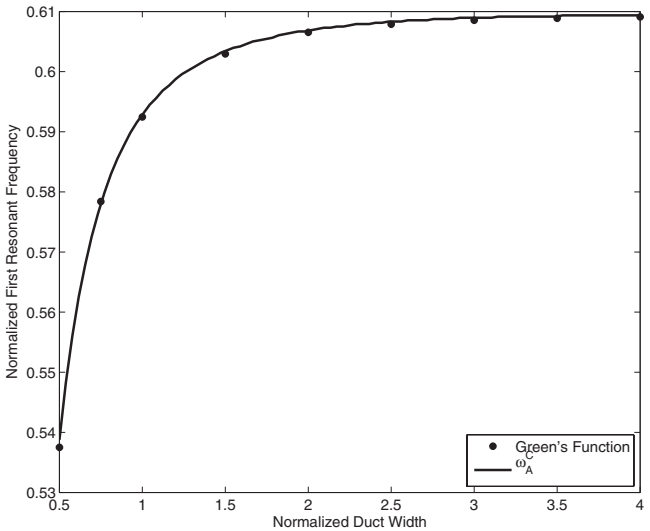


FIG. 5. The comparison of the approximation ω_A^C with the exact results calculated in Green's function formulation when varying the duct width. The other dimensions of the system are $L^*=0.032$ m, $\epsilon_H=1.2$, $L_p=0.5$, $B_p=0.5$, and $t_p=0.016$.

2. Accuracy of the sensitivity

Another parameter to characterize the transducer performance is the sensitivity. As might be expected, Fig. 7 shows the transducer sensitivities predicted by the low frequency sensitivity approximation [Eq. (18)] and the Green's function formulation are the same. The sensitivity from the full formulation is obtained by sweeping over frequencies at each plate length and taking the sensitivity value where the trans-

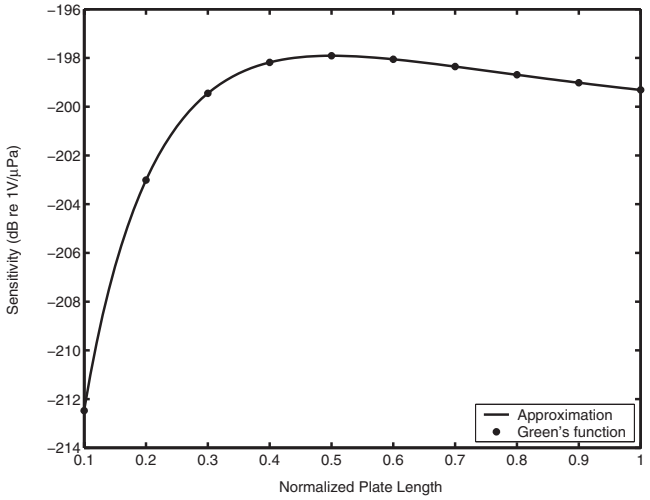


FIG. 7. The comparison of the transducer sensitivity approximation with the exact results calculated in Green's function formulation. The geometry and material properties are the same as in Fig. 3.

ducer has a flat frequency response. The sensitivity approximation with 50 plate modes uses less than 10^{-4} s to calculate the sensitivity, while Green's function formulation needs 35 s for one frequency.

B. Dimensional parameters affecting low frequency sensitivity

In Eq. (18), the transducer sensitivity at low frequencies is shown to be linearly proportional to a term $(t_e^* e_{31}^*/\epsilon_{33}^*) \times (t_p/D)$. All the quantities in that term are related to the material constants and normalized thickness of the laminate plate, which are fixed in this discussion. The rest of the terms in Eq. (18) are functions of the plate length and width only. The dependence of the transducer sensitivity on these two physical dimensions is shown in Fig. 8(a). Each contour line represents a constant voltage sensitivity for the combinations of the plate length and width. Figure 8(b) shows the change of the transducer sensitivity with the plate length and width, respectively. The curves are taken from Fig. 8(a) by fixing the plate width B_p (top figure) or plate length L_p (bottom figure). The sensitivity quickly rises with the plate size before it starts to saturate. Ideally, a large plate is preferred to produce a high sensitivity, but the bandwidth of the transducer is negatively impacted by increasing the size, as we will see later in this section.

C. Dimensional parameters affecting the first resonant frequency

All the approximations derived in Sec. III B show that the normalized resonant frequency ω is a function of nondimensional geometrical variables such as the plate length L_p, width B_p, and duct width ϵ_B and height ϵ_H. We notice that ω does not depend on the duct length L^*, but its dimensional analog ω^* is inversely proportional to L^* through the relation $\omega^* = \omega c^*/L^*$ [revisit Eq. (1)]. Therefore, a high dimensional resonant frequency ω^* can be achieved by two means: Decreasing the length of the duct L^* or increasing the normalized resonant frequency ω. Figure 9 shows the dependence of the normalized first resonant frequency on the plate dimensions. These example calculations employ an all-mode compressible approximation [Eq. (26)] with duct dimensions fixed ($\epsilon_B = 0.8$, $\epsilon_H = 1.2$). Each contour line in Fig. 9(a) represents a constant resonant frequency for the combinations of the plate length and width, so that every point along the contour line has the same value. Figure 9(b) shows the change of the resonant frequency with the plate length and width, respectively. Reducing the plate size increases the normalized fundamental resonant frequency of the system until it reaches an asymptotic value $\pi/2$, which is also the normalized resonant frequency of the rigid-walled duct.

Interestingly, the first resonant frequency of the transducer system never exceeds $\pi/2$, which is mainly determined by the boundary conditions at $x=0$ and $x=L$. Whether or not the system resonance can reach that limit is determined by the plate dimensions. Increasing the duct width or height does not change the limit of the first resonant frequency. Hence, the influence of the duct width on the resonant frequency is not as prominent as that of the duct length, as shown in Fig. 5. The narrower plate and the wider duct produce a higher resonant frequency of the system. The resonant frequency also increases with the duct height, and the pattern is shown in Fig. 6. For a coupled system, increasing the duct height reduces fluid mass loading on the plate, thus increasing the resonant frequency.

D. The performance of the tapered-plate transducer

For a tapered plate, the structure modes are coupled because the orthogonality of the selected structural mode shapes no longer holds. To directly apply the approximations introduced in Sec. III to the tapered-plate transducer, we have to make the further assumption that structural intermodal coupling can be neglected as well. However, we found that such a strong assumption leads to inaccurate results in both the sensitivity and bandwidth predictions when compared to the results from Green's function formulation. A better approximation for the tapered-plate transducer is to incorporate structure modal coupling into the calculation.

1. Sensitivity approximation

In the sensitivity approximation, Eq. (17) is still valid for a tapered-plate transducer. Substituting Eq. (17) into the plate equation yields a matrix equation to solve for the plate displacement. Due to structure intermodal coupling, we do not have an explicit solution for the plate displacement, and therefore, voltage sensitivity cannot be written as simple as Eq. (18). However, the first three structural modes still contribute to most of the voltage sensitivity. Using only those modes is sufficient in the sensitivity approximation and the order of the matrix equation is greatly reduced as well.

2. Resonant frequency approximation

In the first resonant frequency approximation, the matrix $\boldsymbol{V}^I$ in Eq. (19) becomes a full matrix due to lack of orthogonality in the structure modes. As a result, all the resonant frequency approximations which utilize the diagonal property of the matrix $\boldsymbol{V}^I$ cannot be directly extended to apply to the tapered-plate transducer. We are left with calculating the first resonant frequency of a tapered-plate transducer by using Eq. (19). Alternatively, we can first find the orthogonal modes for the tapered plate and then use these modes to expand the plate displacement. The orthogonality of the modes will ensure the diagonality of the matrix $\boldsymbol{V}^I$.

3. Accuracy and comparison

Figure 10 shows the approximate resonant frequency (a) and low frequency sensitivity (b) of the tapered-plate transducer as the plate length changes. The resonant frequency is calculated under the incompressible assumption by using Eq. (14). Any frequency greater than $\pi/2$ is replaced by $\pi/2$ because of fluid compressibility. This approximation slightly overpredicts the resonant frequency of the transducer and the maximum error is 9% by comparing the results at several plate lengths to Green's function. To illustrate the advantage of using the tapered plate, Fig. 10(b) also shows the sensitivity of the constant-width plate transducers in comparison. Note that the equivalent width of the constant-width plate is

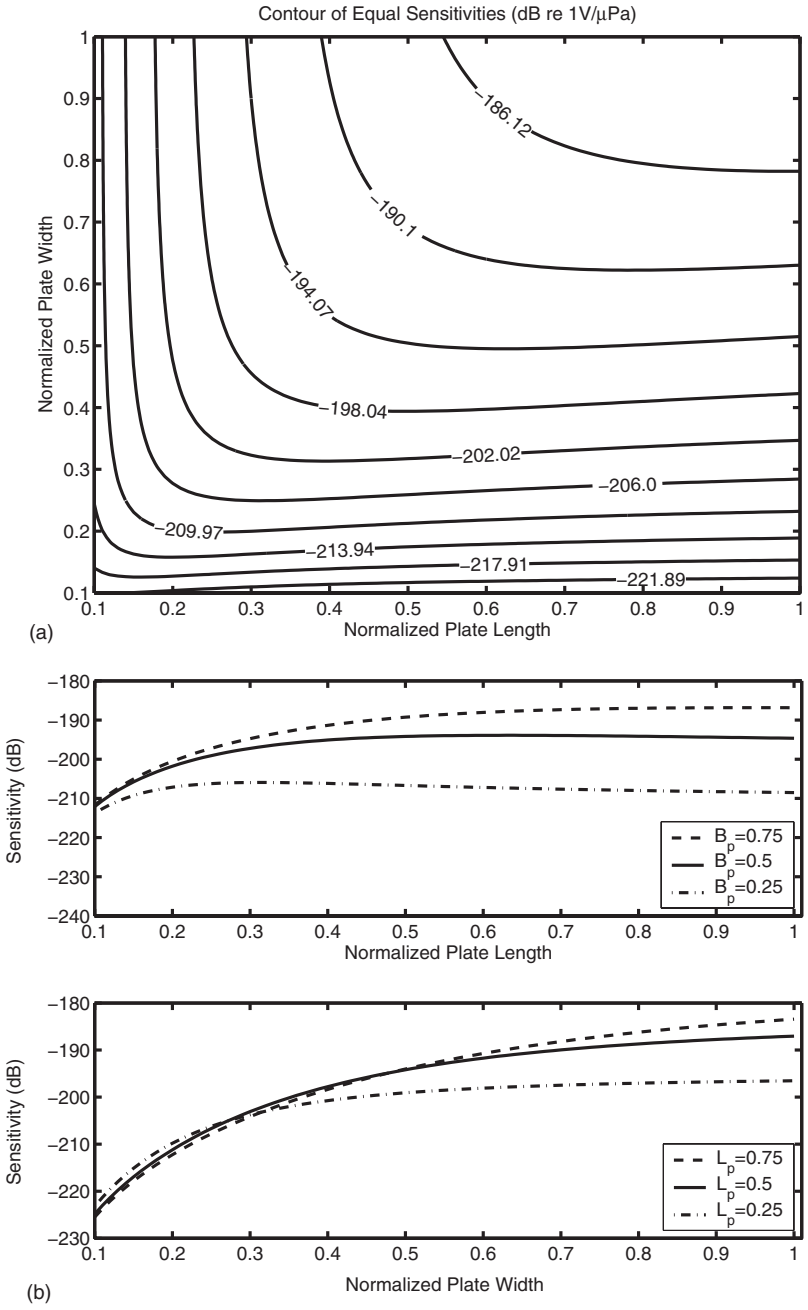


FIG. 8. The dependence of the transducer voltage sensitivity on the plate length and width. The plate normalized thickness is fixed to 0.016. (a) The contour lines of equal sensitivities for different plate lengths and widths. (b) Slices through constant plate width B_p and length L_p, respectively. Top: The dependence of the voltage sensitivity on the plate length for three different plate widths; bottom: The dependence of the voltage sensitivity on the plate width for three different plate lengths.

obtained by equating the area of the tapered plate with the constant-width plate. When the normalized plate length is between 0.3 and 0.7, the tapered-plate transducer has both a higher sensitivity and a broader bandwidth than a constant-width plate transducer with equivalent width. It is also interesting to find that the sensitivity and the first resonant frequency of the tapered-plate transducer are bounded by those of two constant-width plate transducers, the widths of which are the same as the widths of the tapered plate at its narrow and wide ends.

E. Overall design considerations

From the previous discussion of the dependence of the sensitivity and first resonant frequency on the transducer geometrical variables, we can conclude that the designs to improve the transducer sensitivity always reduce the bandwidth, and vice versa. Using Figs. 8 and 9, an optimal transducer dimension can be obtained by finding the range of normalized plate lengths and widths at which both the sensitivity and bandwidth are relatively large. For example, when the plate normalized length is 0.5 and the normalized width is 0.3, Fig. 8 shows that the approximate transducer sensitivity is around −202 dB (1 V/μPa) and Fig. 9 gives a normalized resonant frequency of 1.2, which corresponds to a 9 kHz bandwidth (using $\omega^*=\omega c^*/L^*$ where $L^*=0.032$ m and $c^*=1500$ m/s). At that plate length, a tapered-plate transducer has better overall performance than a constant-width plate transducer, according to Fig. 10. By substituting these plate dimension values into Green's function model, we obtain a similar frequency response (Fig. 11) as shown in the cartoon (Fig. 2). The sensitivity and bandwidth approximations match well with exact results. The tapered-plate transducer gives a sensitivity of around −200 dB (1 V/μPa)

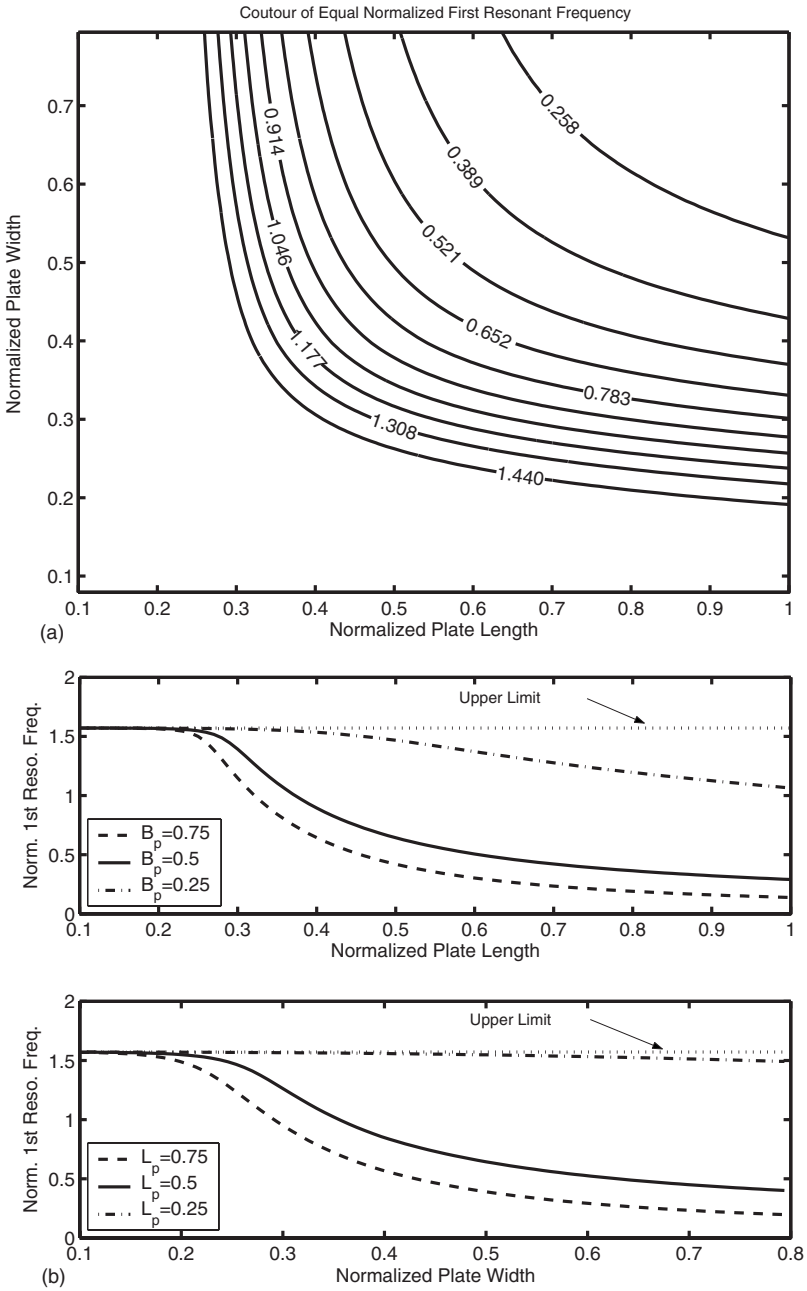


FIG. 9. The dependence of the first resonant frequency of the transducer on the plate length and width. The other dimensions are fixed: $\epsilon_B=0.8$, $\epsilon_H=1.2$, and $t_p=0.016$. The fluid is compressible. (a) The contour lines of equal resonant frequencies for different combinations of the plate length and width. (b) Slices through the constant plate width B_p and length L_p, respectively. Top: the dependence of the first resonant frequency on the plate length for three different plate widths; bottom: the dependence of the first system resonant frequency on the plate width for three different plate lengths.

in the 10 kHz frequency range, with both values higher than those of a constant-width plate transducer with equivalent width. The responses of the other two constant-width plate transducers with maximum and minimum widths provide the limits of the bandwidth and sensitivity for the tapered-plate transducers in the same width range. With the use of our quantitative approximations of transducer sensitivity and first resonant frequency, the transducer performance can be efficiently optimized using Figs. 8–10.

V. CONCLUSIONS

The analysis of a new trapped fluid electroacoustic transducer is introduced in this paper. The fluid-structure coupled system is modeled using a boundary integral method. Exact and approximate solutions for the structure acoustic response of an enclosed fluid space are developed and applied to the design of a trapped-fluid transducer. The frequency response of the transducer, shown in Fig. 11, consists of a low frequency region of nearly constant sensitivity terminated by the first resonance frequency. We show that, as expected, the low frequency sensitivity is well approximated by a quasistatic approximation [Eq. (16)]. We show that the sensitivity depends on the compliance of the plate and the plate geometry. Through an analysis of the poles and zeros of the approximate characteristic equation for the fluid-structure system [Eq. (27)], we show that the upper limit of the resonant frequency is controlled by the length of the duct (L^*) and fluid compressibility. When plate stiffness becomes important, the first resonant frequency is controlled by the mass loading of the fluid and the plate stiffness. Understanding how the plate stiffness and the geometry of the system affect the sensitivity and the resonant frequency enables, through the use of our quantitative approximations, a rapid exploration of the design space to improve the sensitivity without

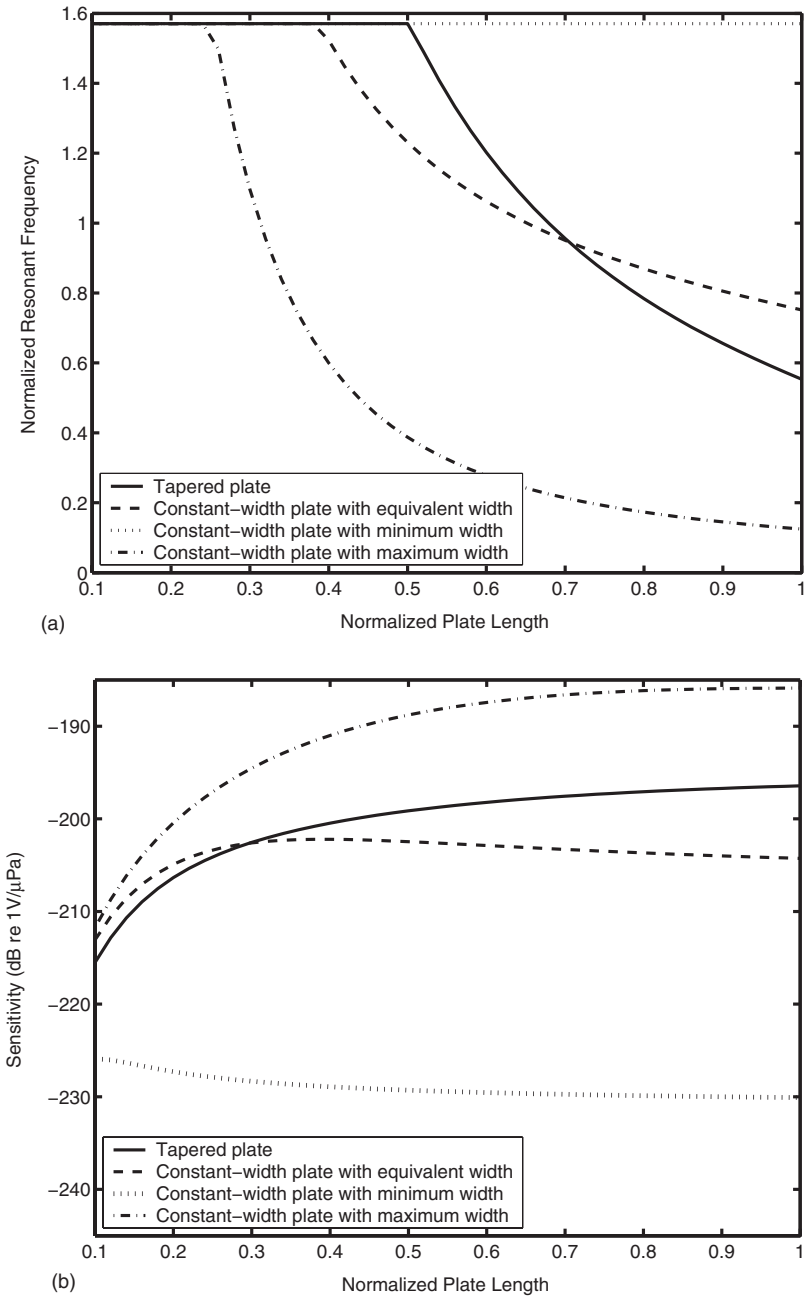


FIG. 10. The comparison of the characteristics of the tapered-plate transducer and constant-width plate transducer. (a) The dependence of the resonant frequency on the plate length; (b) The dependence of the sensitivity on the plate length. Duct size: L=0.032 m, ϵ_B=0.8, ϵ_H=1.2; plate size: t_p=0.016, B_p=0.08–0.8 (the width is exponentially tapered).

significantly compromising the bandwidth. One counterintuitive result is that increasing the duct height increases the resonant frequency of the system by reducing the mass loading of the fluid (of course, this has practical limits in terms of size and non-ideal loading that would be experimentally seen). We also show that the tapered-plate transducer gives better performance than the constant-width plate transducer. The approximations on the constant-width plate transducer can be used to predict the upper and lower bounds of the sensitivity and bandwidth of the tapered-plate transducer. The optimized transducer achieves a sensitivity of −200 dB (1 V/μPa) in the 10 kHz frequency range by using the sensitivity and bandwidth approximation plots (Figs. 8–10), and the performance will approximately maintain the same when operated underwater and in air. This sensitivity is similar to the current state of art piezoelectric hydrophones such as the symbal PZT flextensional hydrophone (13 mm device with a sensitivity of −205 dB in 20–65 kHz bandwidth)[16] and PZT hollow sphere hydrophone (1 mm device with −215 dB sensitivity and 5–90 kHz bandwidth)[17] but the bandwidth of our device is smaller. Apart from the dimension of the system, the practical transducer sensitivity could also be influenced by the electronics of the sensing circuits and the noise generated from the readout circuitry and the mechanical device. When the transducer functions as a hydrophone, the hydrostatic pressure balancing technique should also be investigated in order to fully evaluate the transducer performance.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research.

[1]R. D. White, L. Cheng, and K. Grosh, "Capacitively sensed microma-

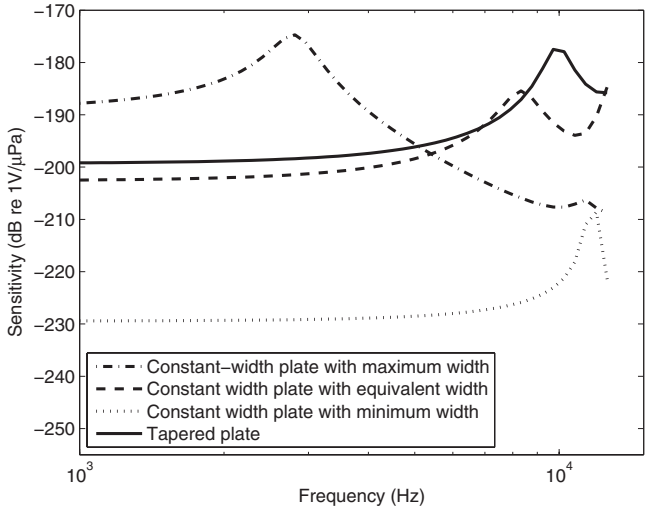


FIG. 11. Optimized performance of the tapered-plate transducer in comparison with the constant-width plate transducer. Duct size: $L=0.032$ m, $\epsilon_B=0.8$, $\epsilon_H=1.2$; plate size: $L_p=0.5$, $t_p=0.016$, $B_p=0.08-0.8$ (the width is exponentially tapered). 50 x modes, 50 y modes, and 100 z modes are used in Green's function formulation.

chined hydrophone with viscous fluid structure coupling," Proc. SPIE **5718**, 89–100 (2005).
[2]J. Bernstein, "A micromachined condenser hydrophone," *Solid-State Sensor and Actuator Workshop, 1992, Fifth Technical Digest* (IEEE, New York, 1992), pp. 161–165.
[3]S. Ramamoorthy, K. Grosh, and J. M. Dodson, "A theoretical study of structural acoustic silencers for hydraulic systems," J. Acoust. Soc. Am. **111**, 2097–2108 (2002).
[4]S. Ramamoorthy, K. Grosh, and T. G. Nawar, "Structural acoustic silencers–design and experiment," J. Acoust. Soc. Am. **114**, 2812–2824 (2003).
[5]L. X. Huang, "A theoretical study of duct noise control by flexible panels," J. Acoust. Soc. Am. **106**, 1801–1809 (1999).
[6]L. Huang, Y. S. Choy, R. M. C. So, and T. L. Chong, "Experimental study of sound propagation in a flexible duct," J. Acoust. Soc. Am. **108**, 624–631 (2000).
[7]L. Huang, "A theoretical study of passive control of duct noise using panels of varying compliance," J. Acoust. Soc. Am. **109**, 2805–2814 (2001).
[8]J. B. Lawrie and R. Kirby, "Mode-matching without root-finding: Application to a dissipative silencer," J. Acoust. Soc. Am. **119**, 2050–2061 (2006).
[9]P. Dallos, "Introduction," in *The Cochlea*, edited by P. Dallos, A. Popper, and R. Fay (Springer-Verlag, New York, 1996), pp. 1–43.
[10]G. Zhou, L. Bintz, D. Z. Anderson, and K. E. Bright, "A life-sized physical model of the human cochlea with optical holographic readout," J. Acoust. Soc. Am. **93**, 1516–1523 (1993).
[11]T. P. Lechner, "A hydromechanical model of the cochlea with nonlinear feedback using pvf(2) bending transducers," Hear. Res. **66**, 202–212 (1993).
[12]R. D. White and K. Grosh, "Microengineered hydromechanical cochlear model," Proc. Natl. Acad. Sci. U.S.A. **102**, 1296–1301 (2005).
[13]C. R. Steele and L. A. Taber, "Comparison of WKB calculations and experimental results for three-dimensional cochlear models," J. Acoust. Soc. Am. **65**, 1007–1018 (1979).
[14]M. Junger and D. Feit, *Sound, Structures and Their Interaction* (MIT Press, Cambridge, 1986).
[15]C. K. Lee, "Theory of laminated piezoelectric plates for the design of distributed sensors actuators. 1. Governing equations and reciprocal relationships," J. Acoust. Soc. Am. **87**, 1144–1158 (1990).
[16]J. Zhang, A. Hladky-Hennion, W. Hughes, and R. Newnham, "Modeling and underwater characterization of cymbal transducers and arrays," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **48**, 560–568 (2001).
[17]S. Alkoy, J. Cochran, and R. Newnham, "Miniature hydrophones from hollow ceramic spheres," in Proceedings of the 11th IEEE Symposium on Applications of Ferroelectrics, 1998, Montreux, Switzerland, August 24–27, pp. 345–348.

Modeling of piezoelectric transducers with combined pseudospectral and finite-difference methods

E. Filoux,[a] S. Callé, D. Certon, M. Lethiecq, and F. Levassort
Université Francois-Rabelais, INSERM U930-CNRS FRE 2448, UFR de Médecine, 10 Bd Tonnellé, BP 3223, 37032 Tours Cedex 1, France

(Received 6 July 2007; revised 10 March 2008; accepted 10 March 2008)

A new hybrid finite-difference (FD) and pseudospectral (PS) method adapted to the modeling of piezoelectric transducers (PZTs) is presented. The time-dependent equations of propagation are solved using the PS method while the electric field induced in the piezoelectric material is determined through a FD representation. The purpose of this combination is to keep the advantages of both methods in one model: the adaptability of FD representation to model piezoelectric elements with various geometries and materials, and the low number of nodes per wavelength required by the PS method. This approach is implemented to obtain an accurate algorithm to simulate the propagation of acoustic waves over large distances, directly coupled to the calculation of the electric field created inside the piezoelectric material, which is difficult with classical algorithms. These operations are computed using variables located on spatially and temporally staggered grids, which attenuate Gibbs phenomenon and increase the algorithm's accuracy. The two-dimensional modeling of a PZT plate excited by a 50 MHz sinusoidal electrical signal is performed. The results are successfully compared to those obtained using the finite-element (FE) algorithm of ATILA™ software with configurations spatially and temporally adapted to the FE requirements. The cost efficiency of the FD-PS time-domain method is quantified and verified.
 [DOI: 10.1121/1.2903876]

PACS number(s): 43.38.Fx, 43.58.Ta [AJZ] Pages: 4165–4173

I. INTRODUCTION

Pseudospectral (PS) methods were introduced by Kreiss and Oliger.[1] Consisting in calculating the derivatives (in our case, spatial derivatives) in the Fourier domain, they were first used for seismic waves propagation modeling.[2,3] Wojcik *et al.*[4] then applied the PS method to acoustic wave propagation. The main advantages of the PS numerical algorithm compared to other algorithms [such as finite-difference (FD) or finite element (FE) methods] is that it has a better numerical stability and does not require a large number of nodes[2,5,6] per wavelength. For waves propagating in fairly inhomogeneous media, it was demonstrated that the Fourier pseudospectral method requires only two nodes per minimum wavelength to achieve exact spatial derivatives.[6,7]

The pseudospectral time-domain (PSTD) algorithm, which solves time-dependent partial differential equations, often combines conventional Fourier pseudospectral method and perfectly matched layers (PMLs), introduced by Berenger[8] in the case of electromagnetic waves propagation. PML absorbing boundary conditions are matched absorbing conditions which avoid reflections on the numerical grid boundaries and counter the wraparound effect from the fast Fourier transform (FFT).

A specific algorithm based on a PS method with PML boundary conditions was previously developed for the propagation of elastic waves in heterogeneous media like biological tissues[9] or metallic materials,[10] and for the calculation of the radiation pattern of lens-focused transducers.[11] The aim of this paper is to use and extend this theory to piezoelectric materials, integrating the electrical excitation in order to simulate both the piezoelectric transducer vibration and the resulting wave propagation in the surrounding media using a single model, while keeping the different advantages described above.

Various numerical models have already been developed, but only a few studies have been performed on the simulation of piezoelectric media using FD methods. Lloyd and Redwood[12,13] first resolved piezoelectric equations using a FD formulation. Later, Kostek and Randall[14] developed a FDTD model for a cylindrical piezoelectric transducer in a borehole and studied its radiation into a fluid medium. In the same way, Yamada and Sato[15] applied another FDTD formulation to the dynamic analysis of two-dimensional (2D) piezoelectric equations. More recently, Smith and Ren[16] described acoustic wave propagation through piezoelectric crystals using a FDTD method, and this model has been extended with PML boundary conditions.[17]

The FE method is widely used, and it is also the most flexible method with respect to arbitrary geometries and coupling to a surrounding fluid medium. A lot of work has been done on the FE modeling of piezoelectric transducers, including the radiated field from such transducers, and there are a large number of different approaches for the modeling of a piezoelectric transducer in a fluid medium. The FE method was applied to piezoelectric media by several authors in the late 1960s and early 1970s.[18–20] Until the 1980s, most piezoelectric FE analyses were either modal or time harmonic. With the advent of faster computers with more

[a]Author to whom correspondence should be addressed. Electronic mail: erwan.filoux@gmail.com

memory, it became possible to perform relatively large-scale transient analyses of piezoelectric transducers. Since the end of the 80s, piezoelectric elements were included in different commercial FE programs.[21–25]

However, both FD and FE methods are more limited than the PS method by numerical dispersion which prevents the modeling of waves propagating over large distances. This is why, so far, piezoelectricity and acoustic propagation were simulated using two different models, one after the other. Moreover, absorbing conditions are not always simple or available for these different methods. But, contrary to FD methods which approximate the derivatives of a function by local arguments, spectral methods are global. Thus, the PS algorithm strictly requires to have a smooth excitation source in both time and space to avoid Gibbs phenomena and is difficult to use with irregular or inhomogeneous domains. This can appear in the simulation of complicated structures which requires an accurate modeling of both small and large objects (compared to the minimum wavelength). For the piezoelectric material, quasistatic approximation is assumed. This leads to the resolution of an equation with a second order derivative which is not time dependent. In this last case, PS methods are not adapted to the corresponding resolution. This is the reason why, in order to retain the advantages of both FD and PS methods, a combination of these two algorithms is used in our model. In this new hybrid FD-PSTD model, the FD algorithm is used to simulate the electrical potentials inside the piezoelectric media, which can present various geometries and materials. The PS algorithm is dedicated to model the acoustic waves propagating in the whole structure, resulting in the accurate calculation of both the electric and acoustic fields, their coupling, and the mechanical propagation over large distances, which is difficult with some more classical algorithms.

The following section will describe the basis of the theory with the governing equations and the detail of the main steps of the algorithm. The implementation of the model, and in particular the description of the spatial and temporal grids used for the pseudospectral and finite-difference methods, is exposed in Sec. III. For the finite-difference method, particular attention is paid to calculation differences during and after the electrical excitation period. Finally, several 2D configurations of a piezoelectric element immersed in water have been designed and the pressure field generated in the propagation media has been simulated in Sec. IV. These results are compared to those obtained for identical configurations with a FE method (ATILA™ software).

II. THEORY

A. Governing equations

Piezoelectric transducers modeling is based on both mechanical and electromagnetic equations. Assuming that Einstein's summation convention is used, the propagation of acoustic waves is described by the following system:[26]

$$\frac{\partial v_i}{\partial t} = \frac{1}{\rho} \cdot \frac{\partial T_{ij}}{\partial x_j}, \tag{1}$$

$$\frac{\partial T_{ij}}{\partial t} = c^E_{ijkl} \cdot \frac{\partial v_k}{\partial x_l}, \tag{2}$$

and their generation can be obtained by the combination of Maxwell's and piezoelectricity equations:

$$\frac{\partial T_{jk}}{\partial t} = c^E_{jklm} \cdot \frac{\partial v_l}{\partial x_m} - e_{ijk} \cdot \frac{\partial E_i}{\partial t}, \tag{3}$$

$$\frac{\partial D_i}{\partial t} = \epsilon^S_{ij} \cdot \frac{\partial E_j}{\partial t} + e_{ilm} \cdot \frac{\partial v_l}{\partial x_m}, \tag{4}$$

where

v is the particle velocity vector,
T is the stress tensor,
D is the electric displacement vector,
E is the electric field vector,
ρ is the material density,
$\mathbf{c}^E$ is the stiffened stiffness tensor,
e is the piezoelectric stress tensor,
ϵ^S is the dielectric permittivity tensor.

As usual for the modeling of piezoelectric media, considering the fact that electromagnetic evolution is nearly instantaneous compared to the slow propagation of mechanical waves, the quasistatic approximation is used. Thus, the description of the electromagnetic fields is reduced to Gauss' law and we can write

$$div\ \mathbf{D} = \rho_e. \tag{5}$$

Given the fact that there is no free charge inside the piezoelectric media (ρ_e=0), a direct relationship can be established between displacements and potentials. The combination of Eqs. (4) and (5) finally gives

$$\frac{\partial}{\partial x_i}\left(\epsilon^S_{ij} \cdot \frac{\partial \dot{\phi}}{\partial x_j}\right) = \frac{\partial}{\partial x_i}\left(e_{ilm} \cdot \frac{\partial v_l}{\partial x_m}\right), \tag{6}$$

where ϕ is the electric potential and $\dot{\phi}=\partial\phi/\partial t$. Equations (1), (3), and (6) form the system to be solved in order to simulate the generation and the propagation of acoustic waves in piezoelectric media. These equations do not take into account mechanical loss, but the algorithm can be improved by modeling, in the Fourier domain, complex stiffness coefficients c^{*E}_{jklm} as follows:

$$c^{*E}_{jklm} = c^E_{jklm}(1 + j\delta_m),$$

with δ_m the mechanical attenuation coefficient.

B. Usual operations

The different derivatives with respect to space are calculated with the pseudospectral (PS) or the finite-difference (FD) method. In the first case we use the fast Fourier transform (FFT) to quantify the spatial evolution of a component along one axis.[27] In the second case the spatial derivation is approximated by

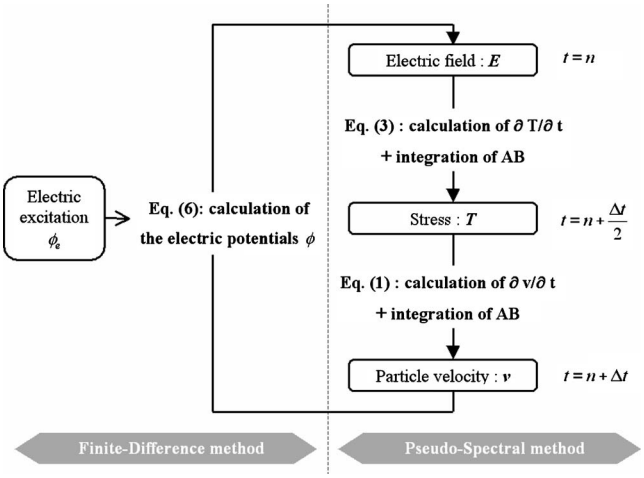


FIG. 1. Scheme of the algorithm.

$$\left(\frac{\partial y}{\partial x_i}\right)_{x_i=k} = \frac{y(k+1/2) - y(k-1/2)}{\Delta x_i},$$

where Δx_i is the spatial increment along the x_i axis.[28] The temporal integration is performed using the fourth order Adams-Bashforth's (AB) relationship for time-staggered grids:[29]

$$y(t+\Delta t) = y(t) + \frac{\Delta t}{24}\cdot\left[26\cdot\frac{\partial y}{\partial t}\left(t+\frac{\Delta t}{2}\right) - 5\cdot\frac{\partial y}{\partial t}\left(t-\frac{\Delta t}{2}\right) + 4\cdot\frac{\partial y}{\partial t}\left(t-\frac{3\Delta t}{2}\right) - \frac{\partial y}{\partial t}\left(t-\frac{5\Delta t}{2}\right)\right], \quad (7)$$

where Δt is the temporal increment. The use of time-staggered grids enhances the temporal resolution of the model.

All these operations are performed on separated variables, as required by the use of the perfectly matched layers (PMLs) introduced by Berenger.[8] Adapted to elastodynamics by Chew and Liu,[30] the PMLs are absorbing boundary conditions which avoid the incident waves to be reflected or transmitted at the borders of the grid. They are based on a stretched-coordinates formulation of the governing equations. For example, Yuan *et al.*,[31] who proposed to combine a second order FD time domain simulation with PML for the simulation of acoustic waves propagation, demonstrated that an eight-node PML can reduce reflections by 80 dB. The separation of each component of stress and velocity, according to its influence on the propagation of mechanical waves along each axis, enables the absorption of waves with any angle of incidence.

C. Summarized steps of the algorithm

In this paper, the study is focused on the emission mode of the piezoelectric resonator. Thus, at the beginning of the simulation, an electric field is created inside the piezoelectric material by an initial electrical excitation (a half sine wave signal at frequency $1/T$ applied from $t=0$ to $t=T/2$). Then, one iteration of the algorithm is performed in three main steps, as presented in Fig. 1. The first step consists in simulating the effect of the electric field on the mechanical parameters. This is why Eq. (3) is used to calculate the temporal derivatives of stress from the electric field and the acoustic velocity known at $t=n$. These derivatives of stress are integrated through AB's method to obtain the stress field $\mathbf{T}$ at $t=n+\Delta t/2$ in the entire computational domain. Then, as a second step, the temporal derivatives of acoustic velocity are obtained by solving Eq. (1), and another integration of AB gives the new acoustic velocity field $\mathbf{v}$ at $t=n+\Delta t$. So far the equations in these two steps are solved using the PS method (see Sec. III A). The third step consists in solving Eq. (6) using the FD method, which enables the electric potentials inside the piezoelectric material to be calculated (see Sec. III B). From a different point of view, one could say that in fact the evolution of the stress field is calculated in the entire computational domain as if there was no piezoelectric effect, which can be summed up in the resolution of equations (1) and (2) [the latter being equivalent to Eq. (3) with $\mathbf{e}=\mathbf{0}$] with the PS method to model simply elastic wave propagation. Only then the stress field inside the piezoelectric media is corrected through Eq. (3) by taking into account the electric field calculated by the FD method.

III. IMPLEMENTATION

A. Pseudospectral method

The PS method is used to calculate the derivatives of stress and velocity with respect to space in the system formed by Eqs. (1), (3), and (6). To obtain, for instance, the first derivative of T_{ij} along x_j axis in the time domain, the following operation is performed:

$$\frac{\partial T_{ij}}{\partial x_j} = \mathrm{FFT}^{-1}[(-j\,.\,k_j)\,.\,\mathrm{FFT}(T_{ij})],$$

where the wave number k_j is the vector:

$$\left(0\ \frac{1}{N}\ \frac{2}{N}\ \cdots\ \frac{1}{2}-\frac{1}{N}\ \frac{-1}{2}\ \cdots\ \frac{-1}{N}\right)\cdot j\frac{2\pi}{\Delta x_j},$$

with N the number of nodes along x_j axis on the grid of T_{ij}. The whole simulation is optimized by a particular organization of staggered grids. Indeed, if a variable b is defined by the derivative of another variable a, through a relation such as $\partial a/\partial t = f(b)$, then these two variables cannot be located on the same grid.[29] This is why, as exposed in Fig. 2, both spatial and temporal staggerings are used to reduce the numerical artifacts introduced by the FFT operations[32,33] and to improve the algorithm's accuracy, respectively. The staggered grid implementation is based on a particular arrangement of the physical variables and is detailed in Fig. 3.

B. Finite-difference method

As one can see on the scheme of the algorithm (Fig. 1) the PS method is used to solve Eqs. (1) and (3). But, contrary to these two equations, the derivatives of potentials in Eq. (6) are not expressed separately as functions of variables which could be approximated by a PS operation. Moreover, staggered grids are useful to calculate first order derivatives with the PS method, but they are not adapted to the calculation of second order derivatives.[34] This is why the PS method can-

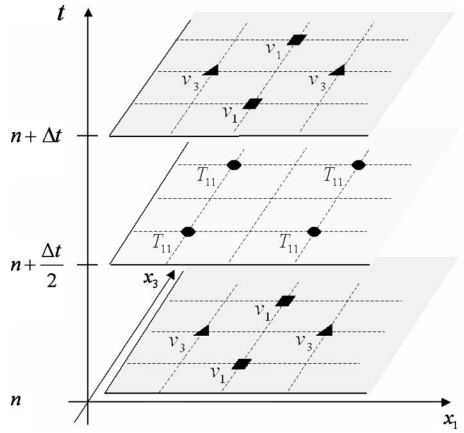


FIG. 2. Scheme of the 2D spatially and temporally staggered grids used by the algorithm, with the corresponding variables.

not be used and Eq. (6) must be solved through a FD method. The purpose of this operation is to determine the potentials inside the piezoelectric material, where the electromechanical interaction is active. These potentials are represented in Fig. 4. The ones accounting for the electrodes are different from the others because they are known—and not calculated—during all the excitation period. The unknown temporal derivatives of potentials are grouped in a vector $\mathbf{X}=(\dot{\phi}_{ij})_{i,j}$. After the excitation period, this vector is completed with the potentials of the electrodes because their values are no longer imposed by the initial electrical excitation. This vector is calculated at each time step by solving Eq. (6). In 2D Cartesian coordinates and in the case of a transversally isotrope piezoelectric material polarized in x_3 direction, Eq. (6) becomes

$$\epsilon_{11}\cdot\frac{\partial^2\dot{\phi}}{\partial x_1^2}+\epsilon_{33}\cdot\frac{\partial^2\dot{\phi}}{\partial x_3^2}+\frac{\partial\epsilon_{11}}{\partial x_1}\cdot\frac{\partial\dot{\phi}}{\partial x_1}+\frac{\partial\epsilon_{33}}{\partial x_3}\cdot\frac{\partial\dot{\phi}}{\partial x_3}=\frac{\partial}{\partial x_1}\left[e_{15}\cdot\left(\frac{\partial v_3}{\partial x_1}+\frac{\partial v_1}{\partial x_3}\right)\right]+\frac{\partial}{\partial x_3}\left[e_{31}\cdot\frac{\partial v_1}{\partial x_1}+e_{33}\cdot\frac{\partial v_3}{\partial x_3}\right]. \tag{8}$$

The left-hand side of Eq. (8) is approximated by a FD representation. The right-hand side, written using the condensed tensor notation, was previously calculated with the PS

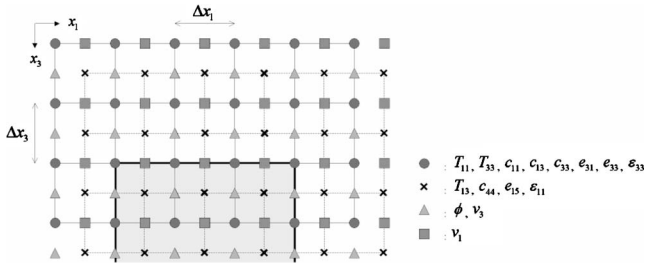


FIG. 3. Spatial arrangement of physical variables on staggered grids. Circle: reference grid defining the piezoelectric media; Square, Triangle, Cross: grids shifted along x_1 axis, x_3 axis, and both axes. The gray-colored piezoelectric element is in a surrounding media (such as water or a solid inert material, for instance). Outside this active area, the piezoelectric tensor **e** is set to zero.

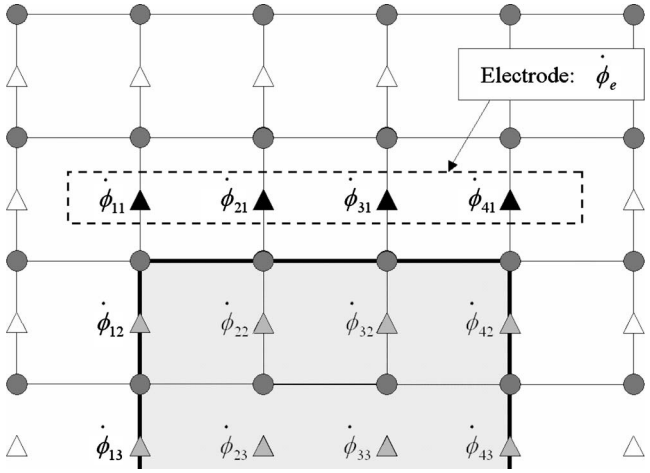


FIG. 4. Representation of the calculated temporal derivatives of potentials ϕ_{ij} inside the piezoelectric media and the upper full-covering electrode.

method and will be noted $R_{j,k}$, with j and k the coordinates of the nodes along x_1 and x_3 axes, respectively. The discretized variables are expressed on the grid of ϕ and a spatial average is determined when they are not defined. After a particular arrangement of the variables the following equation is obtained:

$$\left(\frac{\tilde{\epsilon}^1_{11}}{(\Delta x_1)^2}-\frac{\Delta\epsilon^1_{11}}{2\,.\,\Delta x_1}\right)_{j,k}\cdot\dot{\phi}_{j-1,k}+\left(\frac{\tilde{\epsilon}^1_{11}}{(\Delta x_1)^2}+\frac{\Delta\epsilon^1_{11}}{2\,.\,\Delta x_1}\right)_{j,k}\cdot\dot{\phi}_{j+1,k}+\left(\frac{\tilde{\epsilon}^3_{33}}{(\Delta x_3)^2}-\frac{\Delta\epsilon^3_{33}}{2\,.\,\Delta x_3}\right)_{j,k}\cdot\dot{\phi}_{j,k-1}+\left(\frac{\tilde{\epsilon}^3_{33}}{(\Delta x_3)^2}+\frac{\Delta\epsilon^3_{33}}{2\,.\,\Delta x_3}\right)_{j,k}\cdot\dot{\phi}_{j,k+1}+\left(-2\cdot\frac{\tilde{\epsilon}^1_{11}}{(\Delta x_1)^2}-2\cdot\frac{\tilde{\epsilon}^3_{33}}{(\Delta x_3)^2}\right)_{j,k}\cdot\dot{\phi}_{j,k}=R_{j,k}, \tag{9}$$

with:

- $(j,k)\in[1,n]\times[1,p]$,
- Δx_i the space increment along x_i axis,
- $\tilde{\epsilon}^i_{pq}$ the average value of ϵ_{pq} along x_i axis,
- $\Delta\epsilon^i_{pq}$ the derivative of ϵ_{pq} with respect to x_i axis.

This equation has to be solved at each time step and for each point (j,k) on the grid of $\dot{\phi}$. Then, for each time step, a system of linear equations governing the potentials inside the entire piezoelectric media are obtained, which can be written:

$$\mathbf{Q}\,.\,\mathbf{X}=\mathbf{R},$$

with **Q** a block tri-diagonal matrix that will have to be inversed and $\mathbf{R}=(R_{j,k})_{j,k}$. Equation (9) can be rewritten in a simple way:

$$A_{j,k}\cdot P_{j-1,k}+B_{j,k}\cdot P_{j+1,k}+C_{j,k}\cdot P_{j,k-1}+D_{j,k}\cdot P_{j,k+1}+E_{j,k}\cdot P_{j,k}=R_{j,k}, \tag{10}$$

with $P_{j,k}$ the temporal first derivative of the electric potentials ϕ and $A_{j,k}$, $B_{j,k}$, $C_{j,k}$, $D_{j,k}$, $E_{j,k}$ the corresponding coefficients. A representation of matrix **Q** can be observed in Fig.

	P11	P21	P31	P41	P51	P12	P22	P32	P42	P52	P13	P23	P33	P43	P53	P14	P24	P34	P44	P54
P11	E11	B11				D11														
P21	A21	E21	B21				D21													
P31		A31	E31	B31				D31												
P41			A41	E41	B41				D41											
P51				A51	E51					D51										
P12	C12					E12	B12				D12									
P22		C22				A22	E22	B22				D22								
P32			C32				A32	E32	B32				D32							
P42				C42				A42	E42	B42				D42						
P52					C52				A52	E52					D52					
P13						C13					E13	B13				D13				
P23							C23				A23	E23	B23				D23			
P33								C33				A33	E33	B33				D33		
P43									C43				A43	E43	B43				D43	
P53										C53				A53	E53					D53
P14											C14					E14	B14			
P24												C24				A24	E24	B24		
P34													C34				A34	E34	E34	
P44														C44				A44	E44	B44
P54															C54				A54	E54

FIG. 5. Representation of the block tri-diagonal matrix **Q** in the case of a 4×5 nodes piezoelectric area. Each line corresponds to an equation of the linear system. The notation P*IJ* represents the position of the derivated potentials $P_{j,k}$. The left gray column regroups the positions of the nodes where each equation is solved. The upper gray line regroups the positions of the nodes where the derivated potentials $P_{j,k}$ are multiplied to the corresponding coefficients below. The empty white boxes correspond to a null coefficient.

5. Equation (10) is not exactly the same during and after the electrical excitation period. The potential of one electrode will be kept at zero, being the ground of the transducer. When the potential of the excited electrode is imposed (ϕ_e), it is passed to the right-hand side of the equation and becomes a component of vector **R**. At the end of the excitation period, the potential of this electrode has to be calculated and becomes a component of the unknown vector **X**. This is why different vectors and matrices are used during these two periods.

C. Definition of matrix Q

As the definition of matrix **Q** is the heart of the FD method, some details about the hypotheses taken into account for the computation of this matrix are presented. One of the most important hypothesis which transforms the system of equations, and consequently-matrix **Q**, is the high contrast between the permittivity of the piezoelectric material and those of the surrounding media. This enables the electric field to be neglected outside the piezoelectric medium. Moreover, the electrical potential is described as continuous at the borders of the resonator. These approximations are used to define matrix **Q**. Indeed, in order to solve Eq. (9) at a node (j,k), one must define the potential at coordinates (j,k) and the four ones surrounding it at coordinates $(j-1,k)$, $(j+1,k)$, $(j,k-1)$ and $(j,k+1)$. As one can see in Fig. 6, a problem appears when Eq. (9) requires some potentials that are not defined, being outside of the piezoelectric area. But, according to the condition of continuity, the undefined potential is equal to the closest one defined inside the piezoelectric area. Then some coefficients of Eq. (9) are factorized and, this operation being repeated along all the borders of the area, the corresponding components of matrix **Q** are significantly modified. A second important hypothesis is the equipotentiality of all the nodes defining each electrode. It decreases the degree of freedom of the system, grouping the variables into a single one for each electrode. While the first hypothesis changes the values of the components of matrix **Q**, the second one modifies its structure. Finally, **Q** is a square matrix with $n\times n$ dimensions, where n is the number of potentials (nodes) which have to be calculated minus the ones accounting for the electrodes. The value of n is also increased by the number of electrodes whose potential is unknown.

Once matrix **Q** is determined and inversed, the algorithm will be able to calculate the new value of the electric field inside the piezoelectric material and another iteration of the simulation will start.

IV. RESULTS

The numerical results of the FD-PSTD method are used to compare the efficiency of this algorithm to that of the FE algorithm of ATILA™ software, and to validate the advantages of the exposed method. Thus the simulations are performed using the 2D cut of a thin piezoelectric plate, covered on both sides with electrodes excited by a 50 MHz half sine wave driving voltage. The piezoelectric element is a standard soft PZT resonator (Ferroperm Piezoceramics Pz27) poled along the x_3 direction with a 50 MHz thickness resonant frequency. As the piezoelectric resonator is laid in the plane containing x_1 and x_3 axes (the width of the plate along x_2 axis is considered infinite), its characteristics of interest are sum-

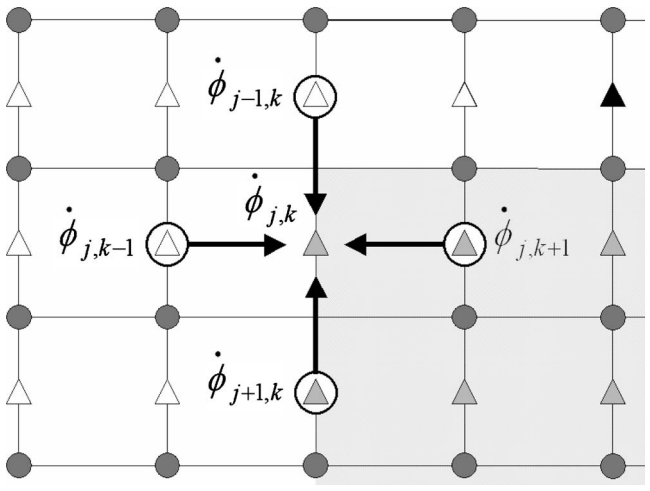


FIG. 6. Spatial arrangement of the potentials ϕ required to solve Eq. (9). Special case of a resolution at a node (j,k) in the upper left corner of the piezoelectric area. The white triangles are undefined potentials, contrary to the gray ones. The black triangle corresponds to the last potential defining the upper electrode, which is imposed and not calculated during the electrical excitation period.

TABLE I. Material constants of the PZT piezoceramic (see Ref. 35): stiffness at constant electric field (GPa), dielectric permittivity and piezoelectric stress (C/m^2).

c_{11}^E	c_{13}^E	c_{33}^E	c_{44}^E	$\epsilon_{11}^S/\epsilon_0$	$\epsilon_{33}^S/\epsilon_0$	e_{15}	e_{31}	e_{33}
147	93	113	23	1130	914	11.6	−3.1	16

marized in Table I. It has a length-to-thickness ratio (L/Th) of approximately 14 which is sufficient to keep the thickness mode predominant. It is immersed in water and the configuration of the simulations is detailed in Table II.

As a first verification of the method, this configuration is used with the FE and FD-PSTD algorithms to calculate the acoustic field generated in water. In both cases the models simulate materials without losses and the displacements along the transversal axis are disabled so as to focus on the pure thickness mode. The model being equivalent to a one-dimensional system, the pressure field is then observed at a fluid node just in front of the middle of the piezoelectric plate. The temporal evolutions of the pressure obtained with the FD-PSTD and FE methods using a grid of $\lambda_{min}/12$ width square cells are satisfactorily compared in Fig. 7. The minimal wavelength λ_{min} is approximately 30 μm as it corresponds to an acoustic wave at 50 MHz propagating in water with a velocity of 1490 m/s.

In the case of a two-dimensional model, where transversal displacements are simulated, the results are very sensitive to the accuracy of the meshing. Thus, to compare the results given by each method, identical meshings were used for the FE and FD-PSTD algorithms. However, as presented in Fig. 3, we use staggered grids contrary to the FE algorithm whose variables are all located on one grid. That is why some parameters such as pressure or electrical potential are not observed exactly at the same location with each model. In order to optimize the simulations, we searched for the largest common dimension of cells that still lead to accurate results for both methods. To do this, several simulations were performed with increasing dimensions of the cells. It appeared that the results obtained by ATILA™ software remained identical with cells of dimensions as large as $\lambda_{min}/12$ (which corresponds approximately to a $\lambda_{piezo}/36$ meshing in the piezoelectric medium as the grids are uniform), and were similar to those of the FD-PSTD simulations. With larger dimensions of the cells, the results given by the FE analysis present significant differences. Thus, using a meshing of $\lambda_{min}/12$ width square cells, a comparison of the electrical and mechanical parameters calculated by the two methods can be performed. The two-dimensionnal evolution of pressure in

TABLE II. Configuration used for the simulation of the piezoelectric plate immersed in water.

e_p (μm)	L/Th	ρ (kg/m^3)	v_p (m/s)	v_w (m/s)	f_e (MHz)	λ_p (μm)	λ_w (μm)
43.34	14	7700	4334	1490	50	86.7	29.8

e_p: thickness; L/Th: length-to-thickness ratio; ρ: density; v_p, v_w: longitudinal wave velocity (m/s) in the piezoelectric material and in water; f_e: excitation frequency; λ_p, λ_w: minimum acoustic wavelength in the piezoelectric material and in water.

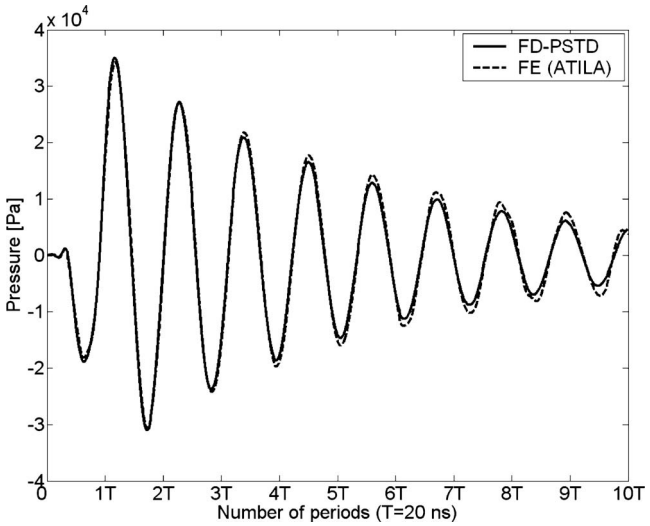


FIG. 7. Pressure curve in water, 10 μm in front of the piezoelectric element, with a meshing of $\lambda_{min}/12$ width square cells.

Fig. 8 can be compared to the one-dimensionnal case (Fig. 7). With a velocity around 2500 m/s along x_1 axis in Pz27, the transversal vibration reaches the central axis of emission after 0.12 μs ($6T$) of propagation as expected. The evolution of axial displacements in Fig. 9 shows that this effect is also observable a little sooner in the piezoelectric medium, 10 μm under the center of the upper electrode. Concerning the electric field, one can see in Fig. 10 the evolution of the potentials along the central axis of the piezoelectric resonator calculated with each method. To better compare these two plots, the amplitude of the voltage obtained at the same previous node, 10 μm under the upper electrode, is detailed in Fig. 11. Though the amplitude obtained with the FE analysis is sometimes higher than the FD-PSTD one, the results of these two methods are in good agreement.

To evaluate the efficiency of the FD-PSTD algorithm, we could not simply compare the processing time because the FE software used for this study did not enable the use of

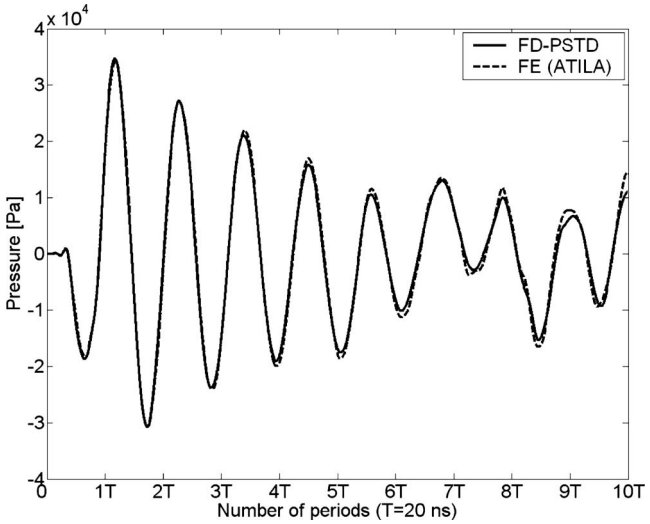


FIG. 8. Evolution of pressure in water, 10 μm in front of the center of the piezoelectric element, with a meshing of $\lambda_{min}/12$ width square cells, taking into account the transversal displacements.

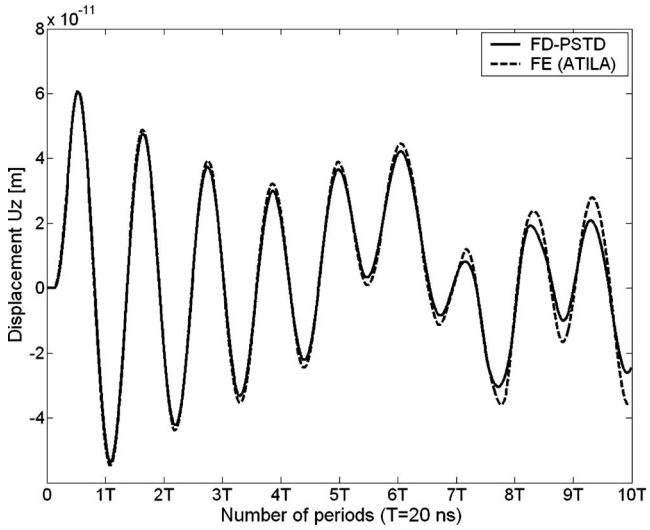


FIG. 9. Evolution of axial displacement on the center axis of the piezoelectric medium, 10 μm under the upper electrode.

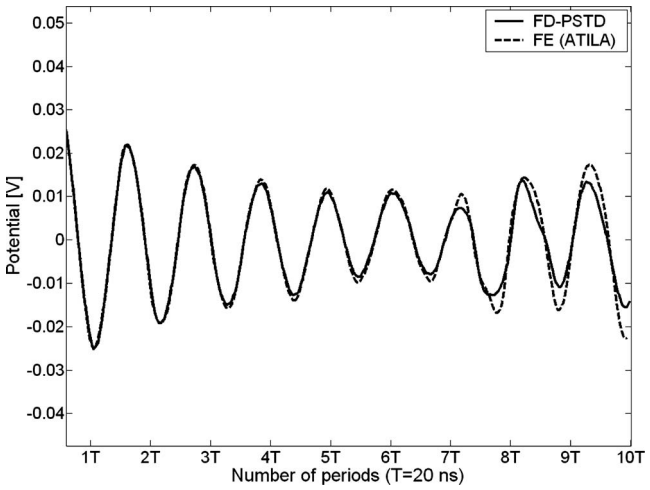


FIG. 11. Evolution of the potential located on the central axis of the piezoelectric medium, 10 μm under the upper electrode.

PML conditions at the borders of the computational domain. To avoid unwanted reflections, the computational space had to be extended as far as the propagating waves go, which made the FE simulations much more time consuming than the FD-PSTD ones. Nevertheless it is possible to study and compare the accuracy of the results as a function of the meshing. By decreasing the number of nodes per minimum wavelength, the sturdiness of the FE method and of the new algorithm can be tested. This is why the same two-dimensional simulations are performed using meshings of $\lambda_{min}/10$, $\lambda_{min}/8$ and $\lambda_{min}/6$ width square cells, to observe the evolution of the numerical error in the results obtained with the FE (Fig. 12) and FD-PSTD (Fig. 13) methods. As one can see, while the meshing pitch increases, the numerical error made by the FE algorithm increases much faster than with the FD-PSTD algorithm. A part of the differences observed between the curves comes from the difference in nodes location on the grids. It is difficult to estimate the proportion of each error in the variations observed between the right curve ($\lambda_{min}/12$) and the one obtained with a meshing of $\lambda_{min}/6$. But, contrary to the FD-PSTD case, it is obvious that the differences obtained with the FE analysis are not only due to the spatial mismatch of the nodes. This comparison shows that the FD-PSTD method is less sensitive to

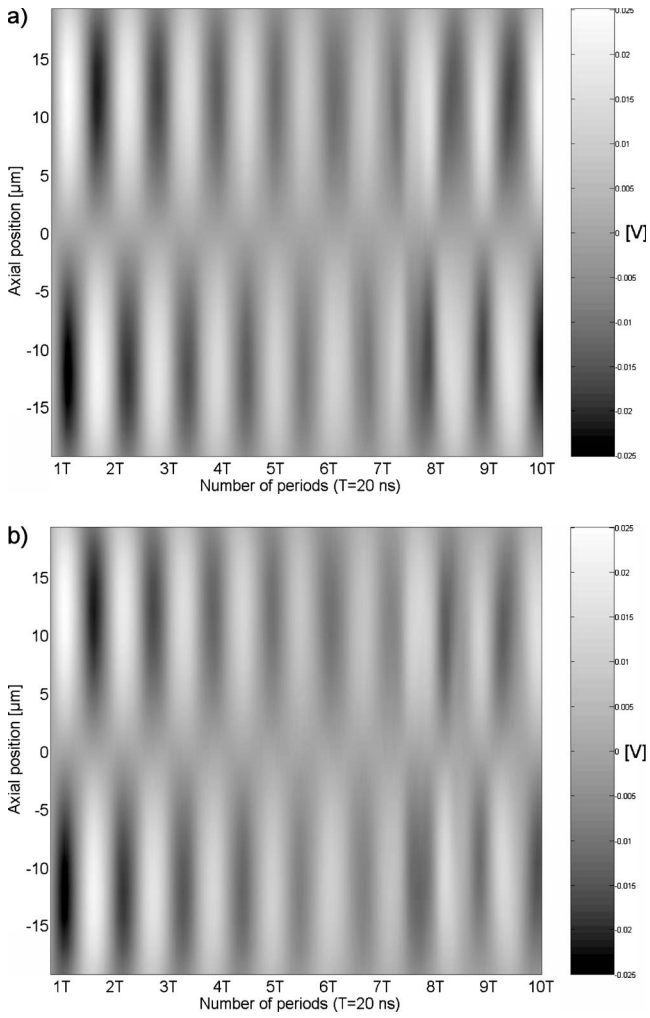


FIG. 10. Evolution of the potentials along x_3 axis at the center of the piezoelectric medium obtained with the (a) FE and (b) FD-PSTD methods.

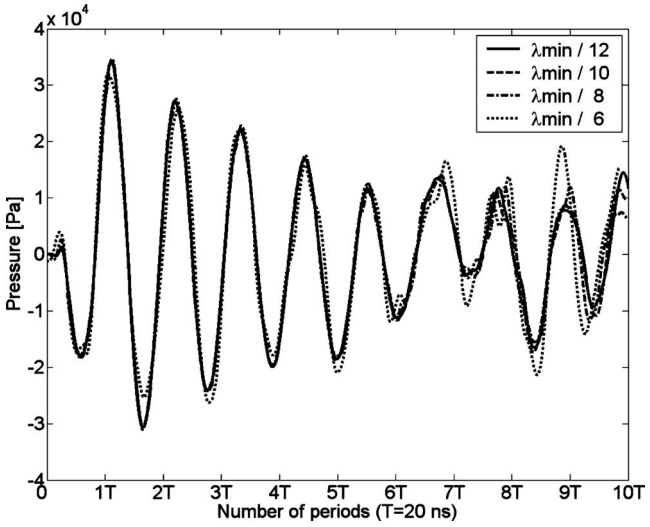


FIG. 12. Evolution of the pressure in front of the center of the piezoelectric element for different meshings using the FE algorithm of ATILA.

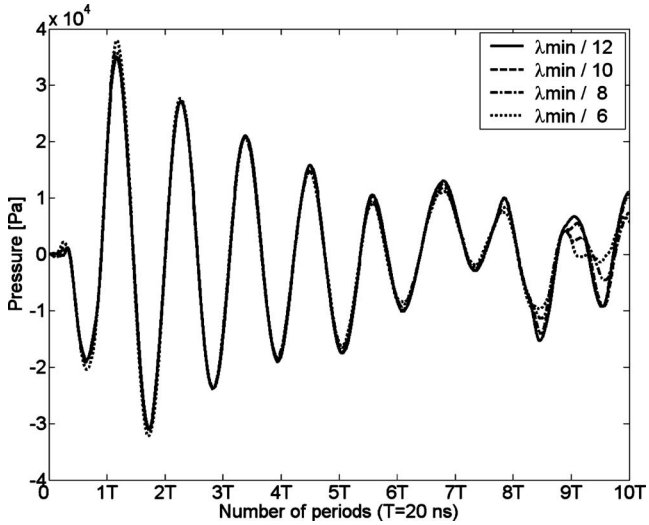


FIG. 13. Evolution of the pressure in front of the center of the piezoelectric element for different meshings using the FD-PSTD algorithm.

the dimensions of the meshing. Nevertheless, it is limited by the second order approximation of the electric field in the resonator. But even with a meshing of $\lambda_{min}/6$, the electric field is calculated with approximately 18 nodes per wavelength and the result remains reasonably accurate. Finally, the error made with the FD-PSTD method when using a meshing of $\lambda_{min}/10$ or $\lambda_{min}/8$ is hardly noticeable, contrary to the one obtained with the FE analysis using these meshings. Thus, the FD-PSTD method requires fewer nodes per minimum wavelength than the FE algorithm to converge, which allows the use of widened meshings and consequently saves processing time.

V. CONCLUSION

A new hybrid FD-PSTD algorithm has been developed to model a piezoelectric resonator in a fluid surrounding media (water in this paper). Its purpose is to efficiently calculate the vibration of an electrically excited piezoelectric material and to model the acoustic waves propagating inside the resonator and in the surrounding media. One of the problems with transducer modeling is that the acoustic waves propagation has to be simulated over large-scale objects, which is much more efficiently achieved with the use of the PSTD method rather than the FDTD. But the PSTD algorithm is not adapted to the resolution of the piezoelectric constitutive equations, contrary to the FDTD algorithm. That is why the combination of these two methods is relevant to solve both the equations of acoustic propagation and piezoelectricity, and to meet the simulation requirements in terms of objects scaling. The results of this hybrid method are in good agreement with those of FE model, for the two-dimensional configuration of a piezoelectric resonator in water, with a 50 MHz thickness resonant frequency. Moreover, this method keeps the main advantage of PS algorithms which require fewer nodes per minimum wavelength than other FE or FD algorithms to converge.

With the validation of this first version of the hybrid model, further improvements are under consideration. The next step will be the computation of losses, in particular the mechanical losses which need to be taken into account in a future model. Then, backing and front matching layers will be introduced in the simulation to test complete structures of ultrasonic transducers.

[1]H. O. Kreiss and J. Oliger, "Comparison of accurate methods for the integration of hyperbolic equations," Tellus **24**, 199–215 (1972).
[2]D. Kosloff and E. Baysal, "Forward modeling by a Fourier method," Geophysics **47**, 1402–1412 (1982).
[3]D. Kosloff, M. Reshef, and D. Loewenthal, "Elastic wave calculations by the Fourier method," Bull. Seismol. Soc. Am. **74**, 875–891 (1984).
[4]G. Wojcik, B. Fornberg, R. Waag, L. Carcione, J. Mould, L. Nikodym, and T. A. Driscoll, "Pseudospectral methods for large-scale bioacoustic models," *IEEE International Ultrasonics Symposium*, Toronto, Canada (1997).
[5]C. R. Daudt, L. W. Braile, R. L. Nowack, C. S. Chang, and D. Loewenthal, "A comparison of finite difference and Fourier method calculations of synthetic seismograms," Bull. Seismol. Soc. Am. **79**, 1210–1230 (1989).
[6]B. Fornberg, "The pseudospectral method: Comparisons with finite differences for the elastic wave equation," Geophysics **52**, 483–501 (1987).
[7]Q. H. Liu, "The PSTD algorithm: A time-domain method requiring only two cells per wavelength," Microwave Opt. Technol. Lett. **15**, 158–165 (1997).
[8]J. P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves," J. Comput. Phys. **114**, 185–200 (1994).
[9]S. Callé, J. P. Remenieras, M. Elkateb, and F. Patat, "Shear wave elastography: Modeling of the shear wave propagation in heterogeneous tissue by pseudospecral method," *IEEE International Ultrasonics Symposium*, Montreal, Canada (2004).
[10]T. Goursolle, S. Callé, S. Dos Santos, and O. Bou Matar, "A 2D pseudospectral model of nonlinear elastic wave spectroscopy applied with time reversal," J. Acoust. Soc. Am. **122**, 3220–3229 (2007).
[11]C. Batifol, S. Callé, P. Marechal, M. Lethiecq, and F. Levassort, "Formulation and validation of Berenger's PML absorbing boundary for the FDTD simulation acoustic scattering," *IEEE International Ultrasonics Symposium*, Rotterdam, Holland (2005).
[12]P. Lloyd and M. Redwood, "Finite-difference method for the investigation of the vibrations of solids and the evaluation of the equivalent-circuit characteristics of piezoelectric resonators. I and II," J. Acoust. Soc. Am. **39**, 346–361 (1966).
[13]P. Lloyd and M. Redwood, "Finite-difference method for the investigation of the vibrations of solids and the evaluation of the equivalent-circuit characteristics of piezoelectric resonators. III," J. Acoust. Soc. Am. **40**, 82–85 (1966).
[14]S. Kostek and C. J. Randall, "Modeling of a piezoelectric transducer and its application to full waveform acoustic logging," J. Acoust. Soc. Am. **95**, 109–122 (1994).
[15]Y. Yamada and M. Sato, "Application of extended finite-difference method to two-dimensional dynamic analysis of a piezoelectric vibrator," Jpn. J. Appl. Phys., Part 1 **37**, 255–256 (1998).
[16]P. M. Smith and W. Ren, "Finite-difference time-domain techniques for SAW device analysis," *IEEE International Ultrasonics Symposium*, Munich, Germany (2002).
[17]F. Chagla and P. M. Smith, "Finite difference time domain methods for piezoelectric crystals," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 1895–1901 (2006).
[18]H. Allik and T. J. R. Hughes, "Finite element method for piezoelectric vibration," Int. J. Numer. Methods Eng. **2**, 151–157 (1970).
[19]Y. Kagawa and G. M. L. Gladwell, "Finite element analysis of flexure-type vibrators with electrostrictive transducers," IEEE Trans. Sonics Ultrason. **SU-17**, 41–49 (1970).
[20]G. H. Schmidt, "Application of the finite element method to the extensional vibrations of piezoelectric plates," in *Conference on Mathematics of Finite Elements and Applications*, edited by J. R. Whiteman (Academic, New York, 1972), pp. 351–361.
[21]"Ansys revision 5.0. Technical description of capabilities," Swanson Analysis Systems, Inc. (1992).
[22]B. Dubus, J. C. Debus, J. N. Decarpigny, and D. Boucher, "Analysis of mechanical limitations of high power piezoelectric transducers using finite element modeling," Ultrasonics **29**, 201–207 (1991).
[23]R. L. Goldberg, M. J. Jurgens, D. M. Mills, C. S. Henriquez, D. Vaughan, and S. W. Smith, "Modeling of piezoelectric multilayer ceramics using

finite element analysis," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 1204–1214 (1997).
[24]*Abaqus/Standard Verification Manual Version 5.4* (Hibbit, Karlsson and Sorensen, 1994).
[25]H. F. Du Toit, P. L. George, P. Laug, D. Steer, and M. Vidrascu, *An Introduction to MOD-ULEF: MODULEF User Guide No. 1* (INRIA, France, 1991).
[26]B. A. Auld, *Acoustic Fields and Waves in Solids* (Wiley, New York, 1990).
[27]B. Fornberg and D. M. Sloan, "A review of pseudospectral methods for solving partial differential equations," Acta Numerica 203–267 (1994).
[28]A. Taflove and S. C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, 3rd ed. (Artech House, Norwood, USA, 2005).
[29]M. Ghrist, B. Fornberg, and T. A. Driscoll, "Staggered time integrators for wave equations," SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **38**, 718–741 (2000).
[30]W. C. Chew and Q. H. Liu, "Perfectly matched layers for elastodynamics: A new absorbing boundary condition," J. Comput. Acoust. **4**, 72–79 (1996).
[31]X. Yuan, D. Borup, J. W. Wiskin, M. Berggren, R. Eidens, and S. A. Johnson, "Formulation and validation of Berenger's PML absorbing boundary for the FDTD simulation acoustic scattering," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 816–822 (1997).
[32]T. Ozdenvar and G. A. McMechan, "Causes and reduction of numerical artifacts in pseudospectral wavefield extrapolation," Geophys. J. Int. **126**, 819–828 (1996).
[33]H. W. Chen, "Staggered-grid pseudospectral viscoacoustic wave field simulation in two-dimensional media," J. Acoust. Soc. Am. **100**, 120–131 (1996).
[34]B. Fornberg, "High-order-finite-difference and the pseudospectral method on staggered-grids," SIAM (Soc. Ind. Appl. Math.) J. Numer. Anal. **27**, 904–918 (1990).
[35]URL: http://www.ferroperm-piezo.com/ (Last viewed 06/12/2007).

Customization of the acoustic field produced by a piezoelectric array through interelement delays

Parag V. Chitnis,[a] Paul E. Barbone, and Robin O. Cleveland
Department of Aerospace and Mechanical Engineering, Boston University, 110 Cummington St., Boston, Massachusetts 02215

(Received 27 September 2007; revised 31 March 2008; accepted 31 March 2008)

A method for producing a prescribed acoustic pressure field from a piezoelectric array was investigated. The array consisted of 170 elements placed on the inner surface of a 15 cm radius spherical cap. Each element was independently driven by using individual pulsers each capable of generating 1.2 kV. Acoustic field customization was achieved by independently controlling the time when each element was excited. The set of time delays necessary to produce a particular acoustic field was determined by using an optimization scheme. The acoustic field at the focal plane was simulated by using the angular spectrum method, and the optimization searched for the time delays that minimized the least squared difference between the magnitudes of the simulated and desired pressure fields. The acoustic field was shaped in two different ways: the −6 dB focal width was increased to different desired widths and the ring-shaped pressure distributions of various prescribed diameters were produced. For both cases, the set of delays resulting from the respective optimization schemes were confirmed to yield the desired pressure distributions by using simulations and measurements. The simulations, however, predicted peak positive pressures roughly half those obtained from the measurements, which was attributed to the exclusion of nonlinearity in the simulations. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912448]

PACS number(s): 43.38.Hz, 43.60.Fg [AJZ] Pages: 4174–4185

I. INTRODUCTION

This work was motivated by an application in shock wave lithotripsy (SWL). SWL has been clinically used to treat kidney stones since 1980,[1,2] but there is still debate over the process by which the SWs comminute kidney stones. There are two main categories of mechanisms deemed plausible for stone breakage: direct stress effects, such as spallation,[1,3] squeezing,[4] shear,[5,6] and cavitation effects consisting of microjets,[7,8] and secondary SWs emitted from collapsing bubbles.[9–12]

Studies have shown that altering the temporal wave profile can affect cavitation,[13–19] thus influencing the cavitation related mechanisms of therapeutic ultrasound. Conceivably, altering the spatial distribution could influence the direct stress related mechanisms. For instance, spallation depends primarily on the portion of the SW that enters the stone.[1,3] Similarly, mechanisms such as squeezing and shear depend more on the pressure at the periphery of the stone.[4–6] Therefore, stress related mechanisms might be manipulated by altering the spatial pressure distribution.

To date, the primary focus of customizing the acoustic field has been to electronically steer the focus to track the kidney stone during treatment.[20,21] Both studies employed a spherical transducer array with high voltage pulser circuits for each individual element and achieved efficient focusing in an ellipsoidal region of about 4 cm in diameter and 6 cm in length. The present work employs a similar spherical transducer and pulse generator system to customize spatial pressure distribution.

The modification of spatial pressure distribution has not been widely implemented in lithotripsy. One notable example where the spatial pressure distribution was altered is the wide-focus lithotripter developed by Eisenmenger *et al.*[22] It employs an electromagnetic source and generates a relatively low pressure amplitude of 20 MPa but a relatively wide focal width of 18 mm (two to four times the diameter of a typical stone). This "wide-focus" lithotripter was reported to trigger the squeezing mechanism[4] leading to successful treatment while reducing patient discomfort. In another instance, Sapozhnikov *et al.*[23] treated cylindrical stones with an electrohydraulic lithotripter and modified the incident acoustic field by using baffles. They implemented various configurations of baffles to selectively block specific waves responsible for different stress related fracture mechanisms such as spallation, shear, or squeezing. They determined that neither spallation nor squeezing is solely responsible for fracture, and that shear stresses induced at the periphery of the proximal face of the stone play a dominant role in comminution.

The focus of the present work was to develop a system capable of producing a predetermined acoustic field by solving an inverse problem for a piezoelectric array. The method utilized in this study to customize the acoustic output of the piezoelectric lithotripter array is loosely based on the approach by Tanter *et al.*[24] There, it was demonstrated that an arbitrary pressure-time profile and spatial pressure distribution can be achieved by implementing a spatiotemporal inverse filter. Though they presented a viable solution of the

[a]Author to whom correspondence should be addressed. Tel.: 617-358-4658. FAX: 617-353-5866. Electronic mail: pchitnis@bu.edu

 0001-4966/2008/123(6)/4174/12/$23.00

inverse problem, the physical implementation of a spatiotemporal inverse filter is instrumentation intensive as it requires the ability to excite each element with an arbitrary signal. This is particularly challenging for the high amplitudes required in lithotripsy applications. The hardware available for this study (detailed description to follow) consisted of independent high voltage pulse discharge circuits for each element, where the only control parameter for each element was the time that the element was excited. The challenge was to find the appropriate timing for each element that best matches the desired acoustic field. We employed an inverse problem approach which resulted in a satisfactory approximation to the desired spatial pressure distribution.

II. THEORY

The theory is presented in three parts. First, the angular spectrum approach which is employed as the forward model is described. Second, the practical implementation of the angular spectrum method to the array source is defined, and, finally, the error function employed in the optimization routine is described.

A. Forward model

The forward model employed the angular spectrum method which models diffractive wave propagation from one plane to any other parallel plane.[25–27] The salient features of the angular spectrum theory will be presented in this section. Readers may refer to Ref. 27 for further details. The pressure can be described in the frequency domain by taking a temporal Fourier transform which is defined here as

$$P(x,y,z,\omega) = \int_{-\infty}^{\infty} p(x,y,z,t)e^{i\omega t}dt. \tag{1}$$

Given $P(x,y,z,\omega)$ on some $z=z_0$ plane, the angular spectrum is defined as the spatial Fourier transform of P with respect to x and y, defined here as

$$\tilde{P}(k_x,k_y,z_0,\omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x,y,z_0,\omega)e^{-i(k_x x+k_y y)}dxdy. \tag{2}$$

Since $P(x,y,z,\omega)$ obeys the Helmholtz equation, the angular spectrum at any other plane of interest $z=z_1$ can be obtained by multiplying the initial angular spectrum by a k-space propagation operator,

$$\tilde{P}(k_x,k_y,z_1,\omega) = \tilde{P}(k_x,k_y,z_0,\omega)e^{i(z_1-z_0)\sqrt{k^2-k_x^2-k_y^2}}, \tag{3}$$

where k is the free-space wave number. The relationship described as a product in k-space [Eq. (3)] can be written as a convolution in the space domain,

$$P(x,y,z_1,\omega) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(\acute{x},\acute{y},z_0,\omega)G(x-\acute{x},y-\acute{y}_1,z_1 - z_0)d\acute{x}d\acute{y}, \tag{4}$$

where $G(x,y,z)$ is the inverse spatial Fourier transform of the propagation operator and can be explicitly written[28] as

$$G(x,y,z) = \frac{e^{ikr}}{2\pi r}\left(\frac{1}{r} - ik\right)\frac{z}{r}, \tag{5}$$

where $r=\sqrt{x^2+y^2+z^2}$. For the purposes in this work, the angular spectrum method was symbolically represented in the form given in Eqs. (4) and (5). Numerically, it was implemented by using the MATLAB convolution algorithm which employs two dimensional (2D) Fourier transforms.

B. Wave propagation model for the piezoelectric lithotripter array

The array was considered as a source with L elements and focal axis along the z direction and the geometric focus taken to be the origin. The angular spectrum technique is not easily adapted to model a source condition prescribed on a curved surface. Furthermore, a large area needs to be discretized in order to obtain the source condition at the face of the array, thus making the model computationally cumbersome. Therefore, the approach employed in this study involved individually measuring the impulse response of each element on a prefocal plane ($z_0=-20$ mm) and then propagating each measurement forward to the focal plane by using the angular spectrum method. The pressure at any point (x,y) on the $z=z_0$ plane can then be defined as

$$p(x,y,z_0,t) = \sum_{l=1}^{L} h_l(x,y,z_0,t-\Delta t_l), \tag{6}$$

where $h_l(x,y,z_0,t)$ is the impulse response of element l measured at (x,y,z_0) and Δt_l is the delay for the lth element that will be used to control the acoustic field. Equation (6) can be rewritten in the frequency domain as

$$P(x,y,z_0,\omega) = \sum_{l=1}^{L} H_l(x,y,z_0,\omega)e^{i\Delta t_l\omega}. \tag{7}$$

The pressure field described in Eq. (7) is then propagated to the focal plane by using the angular spectrum approach. The angular spectrum operator commutes with the summation and time delay factor yielding a field at $z=z_1$ which can be expressed as

$$P(x,y,z_1,\omega) = \sum_{l=1}^{L} e^{i\Delta t_l\omega}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H_l(\acute{x},\acute{y},z_0,\omega)G(x-\acute{x},y - \acute{y},z_1-z_0)d\acute{x}d\acute{y}. \tag{8}$$

Numerically, the equivalent is achieved in three steps. First, the impulse response field for each element, $H_l(z_0,\omega)$ for lth element, is propagated to the plane of interest $(z=z_1)$ by using Eq. (4). Second, the propagated impulse response field $H_l(z_1,\omega)$ is multiplied by $e^{i\Delta t_l\omega}$ which describes the time delay applied to each element. Finally, the cumulative pressure field resulting from the array is simulated by invoking the linear superposition of pressure and summing the products of the propagated impulse response and the delay term for all the elements. This procedure captures both the diffraction from each element and the interelement delay.

C. Optimization scheme

The optimization scheme was implemented with the goal of determining the delays of each element so that the difference between a prescribed pressure field $\hat{P}(x,y,z,\omega)$ and a simulated pressure field $P(x,y,z,\omega)$ was minimized. The prescribed pressure field was defined on a plane at $z=z_1$ over M control points $\hat{P}(x_m,y_m,z_1,\omega)$. The simulated pressure at the $z=z_1$ plane resulting from a set of delays Δt_l is given by Eq. (8), which can be rewritten for each point m in the control plane as the following sum over L elements:

$$P_m(\omega)=\sum_{l=1}^{L}\hat{H}_{ml}(\omega)e^{i\Delta t_l\omega}, \tag{9}$$

where $\hat{H}_{ml}(\omega)=H(x,y,z_0,\omega)*_{x,y}G(x,y,z_1-z_0)|_{x_m,y_m}$ and H_{ml} represents the transfer function from the lth element to the mth control point. The simulated pressure at all M points can be represented by using M equations which can be written in matrix form as

$$\mathbf{P}(\omega)=\hat{\mathbf{H}}(\omega)\mathbf{D}(\omega), \tag{10}$$

where $D_l(\omega)=e^{i\Delta t_l\omega}$, $l=1,\ldots,L$. The goal is to choose each Δt_l, such that the simulated pressure $\mathbf{P}(\omega)$ best matches the prescribed pressure $\hat{\mathbf{P}}(\omega)$.

One optimization approach is to employ a direct inverse filter, *à la* Tanter *et al.*,[24] in which Eq. (10) can be solved for $\mathbf{D}(\omega)=[\hat{\mathbf{H}}(\omega)]^{-1}\hat{\mathbf{P}}(\omega)$. Implementation of the inverse solution obtained in this manner results in a $\mathbf{D}(\omega)$ with varying amplitude and phase. The high voltage pulsers used to drive the transducer array, however, produce a pulse of fixed amplitude and shape; that is, the hardware imposes a constraint of $|D_l(\omega)|=1$. Further, because the only control available is over the time delay Δt_l of an element, the phase must be linear in frequency of the form $\arg[D_l(\omega)]=\Delta t_l\omega$.

Therefore, in this work an alternate method involving nonlinear regression of squared error was used to solve for optimal delays. Because our motivation is to control the pressure amplitude, we choose the objective function to be

$$|\hat{P}(\omega_n)|=|P(\omega_n)|. \tag{11}$$

That is, we do not specify any phase variation. The impact of this is considered in the Discussion. The error function to be minimized was the square error summed over all frequencies and all spatial control points,

$$\Pi(\Delta t_l)=\sum_{n=1}^{N}\left\{\sum_{m=1}^{M}W_m[|\hat{\mathbf{P}}_\mathbf{m}(\omega_n)|-|\mathbf{P}_\mathbf{m}(\omega_n)|]^2\right\}, \tag{12}$$

where W_m was a weighting function defined across the control points to allocate sensitivity of the optimization routine to the error at the mth point. The weighting function was independent of frequency and was dependent on space alone. The optimization was performed by using the optimization toolbox in MATLAB (Release 14, Mathworks, Natick, MA). The optimization function used was "lsqnonlin" which employs the Levenberg–Marquardt algorithm[29] for determining the set of delays that minimizes the error function [Eq. (12)].

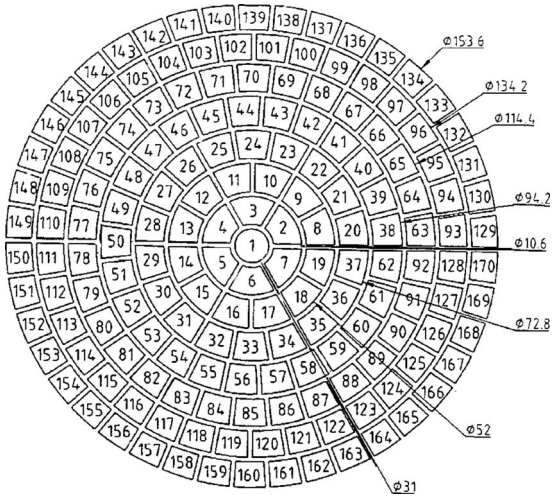


FIG. 1. A diagram of transducer elements in the spherical bowl and their numerical identification.

The following parameters were used for all the optimization runs. The maximum number of functional evaluations was 32 600 and maximum number of iterations was 500. The tolerance for functional evaluation was 10^{-4} and the tolerances for Δt_l were 1 ns for the wide-focus optimization and 10 ns for the ring focus optimization. In all cases, the optimization was terminated when the magnitude of the search direction was less than the tolerance for Δt_l.

III. EXPERIMENTAL SETUP

The instrumentation and experimental methods will be presented in this section. The section addresses three topics: the description of the source, the instrumentation for acoustic pressure measurements, and methods for customizing the acoustic field.

A. Piezoelectric lithotripter array

The piezoelectric lithotripter array used in this study comprised of a 170 element focused array (Imasonic S. A., Besancon, France). The elements were placed on a spherical cap with a radius of curvature of 150 mm and an aperture diameter of 154 mm. The nominal center frequency of each element was 600 kHz ($Q=2$). The transducer was constructed from a "1–3" piezocomposite material which consists of thin ceramic rods encapsulated within a polymer matrix.

The layout of the 170 elements (referred to as e_1 through e_{170}) is shown in Fig. 1. The central element (e_1) had circular cross section, and the other elements were trapezoidal in shape placed on seven axisymmetric rings. Each element had a surface area of 88 mm^2. The maximum permissible driving voltage was 6 kV. The average electrical impedance of an element when the array was in water at the center frequency (600 kHz) was $Z=313-j504\ \Omega$.

The piezoelectric array was driven by a 170 channel high voltage pulser (Gammell Applied Technologies, LLC, Exmore, VA). The circuit details are described in Ref. 30, and the version used here was the nondoubling circuit. Briefly, each channel of the pulser employed a 220 nF ca-

pacitor which can be charged to 1200 V. A high voltage insulated gate bipolar transistor (IGBT) was used to discharge the capacitor through the transducer element. Each channel was independently controlled by a TTL-level pulse generated by a 196 line pulse-pattern generator (81104A, Agilent, Palo Alto, CA). Both the duration and the timing of the TTL pulse for each channel could be independently adjusted with a 100 ns resolution. A single high voltage power supply (N02HP30, Acopian, Easton, PA, USA) was used to charge the capacitors. The 1200 V limit was governed by the breakdown voltage of the IGBT.

The transducer was mounted into the wall of a cast-acryllic test tank of dimensions of $0.5\times0.9\times0.5$ m^3. The tank was fitted with a filtration and degassing system. A 5 μm subparticle filter (02913-30, Cole-Parmer, Vernon Hills, IL) along with a magnetic nonsubmersible pump was used for filtering the water. A pinhead degassing system[31] was used for degassing the water. The dissolved gas content of the water in the test tank was maintained at approximately 40% of gas-saturated water. Positioning in the test tank was accomplished by using a computer controlled three axis positioner (Velmex, Inc., Bloomfield, NY) with a nominal spatial resolution of 5 μm.

B. Acoustic measurements

A polyvinylidene fluoride (PVDF) membrane hydrophone (model 0200, Precision Acoustics, Dorchester, England) with a 0.2 mm diameter active area and a bandwidth of 30 MHz was used for acoustic measurements where only one element was excited at a time. The hydrophone, referred to as the PA hydrophone hereon, used in conjunction with a 20 dB preamplifier (50 Ω termination), had a sensitivity of 33.3 MPa/V (flat within 2 dB over a frequency range of 0.1–20 MHz). The PA hydrophone signal was passed through a tunable high pass filter (model 3940, Kronhite, Avon, MA) with a cutoff frequency of 10 kHz and gain of 20 dB with a nominal frequency range of 0–10 MHz to further improve the signal to noise ratio. The input impedance of the high pass filter was 1 MΩ and output impedance was 50 Ω. A 50 Ω termination was attached to the input of the high pass filter to match the output impedance of the hydrophone's preamplifier.

Shock wave measurements were performed by using a fiber optic probe hydrophone (FOPH) or a second PVDF membrane hydrophone. The FOPH (model 500, R. P. Acoustics, Leutenbach, Germany) had a bandwidth of 0.1–30 MHz and a receiving aperture of 100 μm. This hydrophone measures a change in refractive index of water which can be related to the pressure in the water.[32] The PVDF hydrophone was developed at University of Washington,[33] referred to as UW hydrophone hereon, and had a bandwidth of 0.1–20 MHz, sensitivity of 35.8 MPa/V, and a receiving aperture size of 0.5 mm. Individual waveforms were sampled at a rate of 100 MHz on a digital oscilloscope (Waverunner 6000A, Lecroy, Chestnut Ridge, NY) with 8 bit resolution and 200 MHz low-pass filter and transferred to a computer for postprocessing by using MATLAB (Mathworks, Natick, MA).

C. Customization of acoustic output

The response $h(x,y,z_0,t)$ was obtained for each element at $z_0=-20$, that is, 20 mm prefocal, on a 50×50 mm^2 grid with internode distance of 0.5 mm (10 000 points) and then zero padded to 80×80 mm^2 (25 600 points) to reduce wrap-around errors that may occur in evaluation of convolution by using fast Fourier transform. The grid size of 50×50 mm^2 was the minimum size necessary to faithfully include all the features of the impulse response from all the elements. It will be shown in Sec. IV A that the majority of the energy (97%) of the pulse produced by each element lies in the 0–1.5 MHz band. Therefore, the angular spectrum approach was implemented within this band. The internode distance of 0.5 mm was the maximum possible distance that satisfied the Nyquist sampling criterion of $\lambda/2$.

The optimization scheme required an initial guess for the delays. For inversion problems that are not robust, careful selection of initial guess for the delays is crucial to converge to the global minimum. To further facilitate convergence of the optimization routine toward the desired solution, a weighting function was used to allocate a higher sensitivity of the error function [Eq. (12)] at key locations of interest. The choice of initial condition and weighting function varied depending on whether a broad focus or a ring-shaped focus was desired.

For the prescribed acoustic fields of increased −6 dB focal widths, a constant delay (1 or 3 μs) was assigned to all elements as the initial state for the optimization routine. This was done to prevent negative results for time delays. The weighting function used for this series of optimizations placed a higher emphasis on points along the circle describing the desired focal width as shown in Fig. 2(a).

The constant delay condition as the initial state for the optimization failed to produce the ring-shaped pressure field. Apparently, the optimization method used was unable to overcome the natural tendency of a spherically focused array to produce a peak pressure at the geometric focus. In order to better guide the solution, the initial delays were selected in order to focus groups of elements onto various points on desired ring. Eight points were chosen on the ring and the elements were divided into eight corresponding groups. Elements 1, 9, 17,…,169 comprised the first group, elements 2, 10, 17,…,170 were in the second group, and so on. The initial delays were set so that each group focused on its specified point on the desired ring. Since the ring-shaped pressure distribution critically depends on reducing the pressure at the geometric focus, its weighting function gave highest emphasis at the geometric focus extending out to the ring diameter in the form a 2D Gaussian function, as shown in Fig. 2(b).

IV. RESULTS

The results will be presented in three sections. First, the acoustic output of a single element of the piezoelectric lithotripter array is characterized. Second, the acoustic output of the entire piezoelectric lithotripter array without implement-

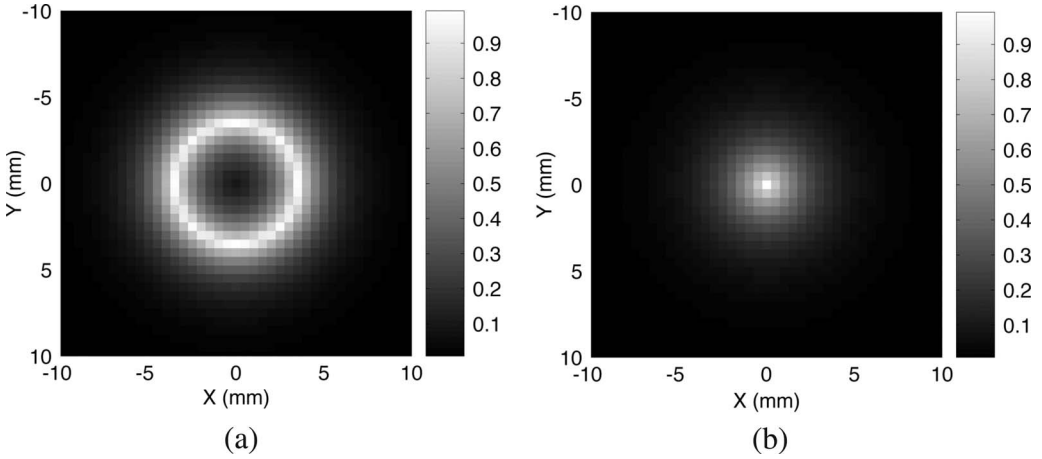


FIG. 2. The weight function employed to control the emphasis given to error at each node on the optimization plane for customizing the spatial distribution of pressure. (a) For the wide focus, a ringlike weighting function was used. (b) For the ring focus, the weighting function was highest on the axis.

ing any customization of the acoustic output is described. Third, results from customizing the spatial pressure distribution will be shown.

A. Single element characterization

A set of ten waveforms obtained from a single element (e_1) driven at 1.2 kV pulse repetition frequency (PRF) of 1 Hz is shown in Fig. 3(a). The received signal was highly reproducible with a temporal jitter of less than 30 ns and is characteristic of waveforms produced by all of the elements. The maximum temporal jitter between any two elements was determined to be less than 50 ns. The peak positive pressure produced from a single element was observed to be around 0.2 MPa at the focus of the piezoelectric array. Figure 3(b) shows the frequency spectrum of the pressure measurement obtained from e_1 at the focus. The emitted signal from the transducer element is a wide-band signal with a peak close to the specified center frequency of 0.6 MHz. The majority of the energy (97%) lies below 1.5 MHz and so the inversion only used signal content in this band. The discrepancy in the peak pressures due to this was less than 15%.

The map of the acoustic response of each element at a prefocal plane ($z_1=-20$ mm) is necessary for implementing the angular spectrum method in the wave propagation model. Formally, one should measure the field for each element in the array, however, for each element this requires measuring 10 000 waveforms, for a total of 1.7×10^6 measurements. We note that because each element has the same area and measurements were obtained in the far field, one could theoretically determine all the data from measuring one element. However, we found variations in field maps from ring to ring in excess of 20%. Therefore, measurements from one element per ring were used for initialization. The response of any other element on that ring was obtained by a rotation based on the difference in angle of the two elements. For instance, the map of $P+$ of e_{170} on the axisymmetric plane $z=z_0$ can be obtained by rotating the response of e_{135} measured on the same plane by 60°, as shown in Fig. 4. The rotation of the field was performed by using the MATLAB function "imrotate"with the linear interpolation scheme. Measurements from multiple elements on multiple rings were conducted to confirm that the geometric transformation described above was sufficient to accurately predict the response of any element on a ring. Thus, the response of all 170 elements could be determined from eight measurements.

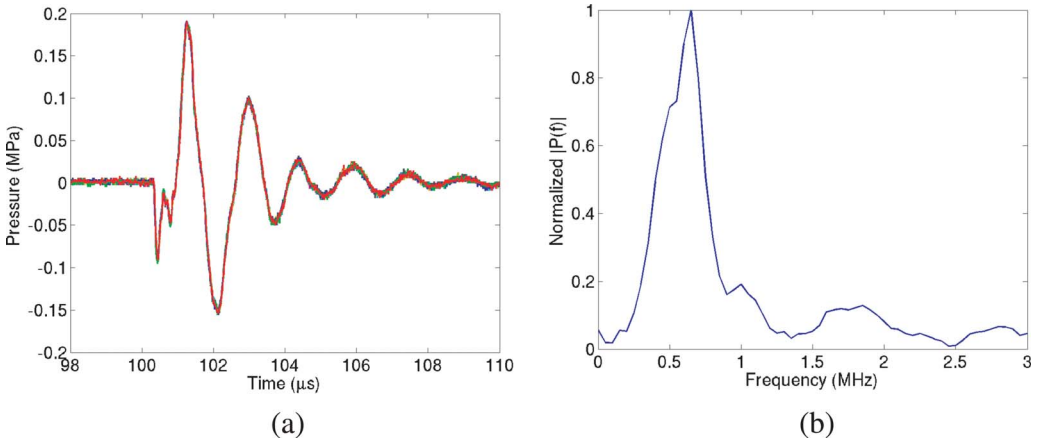


FIG. 3. (Color online) (a) Overlay of ten consecutive waveforms obtained from e_1 driven with 1.2 kV pulse lasting 0.5 μs at a PRF of 1 Hz. (b) Amplitude spectrum of the pressure waveform shown in (a).

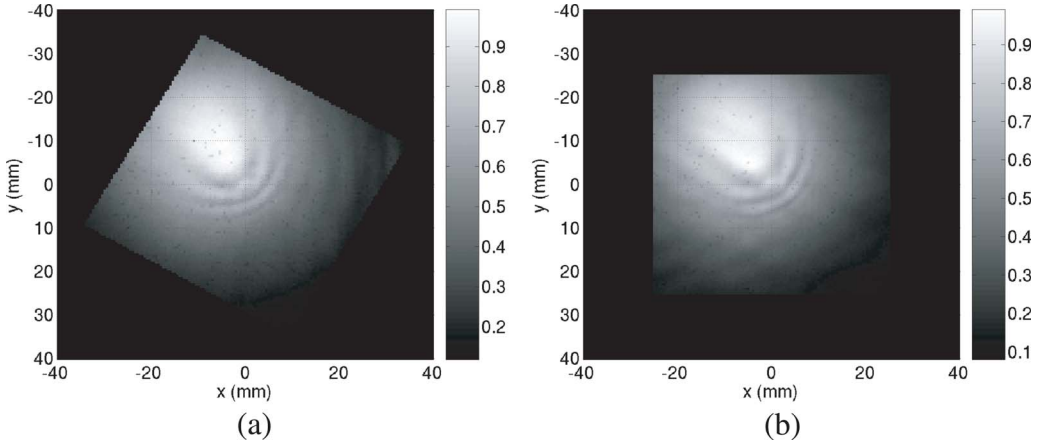


FIG. 4. (Color online) (a) Normalized impulse response of e_{135} measured 20 mm prefocal and rotated by 60°. (b) Normalized impulse response of e_{170} measured 20 mm prefocal.

B. Piezoelectric lithotripter array characterization

The piezoelectric lithotripter array was first characterized while all elements were synchronously driven, that is, without manipulating the acoustic output. Figure 5 shows the waveform produced by driving the array with a 1.2 kV pulse. The waveform shown was obtained by first measuring ten waveforms with the PVDF hydrophone, aligning their main shock fronts, and then averaging the waveforms. Unlike the more typical lithotripsy waveforms generated by electrohydraulic and electromagnetic devices,[34] the waveform has four distinct features. First, a negative precursor of 0.5 μs long and a negative pressure of $P_1-=-10$ MPa followed by the main shock front with a positive peak $P_1+=68$ MPa. The rise time based on time difference between 10% of P_1+ and 90% of P_1+ was 30 ns and was limited by the 20 MHz bandwidth of the UW hydrophone. The duration of the main positive phase was $t_{1+}=0.4$ μs. The negative phase following the main positive peak was 1 μs long with peak negative pressure $P_2-=-15$ MPa. The second shock front had a peak positive pressure $P_2+=36$ MPa. The entire shock wave pulse had a duration $t=5$ μs measured as the time between when the pressure first exceeds 10% of P_1+ and when the pressure last exceeds 10% of P_1+. Figure 6 shows the mean P_{1+} and P_{2-} as a function of charging voltage. The error bars represent the standard deviation of ten measurements. Measurements indicated a monotonic increase in pressure with increase in the driving voltage.

The spatial pressure distribution of the acoustic field produced by the array was also mapped. Since the maximum operating voltage for the IGBTs used in the high voltage drivers was 1.2 kV, pressure maps were acquired while driving the piezoelectric array lithotripter at 1 kV to prevent overload from any voltage fluctuations. Pressure measurements were acquired with the FOPH as its 100 μm fiber diameter provided better spatial resolution than the UW hydrophone, albeit with a modest decrease in signal to noise ratio. The equivalency between the measurements acquired via both the hydrophones was verified by acquiring a waveform measurements at the focus. Figure 7 shows the average of ten waveforms acquired by using both hydrophones at the focus when the array was driven by using a 1.2 kV pulse. All prominent features of the waveform and P_{1+} are in good agreement with each other. The pressure measurements along all three axes are shown in Fig. 8. The top curve in each figure shows P_{1+} and the lower curve P_{2-} with respect to displacement from the focus. The −6 dB isobar of the peak positive pressure of the main shock front P_{1+} was used to describe the focal region which was thus determined to be $1.4\times1.9\times15$ mm^3.

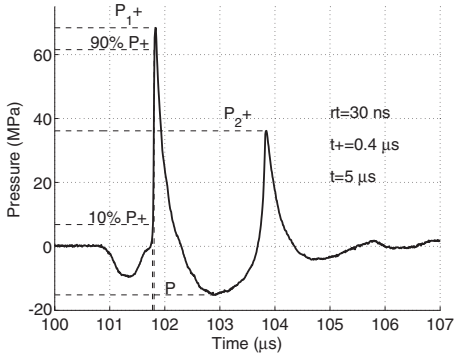


FIG. 5. A characteristic waveform produced by the piezoelectric lithotripter array when all elements were synchronously driven. The displayed waveform is an average of ten waveforms recorded consecutively at a PRF of 1 Hz.

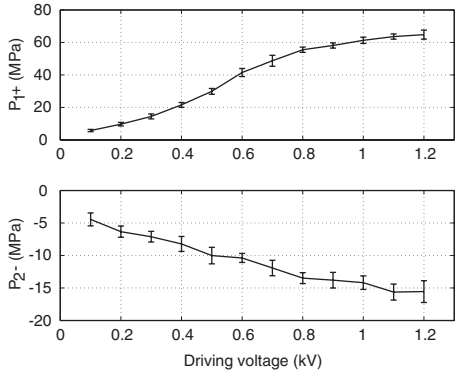


FIG. 6. Peak positive pressure (P_{1+}) and peak negative pressure (P_{2-}) as a function of driving voltage. Peak pressures monotonically increased with increase in driving voltage.

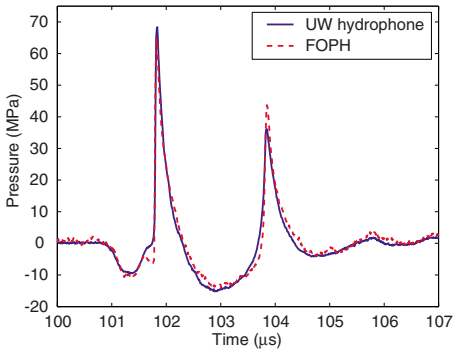


FIG. 7. (Color online) Comparison of waveforms measured by the FOPH and UW hydrophone at the focus while synchronously driving the array with a 1.2 kV pulse. Waveforms are the average of ten measurements.

C. Customization of spatial pressure distribution

The optimization routine was employed by using prescribed fields of different focal widths defined by −6 dB of P_{1+}. The optimal set of delays for the array, for each focal width, was first investigated by using the forward propagation model discussed in the previous section and then programmed into the pulse pattern generator and the resulting pressure field measured.

Figure 9 shows the simulated pressure distribution (based on a driving voltage of 1.2 kV) of P_{1+} along the lateral axis X obtained from the optimization routine for varying focal widths. The displayed results were obtained from the linear forward wave propagation simulations by using the set of optimal delays determined for each of the desired −6 dB focal widths. As the only control parameter used for optimization is the time delay for each element and the optimization is based on minimization of least squares errors, the resulting simulated pressure fields do not exactly match the prescribed pressure field. However, an increase in the desired acoustic focal width resulted in an increase in the focal width of the simulated pressure obtained by using the optimal time delays. Prescribed focal widths of 3, 4, and 5 mm resulted in simulated focal widths of 3.4, 3.6, and 5.1 mm. As one might expect the "broadening" of the focal width results in a decrease of the peak pressure. The simulated focal width obtained by setting the delays to zero was 2.9 mm, which is different from the measured focal width reported in the previous section. The discrepancy can be attributed to the fact that the simulation is linear and therefore does not include the higher harmonics generated due to the nonlinear effects which subsequently lead to tighter focusing of the acoustic field.

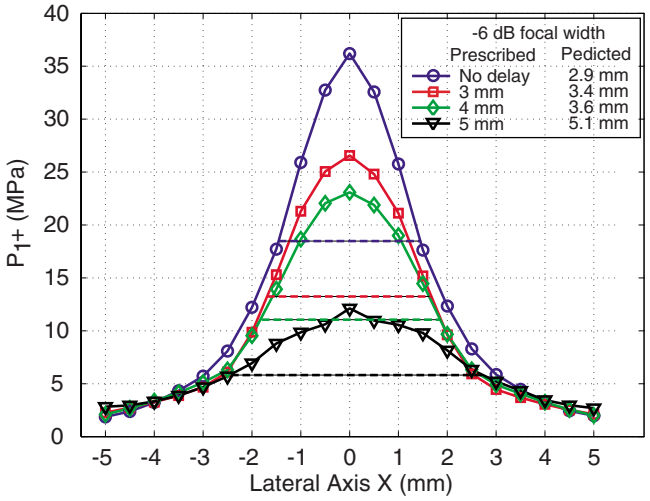


FIG. 9. (Color online) Simulated distribution of P_{1+} along the lateral axis X produced from the optimization routine for prescribed focal widths of 3, 4, and 5 mm. The optimal delays were predicted to produce focal widths of 3.4, 3.6, and 5.1 mm. The predicted focal width when all elements were synchronously fired was 2.9 mm.

The effectiveness of the lsqnonlin function in obtaining the optimal delays for achieving a pressure field of predetermined focal width was investigated via behavior of the error residual and the magnitude of the directional derivative obtained for each iteration during the optimization routine shown in Fig. 10. For the prescribed focal width of 5 mm, the optimization routine effectively reduced the residual and the magnitude of directional derivative, which indicates that the optimization routine approached a primary minimum. The derivative exhibited a significant reduction in oscillations through the course of optimization, indicating that the optimization surface was smooth and that the optimization might be converging toward a unique solution. Similar performance was achieved for the other focal widths.

The lateral pressure distributions measured upon implementing the delay sets corresponding to each prescribed −6 dB focal width are shown in Fig. 11. The driving voltage for each run was adjusted so that magnitude of P_{1+} at $X=0$ was roughly 40 MPa for all focal widths with the long term goal of investigating the dependence of stone fragmentation

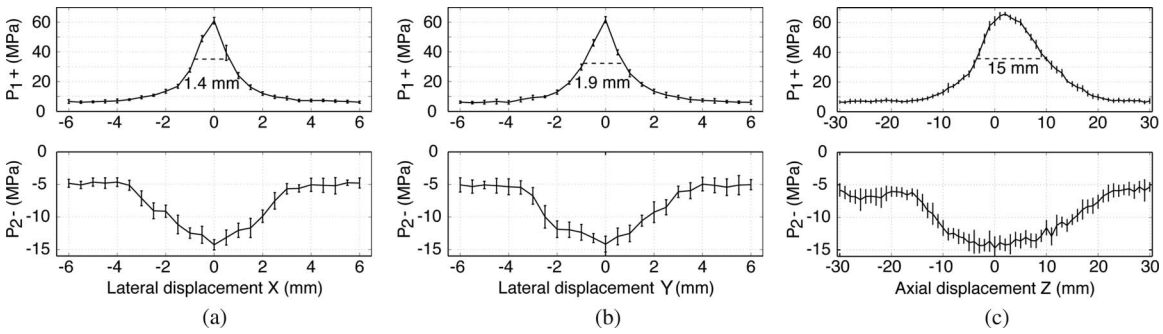


FIG. 8. Peak pressure measurements along the X, Y, and Z (acoustic) axes. The −6 dB focal region based on the primary peak positive pressure was 15 mm long and between 1.4 and 1.9 mm wide.

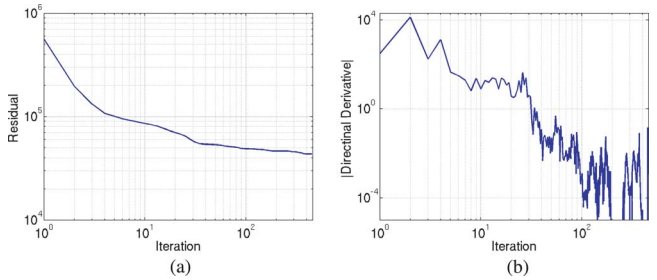


FIG. 10. (Color online) Parameters for monitoring the optimization routine for the prescribed −6 dB focal width of 5 mm. (a) Error residual. (b) Magnitude of the directional derivative.

on focal width alone. The driving voltages for 5, 4, and 3 mm focal widths were 1200, 850, and 800 V, respectively. For the case where the elements were synchronously driven, the field with a maximum P_{1+}=40 MPa was obtained by driving the elements at 575 V.

Similar to the wave propagation simulations, the measured focal widths increased with increase in prescribed focal width. Prescribed focal widths of 3, 4, and 5 mm resulted in simulated focal widths of 2.8, 3.1, and 3.9 mm. The focal width measured when all elements were synchronously driven was 1.7 mm. The focal width of 3.9 mm was the maximum possible focal width we were able to obtain while maintaining P_{1+}=40 MPa. As shown in the previous section, the focal width when all elements are driven at 1 kV with zero delay was 1.4 mm along the X axis. The optimization routine scaled the focal width by a factor of over 2 with a 33% reduction in the magnitude of P_{1+}.

Both the predictions and measurements show that the increase in focal width was achieved with a loss of peak pressure. This was investigated further by scanning the hydrophone in the axial direction. Figure 12 shows that peak pressure has been shifted from the geometrical focus to a point of 8 mm closer to the transducer. That is, the optimization routine's solution was to shift the focus closer to the transducer and exploit the postfocal divergence to create a broad focus.

The optimization routine was also implemented for obtaining a ring-shaped pressure distribution with a diameter of 4 or 7 mm. Figure 13 shows simulated and measured spatial pressure distribution based on P_{1+} and P_{2-} for both cases.

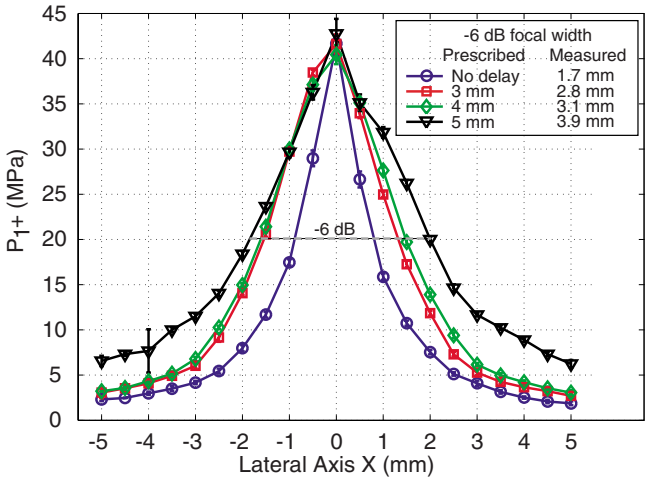


FIG. 11. (Color online) Measured distribution of P_{1+} along the lateral axis X produced from the optimization routine for desired focal widths of 3, 4, and 5 mm. The optimal delays produced focal widths of 2.8, 3.1, and 3.9 mm. The focal width when all elements were synchronously fired was 1.7 mm.

The optimization routine was successful in producing a ring-shaped focal field. The simulated field, obtained by using the linear wave propagation model, and the measured field, when the array was driven by using the set of optimal delays determined for each ring diameter, were in good agreement. The discrepancy in the magnitude of P_{1+} can be attributed to the neglect of nonlinear steepening in the angular spectrum approach.

The error residual and magnitude of the directional derivative for both the 4 and 7 mm rings are shown in Fig. 14. The residual exhibited a modest decrease unlike the case for the optimization results for wider focal width shown in Fig. 10. Even so, the set of delays obtained produced an acceptable ring-shaped pressure distribution. The derivative, on the other hand, shows high oscillations around a large magnitude. This indicates that optimization surface is quite rough, and that it is likely that many different sets of delays could produce tolerably good ring-shaped pressure distributions.

Figure 15 shows the measured distribution of P_{1+} along the lateral axis X for ring diameters of 4 and 7 mm (driving voltage of 1.2 kV) compared to that obtained when elements are synchronously driven at 1 kV. The "no delay"waveform

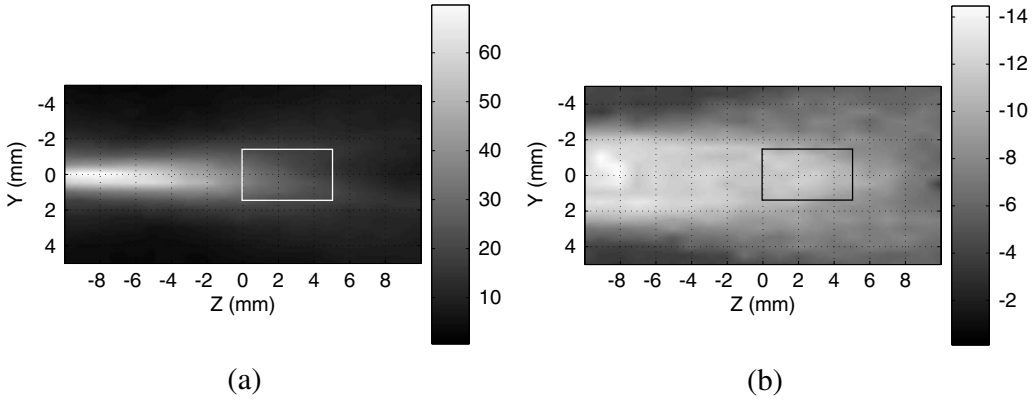


FIG. 12. Axial pressure maps for −6 dB focal width of 3.9 mm from measurements acquired by using the PVDF hydrophone. (a) Map of P_{1+}. (b) Map of P_{2-}. The rectangle denotes the spatial extent of a typical stone. Gray scale represents pressure in MPa.

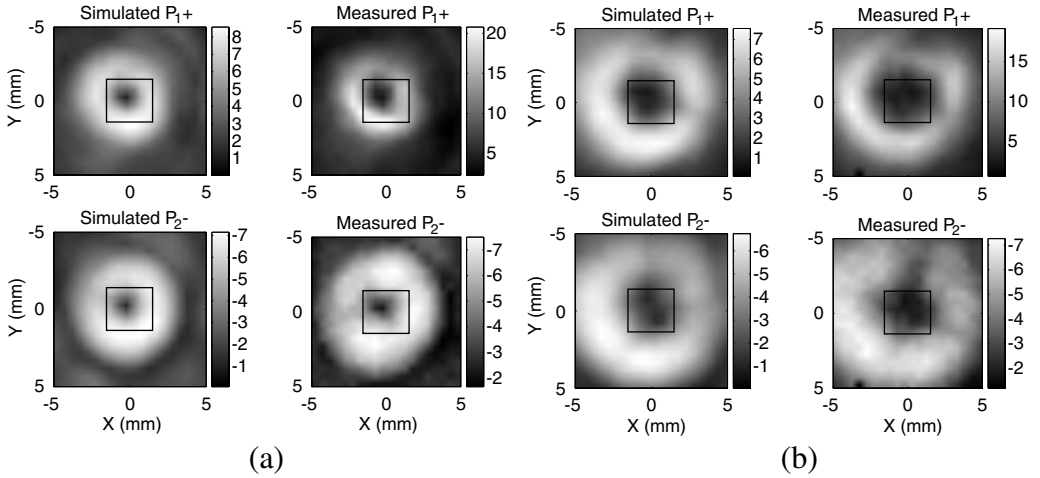


FIG. 13. The results from optimization routine aimed at obtaining a ring-shaped pressure distribution. Left frames are field maps obtained by using the linear propagation model. Right frames are field maps from measurements acquired by using the PVDF hydrophone. (a) Prescribed ring diameter of 4 mm. (b) Prescribed ring diameter of 7 mm. Gray scale represents pressure in MPa.

had a maximum pressure in excess of 60 MPa. In contrast, both ring pressure distributions had a pressure minimum on axis with a peak less than 5 MPa. The ring focal field peaked off axis at an amplitude higher than the no delay field. However, the peak pressures did not exceed 20 MPa.

The relevant parameters from the ring optimization are shown in Table I. Both the simulations and the measurements were obtained at a spatial resolution of 0.5 mm. The ring diameter was determined along both lateral axes based on the maximum value of P_{1+} in the focal plane. Both the simulated and measured ring diameters resulting from the optimization routine were consistent with the prescribed ring diameter. The ring thickness based on −6 dB of P_{1+} (local peak) was between 2 and 4 mm in all the cases. The magnitude of P_{1+} at the focus obtained by using the angular spectrum simulation was around 1 MPa and that observed through measurements was 4 MPa.

The motivation of the ring focus was to produce a wavefront that traveled around the periphery of a kidney stone to efficiently couple into shear waves in the stone. This requires that the distribution remain ring shaped over an axial extent of about 10 mm. The optimization routine invoked to obtain the delays necessary to produce a ring-shaped pressure field prescribed the pressure only at one plane. Figure 16 shows the distribution of P_1+ and P_2- over the axial plane for acoustic field with a ring of 4 and 7 mm in diameter. The measurements indicated that the optimization routine for producing ring acoustic fields does indeed reduce the pressure along the acoustic axis and the ring shape is maintained over typical length of a stone.

Sample waveforms at four points that lie on the peak of the ring pressure field are displayed in Figs. 17(a). Since the ring pressure distribution is not exactly symmetrical, four sampling locations were chosen from the points that constitute the maximum ring pressure. The measured waveforms indicate that different portions of the ring consist of different waveform shapes. The measurements indicate that the time of arrival of the waveforms is not uniform at all locations on the focal plane. This can be attributed to the optimization routine with only time delays as the control parameter in which the pressure distribution was the only constraint and no constraint was placed on phase. The difference in arrival times of pulses at different locations on the ring presents a need for including an arrival time constraint in the optimization routine for the ring pressure distribution (which was beyond the scope of this work). The variation in the arrival

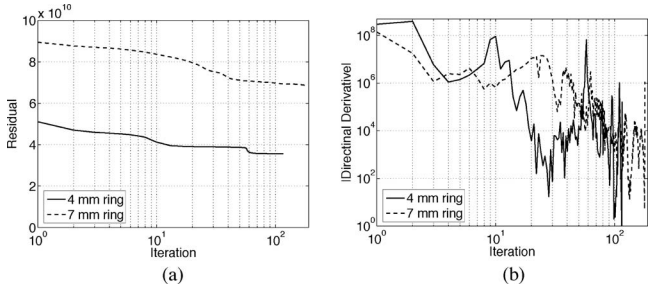


FIG. 14. Optimization results for ring-shaped pressure fields of 4 and 7 mm in diameter. (a) Error residual. (b) Directional derivative. During the course of the optimization, the residue, and the derivative exhibited a modest decrease, indicating that optimization runs aimed at producing a ring focus might have encountered a local minimum.

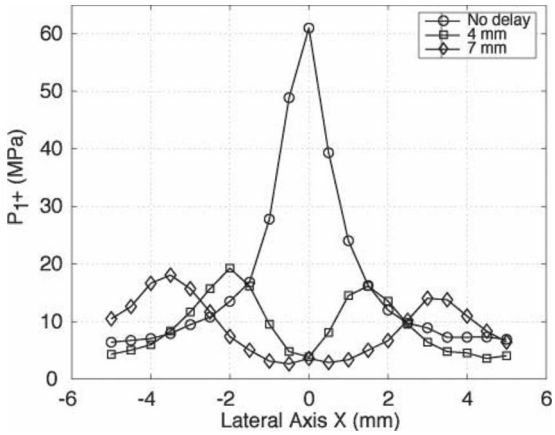


FIG. 15. Comparison between pressure distribution (P_{1+}) along a lateral axis (X) obtained from measurements across the focal plane for acoustic field obtained by synchronously driving elements and the ring pressure distribution of diameters of 4 and 7 mm.

TABLE I. Parameters from ring field optimization.

	Simulated		Measured	
Prescribed diameter (mm)	4	7	4	7
X diameter (mm)	3.5	6	3.5	6.5
Y diameter (mm)	3.5	6	3.5	6
Mean ring thickness (mm)	≈3	≈3	≈2	≈3
P_{1+} at focus (MPa)	1	1	4	4

time in space is illustrated in maps of the arrival time of waveform based on the largest peak (typically P_1+) in Fig. 17(b). The map indicates that the arrival times varied by up to 2 μs along the region that corresponded to the ring of high pressure. The discrete edges correspond to locations where peak pressure changes from P_{1+} to P_{2+}.

V. DISCUSSION

The work presented describes an approach with which a piezoelectric array could generate a prescribed acoustic field with limited control over the drive signal. The algorithm was able to successfully broaden the focus and produce ring-shaped focii. The acoustic fields produced by implementing the delays determined by the optimization routine were not an exact match to the prescribed acoustic fields. This was, in part, because the delays for the elements of the array were the only control parameter in our optimization and, in part, because the numerical propagation model was linear and could not capture the nonlinear effects. Also, in both scenarios, the resulting pressure distributions were achieved at the price of reduced peak pressure.

In the case of the wide focal region, the optimization function electronically shifted the focus toward the source to achieve the objective. This is equivalent to placing a regular acoustic lens that shifts the focus and satisfies the requirement for broader pressure region in the focal plane but clearly results in large acoustic pressures in off-target locations which may not be desirable, particularly in clinical applications. To prevent this from happening, further constraints along the acoustic axis would need to be included in the optimization, e.g., a penalty if the pressure at other axial locations is greater than the pressure at the focus. In the broadest case, one could specify the pressure field in an entire volume around the geometrical focus, rather than just one plane, and optimize for that case. The broadening of the focal region can also be achieved in a more *ad hoc* manner by adjusting the delays to emulate a slightly heterogeneous phase screen to broaden the focus without shifting it. However, the motivation for this work was to develop a methodology that would enable automatic evaluation of the optimal set of delays to best match a prescribed spatial pressure distribution, and therefore the *ad hoc* methods were not pursued.

The optimization function for obtaining the ring acoustic field was not as well behaved as that for obtaining a wider focal field, and convergence to a ring shape depended more heavily on the initial guess for the delays and the weighting

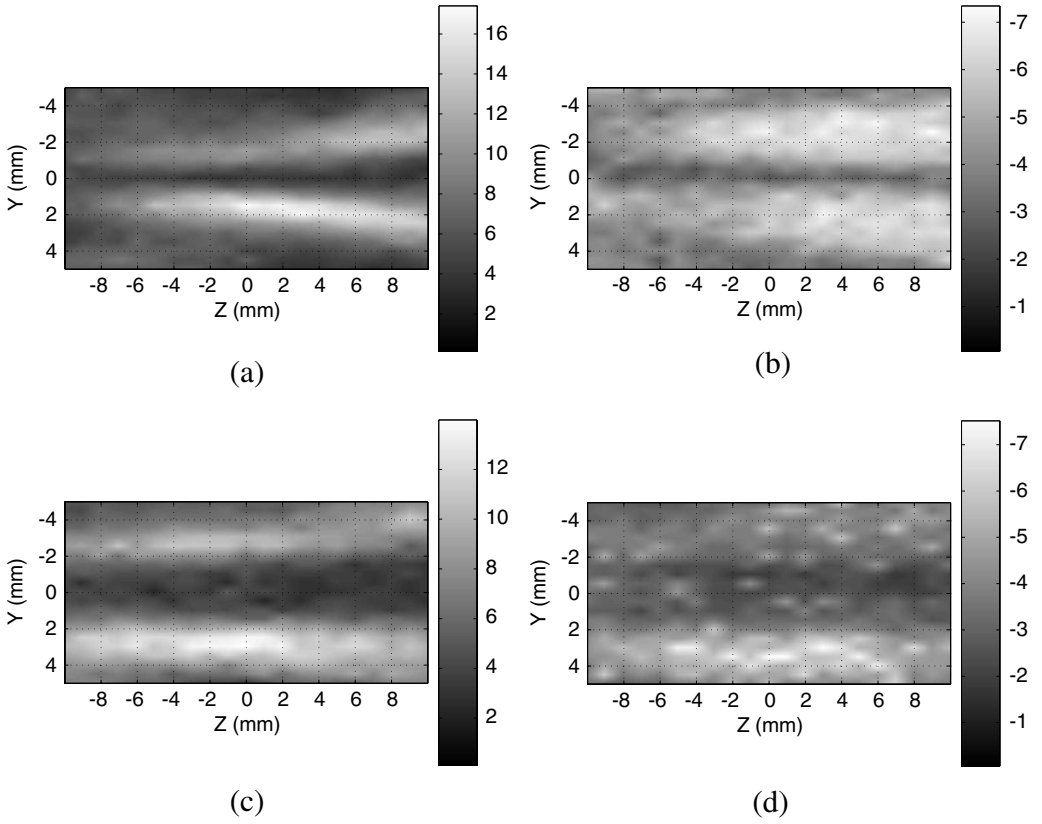


FIG. 16. Axial pressure maps for ring-shaped acoustic fields from measurements acquired by using the PVDF hydrophone. (a) Map of P_{1+} for the 4 mm ring. (b) Map of P_{2-} for the 4 mm ring. (c) Map of P_{1+} for the 7 mm ring. (d) Map of P_{2-} for the 7 mm ring. Gray scale represents pressure in MPa.

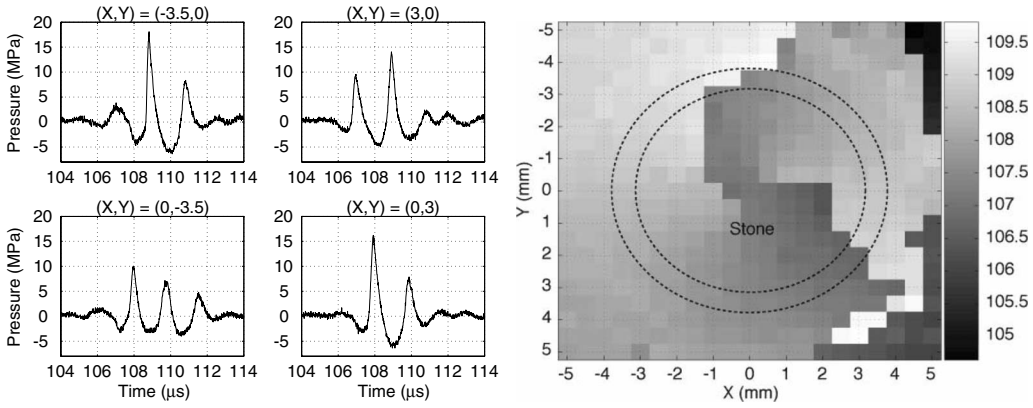


FIG. 17. (a) Sample waveforms measured at the focal plane when the array was driven by using the optimal delays for the 7 mm ring. The measurements indicate spatial variation in the time of arrival of the waveforms. (b) Map of arrival time of waveforms based on the largest peak for the 7 mm ring. The gray scale represents time in μs. The region between the two circles denotes the ring of high pressure and differences in delay of up to 2 μs can be seen. The discrete edges correspond to locations where peak pressure changes from P_1+ to P_2+.

function for the error. Further, the arrival times were not uniform around the ring (recall Fig. 17). Additional constraints that force the arrival times for the pulses along the ring to be uniform might also lead to a more robust convergence to the optimal set of delays for producing the ring acoustic field. We modified the error function by incorporating an additional term of the form $|\hat{P}-P|^2$ which incorporates phase difference. However, despite efforts with different weighting strategies, this modified error function resulted in a very modest reduction of phase error for the case of ring-shaped pressure field. Furthermore, this strategy did not lead to a significant reduction in the amplitude error. It appears that a more sophisticated optimization algorithm in conjunction with a modified error function is necessary for approaching a global minimum which might then lead to a ring-shaped pressure field where wavefronts on the ring simultaneously arrive. On the other hand, it may be impossible to obtain certain spatial distributions within the constraints of this device. For a system with expanded capabilities, such as independent control over both time delay and amplitude for each element, an inverse filter approach to manipulate acoustic vortices[35] might provide better means for producing a ring-shaped pressure distribution. A system that allows control over the waveform shape of the driving signal for each element can be used for implementing a time reversal approach[36] which could be exploited for customizing the acoustic field and automatically focusing the acoustic energy at the stone.

The model accurately predicted the relative spatial pressure distribution for both the wide-focus and ring-focus customizations. The main discrepancy was in the positive peaks and was attributed to the fact that the forward propagation model neglected nonlinear effects and thus failed to include the enhancement of the magnitude due to nonlinear distortion inherent in measurements. The fact that this discrepancy was present in the acoustic field maps derived from the peak positive pressure but not from that derived from the peak negative pressure supports this conjecture. Given the large number of iterations required by the optimization routine, incorporating a fully nonlinear propagation model is computationally infeasible at present. However, the use of a linear model provided an accurate enough model to modify the spatial distribution of the pressure field.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health through Grant No. DK43881. We thank Dr. A. P. Evan and Dr. J. A. McAteer for providing access to the fiber optic probe hydrophone, Dr. M. R. Bailey for the PVDF hydrophone, and Dr. R. A. Roy for helpful discussions and guidance.

[1]C. Chaussy, W. Brendel, and E. Schiemdt, "Extracorporeally induced destruction of kidney stones by shock waves," Lancet **2**, 1265–1268 (1980).
[2]M. A. Averkiou and L. A. Crum, in "Cavitation: Its role in stone comminution and renal injury," *New Developments in the Management of Urolithiasis*, edited by J. E. Lingeman and G. M. Preminger (Igaku-Shoin, New York, 1996), pp. 21–40.
[3]D. Jocham, C. Chaussy, and E. Schmiedt, "Extracorporeal shock wave lithotripsy," Urol. Int. **41**, 357–368 (1986).
[4]W. Eisenmenger, "The mechanisms of stone fragmentation in ESWL," Ultrasound Med. Biol. **27**, 683–693 (2001).
[5]S. M. Gracewski, G. Dahake, Z. Ding, S. J. Burns, and E. C. Everbach, "Internal stress wave measurements in solids subjected to the lithotripter pulses," J. Acoust. Soc. Am. **94**, 652–661 (1993).
[6]R. O. Cleveland and O. A. Sapozhnikov, "Modeling elastic wave propagation in kidney stones with application to shock wave lithotripsy," J. Acoust. Soc. Am. **118**, 2667–2676 (2005).
[7]A. J. Coleman, J. E. Saunders, L. A. Crum, and M. Dyson, "Acoustic cavitation generated by an extracorporeal shock wave lithotripter," Ultrasound Med. Biol. **13**, 69–76 (1990).
[8]L. A. Crum, "Cavitation micro-jets as a contributory mechanism for renal calculi disintegration in ESWL," J. Urol. (Baltimore) **140**, 1587–1590 (1988).
[9]A. Vogel and W. Lauterborn, "Acoustic transient generation by laser-produced cavitation bubbles near solid boundaries," J. Acoust. Soc. Am. **84**, 719–731 (1988).
[10]G. Delacretaz, K. Rink, G. Pittomvils, J. P. Lafaut, H. Vandeursen, and R. Boving, "Importance of the implosion of eswl-induced cavitation bubbles," Ultrasound Med. Biol. **21**, 97–103 (1995).
[11]X. Xi and P. Zhong, "Dynamic photoelastic study of the transient stress fields in solids during shock wave lithotripsy," J. Acoust. Soc. Am. **109**, 1226–1239 (2001).
[12]P. V. Chitnis and R. O. Cleveland, "Quantitative measurements of acoustic emissions from cavitation at the surface of a stone in response to a lithotripter shock wave," J. Acoust. Soc. Am. **119**, 1929–1932 (2006).
[13]R. E. Riedlinger, "Self-focusing piezoelectric high-power sound-pulser for

painless disintegration of urinary canal," *Ultrasonics International 87 Conference Proceedings*, pp. 220–225.
[14]P. Lewin, J. Chapelon, J. Mestas, A. Birer, and D. Cathignol, "A novel method to control $P+/P-$ ratio of the shock wave pulses used in the extracorporeal piezoelectric lithotripsy (PEL)," Ultrasound Med. Biol. **16**, 473–488 (1990).
[15]M. R. Bailey, D. T. Blackstock, R. O. Cleveland, and L. A. Crum, "Comparison of electrohydraulic lithotripters with rigid and pressure-release ellipsoidal reflectors. I. acoustic fields," J. Acoust. Soc. Am. **104**, 2517–2524 (1998).
[16]M. R. Bailey, D. T. Blackstock, R. O. Cleveland, and L. A. Crum, "Comparison of electrohydraulic lithotripters with rigid and pressure-release ellipsoidal reflectors. II. cavitation fields," J. Acoust. Soc. Am. **106**, 1149–1160 (1999).
[17]P. Zhong and Y. Zhou, "Suppression of large intraluminal bubble expansion in shock wave lithotripsy without compromising stone comminution: Methodology and *in vitro* experiments," J. Acoust. Soc. Am. **110**, 3283–3291 (2001).
[18]X. Xi and P. Zhong, "Improvement of stone fragmentation during shock-wave lithotripsy using a combined EH/PEAA shock wave generator-*in vitro* experiments," Ultrasound Med. Biol. **26**, 457–467 (2000).
[19]D. L. Sokolov, M. R. Bailey, and L. A. Crum, "Use of a dual-pulse lithotripter to generate a localized and intensified cavitation field," J. Acoust. Soc. Am. **110**, 1685–1695 (2001).
[20]D. Cathignol, A. Birer, S. Nachef, and J. Chapelon, "Electronic beam steering of shock waves," Ultrasound Med. Biol. **21**, 365–377 (1995).
[21]J. Tavakkoli, A. Birer, A. Arefiev, F. Prat, J. Chapelon, and D. Cathignol, "A piezocomposite shock wave generator with electronic focusing capability: Application for producing cavitation-induced lesions in rabbit liver," Ultrasound Med. Biol. **23**, 107–115 (1997).
[22]W. Eisenmenger, X. Du, C. Tang, S. Zhao, Y. Wang, F. Rong, D. Dai, M. Guan, and A. Qi, "The first clinical results of the "wide-focus and low-pressure" ESWL," Ultrasound Med. Biol. **28**, 769–774 (2002).
[23]O. A. Sapozhnikov, A. D. Maxwell, B. MacConaghy, and M. R. Bailey, "A mechanisitic analysis of stone fracture in lithotripsy," J. Acoust. Soc. Am. **121**, 1190–1202 (2007).
[24]M. Tanter, J. Aubry, J. Gerber, J. Thomas, and M. Fink, "Optimal focusing by spatiotemporal inverse filter. I. Basic principles," J. Acoust. Soc. Am. **110**, 37–47 (2001).
[25]P. Stephanishen and K. Benjamin, "Forward and backward projection of acoustic fields using FFT methods," J. Acoust. Soc. Am. **71**, 803–812 (1982).
[26]M. Schafer and P. Lewin, "Transducer characterization using the angular spectrum method," J. Acoust. Soc. Am. **85**, 2202–2214 (1989).
[27]D. Liu and R. Waag, "Propagation and backpropagation for ultrasonic wavefront design," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 1–13 (1997).
[28]P. Stephanishen, "The relationship between the impulse response and angular spectrum methods to evaluate acoustic transient fields," J. Acoust. Soc. Am. **90**, 2794–2798 (1991).
[29]D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," SIAM J. Appl. Math. **11**, 431–441 (1963).
[30]P. M. Gammell and G. R. Harris, "IGBT based kilovoltage pulsers for ultrasound measurement applications," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 1722–1728 (2003).
[31]A. R. Kaiser, C. A. Cain, E. Y. Hwang, J. B. Fowlkes, and R. J. Jeffers, "A cost effective degassing system for use in ultrasonic measurements: The multiple pinhole degassing (mpd) system," J. Acoust. Soc. Am. **99**, 3857–3859 (1996).
[32]J. Staudenraus and W. Eisenmenger, "Fiber-optic probe hydrophone for ultrasonic and shock-wave measurements in water," Ultrasonics **31**, 267–273 (1993).
[33]A. D. Maxwell, O. A. Sapozhnikov, and M. R. Bailey, "A new PVDF membrane hydrophone for measurement of medical shock waves," *IEEE Ultrasonics Symposium*, Vancouver, BC Canada, October, 2006, pp. 1608–1611.
[34]A. J. Coleman and J. E. Saunders, "A survey of the acoustic output of commercial extracorporeal shock wave lithotripters," Ultrasound Med. Biol. **15**, 213–227 (1989).
[35]R. Marchiano and J. Thomas, "Synthesis and analysis of linear and nonlinear acoustical vortices," Phys. Rev. E **71**, 066616 (2005).
[36]J. Thomas, F. Wu, and M. Fink, "Time reversal focusing applied to lithotripsy," Ultrason. Imaging **18**, 106–121 (1996).

Identification of some perceptual dimensions underlying loudspeaker dissimilarities

Mathieu Lavandier,[a] Sabine Meunier,[b] and Philippe Herzog[c]
Laboratoire de Mécanique et d'Acoustique, CNRS UPR 7051, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

(Received 19 July 2007; revised 9 January 2008; accepted 7 April 2008)

This study investigated the dimensions underlying perceived differences between loudspeakers. Listeners compared the sound reproduction of 12 loudspeakers in a room, using three musical excerpts. For the loudspeakers to be compared one just after the other in exactly the same conditions, the sounds radiated by the loudspeakers were recorded in a listening room, and the recorded sounds were submitted to paired comparisons using headphones. The resulting perceptual dissimilarities were analyzed by using a multidimensional scaling technique, revealing two main perceptual dimensions used by listeners to discriminate the loudspeakers. These dimensions were identical for the three musical excerpts. As the signals heard by listeners were directly accessible, they were used to define acoustical attributes describing the perceptual dimensions. Instead of arbitrarily choosing one acoustical analysis to define these attributes, several analyses were compared. The temporal, spectral, and time-frequency domains were investigated, and different auditory models were tested. These auditory models allowed the best description of the differences perceived by listeners, and were used to define two acoustical attributes describing our perceptual dimensions: the bass/treble balance and the medium emergence. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2916688]

PACS number(s): 43.38.Md, 43.66.Lj, 43.20.Ye, 43.58.Kr [AJZ] Pages: 4186–4198

I. INTRODUCTION

It is important to know why different loudspeakers sound differently both in terms of loudspeaker measurement and design and in terms of fundamentals of perception. Manufacturers have to control the acoustical parameters which can induce perceptual differences in the loudspeaker reproduction. Researchers interested in auditory perception try to identify the relevant acoustical criteria for discriminating complex sounds.

The perceptual characteristics of reproduced sound and their link with acoustical measurements were investigated by various authors, starting from the hypothesis that "perceived sound quality is constituted by a (limited) number of separate perceptual dimensions, and that it would be possible to give a perceptual description of sound-reproducing systems by stating their positions in such dimensions," Gabrielsson and Sjögren (1979). Multidimensional scaling techniques can be used to reveal the perceptual dimensions underlying simple dissimilarity evaluations by listeners [Borg and Groenen (1997); Susini *et al.* (1999)]. They give access to the characteristics used to discriminate the stimuli, without listeners having to know or name what they are experiencing while listening. The differential semantic is another technique used to highlight perceptual dimensions [Gabrielsson and Sjögren (1979); Guski (1997)]. It involves absolute evaluations of loudspeakers on chosen semantic scales labeled by adjectives. The number of scales is then reduced to independent factors by using factor analysis [Dillon and Goldstein (1984)].

The recommendations concerning listening tests on loudspeakers highlight three main categories of characteristics associated with the evaluation of sound reproduction: the timbre-related accuracy, sometimes called sound quality or fidelity and involving monophonic reproduction, the spatial quality involving stereophony and multichannel reproductions, and the dynamics and distortion characteristics related to the behavior of loudspeakers at different levels of solicitation [AES20-1996 (1996); IEC Publication 60268-13 (1998)]. Whereas recent studies were concerned with spatial quality and multichannel reproductions [Choisel and Wickelmaier (2007); Guastavino and Katz (2004); Rumsey *et al.* (2005a,b)], the main interest of our study was timbre-related accuracy. Several perceptual dimensions associated with timbre-related accuracy of sound-reproducing systems have been proposed [Bramsløw (2004); Eisler (1966); Gabrielsson *et al.* (1974); Gabrielsson and Sjögren (1979); Klippel (1990); Staffeldt (1974)]. These dimensions can be summarized as "clearness/distinctness," "sharpness/hardness-softness" and "fullness-thinness," "brightness-darkness," "feeling of space" and "spaciousness," "nearness," "bass emphasis" or "treble emphasis," and "disturbing sounds." As noted by Gabrielsson and Sjögren (1979), these dimensions might be redundant and other dimensions might exist.

The perception of the sound radiated by a loudspeaker is greatly influenced by the room in which it is used and by the positions of both loudspeaker and listener [Bech (1994); Olive *et al.* (1994)]. The evaluation of loudspeakers also de-

[a]Electronic mail: lavandiermn@cardiff.ac.uk. Present address: School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT. United Kingdom.
[b]Electronic mail: meunier@lma.cnrs-mrs.fr
[c]Electronic mail: herzog@lma.cnrs-mrs.fr

 0001-4966/2008/123(6)/4186/13/$23.00

pends on the loudness of their reproduction [Illényi and Korpássy (1981)] and on the musical excerpts used [Eisler (1966)]. Loudspeakers should then be compared in the same room, for the same positions of listener and loudspeaker, at the same sound level and with the same musical excerpt [AES20-1996 (1996); IEC Publication 60268-13 (1998)]. To evaluate their relative differences, they also have to be compared one just after the other, due to our short auditory memory [Olive *et al.* (1994)].

Because of the practical difficulty to compare many loudspeakers, the studies mentioned above tested loudspeakers at different positions in the listening room, and only a limited number of loudspeakers were compared in the same test. To our knowledge, a maximum of eight was reported by Klippel (1990). When more loudspeakers were tested in an experiment, only few comparisons were performed [Eisler (1966); Gabrielsson and Sjögren (1979)]. The perceptual dimensions involved in a listening test depend on the number and nature of the loudspeakers under evaluation. Gabrielsson and Lindstrom (1985) also showed that the evaluation of a loudspeaker on a particular scale was dependent on the other loudspeakers being tested in the experiment. To consider dimensions which are characteristic of the perception of sound reproduction in general, and not only on the particular loudspeakers under test, listening tests should involve as many different loudspeakers as possible.

The previous studies on the perception of reproduced sound often faced the difficulty of finding acoustical measurements of loudspeakers explaining the perceptual dimensions revealed by the listening tests. Three main approaches may be distinguished in the literature. The first one consisted in visual comparisons between the shapes of the frequency responses of loudspeakers and their evaluations along a given perceptual scale [Gabrielsson *et al.* (1991); Staffeldt (1974); Toole (1986)]. The link between perceptual and acoustical evaluations was not quantified and relied on the experimenter interpretation of the frequency responses, this interpretation involving many simultaneous criteria [Toole (1986)]. As a consequence, the most suitable measurement explaining perceptual evaluations differed from one author to the other [Olive (2004a)]. "Black box" models constituted a second approach [Bramsløw (2004); Olive (2004b)]. By taking acoustical measurements as inputs, they generate a rating as output, being as close as possible to a corresponding perceptual one. Olive (2004b) developed a multiple regression model predicting the loudspeaker preference of listeners from many different acoustical measurements. He was able to quantitatively compare the acoustical and perceptual ratings of preference, and found a very good agreement between them. However, this model does not individually treat the different perceptual dimensions, neither does it elucidate their link with the different acoustical measurements. A third approach looked for acoustical attributes derived from measurement curves, which would correlate well with perceptual dimensions revealed by listening tests [Klippel (1990); Olive (2004a)]. In order to bridge the gap between standardized measurements and perception, researchers tended to define acoustical measurements which took into account the listening conditions, but this was done to greater or lesser extents and only a posteriori. To find acoustical attributes well correlated with his perceptual dimensions, Klippel (1990) had to resimulate the sound reproductions involved in his listening tests. This simulation was realized by using the original musical excerpts, the statistics of the room, and the anechoic measurements of the loudspeakers.

The aim of the work presented here was to investigate the perceptual space associated with the timbre-related accuracy of loudspeakers in a listening room and to find out the best way to describe the resulting perceptual dimensions by acoustical attributes. Our goal was to compare many loudspeakers in order to obtain perceptual dimensions which would be as much as possible independent of the tested panel. The loudspeakers were used in monophonic reproduction and their relative differences were evaluated. To be able to compare many loudspeakers one just after the other at the same position in the room, we decided to record their sound reproductions and submit the recordings to paired comparisons by using headphones. Previous studies dealing with the perception of reproduced sound compared listening tests realized live or using headphones [Bech (2002); Olive *et al.* (1994); Pedersen and Mäkivirta (2002); Toole (1991)]. Even if the studies had various aims, they all showed that the two types of listening test gave very similar results concerning timbre-related accuracy. We did not intend to evaluate all the potential dissimilarities between the tested loudspeakers. Some of these dissimilarities might not have been captured in the final recordings. Our aim was to study the remaining dominant dissimilarities, with the advantage of being able to compare many loudspeakers at the same position, one just after the other, making sure that the dissimilarities were associated with the loudspeakers and not another experimental parameter.

As our protocol used recordings of the sound reproductions rather than live listening tests, we also had a direct access to the signals compared by listeners. These signals were used to define acoustical attributes describing the perceptual dimensions involved in the listening tests. We chose to base our acoustical approach on a signal analysis directly done on the recordings, rather than considering estimations of the loudspeakers responses, in order to remain as close as possible to the signals judged by listeners. Instead of arbitrarily choosing one signal analysis, several analyses were compared. The most suitable one was identified as the analysis leading to differences between recordings which were the closest to the differences evaluated by listeners.

II. LISTENING TESTS

The sounds radiated by different loudspeakers in a listening room were recorded and the listening tests were conducted by using headphones. As the spatial component of the sound reproduction could not be reliably investigated with such protocol, the loudspeakers were used in monophonic reproduction, as "timbre-related accuracy is much more easily heard in single-loudspeaker listening," AES20-1996 (1996).

TABLE I. Reverberation time RT measured by third-octave bands in the listening room used for the recordings as a function of the central frequency f of the third-octave band.

f (Hz)	200	250	315	400	500	630	800	1000	1250	1600	2000	2500	3150	4000	5000
RT (s.)	0.78	0.80	0.62	0.58	0.49	0.36	0.49	0.46	0.53	0.68	0.55	0.45	0.49	0.59	0.54

A. Stimuli

The listening room used for the recordings was 8.7 m long, 4.5 m wide, and 2.9 m high. The recording microphones and the loudspeaker were placed along the median axis of the room, with the loudspeaker at 1.5 m from the 4.5 m wall and the microphones at 2.20 m in front of the loudspeaker. The microphones were at 1 m from the floor, which approximately corresponds to the point between the medium and the tweeter of the loudspeakers. This room was chosen among a few which were available over the duration of the study, its characteristics being almost compatible with the standard for loudspeaker quality evaluation [IEC Publication 60268-13 (1998)]. Although this may not be considered as mandatory for estimating relative dissimilarities, we felt that keeping listening conditions close to an average living room was consistent with our goal. This room does not fully comply with the cited standard, especially at low frequencies, where its dimensions lead to a noticeable modal behavior. This is typical of many living rooms in contemporary urban flats, and we decided to consider this feature as part of the "usual" listening conditions for our comparisons, although it would not be quite appropriate for absolute quality evaluation. As stated in the standard, the reverberation time may not be measured at frequencies for which the mode density is low, and we therefore did not include such data below 200 Hz, although we checked that no mode resonance was much stronger than average in the frequency range of the tested loudspeakers. The reverberation time was measured at the position of the recording microphones, with the source at the loudspeaker position. Its value was found around 0.5 s at midrange frequencies (Table I).

Twelve single loudspeakers were involved in the recordings. As we were interested in evaluating their perceived differences, we chose loudspeakers leading to a broad range of dissimilarities: loudspeakers which would obviously be very dissimilar, others which should be very similar, and finally loudspeakers which should lead to intermediate levels of dissimilarities. We tried to use a wide range of models coming from different manufacturers. They were commercial units, available in our laboratory or lent by partners and manufacturers. Because we were aware that our comparisons were done in nonstandard conditions, we agreed with our partners not to disclose the actual references of the loudspeakers, which are only designated by their number. Table II gives general characteristics about the tested panel. We chose units in the high-fidelity and small studio monitoring ranges, avoiding low-end or very high prices. We considered this panel as representative of loudspeakers dedicated to music listening for people with actual interest in sound quality, but neither professional nor audiophile. This is consistent with our panel of listeners. We only kept electrodynamic units with widely available acoustic loads (closed or vented boxes) and usual filtering options (wide band to three ways). All loudspeakers were different from each other, except numbers 7 and 8, which correspond to the two units of a stereo pair. They were included in the test to get an example of two very similar reproductions, in order to check the validity of our protocol.

Figure 1 provides an estimation of the frequency response of each loudspeaker, measured in the listening room used for the experiment, with the loudspeakers and the measuring microphone at the positions used for the recordings. The responses were estimated with pink noise, considering the root-mean-square average of 128 fast Fourier transforms (FFTs) of Hanning windowed sections of the signal with 50% overlap. They were calculated as the spectrum at the measurement microphone divided by the spectrum of the input voltage of the loudspeaker. The FFT bins were summed in each third-octave band. The 12 estimated responses were normalized by their value in the 1 kHz band to facilitate their comparison. Figure 1 gives a general view of the loudspeaker panel, but also illustrates the performance of the

TABLE II. General characteristics of the 12 tested loudspeakers.

No.	Use	Box	Speakers	Filter
1	HiFi-medium range	Column, closed	2 boomers+1 wide band+1 tweeter	3 ways
2	Miniature monitor	Compact, closed	1 wide band	1 way
3	Professional monitor (90's)	Shoebox, vented	1 boomer+tweeter	2 ways
4	HiFi-bookshelf	Shoe box, closed	1 boomer+tweeter	2 ways
5	Professional monitor (80's)	Shoebox, vented	1 boomer+1 medium+1 tweeter	3 ways
6	Home-Studio	Shoebox, vented	1 boomer+tweeter (horn)	2 ways
7	HiFi-budget	Shoebox, closed	1 boomer+tweeter	2 ways
8	HiFi-budget	Shoebox, closed	1 boomer+tweeter	2 ways
9	Small monitor	Shoebox, vented	1 coax(boomer+tweeter)	2 ways
10	Small monitor	Shoebox, vented	1 coax(boomer+tweeter)	2 ways
11	HiFi-bookshelf	Shoebox, closed	1 boomer+tweeter	2 ways
12	Miniature monitor	Shoebox, vented	1 wide band	1 way

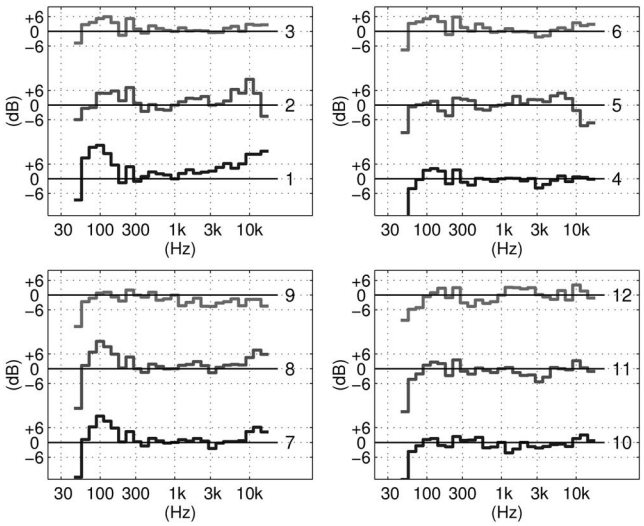


FIG. 1. Frequency responses of the 12 tested loudspeakers (Table II), measured in the configuration used for the recordings.

room: The response of loudspeaker 3 (considered as the higher-end unit) is kept within ±6 dB, but presents a few accidents at low frequencies which are shared with many other loudspeakers, and may be attributed to the room.

The musical excerpts used for the recordings were chosen according to the recommendations concerning listening tests on loudspeakers [AES20-1996 (1996); IEC Publication 60268-13 (1998)]. They were of very different musical styles, involving acoustical instruments rather than synthetic music, and were selected among high-quality commercial recordings. Our monaural stimuli were obtained by considering only one channel of the original stereophonic excerpts. Three musical excerpts were recorded and we only kept a short part of them for the listening tests: McCoy Tyner (Miss Bea, Best of Chesky Jazz, Vol. 2, Chesky CD:68, from 01:01 to 01:05, right channel), Kan'nida (Konsyans, Kyenzenn, Indigo LBLC 2566, from 00:07 to 00:09, left channel), and Vivalid (L'Europa Galante, Fabio Biondi, Opus 111, OPS 30-86, Concerto pour violino e organo in re minore, from 00:08 to 00:13, left channel). Such short excerpts were reported as suitable for the evaluation of perceived differences in timbre-related accuracy [Bech (1995, 1996); Moore and Tan (2003)].

The stereophonic ORTF recording technique was used, with two cardioid microphones placed 17 cm and 110° apart from each other (AKG Blue Line CK-91, SE-300B, preamplifier Tascam MX-4). The musical excerpts were stored on a compact disk and reproduced on the loudspeakers by high-grade CD player and amplifier (Vecteur I-4.2, Vecteur L-3.2, Behringer Ultralink Pro MX882 as a preamplifier for the studio monitoring loudspeakers). The recordings were directly carried out with an audio workstation featuring an RME DIGI9652 sound card and an external Fostex VC-8 A/D converter. The sampling frequency was 44 100 Hz. During the recordings, reproduction levels were roughly adjusted to normal listening conditions. A more accurate loudness equalization was undertaken by the experimenters prior to the listening tests.

B. Procedure

Perceptual dissimilarities between recordings were gathered during three listening tests, one test for each musical excerpt. The listening tests were paired comparisons where the listener had to directly evaluate the dissimilarities between the stimuli. For each test, the 12 recordings were presented by pairs to the listener in random order. The listener had to quantify the overall dissimilarity within each pair by adjusting a cursor on a line whose end points were labeled "very similar" and "very dissimilar," this line being displayed on a computer terminal screen in front of the listener. A numerical value was linearly assigned to the position of the cursor. The resulting perceptual dissimilarity varied from 0 for very similar to 1 for very dissimilar. Participants could listen to each pair of recordings as many times as they wanted before doing their dissimilarity evaluation, but they were encouraged to base this evaluation on their first impression and avoid multiple listenings. Listening tests lasted about 20 min (Kan'nida), 25 min (McCoy Tyner), or 35 min (Vivaldi).

Before the tests, the overall loudness of the recordings was set to the same level of 70 phones, as judged by the experimenters and verified by a loudness estimation model [Zwicker and Fastl (1983, 1999)]. This prevented loudness from creating uninteresting dissimilarities potentially masking more subtle ones. Listening tests were run in an isolated soundproof room by using a Tucker&Davis workstation and Stax SR Lambda Professional headphones. The frequency responses of the recording microphones and headphones were not compensated for, but all comparisons were made with the same headphones and between recordings involving the same microphones.

Twenty-seven listeners (12 women, 15 men) took part in the experiment, each of them participating in the three listening tests in random order. They were between 14 and 53 years old, with an average age of 30. The listeners were otologically normal. Twenty five had hearing thresholds less than 25 dB HL from 250 to 8000 Hz. Two had slight presbyacusia (thresholds between 25 and 40 dB HL above 4 kHz), normal for their age. The results from these two listeners did not differ from the data of the group, indicating that their presbyacusia did not affect their judgments. They were members of the laboratory or students. None of them had a significant previous experience in loudspeaker comparison.

C. Results

For each listening test, the final perceptual dissimilarities were obtained by averaging individual dissimilarities. Individual dissimilarities were not scaled before the averaging, but it was verified that there was no group with different judgement strategies among listeners. This verification was based on a cluster analysis of the similarities between listeners, the similarity between two listeners being evaluated by the correlation of their judgments. Figure 2 presents the dendrogram resulting from this analysis for the listening test involving the musical excerpt McCoy Tyner, which was here arbitrarily chosen as an example. There would be classes of

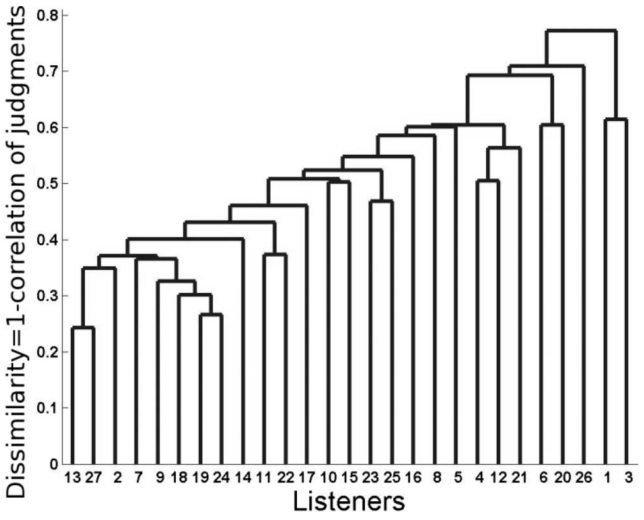


FIG. 2. Dendrogram resulting from the cluster analysis of the similarities between listeners (listening test involving the musical excerpt McCoy Tyner). Each number corresponds to one of the 27 listeners. The height of the node connecting two listeners corresponds to the dissimilarity of their judgments, whereas the position of the listeners on the horizontal axis is irrelevant.

listeners with different judgement strategies, if one could isolate at least two groups of listeners having well correlated judgments within each group, and uncorrelated judgments between groups. No such group appears in Fig. 2. Any partition of the listeners "continuously" depends on the degree of similarity chosen to discriminate them. No classes of listeners were found in any of the three listening tests.

The final mean perceptual dissimilarities between recordings were ranging between 0.22 and 0.83 (McCoy Tyner), 0.13 and 0.90 (Kan'nida), or 0.21 and 0.84 (Vivaldi), indicating that the chosen loudspeakers and musical excerpts led to a broad range of dissimilarities [Lavandier *et al.* (2008)]. These dissimilarities were submitted to multidimensional scaling analysis (MDS). The classical model MDSCAL was used. The algorithm SMACOF, for "scaling by majorizing a complicated function," has been detailed by Borg and Groenen (1997). Stimuli are represented in an Euclidean space of N orthogonal dimensions, the most appropriate value of N being chosen by the experimenter relying on stress measurements. The algorithm does not fix the axes of the multidimensional space. This space can be rotated or dilated; only relative distances between stimuli are fixed. By default, the axes coincide with the directions of maximum variance, determined by a principal component analysis, and the ranking of dimensions is fixed by the amount of variance explained by each dimension.

Interpretation of MDS results requires much caution: Even if the analysis allows to locate stimuli in a space with continuous orthogonal dimensions, the MDS does not mean that such dimensions correspond to actual perceptual attributes. Even heterogeneous stimuli—involving different categories with no perceptual dimension relating them—lead to a continuous MDS representation. Our experiments apparently involved homogeneous stimuli, so it seemed relevant to look for continuous perceptual dimensions [Susini *et al.* (1999)]. Nevertheless, our MDS analyses were always associated with cluster analyses applied to the same dissimilarity data, in order to verify the homogeneity of our stimuli by highlighting the potential presence of classes among them. These cluster analyses never revealed any classes among our recordings.

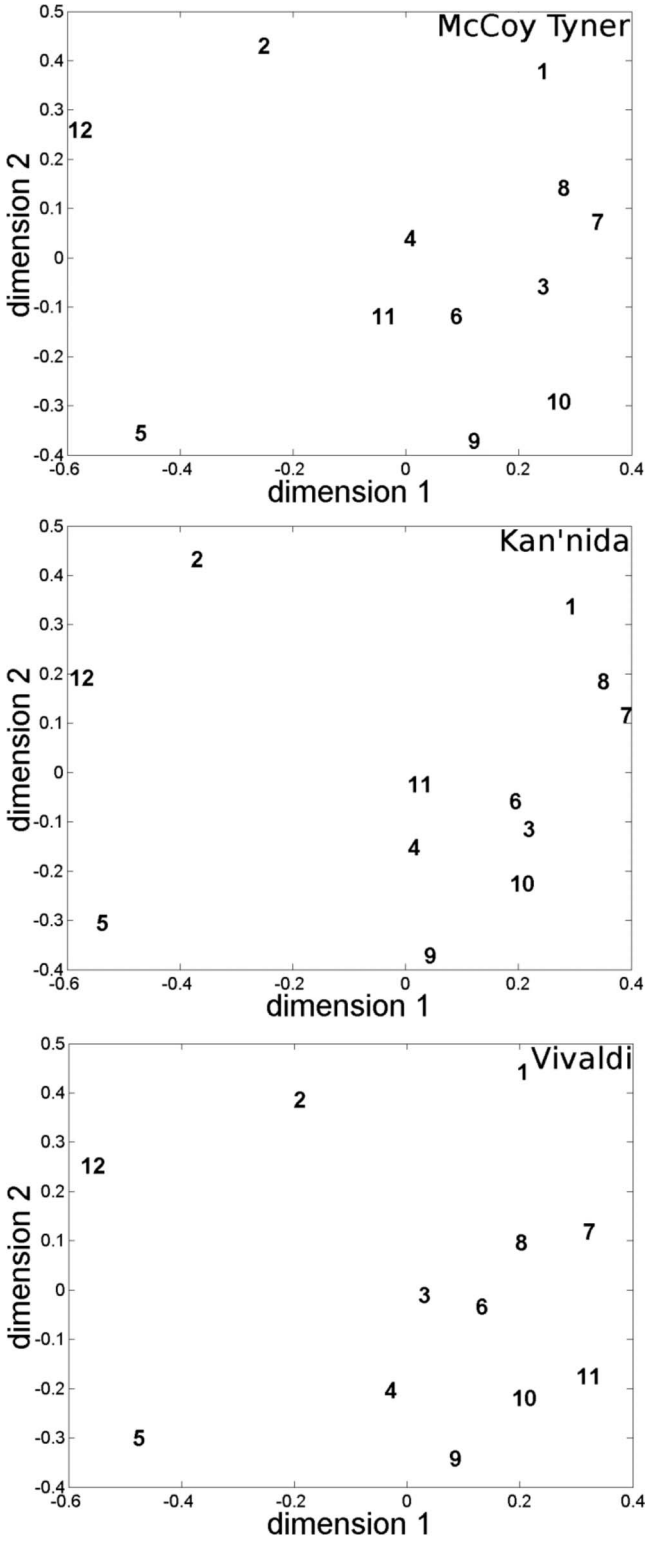


FIG. 3. Two-dimensional spaces resulting from the MDS analysis of the perceptual dissimilarities obtained with each musical excerpt (McCoy Tyner, Kan'nida, and Vivaldi). Each loudspeaker is identified by its number (Table II).

Figure 3 presents the perceptual spaces resulting from

the three listening tests. In each case, a two-dimensional space was obtained. Each number in these spaces corresponds to a loudspeaker (Table II). To be compared, the perceptual spaces were all rotated and dilated to fit as much as possible the space involving the musical excerpt Kan'nida, which was arbitrarily chosen as a reference with its default orientation.

The perceptual spaces obtained with the three musical excerpts were very similar. The influence of the musical excerpt on the evaluation of the perceptual dissimilarities was small, as shown by the rather small changes in the relative positions of loudspeakers between the three spaces. Listeners used the same two main perceptual dimensions to discriminate the recordings during the three listening tests, despite the change of musical excerpt. These two dimensions would then be characteristic of the loudspeakers under evaluation. One can notice that loudspeakers 7 and 8, the two loudspeakers of the same model, stand very close together in the perceptual spaces.

We listened to the recordings while looking at the perceptual spaces and used the free comments of listeners to propose an interpretation of the dimensions. The predominant dimension 1 seemed linked to the bass/treble balance of the recordings. The dimension 2 would be linked to the amount of medium frequencies, and could be associated with the notion of sound clarity. These interpretations were consolidated by acoustical attributes presented in the following.

III. ACOUSTICAL ANALYSES

As our protocol used recordings of the sound reproductions of loudspeakers rather than live listening tests, we had a direct access to the signals compared by listeners. These signals were used to define acoustical attributes describing the perceptual dimensions revealed by the listening tests. First, a suitable signal analysis of the recordings had to be chosen. For example, the recordings could be compared in the time, spectral, or time-frequency domains. Instead of arbitrarily choosing one signal analysis, we decided to compare several analyses and choose the analysis that showed out differences between recordings which were similar to those evaluated by listeners. Several methods of signal analysis were tested, and for each method, a metric evaluating the acoustical dissimilarity between two signals was defined. The acoustical dissimilarities were computed for all recordings and compared to the perceptual dissimilarities, in order to identify the most relevant acoustical discrimination method. As the analyses and metrics were presented in detail in a previous paper [Lavandier *et al.* (2008)], they are only summarized here.

A. Procedure

1. Signal analyses

The temporal, spectral, and time-frequency domains were investigated. Two spectral weightings were tested. Different auditory models were also used. These auditory models are considered here as acoustical analyses and not perceptual ones, because they were used like any other signal analysis technique without any listener being involved.

The spectral domain was investigated by using the discrete Fourier transform of the signals and their power spectral density. The time-frequency plane was obtained by calculating their short-time Fourier transform. We considered the spectral and time-frequency domains with and without phase information. Spectral weightings were applied to the power spectral density of the signals to model the ear sensitivity. They were the A-weighting and a weighting based on the normal equal-loudness contour at 70 phones [British Standard ISO 226 2003 (2003)], which corresponds to the sound level used during our listening tests.

Two auditory models were tested. These models were proposed by Zwicker and Fastl (1999) to calculate loudness, and are based on auditory masking. They were used to analyze the signals in terms of specific loudness which is the density of loudness along the Bark scale modeling the perceptual frequency scale. The first model [Paulus and Zwicker (1972); Zwicker *et al.* (1984)] was originally designed to evaluate the loudness of stationary sounds. It takes into account only frequency masking, and was used to calculate what we called the overall specific loudness of the signals, which is the specific loudness determined over the entire signal taken at once. Two time-frequency patterns of specific loudness were examined. The first one was obtained by applying the first previous model to successive 100 ms sections of the signals. We called the resulting analysis time-varying specific loudness 1. To get the second time-frequency pattern of specific loudness, a second auditory model was used. This model was designed for nonstationary sounds [Zwicker and Fastl (1983, 1999)]. It takes into account both frequency and temporal masking, calculating specific loudness every 10 ms. We called the resulting analysis time-varying specific loudness 2. Finally, another way to analyze the signals by using these loudness models was to average the time-varying specific loudness over time. The temporal means of time-varying specific loudnesses 1 and 2 were considered. Compared to the overall specific loudness, they took into account the fact that auditory masking depends on the spectral content of signals and that this content varies over time for nonstationary musical signals.

2. Acoustical dissimilarities

For each method of analysis, we defined a metric evaluating an overall acoustical dissimilarity between two recordings, giving a single value from the differences between the analyses of these two recordings [Lavandier *et al.* (2008)]. To get a single scalar, dissimilarities calculated along the temporal, spectral, or Bark scales were integrated. As we could not suppose that a sample period or a frequency region was more important than another, this integration was realized by an arithmetic mean, giving the same weight to all dissimilarities.

For the analyses not involving the auditory models, the acoustical dissimilarity between recordings x and y was calculated following the general formula

$$\min\{\langle [Ax(v) - Ay(v)]^2\rangle_v, \langle [Ax(v) + Ay(v)]^2\rangle_v\}, \qquad (1)$$

where Ax and Ay are the considered signal analysis of x and y (waveform, Fourier transform, short-time Fourier trans-

form, modulus of Fourier transform or short-time Fourier transform, or power spectral density with or without weightings), v is the variable associated with this analysis (time, frequency, or time and frequency), $\langle\ \rangle_v$ takes the arithmetic mean over the variable v, and min{ } takes the minimum value. To get a single overall dissimilarity between x and y, dissimilarities calculated at each sample period or each frequency are integrated by taking the arithmetic mean over the considered scale. Because phase information was sometimes considered, a potential phase inversion between the two compared recordings x and y is taken into account by considering y and $-y$, and the smallest dissimilarity is kept.

The acoustical dissimilarity between the overall specific loudnesses or the temporal mean of time-varying specific loudnesses of recordings x and y was calculated following the formula

$$\langle \max\{\mathrm{SL}(x),\mathrm{SL}(y)\}/\min\{\mathrm{SL}(x),\mathrm{SL}(y)\} - 1\rangle_B, \qquad (2)$$

where SL is the specific loudness considered, $\langle\ \rangle_B$ takes the arithmetic mean over the Bark scale, min{ } takes the minimum value, and max{ } takes the maximum value. The dissimilarities were computed by taking the ratio of the specific loudnesses, as two loudnesses should be compared by their ratio not their difference [Zwicker and Fastl (1999)]. We took care of always dividing the largest value by the smallest one, as we would have taken the absolute value for a difference. Before calculating the ratios, a threshold was introduced in order to avoid artificially high dissimilarities where there was no useful signal but noise. The integration of the resulting dissimilarities over the Bark scale was done by taking their arithmetic mean. For the acoustical dissimilarities involving the time-varying specific loudnesses, the instantaneous dissimilarities were calculated following formula (2), and then integrated over time.

Before calculating the different acoustical dissimilarities, our signals were synchronized [Lavandier *et al.* (2008)]. The right and left channels of our stereophonic recordings were separately analyzed. For each musical excerpt and each channel, we determined the different acoustical dissimilarities between the 12 recordings corresponding to the 12 loudspeakers. Acoustical dissimilarities were then compared to the perceptual ones, in order to determine the acoustical analyses allowing to show out the differences heard by listeners.

3. Comparison of acoustical and perceptual dissimilarities

A previous comparison of acoustical and perceptual dissimilarities consisted in directly evaluating the correlation between them [Lavandier *et al.* (2008)]. The dissimilarities between the 12 recordings were gathered in 13 acoustical 66-value dissimilarity vectors—one vector per type of acoustical dissimilarity computed on the signals—and one perceptual 66-value dissimilarity vector. The different acoustical dissimilarity vectors were compared by evaluating their coefficient of correlation with the perceptual dissimilarity vectors. The acoustical analyses keeping phase information led to a correlation of about 0.2, whereas the correlation increased to 0.5 when phase information was not considered. The spectral weightings only improved the correlation for one musical excerpt (Kan'nida). On the other hand, the auditory models greatly improved the correlation between acoustical and perceptual dissimilarities, for all tested signals, with a coefficient of correlation often above 0.8. These correlations indicated that the acoustical dissimilarities involving auditory models contained at least part of the information contained in the perceptual dissimilarities.

A second comparison of the acoustical and perceptual dissimilarities is presented here. Indeed, the acoustical and perceptual dissimilarities could be badly correlated because they are not linked at all, but also because their relation is nonlinear, or because the acoustical dissimilarities contain more or less underlying dimensions than the perceptual ones. The restricting criterion of linearity associated with correlation was abandoned, and the multidimensional nature of our perception of reproduced sound was taken into account. The acoustical dissimilarities were submitted to the same MDS analysis as the perceptual dissimilarities, and the resulting acoustical spaces were visually compared to the perceptual ones shown in Fig. 3. This type of comparison indicates if an acoustical analysis seems appropriate to highlight any of the perceptual dimensions. Even if the acoustical and perceptual spaces are different, it is interesting to know if some acoustical dimensions are relevant regarding perception. Compared to the evaluation of correlation between dissimilarities, it gives a better understanding of the information contained in acoustical and perceptual dissimilarities.

B. Results

Thirteen acoustical spaces resulted from the MDS analysis of the acoustical dissimilarities associated with the 13 different acoustical discrimination methods presented above. To facilitate their comparison, the acoustical spaces were rotated and dilated to fit the same reference as the perceptual spaces. In front of the large number of acoustical spaces obtained, only some examples are presented here. The analyses involving auditory models were the only acoustical analyses leading to two-dimensional spaces being close to the perceptual ones for the three musical excerpts. The other analyses led to one- or two-dimensional spaces, depending on the signals considered. No link could be found between the corresponding acoustical dimensions and the perceptual ones. The two acoustical spaces presented in Fig. 4 result from the analysis of the dissimilarities obtained with the waveform of the recordings (top) and with their power spectral density (bottom). They should be compared to the corresponding perceptual spaces of Fig. 3.

Figure 5 presents one acoustical space based on the overall specific loudness for each musical excerpt. The overall specific loudness led to acoustical spaces close to the perceptual space only for the musical excerpt Kan'nida. For the two other musical excerpts, the global analysis applied on the entire signal taken at once used while calculating overall specific loudness was not sufficient to fully describe the perceptual dissimilarities.

On the other hand, time-varying specific loudnesses and their temporal mean led to acoustical spaces very similar to

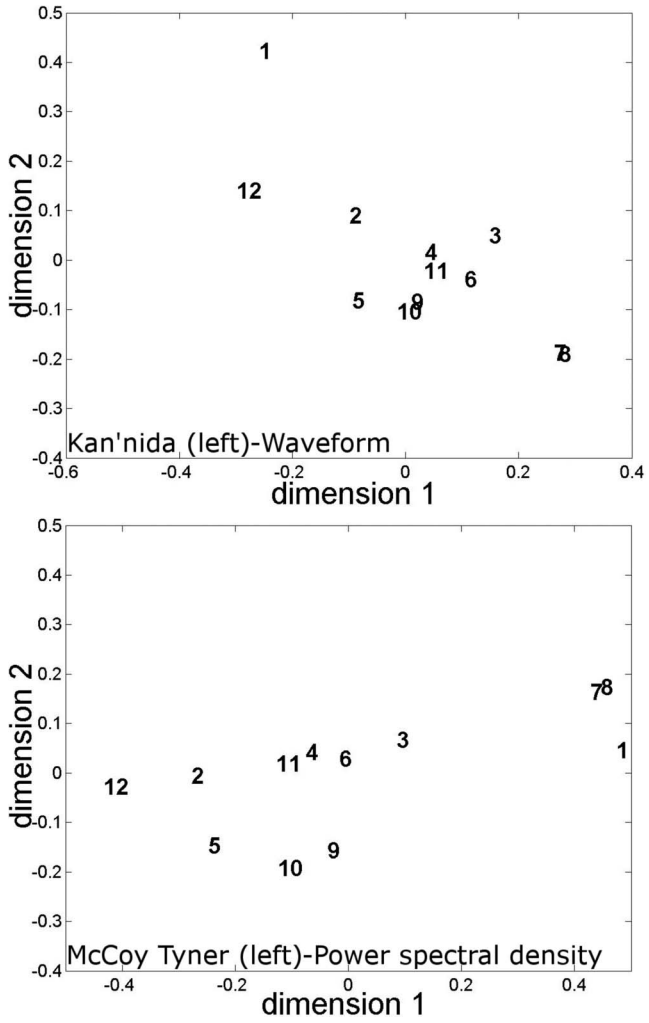


FIG. 4. Examples of acoustical spaces based on the waveform (top) and the power spectral density (bottom). They should be compared to the corresponding perceptual spaces of Fig. 3.

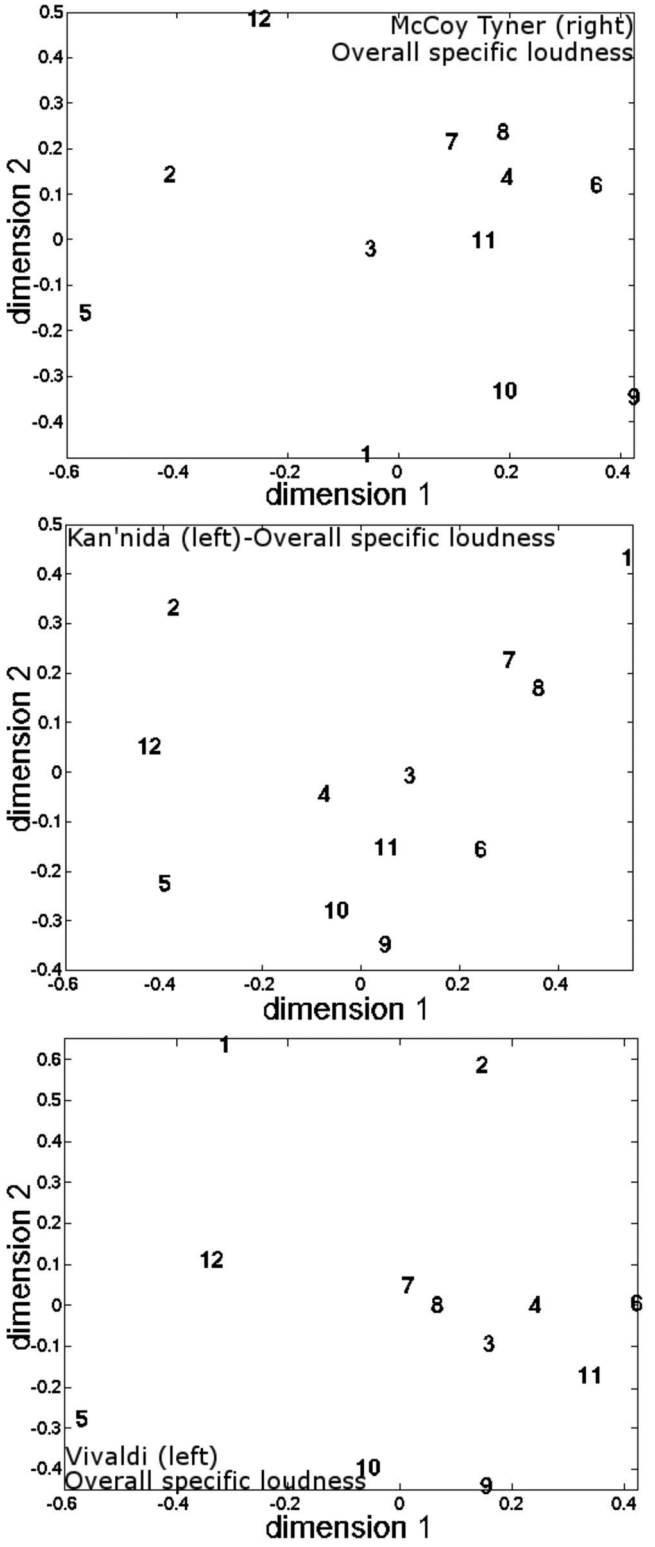


FIG. 5. Examples of acoustical spaces based on the overall specific loudness. They should be compared to the corresponding perceptual spaces of Fig. 3.

the perceptual ones for the three musical excerpts. The results were equivalent for models 1 and 2 and for time-varying specific loudnesses and their temporal means. Figure 6 presents two examples obtained with the signals leading to the best (Kan'nida, left channel) and the worst (Vivaldi, right channel) correspondence between acoustical and perceptual dissimilarities. The time-varying specific loudness 2 was preferred to the time-varying specific loudness 1 because it is based on a more widely used computation [Fastl (1997)]. We decided to present examples based on the temporal mean of time-varying specific loudness rather than on the time-varying specific loudness because the temporal mean significantly reduces the amount of data involved in the dissimilarity evaluation, while giving results which were found very close to those obtained without averaging.

C. Discussion

The comparison of acoustical and perceptual spaces confirmed the results based on the simple correlation between acoustical and perceptual dissimilarities [Lavandier *et al.* (2008)]: The acoustical discriminations by using auditory models were the closest to the perceptual evaluation. Even if a criterion is still needed to quantify the similarity between two spaces, the visual comparison of spaces indicated that the time-varying specific loudnesses and their temporal means were suitable to describe the two perceptual dimensions revealed by the listening tests. The overall specific loudness might, however, not be sufficient to fully describe the perceptual dissimilarities.

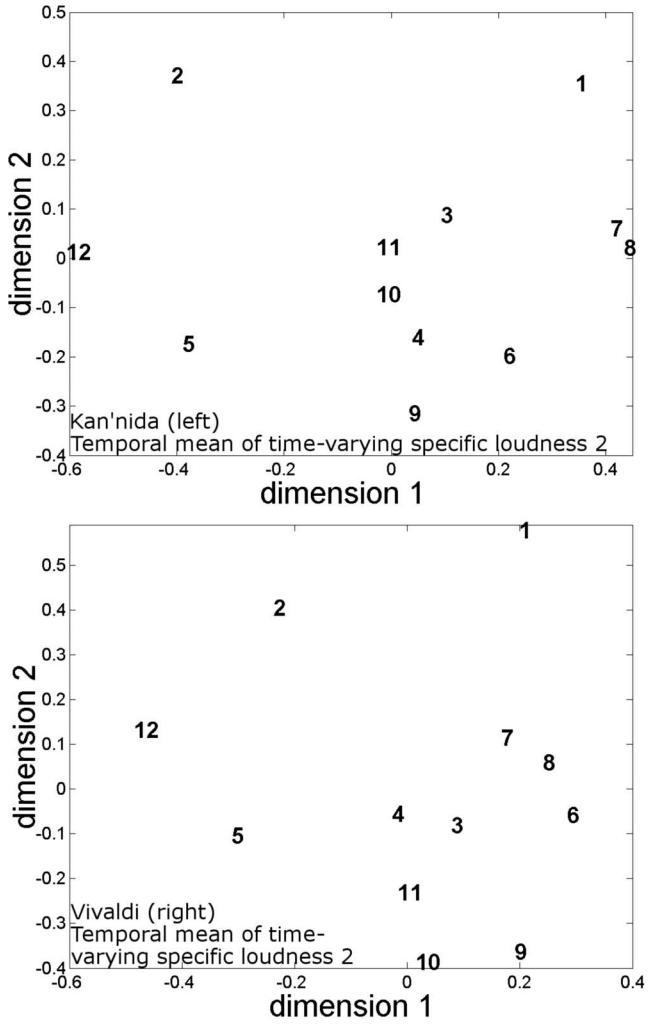


FIG. 6. Examples of acoustical spaces based on the temporal mean of time-varying specific loudness 2. They should be compared to the corresponding perceptual spaces of Fig. 3.

The temporal resolution of the auditory models did not seem crucial regarding the two dimensions involved in our listening tests. The two time-varying specific loudnesses gave equivalent results, although time-varying specific loudness 1 was calculated every 100 ms and time-varying specific loudness 2 was evaluated every 10 ms. Taking into account temporal masking in model 2 did not bring further improvement. This would support the hypothesis used by Bramsløw (2004) who neglected this last effect in his reproduced sound quality evaluation model. Time-varying specific loudnesses and their temporal mean gave equivalent results, suggesting that the two perceptual dimensions were linked to an average impression over the entire musical excerpt [Bramsløw (2004); Gabrielsson *et al.* (1990, 1991); Staffeldt (1984, 1991)].

IV. ACOUSTICAL ATTRIBUTES

We defined acoustical attributes describing the perceptual/acoustical dimensions. The temporal mean of time-varying specific loudness 2 of the stimuli was chosen as a basis for this definition, as it was one of the most relevant analyses for discriminating the recordings. We described the acoustical dimensions as they were very similar to the perceptual dimensions (Figs. 3 and 6). Our first criterion to define acoustical attributes was to look for a monotonic relation between the dimension and the attribute. The criterion of correlation was also considered, even if it is more restrictive as it looks for a linear relationship between dimension and attribute.

A. Bass/treble balance

The first proposed acoustical attribute evaluates the bass/treble balance of the recordings. It is defined as the ratio between the loudnesses in the three first Bark bands (corresponding to frequencies between 20 and 280 Hz) and in the ten last Bark bands (corresponding to frequencies between 1.8 and 15.5 kHz). These two values of loudness are calculated by integrating the temporal mean of time-varying specific loudness 2 over the specified Bark bands. These Bark bands were chosen in order to give the best possible results to describe the first dimension of the spaces resulting from the three musical excerpts.

The relation between bass/treble balance and first dimension of the acoustical space was monotonic for almost all signals, and often linear. The correlation coefficient between attribute and dimension varied between 0.80 and 0.96 for the different musical excerpts, with an average value of 0.90 (Table III). Figure 7 presents two examples: The most favorable one with a linear relation (Kan'nida, left channel) and the least favorable one for which monotony was not preserved (Vivaldi, right channel).

B. Medium emergence

The second proposed acoustical attribute evaluates the emergence of medium frequencies in the recordings. The medium emergence is defined as the loudness in the Bark bands 5–9 (corresponding to frequencies between 355 and 1120 Hz). This loudness is calculated by integrating the temporal mean of time-varying specific loudness 2 over the specified Bark bands. As all recordings were equalized in overall loudness, the attribute evaluates the perceived emergence of this frequency range compared to the rest of the spectrum.

TABLE III. Correlation coefficient between the bass/treble balance and the dimension 1 of the acoustical space based on the temporal mean of time-varying specific loudness 2, for the left and right channels of the recordings and the three musical excerpts.

	Kan'nida		McCoy Tyner		Vivaldi	
Musical excerpt	Left	Right	Left	Right	Left	Right
Correlation	0.96	0.95	0.92	0.83	0.93	0.80

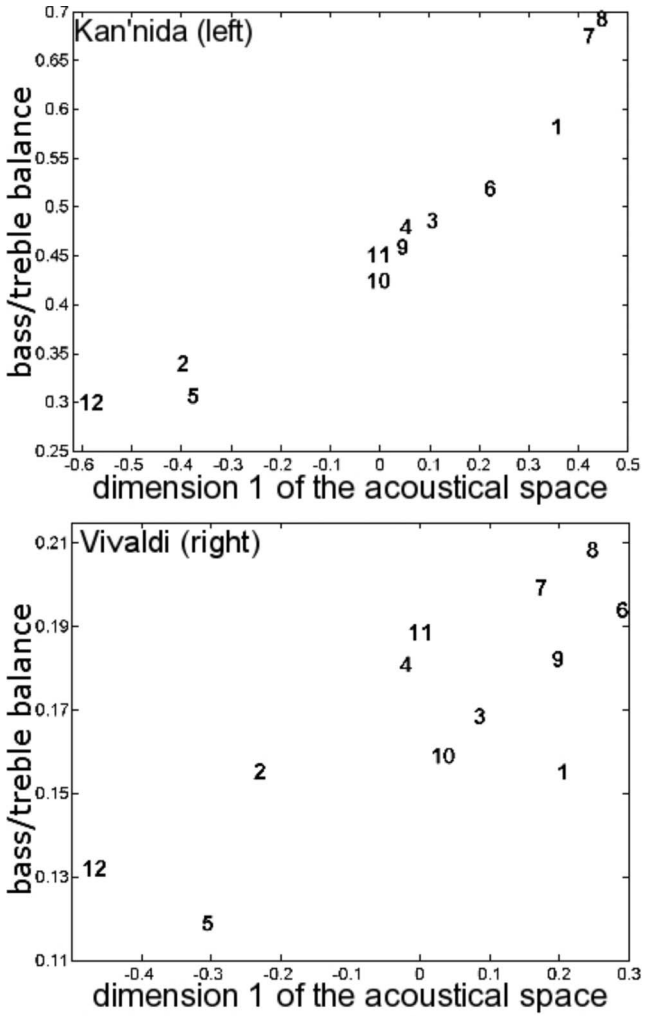


FIG. 7. Bass/treble balance as a function of the dimension 1 of the acoustical space based on the temporal mean of time-varying specific loudness 2. Top: case leading to the best correlation between dimension and attribute (Kan'nida, left channel). Bottom: case leading to the worst correlation between dimension and attribute (Vivaldi, right channel). Each loudspeaker is identified by its number (Table II).

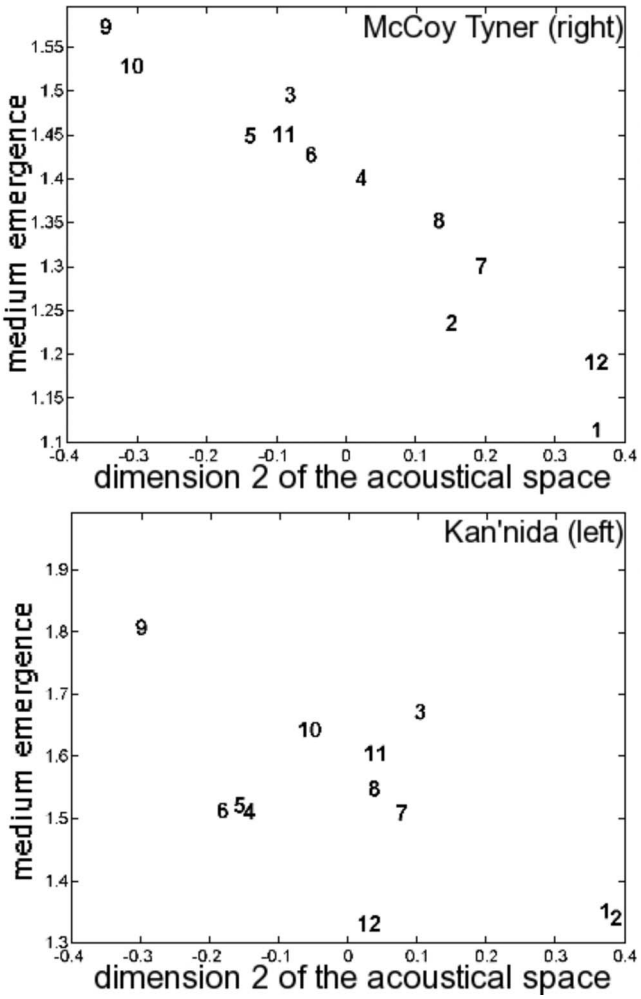


FIG. 8. Medium emergence as a function of the dimension 2 of the acoustical space based on the temporal mean of time-varying specific loudness 2. Top: case leading to the best correlation between dimension and attribute (McCoy Tyner, right channel). Bottom: case leading to the worst correlation between dimension and attribute (Kan'nida, left channel).

The relation between medium emergence and second dimension of the acoustical space was monotonic for almost all recordings, and often linear. The correlation coefficient between attribute and dimension varied between 0.65 and 0.96 for the different musical excerpts, with an average value of 0.87 (Table IV). Figure 8 presents two examples: The most favorable one with a linear relation (McCoy Tyner, right channel) and the least favorable one (Kan'nida, left channel). This last case was an exception among the tested signals, for which the proposed attribute was not sufficiently discriminant.

V. GENERAL DISCUSSION

The dissimilarities between 12 loudspeakers radiating in a listening room were investigated by using three musical excerpts. As our experimental protocol was based on recordings of the sound reproductions rather than live listening tests, loudspeakers could be compared one just after the other at the same position in the room. A MDS analysis of the resulting perceptual dissimilarities revealed two main dimensions used by listeners to discriminate the loudspeakers. Our protocol gave us direct access to the signals heard by listeners, so we were able to use these signals to define acoustical attributes describing the perceptual dimensions.

TABLE IV. Correlation coefficient between the medium emergence and the dimension 2 of the acoustical space based on the temporal mean of time-varying specific loudness 2, for the left and right channels of the recordings and the three musical excerpts.

	Kan'nida		McCoy Tyner		Vivaldi	
Musical excerpt	Left	Right	Left	Right	Left	Right
Correlation	0.65	0.87	0.88	0.96	0.93	0.95

As the perceptual dimensions did not depend on the particular musical excerpt used, we defined general attributes which were appropriate for all types of signal, giving a measure more characteristic of the loudspeakers and less dependent on the particular signal being reproduced. The specific loudness of the recordings was used to define two acoustical attributes well correlated with the perceptual dimensions: the bass/treble balance and the medium emergence.

These dimensions were in good agreement with the literature. The bass/treble balance can be compared to the "balance between bass and treble" or "brightness-darkness" found by Gabrielsson *et al.* (1974) or the "fullness-thinness" highlighted by Gabrielsson and Sjögren (1979). This last dimension was later defined by Toole (1985) as being associated with the amount of low frequencies compared to medium and high frequencies. Gabrielsson *et al.* (1990) linked their dimension "brightness versus dullness" to the levels of bass and treble, and the "fullness" was associated with the amplification of low frequencies. In Gabrielsson *et al.* (1990), the acoustical attributes were directly evaluated on the loudspeaker's frequency responses, but the link between attribute and perceptual dimension was not quantified. It was only based on visual comparisons, that is to say on trends observed by experimenters on frequency responses compared to perceptual evaluations. In the same way, Staffeldt (1974) mentioned "emphasized treble" and "emphasized bass" to interpret his results. The listening tests of Klippel (1990) revealed the same kind of dimension, and he used the specific loudness of his signals to define acoustical attributes describing the results of his listening tests. These attributes relied on different weightings favoring different frequency bands. He obtained attributes well correlated with the dimensions "treble stressing" ("sharpness"), "general bass emphasis" ("volume"), and "low bass emphasis," with correlation coefficients ranging from 0.87 to 0.90. He also described "brightness," presented as an overall evaluation including "general bass emphasis" and "treble stressing," by a linear combination of the two corresponding attributes. This last attribute seems closely related to our bass/treble balance. Our first dimension could also be compared to the "sharpness/fullness" highlighted by Bramsløw (2004), which was interpreted as the ratio between low and high frequencies. For Olive (2004a), the perceived spectral balance was an important factor relative to listener's preferences. A semantic analysis of free comments from his listeners highlighted nine factors, from which he interpreted the two main ones as "bass quality, overall spectral balance, and smoothness" and "midtreble colorations like bright, midpeak (megaphone), and colored." The first factor can be linked to our bass/treble balance, whereas the midtreble colorations can be linked to the second dimension revealed by our listening tests, the medium emergence. This last dimension also seems similar to the "distinctness" ("clarity" and "clearness") of Gabrielsson *et al.* (1974) and the "clearness" of Klippel (1990) or Bramsløw (2004). Our attribute describing the medium emergence involved the frequency range 355–1120 Hz. Bech (1996) indicated that the frequency range 500–2000 Hz was crucial for changes in the restitution of timbre. Gabrielsson *et al.* (1990) associated their dimension "clarity" with an amplification of medium frequencies between 500 and 4000 Hz.

Even if musical excerpts influence the perceptual evaluation of sound reproduction, the dimensions involved in our experiment were interpreted in the same way for the three tested musical excerpts, like in the listening tests of Gabrielsson *et al.* (1974). Our dimensions also remained unchanged when other recording techniques [Lavandier *et al.* (2004)], reproduction modes [Lavandier *et al.* (2005a)], and listening room [Lavandier *et al.* (2005b)] were tested. These perceptual dimensions seem then characteristic of the loudspeakers we intended to measure. The experiment presented here involving 12 loudspeakers was a first attempt to use our protocol to investigate the sound reproduction of loudspeakers in rooms. In order to consider dimensions which are characteristic of the perception of sound reproduction in general, listening tests should involve as many different loudspeakers as possible. To our knowledge, 12 loudspeakers involved in the same paired-comparison test is the maximum of what can be found in the literature. Paired comparisons rapidly limit the number of loudspeakers which can be included in the same test, as the listening test becomes too long for the participants to be able to handle the task. A listening test based on a free classification task allowed us to involve up to 37 loudspeakers [Lavandier *et al.* (2005b)]. The results were promising as four perceptual dimensions were revealed, two of them being very similar to the dimensions presented in this paper. The potential influence of the classification task has yet to be clarified.

Our approach to define acoustical attributes describing the perceptual dimensions consisted in identifying first the best method of analysis for a relevant description of the signals. Whereas the evaluation of the correlation between the different types of acoustical dissimilarities and the perceptual dissimilarities allowed us to quantify the link between measurement and perception [Lavandier *et al.* (2008)], the comparison of acoustical and perceptual spaces took into account the multidimensional nature of our perception of reproduced sound. Pols *et al.* (1969) used a similar approach, comparing acoustical and perceptual spaces resulting from a principal component analysis, to look for a method of analysis relevant to study the perception of vowels. Even if a criterion is still needed to quantify the similarity between two spaces, it seems easier to visually compare spaces with a low number of dimensions rather than loudspeaker responses taken as a whole like in Staffeldt (1974), Toole (1986), or Gabrielsson *et al.* (1991).

The comparison of the acoustical and perceptual spaces led to the same conclusion as the evaluations of the correlation between the acoustical and perceptual dissimilarities: By using an auditory model to discriminate recordings, we got closer to the perceptual dissimilarities, and to our goal aiming at a more relevant acoustical discrimination of loudspeakers. Among the tested acoustical analyses, only specific loudnesses could highlight our perceptual dimensions. Klippel (1990) and Bramsløw (2004) based their acoustical evaluation of sound reproduction on the specific loudness of their signals. They chose this method of analysis a priori,

assuming that it would lead to better results than other analyses. They obtained good results, but did not perform actual comparisons with other methods of analysis. Our results seem to validate their original hypothesis. Staffeldt (1974) tested loudspeaker measurements involving or not an auditory model. Even if he relied only on visual comparisons, using an auditory model led to a better fit between the shape of the measurement curves and the results of the listening tests. It seems crucial to take into account auditory masking in order to reveal the information used by listeners in their evaluation. The reason why measurements of frequency responses usually used to evaluate loudspeaker cannot be easily linked to the results of listening tests might be that these measurements cannot take into account these auditory phenomena.

It is difficult at this stage to precisely state which particular auditory phenomena should be taken into account or not. Thirteen acoustical analyses were tested with their associated metrics, but other signal analyses and metrics need to be investigated to refine the results. For example, the excitation pattern [Moore and Glasberg (1983)] of the recordings could be considered as an intermediate analysis between spectrum and specific loudness. The excitation pattern models the frequency resolution of the ear, without taking into account auditory masking effects. However, even if it remained to be tested, modeling only the frequency resolution of the ear might not be sufficient to fully describe the differences perceived by listeners. If it were the case, the overall specific loudness should have been successful in highlighting these differences. Overall specific loudness led to an acoustical space very similar to the perceptual one with the excerpt Kan'nida, but not with the two other musical excerpts (Fig. 5). The global analysis used while calculating overall specific loudness was not sufficient to accurately describe the perceptual dissimilarities. Our experiment indicated that it might be necessary to take into account the nonstationarity of auditory masking. We are not aware of any other study comparing different auditory models for the purpose of estimating acoustical dissimilarities. Klippel (1990) and Staffeldt (1974) only considered overall specific loudness, while Bramsløw (2004) only used a time-varying specific loudness. Further experiments are required to confirm our results. When Staffeldt (1974, 1984, 1991) directly applied an auditory model on the frequency response of loudspeakers to evaluate their sound reproduction, he did not take into account the fact that loudspeakers reproduce nonstationary musical signals and that the temporal evolution of these signals might influence the perception of listeners. If the importance of taking into account these nonstationary effects is confirmed in the future, the evaluation of loudspeakers by the direct examination of their frequency response might be questioned.

VI. CONCLUSION

This study investigated the perceived characteristics of the sound reproduction of 12 loudspeakers in a listening room. Paired comparisons were realized by using headphones and recordings of the sound reproductions. A MDS analysis of the resulting perceptual dissimilarities revealed two main perceptual dimensions used by listeners to discriminate the loudspeakers. These dimensions were identical for three musical excerpts. The signals heard by the listeners were used to define acoustical attributes describing the perceptual dimensions. Instead of arbitrarily choosing one acoustical analysis, several analyses were compared. Only auditory models analyzing the signals in terms of their specific loudness allowed a good description of the differences perceived by the listeners. These models were used to define two acoustical attributes describing our perceptual dimensions: the bass/treble balance and the medium emergence.

ACKNOWLEDGMENTS

We wish to thank Dr. Jeremy Marozeau for providing us his MDS programme and for our very instructive discussions, Dr. Barrie Edmonds for his comments on an earlier version of this manuscript, the Mosquito Group and Genesis for lending us their loudspeakers, and all the listeners who took part in the experiment. The authors are grateful to the associate editor and the three anonymous reviewers for their helpful comments. The work of Mathieu Lavandier was supported by a grant from the Centre National de la Recherche Scientifique (C.N.R.S.) and the Région Provence-Alpes-Côtes d'Azur.

AES20–1996 (**1996**). "AES recommended practice for professional audio—Subjective evaluation of loudspeakers," J. Audio Eng. Soc. **44**, 382–400.
Bech, S. (**1994**). "Perception of timbre of reproduced sound in small rooms: Influence of room and loudspeaker position," J. Audio Eng. Soc. **42**, 999–1007.
Bech, S. (**1995**). "Timbral aspects of reproduced sound in small rooms. I," J. Acoust. Soc. Am. **97**, 1717–1726.
Bech, S. (**1996**). "Timbral aspects of reproduced sound in small rooms. II," J. Acoust. Soc. Am. **99**, 3539–3549.
Bech, S. (**2002**). "Requirements for low-frequency sound reproduction, Part 1: The audibility of changes in passband amplitude ripple and lower system cutoff frequency and slope," J. Audio Eng. Soc. **50**, 564–580.
Borg, I., and Groenen, P. (**1997**). *Modern Multidimensional Scaling: Theory and Applications* (Springer, New York).
Bramsløw, L. (**2004**). "An objective estimate of the perceived quality of reproduced sound in normal and impaired hearing," Acta. Acust. Acust. **90**, 1007–1018.
British Standard ISO 226 2003 (**2003**). "Acoustics—normal equal-loudness level contours," BSi.
Choisel, S., and Wickelmaier, F. (**2007**). "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," J. Acoust. Soc. Am. **121**, 388–400.
Dillon, W. R., and Goldstein, M. (**1984**). *Multivariate Analysis: Methods and Applications* (Wiley, New York).
Eisler, H. (**1966**). "Measurement of perceived acoustic quality of sound-reproducing systems by means of factor analysis," J. Acoust. Soc. Am. **39**, 484–492.
Fastl, H. (**1997**). "The psychoacoustics of sound-quality evaluation," Acta. Acust. Acust. **83**, 754–764.
Gabrielsson, A., Hagerman, B., Bech-Kristensen, T., and Lundberg, G. (**1990**). "Perceived sound quality of reproductions with different frequency responses and sound level," J. Acoust. Soc. Am. **88**, 1359–1366.
Gabrielsson, A., and Lindstrom, B. (**1985**). "Perceived sound quality of high-fidelity loudspeakers," J. Audio Eng. Soc. **33**, 33–53.
Gabrielsson, A., Lindström, B., and Till, O. (**1991**). "Loudspeaker frequency response and perceived sound quality," J. Acoust. Soc. Am. **90**, 707–719.
Gabrielsson, A., Rosenberg, U., and Sjögren, H. (**1974**). "Judgments and dimension analyses of perceived sound quality of sound-reproducing systems," J. Acoust. Soc. Am. **55**, 854–861.
Gabrielsson, A., and Sjögren, H. (**1979**). "Perceived sound quality of sound-reproducing systems," J. Acoust. Soc. Am. **65**, 1019–1033.

Guastavino, C., and Katz, B. F. G. (**2004**). "Perceptual evaluation of multidimensional spatial audio reproduction," J. Acoust. Soc. Am. **116**, 1105–1115.
Guski, R. (**1997**). "Psychological methods for evaluating sound quality and assessing acoustic information," Acta. Acust. Acust. **83**, 765–774.
IEC Publication 60268-13 (**1998**). "Sound system equipment—Part 13: Listening tests on loudspeakers," International Electrotechnical Commission, Geneva, Swizerland.
Illényi, A., and Korpássy, P. (**1981**). "Correlation between loudness and quality of stereophonic loudspeakers," Acustica **49**, 334–336.
Klippel, W. (**1990**). "Multidimensional relationship between subjective listening impression and objective loudspeaker parameters," Acustica **70**, 45–54.
Lavandier, M., Guyot, B., Meunier, S., and Herzog, P. (**2005a**). "The influence of stereophony on the restitution of timbre by loudspeakers," *Proceedings AES 119th Convention*, Paper No. 6619.
Lavandier, M., Herzog, P., and Meunier, S. (**2004**). "Perceptual and physical evaluation of loudspeakers," *Proceedings AES 117th Convention*, Paper No. 6240.
Lavandier, M., Herzog, P., and Meunier, S. (**2008**). "Comparative measurements of loudspeakers in a listening situation," J. Acoust. Soc. Am. **123**, 77–87.
Lavandier, M., Meunier, S., and Herzog, P. (**2005b**). "Perceptual and physical evaluation of differences among a large panel of loudspeakers," *Forum Acusticum 2005*, Paper No. 430-0, pp. 1689–1694.
Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Moore, B. C. J., and Tan, C. T. (**2003**). "Perceived naturalness of spectrally distorted speech and music," J. Acoust. Soc. Am. **114**, 408–419.
Olive, S. E. (**2004a**). "A multiple regression model for predicting loudspeaker preference using objective measurements: Part 1—Listening test results," *Proceedings AES 116th Convention*, Paper No. 6113.
Olive, S. E. (**2004b**). "A multiple regression model for predicting loudspeaker preference using objective measurements: Part 2—Development of the model," *Proceedings AES 117th Convention*, Paper No. 6190.
Olive, S. E., Schuck, P. L., Sally, S. L., and Bonneville, M. E. (**1994**). "The effect of loudspeaker placement on listener preference ratings," J. Audio Eng. Soc. **42**, 651–669.
Paulus, E., and Zwicker, E. (**1972**). "Programme zur automatischen bestimmung der lautheit aus terzpegeln oder frequenzgruppenpegeln (Computer programs for calculating loudness from third octave band levels or from critical band levels)," Acustica **27**, 253–266.
Pedersen, J. A., and Mäkivirta, A. (**2002**). "Requirements for low-frequency sound reproduction, Part 2: Generation of stimuli and listening system equalization," J. Audio Eng. Soc. **50**, 581–593.
Pols, L. C. W., Van Der Kamp, L. J. T., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**, 458–466.
Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. (**2005a**). "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," J. Acoust. Soc. Am. **118**, 968–976.
Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. (**2005b**). "Relationships between experienced listener ratings of multichannel audio quality and naive listener preferences," J. Acoust. Soc. Am. **117**, 3832–3840.
Staffeldt, H. (**1974**). "Correlation between subjective and objective data for quality loudspeakers," J. Audio Eng. Soc. **22**, 402–415.
Staffeldt, H. (**1984**). "Measurement and prediction of the timbre of sound reproduction," J. Audio Eng. Soc. **32**, 410–414.
Staffeldt, H. (**1991**). "Differences in the perceived quality of loudspeaker sound reproduction caused by the loudspeaker-room-listener interactions," *Proceedings AES 90th Convention*, Paper No. 3046 (H-1).
Susini, P., McAdams, S., and Winsberg, S. (**1999**). "A multidimensional technique for sound quality assessment," Acta. Acust. Acust. **85**, 650–656.
Toole, F. E. (**1985**). "Subjective measurements of loudspeaker: sound quality and listener performance," J. Audio Eng. Soc. **33**, 2–32.
Toole, F. E. (**1986**). "Loudspeaker measurements and their relationship to listener preferences: Part 2," J. Audio Eng. Soc. **34**, 323–348.
Toole, F. E. (**1991**). "Binaural record/reproduction systems and their use in psychoacoustic investigations," *AES 91st Convention*, Paper No. 3179 (L-6).
Zwicker, E., and Fastl, H. (**1983**). "A portable loudness-meter based on ISO 532B," *11th International Congress on Acoustics*, 135–137.
Zwicker, E., and Fastl, H. (**1999**). *Psychoacoustics: Facts and Models* (Springer, New York).
Zwicker, E., Fastl, H., and Dallmayr, C. (**1984**). "BASIC-Program for calculating the loudness of sounds from their 1/3-oct band spectra according to ISO 532 B," Acustica **55**, 63–67.

Vibration activity and mobility of structure-borne sound sources by a reception plate method

B. M. Gibbs,[a] R. Cookson, and N. Qi
Acoustics Research Unit, School of Architecture, University of Liverpool, Liverpool L69 3BX, United Kingdom

(Received 31 May 2007; revised 10 March 2008; accepted 12 March 2008)

This paper considers a practical structure-borne sound source characterization for mechanical installations, which are connected to plate-like structures. It describes a laboratory-based measurement procedure, which will yield single values of source strength in a form transferable to a prediction of the structure-borne sound power generated in the installed condition. It is confirmed that two source quantities are required, corresponding to the source activity and mobility. For the source activity, a high-mobility reception plate method is proposed which yields a single value in the form of the sum of the squared free velocities, over the contact points. A low-mobility reception plate method also is proposed which, in conjunction with the above, yields the source mobility in the form of the average magnitude of the effective mobility, again over the contact points. Experimental case studies are described and the applicability of the laboratory data for prediction and limitations of the approach are discussed. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2904469]

PACS number(s): 43.40.At, 43.40.Sk, 43.58.Bh [KA] Pages: 4199–4209

I. INTRODUCTION

Airborne sound transmission is generally well understood and consultants and manufacturers are able to use relatively straightforward methods and data for the prediction of the sound pressure at distance from a source of known sound power. There is a menu of standard recommended laboratory methods of obtaining airborne sound power: by source substitution, in anechoic or reverberant conditions, and by intensity measurement. In many cases, a single value of sound power is suitably representative of the source as a whole with resultant simplifications in the prediction of sound pressure. This is not the case for structure-borne sources. Reviews by Bodén,[1] Petersson and Gibbs,[2] and Moorhouse[3] have identified the need for a source characterization which will yield data and methods appropriate for transmission prediction. ten Wolde and Gadefelt highlight the difficulties in developing test methods which are simple but at the same time generate data appropriate for prediction.[4] Mondot and Petersson propose the source descriptor on a power basis, which in combination with a coupling function gives the installed power.[5] Verheij has employed reverse and reciprocal methods to quantify source strength and the hierarchy of transmission paths *in situ*.[6] Simplifications are possible by visual inspection of the source point mobility to establish whether it is mass, stiffness or resonance controlled, leading to data reduction.[7]

Despite such work over previous years, the methods proposed have not been taken up by industry and there remains a need for a practical approach. Practical methods normally involve limited data acquisition and are not computationally intensive. Manufacturers view their products as single entities and seek an associated single value of source strength. Again, this is usually possible for airborne sound sources. Test houses and small R&D facilities are geared to measure spatial and spectral average values, typically as octave or third octave band magnitudes, using microphones or accelerometers.

There are significant challenges in seeking a practical approach. The structural dynamics of both the source and receiving systems are required for prediction of transmission in the installed condition (the installed power). The structure-borne transmission is complicated and requires consideration of several contacts and up to six components of excitation at each contact; the source also may be connected to several structural elements all of which will contribute to the total transmission to a greater or lesser degree. The transmissions at the contacts may differ greatly, in which case the concept of a single equivalent excitation becomes tenuous.

However, the potential benefits of simplification warrant further consideration, even if there is loss of accuracy and not all situations are amenable to the approach. This paper considers machines or components which are installed in contact with structures such as homogeneous plates or rib-stiffened plates. Examples are mechanical installations in heavyweight and lightweight buildings. The main motivation for the work reported was to examine how laboratory data, in the form of single equivalent magnitudes, might be used for prediction of installed structure-borne power, which conventionally requires complex-valued data for each contact between machine and supporting structure.[8]

Internal source mechanisms, such as impacts and pressure variations, generate vibrations at the interface between the source and the supporting/connected structure. The map of the internal mechanisms to the external velocities or forces can be termed the source *activity*. Source activity can be expressed as the free velocity v_{sf}, the velocity of the freely suspended source, or the blocked force F_b, the force at the contact with an inert receiver.[4] Direct measurement of

[a]Author to whom correspondence should be addressed. Electronic mail: bmg@liv.ac.uk

free velocity and blocked force usually requires the source to be removed from the installation while operating in otherwise normal conditions.[9] Elliot *et al.* demonstrate that the blocked force of a rigidly connected machine can be obtained indirectly from the contact velocity and contact mobility at each point, irrespective of the dynamic conditions at the contact.[10] Whether or not they can be measured directly, free velocity and blocked force are important concepts and free velocity forms the basis of the following discussion.

For simplicity, initially consider a source transmitting power to a receiver through a single contact and a single component of excitation. The transmitted power P is the real part of the complex power W and can be expressed in terms of the free velocity v_{sf} of the source and the complex source mobility Y_S and receiver mobility Y_R[11] as

$$P = \mathrm{Re}(W) = \frac{1}{2}|v_{sf}|^2 \frac{\mathrm{Re}(Y_R)}{|Y_S + Y_R|^2}. \tag{1}$$

Machines and machine components are connected to supporting structures through multiple points, line and area contacts. At each contact, up to six components of excitation and response are possible (three translations and three rotations) and in general there are dynamic interactions between the different contacts and components. The expression for total power becomes

$$P = \frac{1}{2}(\{v_{sf}\}^T[[[Y_S] + [Y_R]]^T]^{-1}\mathrm{Re}([Y_R])[[Y_S^*] + [Y_R^*]]^{-1}\{v_{sf}^*\}). \tag{2}$$

$\{v_{sf}\}$ is the free velocity vector and $[Y_{S,R}]$ the mobility matrices. For N contacts and six components of excitation, a $6N \times 6N$ mobility matrix is required.

It is possible to preserve the simplicity of the single contact/component case [Eq. (1)] by reference to the concept of the effective mobility.[12,13] Consider a source connected through multiple contacts to a supporting plate structure. In this study, forces perpendicular to the receiver structure only are assumed. This assumption results from studies of the relative importance of moments and forces in the transmission of structure-borne power from machines in buildings. Reciprocal methods have been used to indirectly obtain the contact forces and moments generated by a fan base on a concrete floor.[14] Moments are less important than perpendicular forces when the source is away from structural discontinuities such as floor edges. Moments assume more importance in the proximity of structural discontinuities and have an increasing contribution with increased frequency irrespective of excitation location. For the cases considered, moments at most have an equal contribution to perpendicular forces and in many situations can be neglected. Recent work confirms this for the case of lightweight timber stairs connected to party walls between dwellings.[15]

The total power, from a source S to a receiver R, through all N contacts is

$$P_{SR}^{\mathrm{Total}} = \frac{1}{2}\sum_i^N |v_{sfi}|^2 \frac{\mathrm{Re}(Y_{Ri}^{\Sigma})}{|Y_{Si}^{\Sigma} + Y_{Ri}^{\Sigma}|^2}. \tag{3}$$

The power through the ith contact is obtained from Eq. (1) but where point mobilities are replaced by effective point mobilities. For both the source and receiver, the effective point mobility at the ith contact can be written as

$$Y_i^{\Sigma} = Y_i + \sum \frac{F_j}{F_i} Y_{i,j}, \tag{4}$$

where Y_i is the point mobility at the ith contact, $Y_{i,j}$ is the transfer mobility between the ith and jth contacts and $F_j/(F_i)$ is the ratio of the forces at the jth and ith contact, respectively. Equation (4) reveals a requirement for the force distribution $F_j/(F_i)$ over the contacts.[16] This information is not likely to be available in practice since the installation conditions (location, receiver plate geometry, edge conditions, etc.) will not be known in sufficient detail. In the absence of such information, simplifying assumptions are necessary. The forces can be assumed to be of equal magnitude. The phase difference between forces depends on the vibration behavior of the source. If the source has a rigid body motion, as occurs at low frequencies, then a multi-pole representation is appropriate, including a zero phase difference (bouncing mode) condition. At high frequencies, a resonant behavior is likely for either or both of the source and receiver structures and a random phase difference between contact points can be assumed. Again, in the absence of detailed information, two asymptotic conditions were considered.[12,13] If a zero phase difference is assumed, then Eq. (4) becomes

$$Y_i^{\Sigma} \approx Y_i + \sum Y_{i,j}. \tag{5}$$

If a random phase is assumed then,

$$|Y_i^{\Sigma}|^2 \approx |Y_i|^2 + \sum |Y_{i,j}|^2$$

and

$$\mathrm{Re}(Y_i^{\Sigma}) \approx \mathrm{Re}(Y_i). \tag{6}$$

II. RECEPTION PLATE METHOD

The approach presently proposed stems from recent work concerning test methods for mechanical installations in heavyweight buildings.[17,18] It is based on the reception plate method.[11] A prototype thin reception plate, appropriate for small equipment, previously was proposed by others and is described by Lu *et al.*[19] Further consideration to the reception plate method is given in recent publications.[20–22] The machine under test is attached to a simple plate, under otherwise normal operating conditions. The total structure-borne power transmitted is obtained from the spatial average of the mean square plate velocity $\langle v_R^2 \rangle$, as

$$P_{SR}^{\mathrm{Total}} = \omega \eta_R \ddot{m}_R S_R \langle v_R^2 \rangle. \tag{7}$$

η_R is the total loss factor of the receiving plate of area S_R and $\ddot{m}$ is the mass per unit area.

Alternatively, the total power can be obtained by a source substitution procedure. A shaker is attached to the

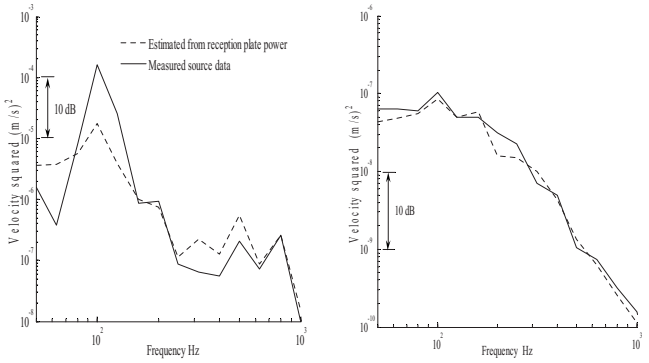


FIG. 1. Sum of squared free velocities of fan (left) and whirlpool bath (right); (after Ref. 8).

plate, through a force transducer and a cross-spectral estimate of the power (input force-contact velocity) recorded, along with the spatial average plate response velocity. The machine under test then is installed and the resultant spatial average plate response velocity again is recorded. The ratio of response velocities equals the ratio of powers of the two sources and the unknown machine power is obtained.[20]

If, at each contact, the reception plate has a high mobility relative to that of the machine under test, i.e., $|Y_{Si}^{\Sigma}| \ll |Y_{Ri}^{\Sigma}|$ for all i, then Eq. (3) becomes

$$P_{SR}^{\text{Total}} \approx \frac{1}{2}\sum_{i}^{N} \text{Re}\left(\frac{1}{Y_{Ri}^{\Sigma}}\right)|v_{sfi}|^2. \tag{8}$$

If the spatial variation in plate effective mobility, about the contact points, is small, then

$$P_{SR}^{\text{Total}} \approx \frac{1}{2}\text{Re}\left(\frac{1}{Y_{R}^{\Sigma}}\right)\sum_{i}^{N} |v_{sfi}|^2, \tag{9}$$

where Y_R^{Σ} is the average effective mobility of the plate. Equating the right hand terms of Eqs. (7) and (9), where terms outside the summation can be obtained from laboratory measurement, yields the sum of the squares of the free velocities at the source contacts $\sum_{i}^{N}|v_{sfi}|^2$.

In a previous study, reception plates were numerically modeled as free (FFFF) thin plates.[8] Two sources, a medium-size fan unit and a whirpool bath, were measured to give source data consisting of four free velocity spectra and 4 $\times$4 point and transfer mobility spectra. Plate point and transfer mobilities were combined with the measured source mobilities and free velocities to yield the reception plate power, according to Eq. (2). In Fig. 1 are shown the sum of the squared free velocities $\sum_{i}^{N}|v_{sfi}|^2$ obtained by the thin reception plate method, for a fan unit and whirlpool bath. Directly measured values also are shown. While the results for the whirlpool bath are promising, there are discrepancies of the order of 10 dB at 100 and 63 Hz for the fan unit. The discrepancies are a result of plate modal behavior and the small distance between contact points (of the order of 350 mm) compared with that of the whirlpool bath (700 mm), giving rise to interference effects.

If the source now is connected to a plate of low mobility where $|Y_{Si}^{\Sigma}| \gg |Y_{Ri}^{\Sigma}|$ for all i, then the total power is,

$$P_{SR}^{\text{Total}} \approx \frac{1}{2}\sum_{i}^{N} \frac{|v_{sfi}|^2}{|Y_{Si}^{\Sigma}|^2} \text{Re}(Y_{Ri}^{\Sigma}). \tag{10}$$

Again, when mounted on a reception plate, the total power is obtained from Eq. (7). Also, assuming a small spatial variation in effective mobility of the reception plate, then Eq. (10) can be expressed as

$$P_{SR}^{\text{Total}} \approx \frac{1}{2}\text{Re}(Y_{R}^{\Sigma})\sum_{i}^{N} \frac{|v_{sfi}|^2}{|Y_{Si}^{\Sigma}|^2}, \tag{11}$$

The structure-borne sound source strength is obtained, based on the blocked force,

$$\sum_{i}^{N} \frac{|v_{sfi}|^2}{|Y_{Si}^{\Sigma}|^2} = \sum_{i}^{N} |F_{bi}|^2. \tag{12}$$

It is possible to obtain a source mobility term from Eqs. (9) and (12) if the effective source mobilities at each contact are assumed equal. Equation (12) becomes,

$$\sum_{i}^{N} \frac{|v_{sfi}|^2}{|Y_{Si}^{\Sigma}|^2} = \frac{1}{|Y_{S}^{\Sigma}|^2}\sum_{i}^{N} |v_{sfi}|^2. \tag{13}$$

From Eqs. (9), (11), and (13) the magnitude of the average effective source mobility $|Y_S^{\Sigma}|$ is obtained. In Fig. 2 are shown the average magnitude of the effective mobility for the same sources, the fan unit, and whirlpool bath, obtained from the thick reception plate method, in combination with the thin reception plate data. Directly measured values again are shown. The discrepancies generally are greater than in Fig. 1. This is to be expected since the errors in the reception plate estimates of the sum of the squared free velocities have been carried into the second stage of the procedure. In addition, the low-mobility (thick) plate has a lower modal density than the thin plate and the influence of contact location and distance between contacts is greater.

III. EXPERIMENTAL CASE STUDY: ELECTRIC MOTOR

An experimental evaluation of the approach is described for two machines. The first was a small electric motor, previously installed in a domestic tumble dryer. In Fig. 3 is shown the motor attached to a high mobility reception plate. The motor previously was freely suspended and the free velocities at the four support points recorded at two operating speeds, 2600 and 2960 rpm. In Fig. 4 are shown the directly measured sums of the squared velocities. The level difference, between running speeds, is about 5 dB. Both curves show a decrease with increased frequency of −10 dB per octave from a maximum at 50 Hz. In Figs. 8 and 9, the free velocities also are shown as one-third octave values for comparison with the reception plate estimates.

A. High-mobility reception plate

The high-mobility reception plate was constructed from 1.5 mm perforated mild steel (with hole diameter 6.35 mm and open area 47%). The plate was of area 2×1 m^2 and was

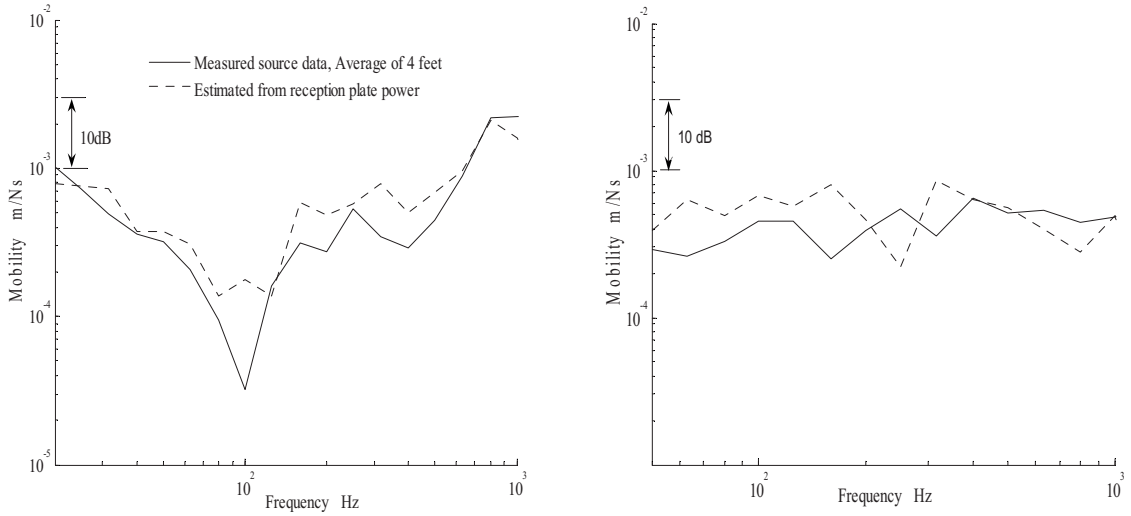


FIG. 2. Average magnitude of effective mobility of fan (left) and whirpool bath (right); (after Ref. 8).

clamped into a steel frame for stability (Fig. 3). To confirm that the required source-receiver mobility condition $|Y_S^{\Sigma}| \ll |Y_R^{\Sigma}|$ applied, the point mobilities at the four motor contact points were recorded using a calibrated force hammer.[22] The plate point mobility was measured at the four contact points with the motor and the mean value obtained. Figure 5 indicates that the level difference between source and receiver point mobility magnitudes is seldom less than 10 dB and the required mobility mismatch condition can be assumed. The measured plate mobility indicated frequency-invariant infinite plate behavior over the frequency range of interest 50 Hz–4 kHz.

The reception plate method is also known as the reverberant plate method[23] and the bending field is normally required to be of high modal density. The modal density of a plate of area S and thickness h is given approximately[11] as

$$n(f) \approx \frac{S}{c_L h}\sqrt{3}, \tag{14}$$

where c_L is longitudinal wave velocity. For the reception plate under consideration, the modal density is 0.53 modes/Hz. Since measured results are presented as one-third octave values, the modal density is considered in a similar manner. The lowest bandwidth considered is centered on 50 Hz, and contains approximately six modes. The 500 Hz band contains 61 modes; the 5k Hz band, 613 modes. The plate bending field therefore was assumed diffuse over the frequency range of interest.

Point and transfer mobilities were measured at the four attachment points on the reception plate. The effective mobilities were calculated for each point, assuming a unit force, zero phase difference, from Eqs. (5), or a unit force, random phase difference, from Eq. (6). In Fig. 6 are shown the resultant mean values of $\mathrm{Re}\{1/(Y_{R,}^{\Sigma})\}$, as one-third octave values, for both phase assumptions. The peak value in the zero phase estimate results from a dip in the complex effective mobility, which occurs when contact distances are of the order of one half the plate bending wave-length [see Eq. (5)]. This is not seen in the random phase estimate [see Eq. (6)] where squared values are summed.

From analogous simple circuit theory,[24] the contact force between a low mobility source and a high mobility receiver is given approximately by

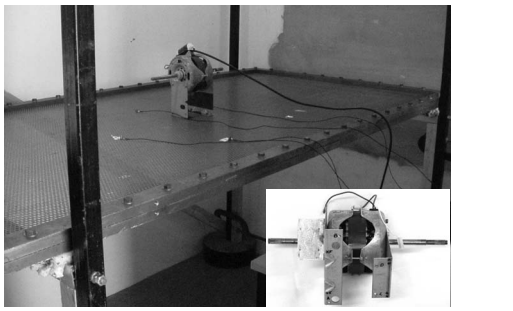


FIG. 3. Electric motor on high mobility reception plate; inset, attachment points.

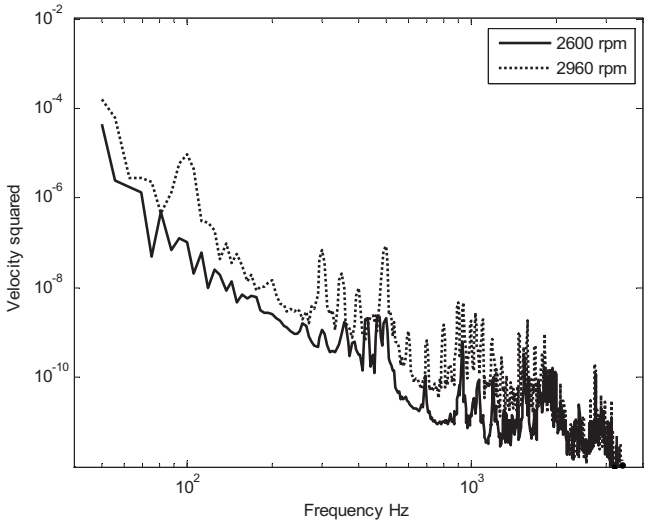


FIG. 4. Directly measured sum of squared free velocities of electric motor at two running speeds.

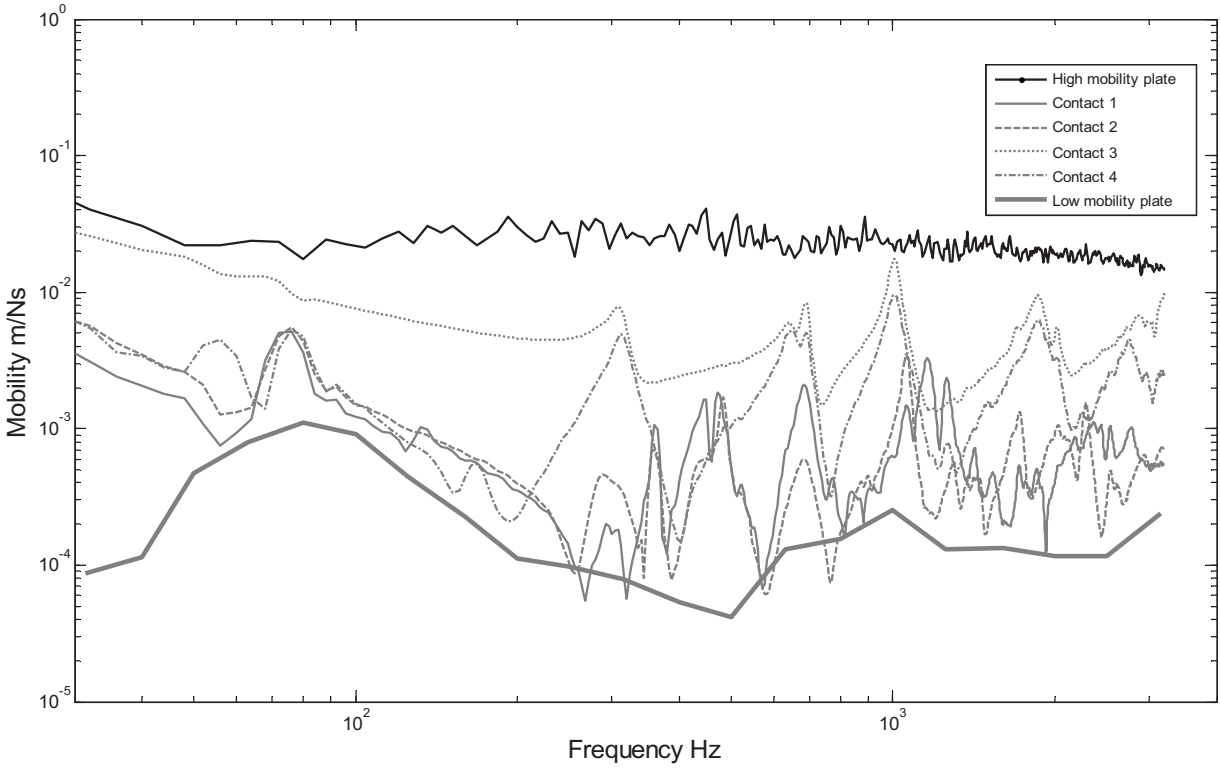


FIG. 5. Magnitude of mean point mobility of the high mobility reception plate and point mobilities at four motor contact points; also shown is the mean point mobility of the low-mobility plate, as one third octave values.

$$F_{\text{contact}} \approx \frac{v_{sf}}{Y_R}. \tag{15}$$

The phase differences between contact forces therefore are determined in equal part by the phase differences between the free velocities at the source contacts and the phase of the reception plate transfer mobilities, again between contacts. Therefore, if either or both have a random phase difference, then this condition applies to the contact forces. In Fig. 7 are shown the phase differences between the free velocities of three contact points on the motor, relative to one contact, for two operating speeds. A zero-phase difference is indicated at frequencies below 300 Hz, at the slower speed. At the faster speed, more internal mechanisms contribute to the source activity, such as bearing noise, and random phase behavior occurs above 100 Hz. Regarding the reception plate transfer mobilities and from previous consideration of the modal density of the reception plate, a random phase difference between forces can be assumed over the whole frequency range of interest.

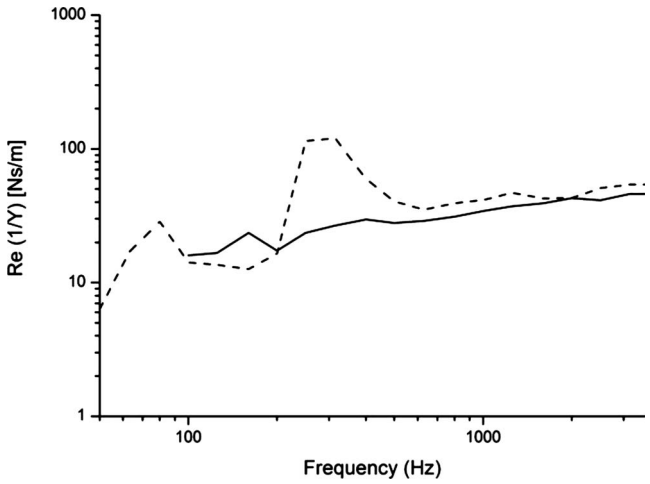


FIG. 6. Measured mean value $\mathrm{Re}\{1/(Y^{\Sigma}_{R'})\}$ of high-mobility reception plate. Dashed line, assuming zero phase difference between contact forces; solid line, assuming random phase difference.

It remained to measure the plate loss factor η_R, which was obtained from measurements of reverberation time T_R, in one-third octave bands, using reverse integration of the impulse response.

With the electric motor attached and in operation, the plate velocity was recorded at five randomly selected positions. The plate power P_{SR} was calculated according to Eq. (7) and the sum of the squared free velocities, across the four support points, was obtained using Eq. (9). Results are shown in Fig. 8 for two motor speeds, with directly measured values for comparison. The difference between directly measured values and reception plate values is within 5 dB for mid frequencies and for the higher motor running speed. Larger discrepancies, of the order of 10 dB, at low frequencies, are expected, again because of low modal density of the reception plate, which corresponds to large spatial and spectral variations. The discrepancy also is greater at the lower running speed. This is likely to be the result of reduced signal to noise and because there are fewer contributing peaks in the velocity spectrum (see Fig. 4). Airborne excitation was assumed not to contribute to the plate motion. This was con-

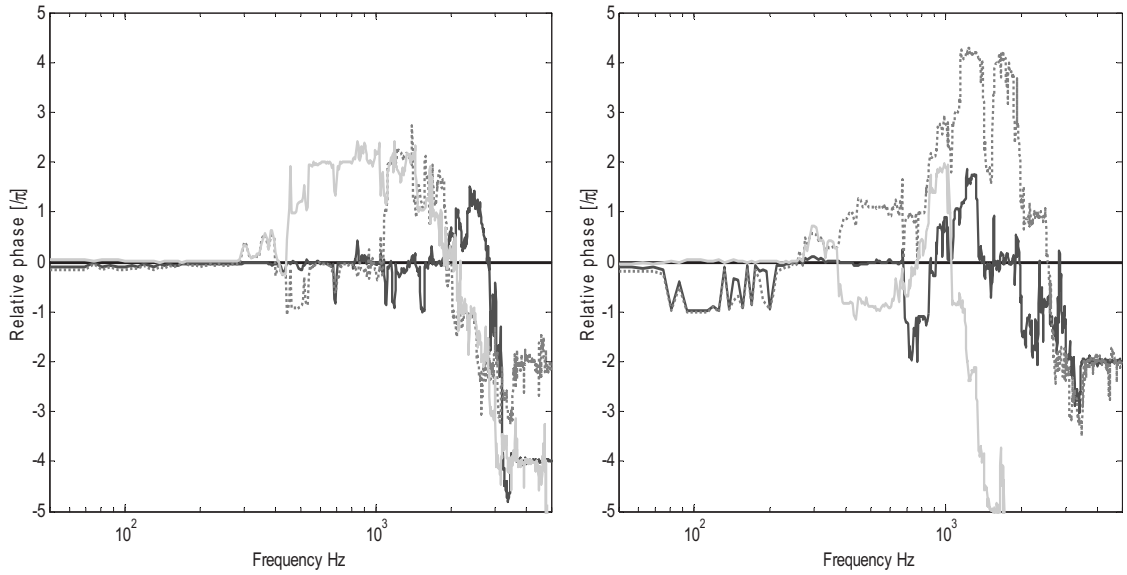


FIG. 7. Phase difference between free velocities at three contact points of electric motor, relative to fourth contact point, at 2600 rpm (left), at 2960 rpm (right).

firmed in preliminary measurements, where the operating motor was isolated from the plate but was positioned in close proximity.

The calculation of $\mathrm{Re}\{1/(Y^{\Sigma}_{R'})\}$ is central to the estimate of source free velocity [see Eq. (9)] and how it is assembled has practical implications regarding how the reception plate should be instrumented and calibrated. Modern frequency analyzers allow the direct measurement of complex frequency response functions in constant percentage bandwidths using digital filters. The advantage is the reduction in data processing. However, the loss of high frequency resolution phase data may cause an unacceptable reduction in the accuracy of the results. Three approaches were considered: (a) complex point and transfer mobilities were measured and converted immediately into 1/3 octave values; (b) complex narrowband values were used to calculate $\mathrm{Re}\{1/(Y^{\Sigma}_{R'})\}$, which then was converted to a 1/3 octave value; (c) mobilities were measured directly as complex 1/3 octave values. Figure 9 shows estimates of the sum of the square free velocities, obtained by the three procedures, for the motor operating at 2960 rpm. The three procedures give comparable agreements with directly measured free velocities and there appears to be little benefit from using narrowband values.

B. Low-mobility reception plate

From consideration of Eqs. (10)–(13), if the source now is connected to a reception plate of low mobility, then the magnitude of the average effective source mobility $|Y^{\Sigma}_{S}|$ is obtained. In Fig. 10 is shown the average effective mobility of the electric motor obtained from directly measured point and transfer mobilities of the freely suspended motor. Both

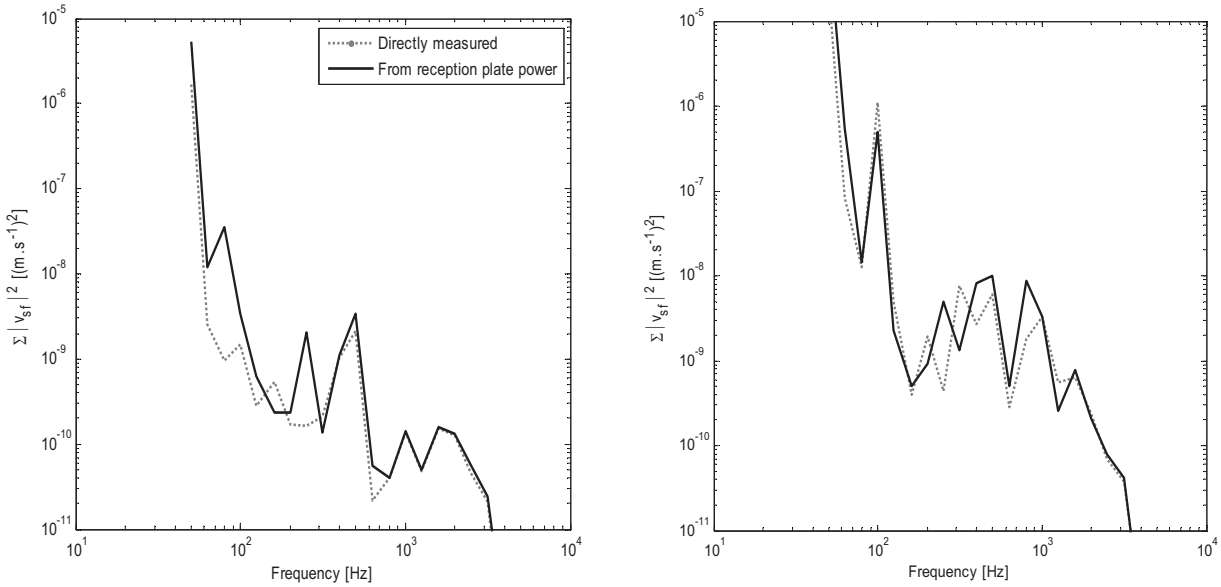


FIG. 8. Reception plate estimate and directly measured sum of squared free velocity of electric motor, at 2600 rpm (left), at 2960 rpm (right).

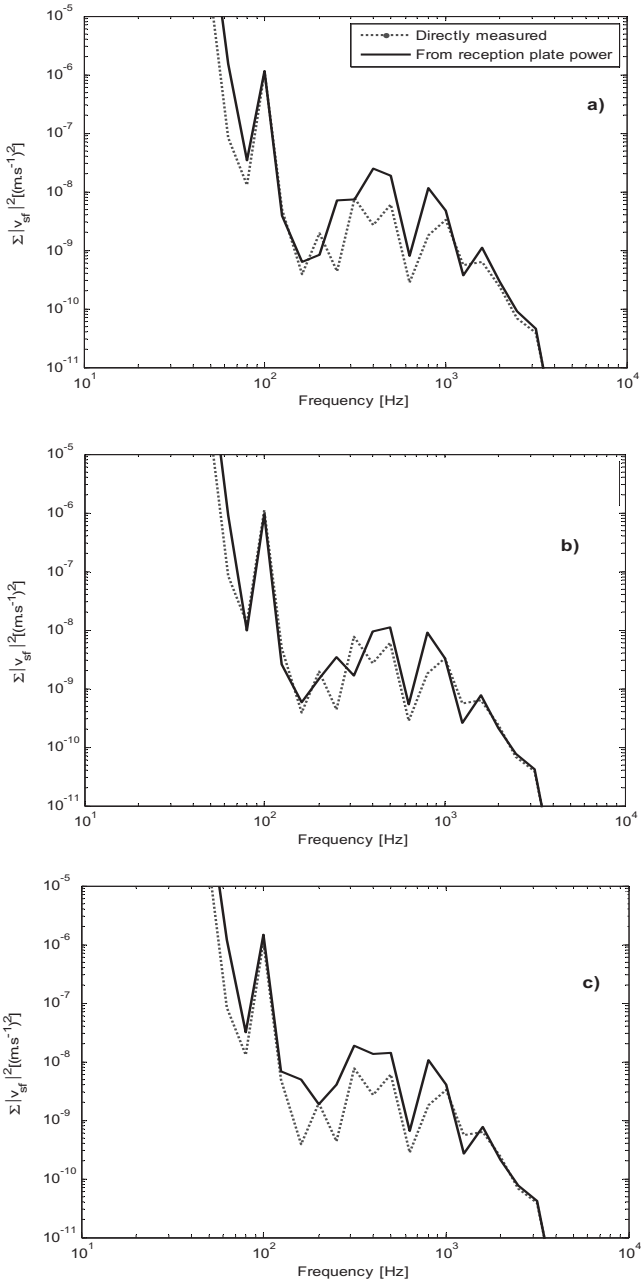


FIG. 9. Estimates of sum of squared free velocity of a motor, at 2960 rpm, from reception plate measurements. Also shown are directly measured values.

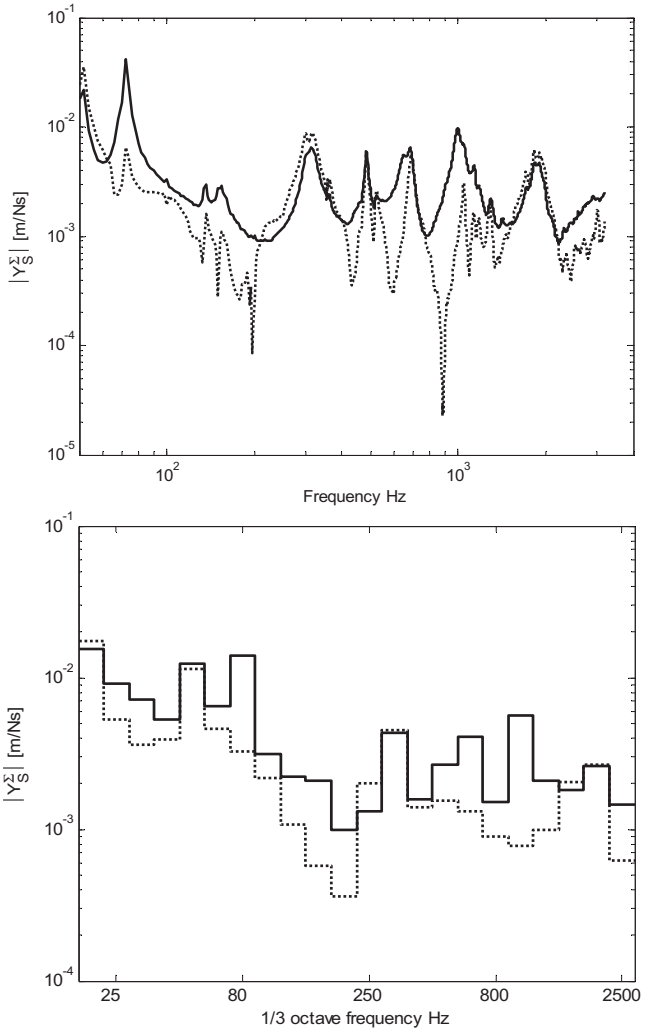


FIG. 10. Magnitude of average effective source mobility of electric motor, as narrowband and one-third octave values; dashed line, assuming zero phase difference between contact forces; solid line, assuming random phase difference.

zero phase and random phase estimates are shown, for completeness, although the random-phase estimate (i.e., uncorrelated forces), in one third octaves, was the target value.

The low-mobility plate was of aluminium with dimensions: 1500 mm × 2120 mm × 19.05 mm thickness. The plate was resiliently supported along the edges by 200 mm foam, which also provided some damping. In Fig. 11 is shown the low-mobility plate with electric motor attached. The motor was located away from the axes of symmetry of the plate.

In Fig. 4 is shown the average plate point mobility for comparison with source point mobility and the condition $|Y_{Si}^{\Sigma}| \gg |Y_{Ri}^{\Sigma}|$ therefore could be assumed. The achieved low-mobility condition is at the expense of the modal density. From Eq. (14) the modal density is given as 0.055 modes/Hz. This corresponds to 0.6 modes in the 50 Hz one-third octave bandwidth; 6.3 modes at 500 Hz; 63.3 modes at 5 kHz. A diffuse bending field condition cannot be assumed at low frequencies and thus Eq. (7) would appear not to apply. However, in recent investigations of laboratory methods for machines in buildings, consideration was given of the structural dynamics of low mobility plates, with a modal behavior, and the estimation of machine power from measurement of plate velocity.[18] It was concluded that

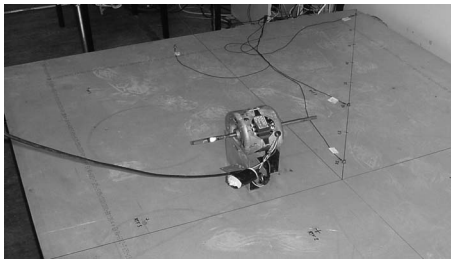


FIG. 11. Electric motor attached to the 19 mm aluminium plate.

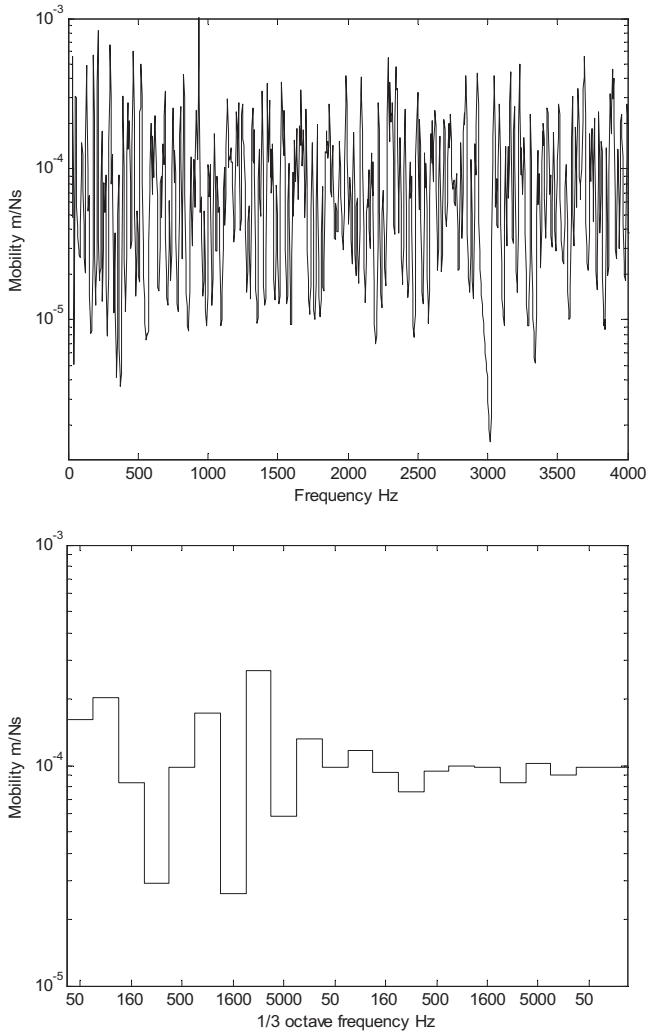


FIG. 12. Real part of effective mobility of 19 mm aluminium plate, assuming random phase difference between contacts, for frequency resolution 2 Hz and in one third octaves.

an estimate of the total structure-borne power was possible by careful sampling of the plate response, which in the case reported was a 100 mm concrete plate of dimensions 2.8 m by 2.0 m.

For calculation of the structure-borne source strength [see Eq. (11)] the spatial average value $\mathrm{Re}(Y_R^{\Sigma})$ of the plate is required. Similar to the discussion of the high mobility reception plate [Eq. (15)] and from analogous simple circuit theory[24] the contact force between a high mobility source and a low mobility receiver now is given by,

$$F_{\mathrm{contact}} \approx \frac{v_{sf}}{Y_s}. \tag{16}$$

Therefore, the phase difference between the free velocities at the source contacts, and of the source transfer mobilities between the same contacts are both influential. Indeed, the quantities are interdependent when the source displays rigid body behavior. This is likely to be the case for frequencies below 100 Hz. Again, over most of the frequency range of interest, a random phase estimate of plate effective mobility could be assumed. In Fig. 12 is shown the real part of the effective mobility of the plate, as narrowband and one-third octave values.

From consideration of Eqs. (10)–(13), the magnitude of the average effective source mobility $|Y_S^{\Sigma}|$ was obtained and is shown in Fig. 13, along with the directly measured value. There is agreement between the directly measured and the indirect reception plate estimates in terms of overall magnitude. There also is some agreement in terms of signature at high frequencies. The discrepancies at mid and low frequencies primarily are the result of the modal characteristics of the low mobility reception plate as discussed earlier.

IV. EXPERIMENTAL CASE STUDY: MEDIUM SIZE CENTRIFUGAL FAN

The experimental procedure was repeated for a medium size centrifugal fan unit shown in Fig. 14. In this case, however, the free velocities at four contact points were obtained directly by resiliently suspending the fan when running in otherwise normal conditions. The target value, the magnitude of the average effective source mobility also was obtained directly, again with the fan freely suspended. The indirectly measured value of source effective mobility was obtained by the procedure described for the small electric motor. The measurements and calculations were with a one third octave resolution throughout. The reception plate estimate and the directly measured value are shown in Fig. 15. The reception plate method gives an overestimate of 2 dB between 100 and 400 Hz, and an underestimate of about 4 dB above 500 Hz.

V. RECEPTION PLATE DATA AND PREDICTED INSTALLED POWER

A question remains on how source data, obtained by a laboratory reception plate method, can be used for prediction of the structure-borne power of the source when installed. In addition to the source data, an estimate of the complex effective receiver mobility Y^{Σ_R} is required for prediction of the installed power [see Eq. (3)]. In recent work on mechanical installations in buildings, it has been demonstrated that both heavyweight homogeneous floors and walls[18] and lightweight inhomogeneous structures[25] display plate-like dynamic behavior which can be described by the characteristic mobility Y_{char} over much of the frequency range of interest. Recent consideration of aircraft structures allows the same conclusion.[21] This statement cannot be viewed as general since supporting structures can be frame constructions. However, even in these cases, a simple estimate of receiver mobility, as a function of the characteristic beam mobility and the characteristic plate mobility, may be appropriate.[11,25]

Returning to plate-like structures, the receiver mobility is estimated from the characteristic mobility,[11]

$$Y_{\mathrm{char}} = \frac{1}{8\sqrt{B'\ddot{m}}}, \tag{17}$$

where B' is the bending stiffness. The characteristic mobility is a pure real value and so the real part and the magnitude of the receiver mobility are obtained simultaneously. The effective receiver mobility can be estimated, in simplest form, by assuming that transfer mobilities are equal to the point mo-

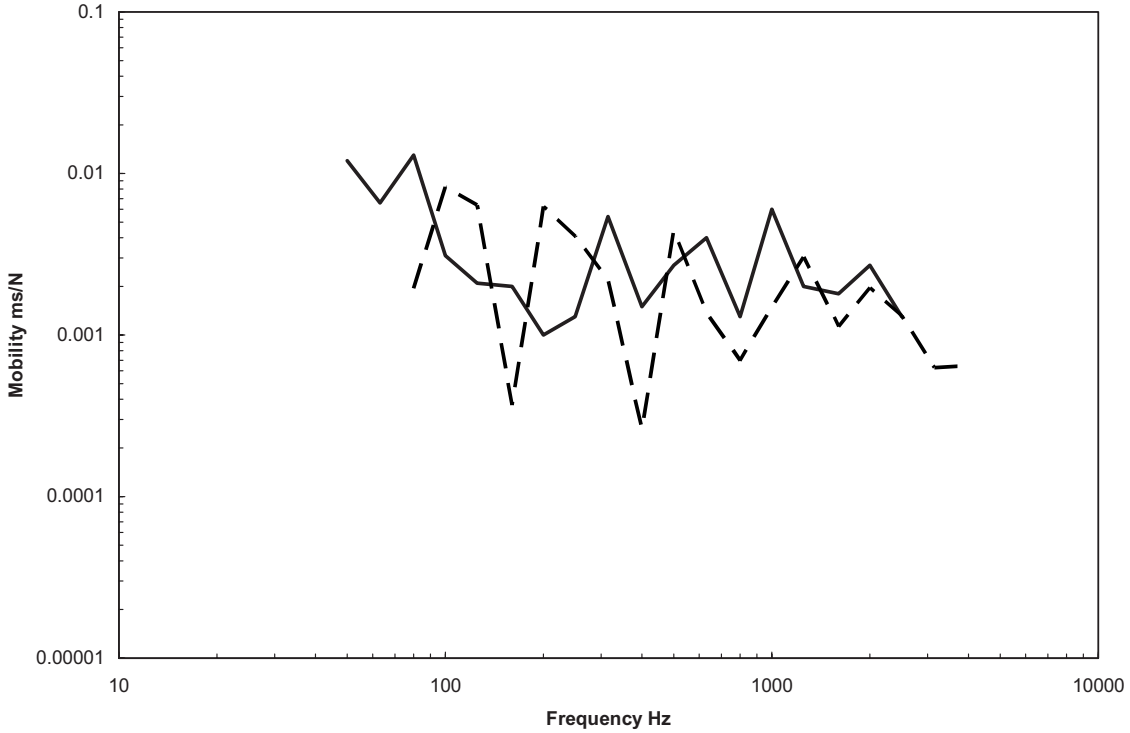


FIG. 13. Average magnitude of source effective mobility of an electric motor; solid line, directly measured; dashed line, reception plate estimate.

bilities at low frequencies and, at high frequencies, transfer mobilities reduce relative to the point mobilities, and the effective mobility converges to the characteristic point mobility.[8]

The remaining issue then concerns the source-receiver mobility condition. Again, reference to Eq. (3) indicates that the phase difference between source and receiver mobilities, contained in $|Y_S^{\Sigma}+Y_R^{\Sigma}|$, will be important, particularly when the mobilities are of the same magnitude.[3,5] However, this problem can be partly circumvented by assuming that the installation source-receiver mobility condition is in one of only three possible states.

If the magnitude of the source mobility significantly exceeds that of the receiver mobility then Eqs. (11) and (12) apply. The source strength, obtained by means of the low-mobility reception plate, can be used in combination with the characteristic receiver mobility for the installed power. This condition is common in heavyweight buildings and the transformation of laboratory data to prediction is relatively straightforward.[17]

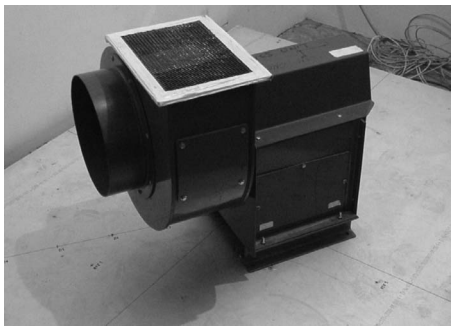


FIG. 14. Medium size centrifugal fan on low-mobility reception plate of 19 mm aluminium.

If the receiver mobility significantly exceeds the source mobility then Eqs. (8) and (9) apply and the obtained source strength can be used in combination with characteristic receiver mobility for the installed power.

If the source and receiver mobilities are of the same magnitude, a matched mobility condition is of relevance. A detailed discussion of the possible forms of matching is given by Moorhouse.[3] In a recent study, the use of a mirror condition was proposed, where the real parts and imaginary parts are, respectively, equal.[8] None of the three assumed possible source-receiver conditions requires source data in complex form.

VI. DISCUSSION AND CONCLUSIONS

A two-stage reception plate method of characterizing structure-borne sound sources is proposed for mechanical installations connected to plate-like receiver structures. The first stage employs a high-mobility plate and yields the sum of the square free velocities over the source contacts. The second stage employs a low-mobility plate and, in conjunction with the above, yields the average magnitude of the effective mobility over the same source contacts.

For resiliently suspended or supported machines with few contacts and where the loaded operating condition can be maintained, the free velocity can be measured directly with the same convenience as the reception plate method, and the source activity expressed as the sum of the squared values. However, the reception plate method offers advantages of simplicity when considering the free velocity at many contact points or when many similar machines have to be tested.

The method has not yet been applied to line and area contacts. In this case, the contacts could be discretized and

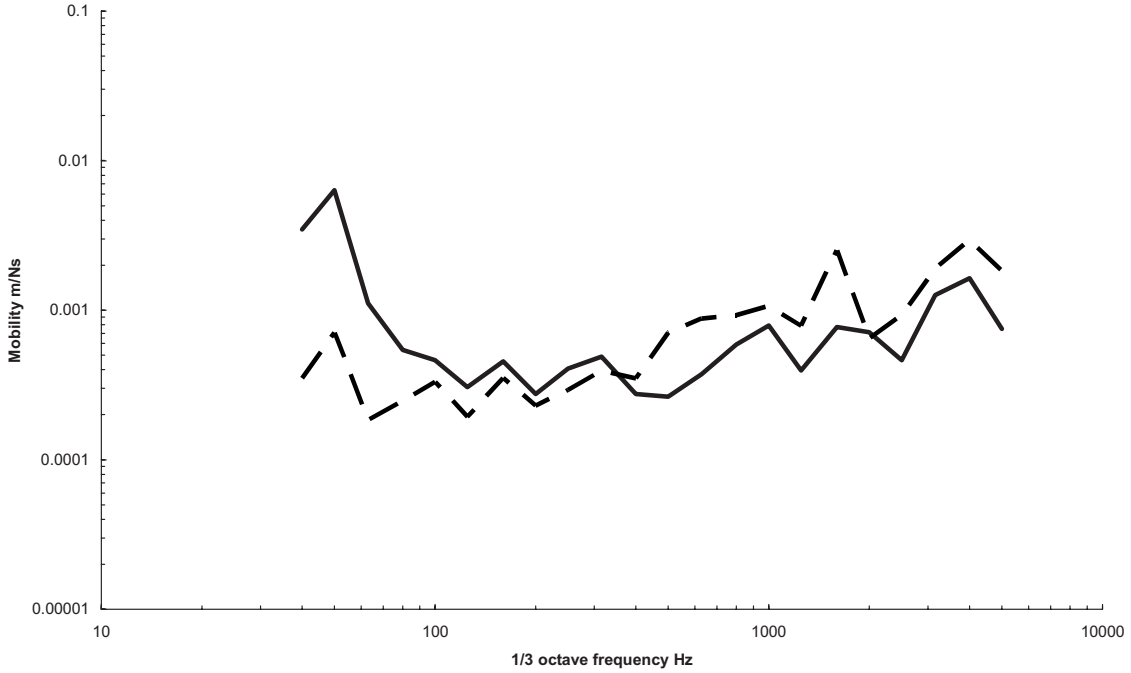


FIG. 15. Average magnitude of effective source mobility of centrifugal fan. Solid line, directly measured; dashed line, reception plate estimate.

values assigned to the linear array or mesh, although the interaction between adjacent cells will assume significance.

Mechanical installations often are in contact with more than one receiving structure (e.g., a machine on a floor with connections to walls). The outlined approach still applies but with the requirement that source activity and mobility must be estimated for each receiver surface in contact.

It is unlikely that receiver mobilities will be measured prior to predicting the structure-borne power from machines to be installed. However, early and often acceptable estimates of receiver effective mobility are possible, for homogeneous plates, based on the characteristic plate mobility (the assumed mobility of an infinite plate of the same thickness and material as the real plate). For the case of ribbed plates, the receiver effective mobility can be estimated based on both characteristic plate and beam mobilities.

So far, perpendicular force excitations have been considered and it remains to include the effect of other components of excitation, particularly moments, and this should be possible by reference to the expanded form of the effective mobility. In addition, the receiving structures so far considered are damped, either existing (e.g., concrete floors bonded into walls or timber-frame walls) or added (e.g., through the edge support of the thick reception plate). The method has yet to be applied to low-loss systems, with highly reactive receivers.

The approach attempts to address the need for a structure-borne characterization of use to practitioners, who work with third octave band magnitudes (or octave values) and for test houses and small R&D facilities, which are geared to measure spatial and spectral averages. The approach contains several significant assumptions concerning the spatial variation of source and receiver mobility and of the contact forces and the concept of the single equivalent value become increasingly tenuous with increase in spatial variation. A single contact power may dominate the total emission from the machine and further work is required on the statistics of power distributions over contacts.

As with most simplified methods, the approach is a trade-off between practical application and accuracy. Indeed this is the case of many standard methods, such as for measurement of sound power, sound absorption, etc., that are based on fairly sweeping assumptions which are not met in many cases. However, they continue to play an important role in acoustics and noise control.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of Moritz Späh, Andreas Mayr, Jochen Scheck, Thomas Alber, and Heinz-Martin Fischer of Stuttgart University of Applied Science, and Andrew Moorhouse, Tomos Evans, and Andrew Elliot of Salford University, UK. The authors also acknowledge the financial support of the Engineering and Physical Sciences Research Council of the U.K.

[1]H. Bodén, "Characterization of fluid-borne and structure-borne sound sources," *Ninth International Congress on Sound and Vibration*, Orlando, FL. (2002), pp. 1–30.
[2]B. A. T. Petersson and B. M. Gibbs, "Towards a structure-borne sound source characterization," Appl. Acoust. **61**, 325–343 (2000).
[3]A. T. Moorhouse, "On the characteristic power of structure-borne sound sources," J. Sound Vib. **248**(3), 441–459 (2001).
[4]T. ten Wolde and T. G. R. Gadefelt, "Development of standard methods for structure-borne sound emission," Noise Control Eng. J. **28**, 5–14 (1987). Note Errata in **28**, 84 (1987).
[5]J. M. Mondot and B. A. T. Petersson, "Characterization of structure-borne sound sources: The source descriptor and coupling function," J. Sound Vib. **114**(3), 507–518 (1987).
[6]J. W. Verheij, "Inverse and reciprocal methods for machinery noise source characterization and sound path quantification," Int. J. Acoust. Vib. **2**(1), 11–20 (1997).
[7]B. M. Gibbs and A. T. Moorhouse, "Case studies of machine bases as structure-borne sound sources in buildings," Int. J. Acoust. Vib. **4**(3), 125–

133 (1999).
[8]B. M. Gibbs, N. Qi, and A. T. Moorhouse, "A practical characterization for vibro-acoustic sources in buildings," Acta. Acust. Acust. **93**, 84–93 (2007).
[9]ISO 9611: 1996 "Characterization of sources of structure-borne sound with respect to sound radiation from connected structures—Measurement of velocity at the contact points of machinery when resiliently mounted."
[10]A. Elliot, A. T. Moorhouse, and G. Pavic, "Characterization of a structure-borne sound source using independent and *in situ* measurement," *Proc. International Congress on Acoustics*, Madrid (2007).
[11]L. Cremer, M. Heckl, and B. A. T. Petersson, *Structure-Borne Sound*, 3rd ed. (Springer, New York, 2005).
[12]B. A. T. Petersson and J. Plunt, "On effective mobilities in the prediction of structure-borne sound transmission between a source and a receiving structure, Part 1: Theoretical background and basic experimental studies," J. Sound Vib. **82**(4), 517–529 (1982).
[13]B. A. T. Petersson and J. Plunt, "On effective mobilities in the prediction of structure-borne sound transmission between a source and a receiving structure, Part 2: Estimation of mobilities," J. Sound Vib. **82**(4), 531–540 (1982).
[14]S. H. Yap and B. M. Gibbs, "Structure-borne sound transmission from machines in buildings, part 2: Indirect measurement of force and moment at the machine-receiver interface of a single point connected system by a reciprocal method," J. Sound Vib. **222**(1), 99–113 (1999).
[15]J. Scheck, H.-M. Fischer, and B. M. Gibbs, "Direct and indirect methods to assess the structure-borne power transmission into receiving structures," *Proc. International Congress on Acoustics*, Madrid, 2007.
[16]R. A. Fulford and B. M. Gibbs, "Structure-borne sound power and source characterization in multi-point-connected systems, Part 1: Case studies for assumed force distributions," J. Sound Vib. **204**(4), 659–677 (1997).
[17]M. Späh, B. M. Gibbs, and H.-M. Fischer, "Measurement of structure-borne sound power of mechanical installations in buildings," *11th International Congress on Sound and Vibration*, St. Petersburg (2004).
[18]M. Späh, H.-M. Fischer, and B. M. Gibbs, "New laboratory for the measurement of structure-borne sound power of sanitary installations," *Forum Acusticum*, Budapest, 1907–1912 (2005).
[19]J. Lu, B. Louvigne, J. B. Pascal, and J. Tourret, "The perforated reception plate; a practical method for the characterization of structure-borne noise emitted by small equipment," *Proc. Internoise 90*, Gothenburg, Sweden, **1**, 217–220 (1990).
[20]M. Ohlrich, L. Friis, S. Aatola, A. Lehtovaara, M. Martikainen, and O. Nuutila, "Round Robin test of techniques for characterizing the structure-borne sound source strength of vibrating machines," *Euronoise 2006*, Tampere, Finland (2006).
[21]H.-Y. Lai, "Alternative test methods for measuring structure-borne sound power," *Inter-Noise 2006*, Honolulu, Hawaii (2006).
[22]R. Cookson and N. Qi, "A reception plate method of measurement of the free velocity of machines in buildings," *13th International Congress on Sound and Vibration*, Vienna (2006).
[23]E. B. Davis, "Characterization of structure-borne noise sources using a reverberant or anechoic plate," *Proc. Inter-Noise 2006*, Honolulu, Hawaii (2006).
[24]P. Gardonio and M. J. Brennan, "Mobility and impedance methods in structural dynamics," in *Advanced Applications in Acoustics, Noise and Vibration*, edited by F. Fahy and J. Walker (Spon Press, London, 2004), Chap. 9.
[25]A. Mayr and B. M. Gibbs, "On the use of an equivalent receiver mobility in lightweight buildings," *Proc. International Congress on Acoustics*, Madrid (2007).

A finite difference analysis of the field present behind an acoustically impenetrable two-layer barrier

Andrew M. Hurrell[a)]
Precision Acoustics Ltd., Hampton Farm Business Park, Higher Bockhampton, Dorchester, Dorset DT2 8QH, United Kingdom

(Received 22 November 2007; revised 27 March 2008; accepted 28 March 2008)

The interaction of an incident sound wave with an acoustically impenetrable two-layer barrier is considered. Of particular interest is the presence of several acoustic wave components in the shadow region of this barrier. A finite difference model capable of simulating this geometry is validated by comparison to the analytical solution for an idealized, hard-soft barrier. A panel comprising a high air-content closed cell foam backed with an elastic (metal) back plate is then examined. The insertion loss of this panel was found to exceed the dynamic range of the measurement system and was thus acoustically impenetrable. Experimental results from such a panel are shown to contain artifacts not present in the diffraction solution, when acoustic waves are incident upon the soft surface. A finite difference analysis of this experimental configuration replicates the presence of the additional field components. Furthermore, the simulated results allow the additional components to be identified as arising from the S_0 and A_0 Lamb modes traveling in the elastic plate. These Lamb mode artifacts are not found to be present in the shadow region when the acoustic waves are incident upon the elastic surface. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912437]

PACS number(s): 43.40.Dx, 43.20.Gp, 43.20.El [KGF] Pages: 4210–4217

I. INTRODUCTION

The process of characterizing the material properties of underwater acoustic panels can often be influenced by undesirable edge effects. The most common example of such behavior is diffraction around the edge of the panel. This problem is well known and can be a significant cause of experimental uncertainty. Picquette[1] proposed a method of eliminating diffraction artifacts in the context of a measurement of reflection coefficient. Subsequently the same author was then able to make direct measurements of edge diffraction effects.[2] However, diffracted field components are not the only artifacts to be seen in the shadow region behind a panel, particularly if it has a multilayer construction. Experimental data demonstrating two other shadow region artifacts, in addition to the diffracted component, will be presented later in this paper.

It could be argued that if multilayer panels are problematic, then it would seem obvious to only make measurements on simple homogeneous samples. However, this is not practicable for the following reason; the majority of underwater acoustic materials are applied to an object to modify either the reflection from or transmission through that object. The combination of coating and substrate behaves as an acoustic transmission line, and therefore the acoustic properties of the backing (typically a metal) need to be included in the coating design process. Clearly then, the measurements also need to take into account the substrate on which the coating is placed, and thus samples are routinely mounted on steel or aluminum backing panels. This gives rise to a scenario where the front surface of the panel is polymeric and thus compliant, while the rear surface is metallic and thus of much higher acoustic impedance. Panels of this nature are often referred to as hard-soft barriers.

Analytical solutions for diffraction by a hard-soft barrier exist, but only under the assumption that the diffracting object behaves like a half-plane with of infinitesimal width, with one perfectly soft and one perfectly hard surfaces; this is described as an ideal hard-soft barrier. Although this simplified geometry has been addressed by Rawlins,[3] this paper will follow the analysis of Kendig and Hayek.[4] However, to consider the generalized two-layer problem, a more comprehensive solution is required, preferably something that can take into account for the finite width of the barrier and the nonideal nature of the materials used. This paper begins by giving an overview of a finite difference model of acoustic propagation. The validity of this model will be confirmed by comparing results obtained from it with the analytical solutions to diffraction by idealized hard-soft half-planes.

The results of an experimental investigation of the shadow region components behind an acoustically impenetrable hard-soft barrier will then be presented. The numerical model will then be used to simulate the same geometry. Insight provided by the simulation then informs the analysis of problem, and the causes of some of the additional shadow region components are identified.

II. A STAGGERED GRID FINITE DIFFERENCE MODEL

The numerical model used in this work is the AFIDS (acoustoelastic finite difference software) suite distributed by AMH Consulting (www.amh-consulting.co.uk). There now follows a brief description of underlying equations and the method that they are implemented within AFIDS.

[a)]Electronic mail: andrew@acoustics.co.uk. URL: www.acoustics.co.uk.

 0001-4966/2008/123(6)/4210/8/$23.00

A. Underlying theory

Propagation within fluid media can be considered as being a simplification to the full elastodynamic case. The following derivation follows the method of Kolsky[5] and can be found in more detail in Hurrell.[6] In its simplest form, the propagation of acoustic waves in elastic media can be reduced to two equations,

$$F = \tau \cdot A = m \cdot \frac{d^2x}{dt^2} \tag{1}$$

and

$$\tau = c:S, \tag{2}$$

where F and x are vectors of length of 3 for force and particle displacement, τ and S are vectors of length of 6 for stress and strain, c is a 6×6 matrix representing the elastic constants, A is area, and m is mass. By assuming isotropic elastic media, the strain variables of Eq. (2) can be simplified to terms involving spatial derivatives of particle velocity. This also suggests that Eq. (1) could be recast to use particle velocity rather than displacement. Restricting the problem to two dimensions produces a concise set of five coupled equations, such that

$$\frac{\partial v_x}{\partial t} = \frac{1}{\rho}\left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y}\right), \tag{3}$$

$$\frac{\partial v_y}{\partial t} = \frac{1}{\rho}\left(\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y}\right) \tag{4}$$

are derived from Eq. (1), while

$$\frac{\partial S_{xx}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_x}{\partial x} + \lambda\frac{\partial v_y}{\partial y}, \tag{5}$$

$$\frac{\partial S_{yy}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_y}{\partial y} + \lambda\frac{\partial v_x}{\partial x}, \tag{6}$$

$$\frac{\partial S_{xy}}{\partial t} = \mu\left(\frac{\partial v_x}{\partial y} + \lambda\frac{\partial v_y}{\partial x}\right) \tag{7}$$

are derived from Eq. (2). Throughout Eqs. (3)–(7), λ and μ are the Lamé coefficients, ρ is density, and the subscripts indicate the direction in which the stress or velocity acts. This set of five equations is noteworthy for several reasons.

(1) All equations have a time derivative on the left-hand side so a time-marching algorithm can be used to solve them.
(2) Expressions for particle velocities are dependent only on spatial derivatives of stress and vice versa. Therefore, calculations of stresses and velocities are completely decoupled.
(3) Densities are the only material parameter used in the calculation of velocities, while only the Lamé constants are used in the calculation of stresses. Furthermore, nowhere within the formulation is there a dependence on the derivatives of material parameters.
(4) If the medium of propagation is a fluid $\mu=0$, and Eqs. (3)–(6) can be condensed to form the familiar acoustic wave equation, while Eq. (7) is zero.

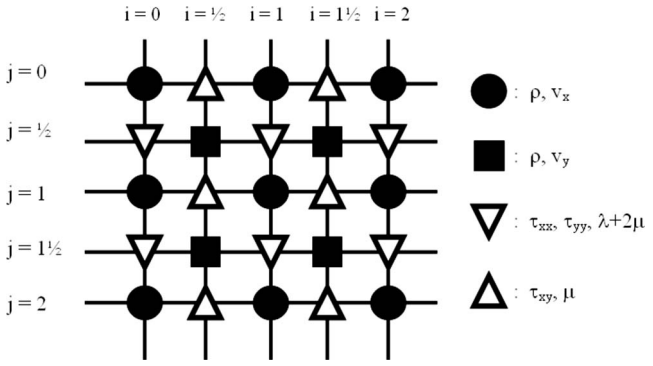


FIG. 1. The spatial staggering of finite difference grids.

B. Finite difference approximations

Numerical calculation of the derivatives within Eqs. (3)–(7) is accomplished by the rearrangement of a truncated Taylor series expansion to produce a finite difference (FD) approximation. This approach is discussed in detail by Virieux[7] and Levander.[8] FD solutions are commonly[9,10] applied by overlaying a single grid onto the computational domain to obtain the discrete nodes at which the values of the field variables are calculated. However, as discussed by Dablain,[11] single grid models suffer from a noticeable limitation—they tend to exhibit significant numerical dispersion as well as anisotropic propagation speed with faster propagation in directions aligned with the grid. To overcome this limitation, a staggered grid formulation was proposed by Madariaga[12] and then enhanced by Virieux[7] and Levander.[8] Interestingly, all three of these publications originate within the geophysics community. Geophysical models often have to cater for highly heterogeneous simulations involving both fluid and elastic or viscoelastic media and are thus also likely to meet the requirements for underwater acoustic modeling. In fact, the only difference between the requirements of the two communities relates to wavelength, with seismic events needing a greater length scale due to their much lower frequency.

The staggered grid model consists of four grids upon which five field variables (three stresses and two particle velocities) are calculated. Each of these grids is spatially offset from one another by half a grid spacing, such that all grids are interleaved as shown in Fig. 1. This provides a convenient method to exploit the decoupled nature of the fundamental equations. The FD method used in this paper computes spatial derivatives by using the fourth order accurate approximations of Levander.[8] Temporal derivatives are computed with second order accurate approximations in a time-marching fashion. As discussed by Levander,[8] the staggering of grids both spatially and temporally produces a model that is numerically stable for a wide range of different media while also being much less prone to numerical dispersion and anisotropy. Additionally, the five variable formulation is now a heterogeneous solution with the same equations applying everywhere within the computational domain. Boundaries between one media and another are simply catered for by changing the value of the material parameters at nodes on either side of the boundary.

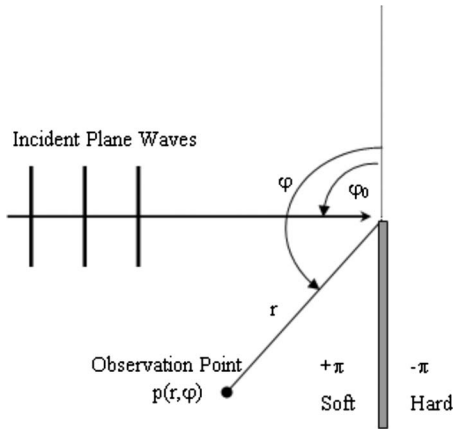


FIG. 2. Geometry of hard-soft barrier.

III. AN IDEAL HARD-SOFT BARRIER

A. Analytical solution

This solution relies upon the ideal hard-soft barrier assumptions described in the Introduction. Kendig and Hayek[4] drew comparisons with results obtained for the hard-hard (i.e., perfectly rigid) and soft-soft (i.e., perfectly compliant) barriers and concluded that the diffraction characteristics of the hard-soft barrier are strongly influenced by the boundary conditions of the face bounding the appropriate half space. They noted that the behavior tends toward that of the perfectly rigid barrier in the half space adjacent to the hard face, yet has behavior more akin to the perfectly compliant barrier in the half space next to the soft face. The geometry for Kendig and Hayek's solution is presented in Fig. 2.

It is found that the pressure amplitude of the diffracted field components p_d can be expressed as

$$p_d = \frac{p_0 e^{-ikr}}{\sqrt{2}}\left[\sin\left(\frac{t_1}{4}\right)\mathrm{Sgn}(c_1)e^{ic_1^2}F(|c_1|) + \cos\left(\frac{t_2}{4}\right)\mathrm{Sgn}(c_2)e^{ic_2^2}F(|c_2|)\right], \tag{8}$$

where p_0 is the incident acoustic pressure, k is the wavenumber, r is the distance to the observation point, and $\mathrm{Sgn}(x)$ is the sign function. The function $F(x)$, defined in terms of the Fresnel integrals $C(x)$ and $S(x)$, is

$$F(x) = 1 - \sqrt{2}e^{i\pi/4}[C(x) - iS(x)]. \tag{9}$$

It should also be noted that in Eq. (8) the c and t values are given by

$$c_{1,2} = \sqrt{2kr}\cos\frac{t_{1,2}}{2}, \quad t_1 = \phi + \phi_0, \quad t_2 = \phi - \phi_0. \tag{10}$$

Calculation of the total acoustic field can be obtained by an appropriate combination of the diffracted field with plane wave components incident upon and reflected from the hard-soft barrier. Following Kendig and Hayek with the necessary corrections, this is achieved by splitting the acoustic field into three regions, such that

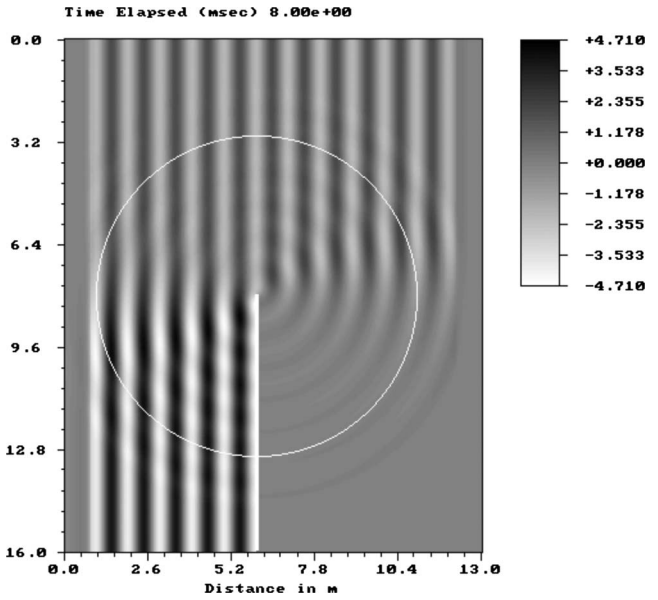


FIG. 3. Snapshot of pressure field around ideal hard-soft barrier (black =compression; white=rarefaction). The white circle indicates the locus of points from which time history data have been acquired.

$$\begin{aligned} p &= p_d + (p_i - p_r)H(\phi_0), \quad \text{I: } \pi - |\phi_0| < \phi < \pi \\ &= p_d + p_i, \quad \text{II: } |\phi_0| - \pi < \phi < \pi - |\phi_0| \\ &= p_d + (p_i + p_r)H(-\phi_0), \quad \text{III: } -\pi < \phi < |\phi_0| - \pi. \end{aligned} \tag{11}$$

Within Eq. (11), the function $H(\,)$ is the Heaviside step function. The notation of Eq. (11) caters for both possible geometries of the hard-soft barrier; specifically, when ϕ_0 is positive, plane waves are incident upon the soft surface, whereas when ϕ_0 is negative, they are incident upon the hard surface. For the example shown within this work, the wavenumber and the observer distance were chosen such that $kr=10$.

B. Numerical solution

The mesh generation program within the AFIDS suite was used to prepare a simulation with the same geometry, as shown in Fig. 2. The computational grid contained 325 $\times$400 grid nodes at a spatial increment of 4 mm and a temporal increment of 10 s. The incident harmonic plane waves had a frequency of 1480 Hz, resulting in a wavelength of 1 m in water and a wave number of 2. Consequently, all output points were placed on a circle (radius of 5 m) centered on the tip of the half-plane to be consistent with kr =10 (used above). A snapshot in time of the acoustic field after 8 ms can be seen in Fig. 3.

To prevent undesirable reflections from the computational boundary affecting the simulation, an absorbing boundary condition known as a perfectly matched layer (PML) was placed upon the bottom and right boundaries. PML boundary conditions were first proposed by Berenger[13] and have been shown to provide a very effective means of absorbing unwanted reflections. Collino and Tsogka[14] provide an excellent example of how to apply the PML to elastodynamic wave modeling.

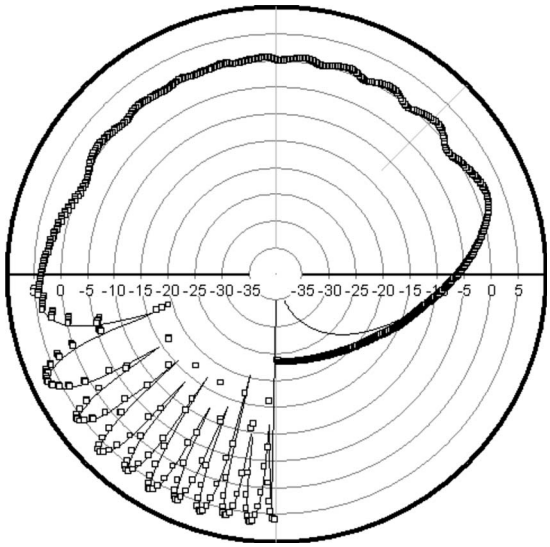


FIG. 4. Comparison of analytical and numerical field amplitudes around an ideal hard-soft barrier with sound incident upon the hard surface.

A comparison of the analytical and FD predictions for the diffracted field produced by the hard-soft barrier, when plane waves are incident onto the hard surface, can be found in Fig. 4. A similar comparison for incidence onto the soft surface is shown in Fig. 5. In both of these figures, the direction of the incident plane waves and the position of the hard-soft barrier mimic that of Fig. 2.

The overall shape of the FD predictions (particularly the complex maxima and minima on the incident side of the barrier) is in good agreement with the analytical predictions. However, in the region immediately behind the panel in Fig. 4, there is some deviation between the two curves, and this probably arises because (1) the FD simulation has a minimum object width of 4 mm (one grid node), whereas the analytical solution assumes the barrier infinitesimal, and (2) the FD simulation can only approximate ideal materials (i.e., it cannot use a material that is infinitely compliant or infinitely rigid).

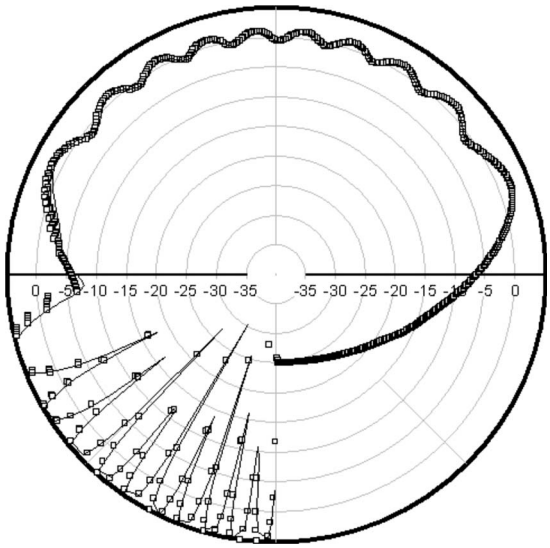


FIG. 5. Comparison of analytic and numerical field amplitudes around an ideal hard-soft barrier with sound incident upon the soft surface.

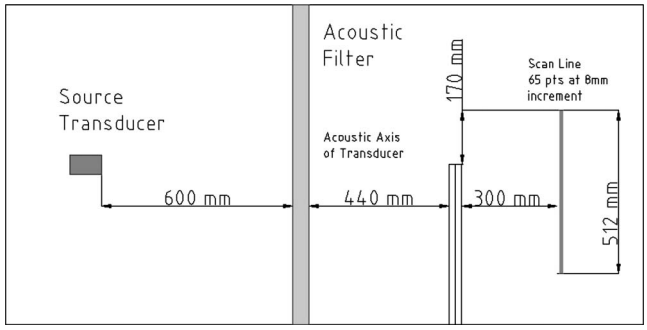


FIG. 6. Comparison of analytic and numerical field amplitudes around an ideal hard-soft barrier with sound incident upon the hard surface.

This latter issue is particularly important since an infinitely soft material would act as a true pressure release and the acoustic pressure at the rear surface of the panel would tend to zero. However, in the numerical case, it is only possible to produce an approximation to an ideally soft material, and thus while the field values become small, they do not tend to zero.

IV. EXPERIMENTAL MEASUREMENTS WITH A TWO-LAYER BARRIER

The formula proposed by Kendig and Hayek is a good starting point for the two-layer boundary problem, however, it relies upon ideal materials and only accounts for diffraction. The author was aware of anecdotal evidence that hard-soft barriers constructed from real materials behave in a somewhat different manner, and it was decided to conduct an experimental investigation of this matter.

An acoustic test panel comprising a layer of Plastazote firmly stuck to a layer of 10-mm-thick aluminum was prepared. Plastazone (Zotefoams, Croydon, UK; www.zotefoams.com) is a very high air content, closed cell cross-linked polyethylene foam. Due to its high void content, this highly compliant material has a very low density, very low acoustic wavespeed, and very high acoustic absorption. For the frequency range used in these experiments, it provides a good approximation to an idealized pressure release material. Furthermore, its high air content also ensures that the impedance mismatch between it and the surrounding water is sufficiently high to prevent any acoustic energy from propagating through the foam layer. The insertion loss (defined as the reduction in amplitude of signal measured by a receiver as a consequence of the inserting the panel between source and receiver) of a 10 mm layer of Plastazote was found to be in excess of the dynamic range of the measurement system (69 dB). The aluminum layer, however, is not an ideally rigid material and exhibits elastic behavior. The test panel was immersed in a water tank according to the geometry described in Fig. 6.

The source transducer was driven, via an ENI 2100L power amplifier, with a 983 kHz sinusoidal carrier (provided by a Philips PM5134 function generator) modulated by a 55 kHz raised cosine bell or haversine (provided by a Philips PM5133 function generator). The source transducer is driven in this manner so that it behaves as a parametric array, as described by Moffet and Mellen.[15] Nonlinear conversion

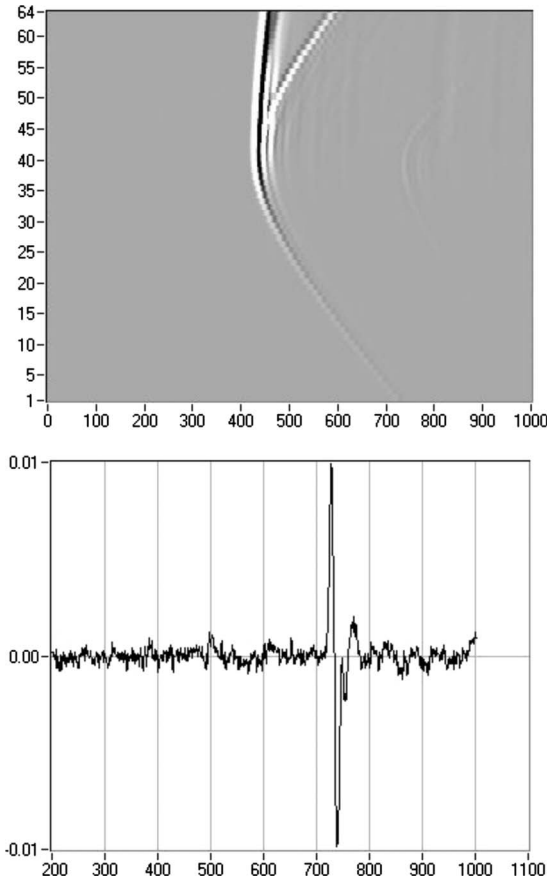


FIG. 7. Measured acoustic field behind an elastic-soft barrier, elastic face toward acoustic source. Upper plot; grayscale representation of complete acoustic field, with measurement location recorded on vertical axis (black =compression; white=rarefaction). Lower plot: detailed time trace from measurement location 1, with hydrophone amplitude in volts on vertical axis. Both plots have arbitrary time units along horizontal axis.

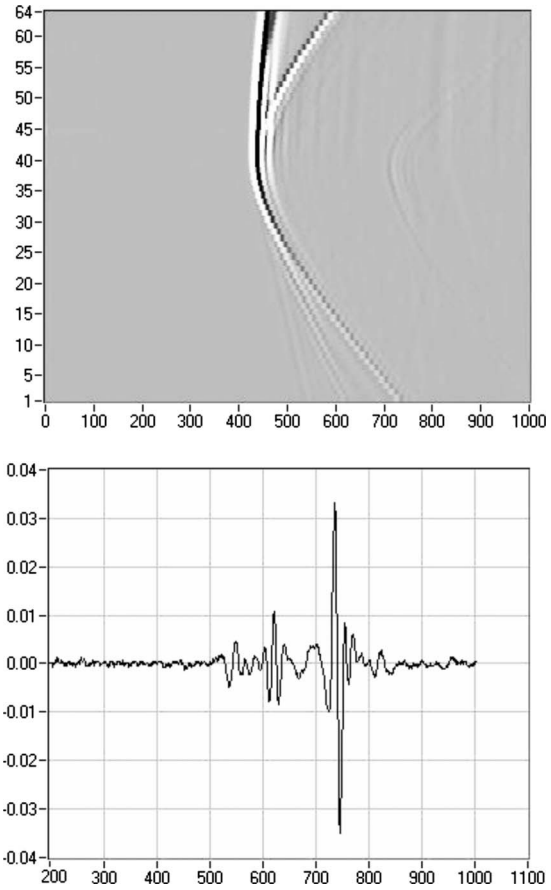


FIG. 8. Measured acoustic field behind an elastic-soft barrier, soft face toward acoustic source. Upper plot; grayscale representation of complete acoustic field, with measurement location recorded on vertical axis(black =compression; white=rarefaction). Lower plot: detailed time trace from measurement location 1 (to give maximum temporal separation of the three pulses) with hydrophone amplitude in volts on vertical axis. Both plots have arbitrary time units along horizontal axis.

within the water results in an acoustic field that has the frequency content that is related to twice the modulating waveform. However, this configuration exhibits directional characteristics related to the transducer's radiation pattern when driven at the frequency of the carrier signal.

Measurements of the acoustic field were made by using a Brüel and Kjær 8103 hydrophone (S/N 1176386) that has a sensitivity of −212.5 dB (1 μV/Pa) over the frequency range of 10–100 kHz. A Brookdeal 9452 precision preamplifier provided an additional 60 dB gain to the received signal before it was acquired by using a LeCroy 9304C 200 MHz digital storage oscilloscope. All of the acquired data were then stored on a personal computer for further postprocessing. Data were acquired in this manner from a range of positions in the acoustic field by means of a stepper motor controlled hydrophone positioning system. A single computer program controlled the positioning and acquisition hardware for fully automated data capture.

An acoustic filter was used in order to prevent significant carrier signal being recorded by the hydrophone. The properties of this filter are such that at 100 kHz (the center frequency of the propagating low frequency component), the insertion loss of the filter is approximately 2 dB. In contrast, the filter's insertion loss at 983 kHz (the carrier frequency) is in excess of 38 dB. Consequently, there is an insertion loss differential of at least 36 dB between low frequencies and the carrier, and the acoustic field beyond the filter is dominated by the low frequency components, with the carrier signal heavily suppressed.

As shown in Fig. 6, the hydrophone was mechanically positioned at 65 locations along a scan line behind and parallel to the back of the test panel. The step distance between each location was 8 mm, resulting in a total scan distance of 512 mm, arranged such that 2/3 of the scan was in the "shadow" region of the test panel, and 1/3 of the scan was unobstructed by the panel. This measurement was conducted for three cases: (1) with the elastic side of the test panel facing the incident wavefront; (2) with the soft side of the test panel facing the incident wavefront; (3)with no panel present.

As would be expected, for configurations 1 and 2 (where the acoustic test panel was present), there was a small but measurable acoustic signal in the shadow region behind the panel. Since a pulsed acoustic waveform was used, it was possible to determine the time of flight, and a comparison was made between the arrival time of the pulse in the shadow region with that of a pulse at the same location without the test panel. This indicated that the shadow region pulse was delayed with respect to a direct path, and the

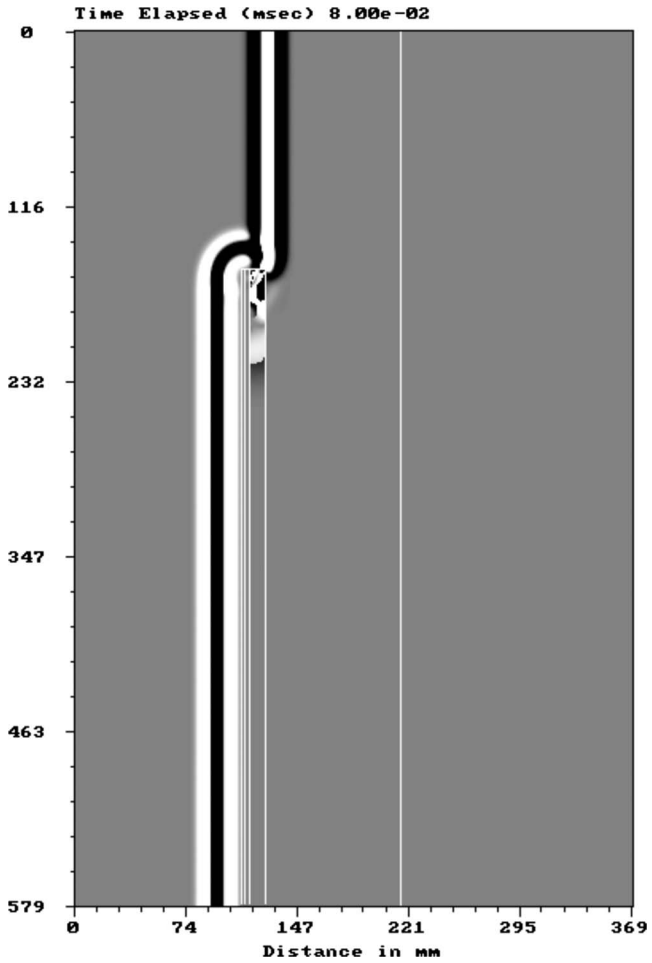


FIG. 9. Pressure field snapshot as the incident plane wave is just beginning to interact with a two-layer barrier.

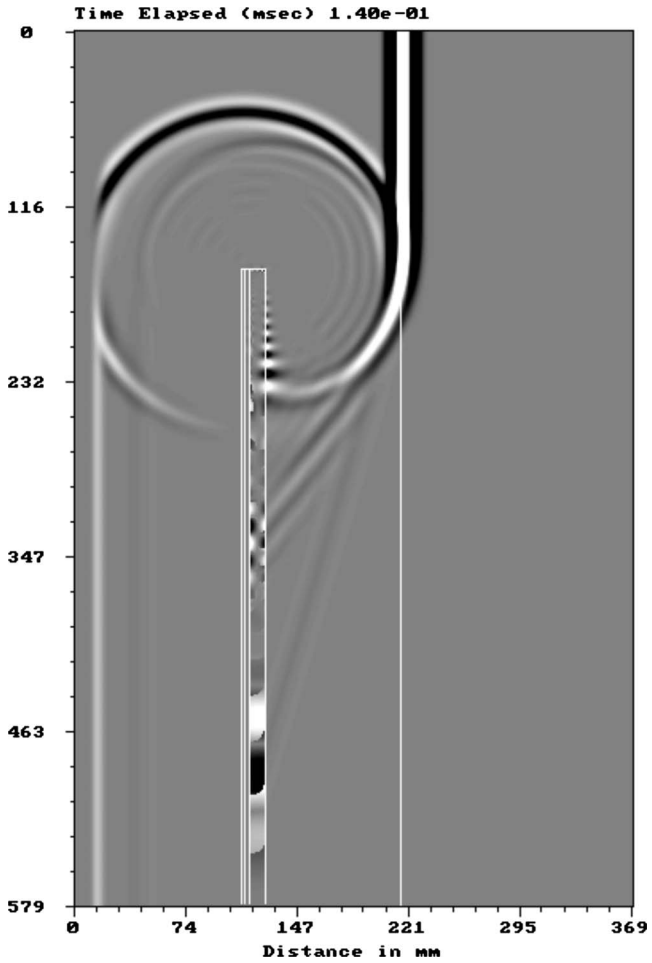


FIG. 10. Pressure field snapshot 60 ms after the incident plane wave encountered the two-layer barrier.

simple calculations revealed the delay to correspond to the additional acoustic path caused by diffraction around the edge of the test panel; this was the classical diffracted pulse. Figure 7 displays the experimental results from measurements by using configuration 1.

Figure 8 displays the similar information as Fig. 7, but for the second configuration. An important feature of the type 2 measurements is that two additional waveforms were seen in the shadow region. Both of these pulses were of lower amplitude than the diffracted pulse, but both arrived earlier than the diffracted signal. Interestingly, these extra pulses are seen to separate in time as the distance from the edge of the panel increased.

Another important difference between the two measurement configurations is that the diffracted pulse is more than three times larger when the elastic layer is at the rear of the panel (configuration 2). When the rear surface of the panel is a soft material, the diffracted wave amplitude is suppressed by the pressure release behavior. In the presence of an elastic material, the diffracted amplitude suffers no such attenuation.

To determine the physical phenomenon responsible for the presence of these additional pulses, it is useful to consider them in detail. The issue of arrival times is of particular interest. There are only two possible explanations for the appearance of these pulses before the diffracted pulse. Either these pulses have traveled a shorter acoustic path than the diffracted pulse or they have propagated with a faster acoustic velocity. If the acoustic path was shorter than the standard diffracted path (via the edge of the panel onto the measurement location), it would be necessary for the pulse to have traveled through the panel at some point. Type 1 measurements showed no such pulses, even though the panel should have the same insertion loss regardless of orientation. For the purposes of this experiment, the insertion loss of the panel is sufficiently high to make transmission through the panel unlikely. Consequently, the pulses must have arrived by a "faster route." Furthermore, the fact that the two additional pulses appear to separate with increasing distance from the panel edge would suggest that they are traveling at different acoustic velocities.

V. FD SIMULATIONS WITH A TWO-LAYER BARRIER

A FD simulation of the experimental configuration discussed in the previous section was prepared. The spatial and temporal increments used were 1.05 mm and 0.1 μs, respectively, and the simulation was modeled over a grid of 350 $\times$ 500 nodes for 3500 timesteps. An impulsive plane wave based on a Ricker wavelet of center frequency of 80 kHz was incident on the barrier from the left-hand edge. Fourier analysis indicates that this waveform has spectral content

similar to the pulse used in the experiment, with comparable pulse duration but slightly different waveform shape. The dimensions of the barrier were exactly as shown in Fig. 6, and the materials were Plastazote for the front face and aluminum for the backing panel. A simulation by using this geometry took 26 min to run on a Athlon XP 3000 PC and used 37 Mbits of memory. Figures 9 and 10 show two snapshots of the acoustic field observed as the simulation proceeds.

Figure 10 is of particular interest, since in addition to the circular diffracted wave, it also shows the existence of two propagating features within the aluminum plate. These waves appear to leak from the rear of the panel and are hence the source of the additional pulses under investigation. The fastest of these travels down the plate with compression and rarefaction phases that are relatively uniform across the width of the plate. This type of behavior would result in displacements of the aluminum surface that are symmetrical about the center line of the plate. In contrast, the features corresponding to the second wavefront are far from uniform across the width of the plate. In fact, wherever a region of compression occurs on one surface, a rarefaction can be seen on the other surface. This behavior is antisymmetric about the center line of the plate. Given that these waves are being guided within the elastic layer, it appears likely that these features are the lowest order symmetric (S_0) and antisymmetric (A_0) Lamb modes.

It is useful to consider the nature of Lamb waves as described by Viktorov,[16] starting with the S_0 mode. For normalized frequency $k_s \cdot d < 2$, where k_s is the shear wave number and d is the thickness of the panel, the phase (c_p) and group (c_g) velocities for the S_0 mode are given by

$$c_p = c_g = c_s\sqrt{3}, \tag{12}$$

where c_s is the shear velocity in the material. At the center frequency of 80 kHz in aluminum, the shear wave number is 161.57. Thus, for a plate of thickness of 10 mm, the normalized frequency is 1.61, and conditions for Eq. (12) are satisfied. The velocity of the S_0 mode calculated by Eq. (12) is found to be 5388 m s^{-1}. The expression for the low frequency limit of the A_0 mode velocity is

$$c_p = \sqrt{\omega d}\sqrt[4]{\frac{E}{3\rho(1-\nu)^2}}, \tag{13}$$

where E is the Young modulus and ν is Poissons ratio. Substitution of the values appropriate to this problem yields a phase velocity for the A_0 mode 3138 m s^{-1}.

As can be seen from Fig. 10, the two plane waves that leak from the aluminum plate propagate at different angles relative to it. Snell's law can then be used in conjunction with the velocities obtained from Eqs. (12) and (13) to predict a direction of propagation that can be compared to the modeled values. If the direction of the waves within the plate is assumed to be along the plate (i.e., at 90° to the normal), then Snell's law can be rewritten as

$$\frac{1480}{\sin\theta_2} = c_L, \tag{14}$$

where c_L is the velocity of the Lamb wave in the plate, θ_2 is the direction of propagation of the leaking Lamb wave in the water relative to the normal, and 1480 is the velocity of acoustic propagation in water expressed in m s^{-1}.

Considering first the S_0 mode, Eqs. (12) and (14) predict that the radiation should leak from the plate at an angle of 15.9° to the surface. The propagation angle for the simulated wavefronts was measured as 16.1. Similarly the A_0 mode direction of propagation predicted by Eqs. (13) and (14) is 28.15°, whereas the value of 28.9° is obtained from the FD simulation. Such a close similarity between predicted and modeled angles of propagation supports the hypothesis that Lamb waves leaking from the elastic panel are responsible for the two additional wavefronts within the shadow region.

It is also useful to compare the experimental time traces to those derived as pressure histories from within the numerical model. Figure 11 contains this sort of comparison from a measurement point that is at perpendicular distances of 100 mm from the rear of the barrier and 340 mm from the barrier edge.

As can be seen from Fig. 11, the two waveforms compare well to each other; three pulses are present, and the arrival time of each pulse is consistent between experimental and simulated results. However, there are a two minor differences that merit further discussion. The first point to note is the difference in relative amplitude of the three pulses. The diffracted pulse is somewhat larger than S_0 and A_0 pulses within the experimental data, whereas the difference is less in the modeled results. Currently, AFIDS is only capable of simulating propagation within elastic, fluid and quasiidealized media, and it is not capable of representing absorption or other loss mechanisms. However, the closed cell foam structure of Plastazote is very lossy, probably due to viscoelastic and porous losses, and will thus absorb a significant amount of the acoustic energy interacting with it. For this reason, the S_0 and A_0 mode amplitudes are likely to be lower in the experimental case than has been numerically predicted.

The second issue is that of pulse shape. The A_0 pulse in the FD trace appears to ring on for a full cycle longer than the same feature in the experimental trace. As has already been discussed, the source wavelet used in the FD simulation had a similar spectral content but different temporal waveshape. However, this is unlikely to be the sole reason for the difference. Once again, the issue of loss mechanisms needs to be considered. As modeled, the A_0 mode is free to ring on for a number of cycles. If a lossy material was in contact with the surface of the aluminum plate, such oscillations would be damped, and the subsequent wave leaking out from the plate would have fewer cycles. Future work to extend the capabilities of the AFIDS suite to include simulation of viscoelastic materials is underway. Following these enhancements, an even closer correlation between simulated and experimental results is expected.

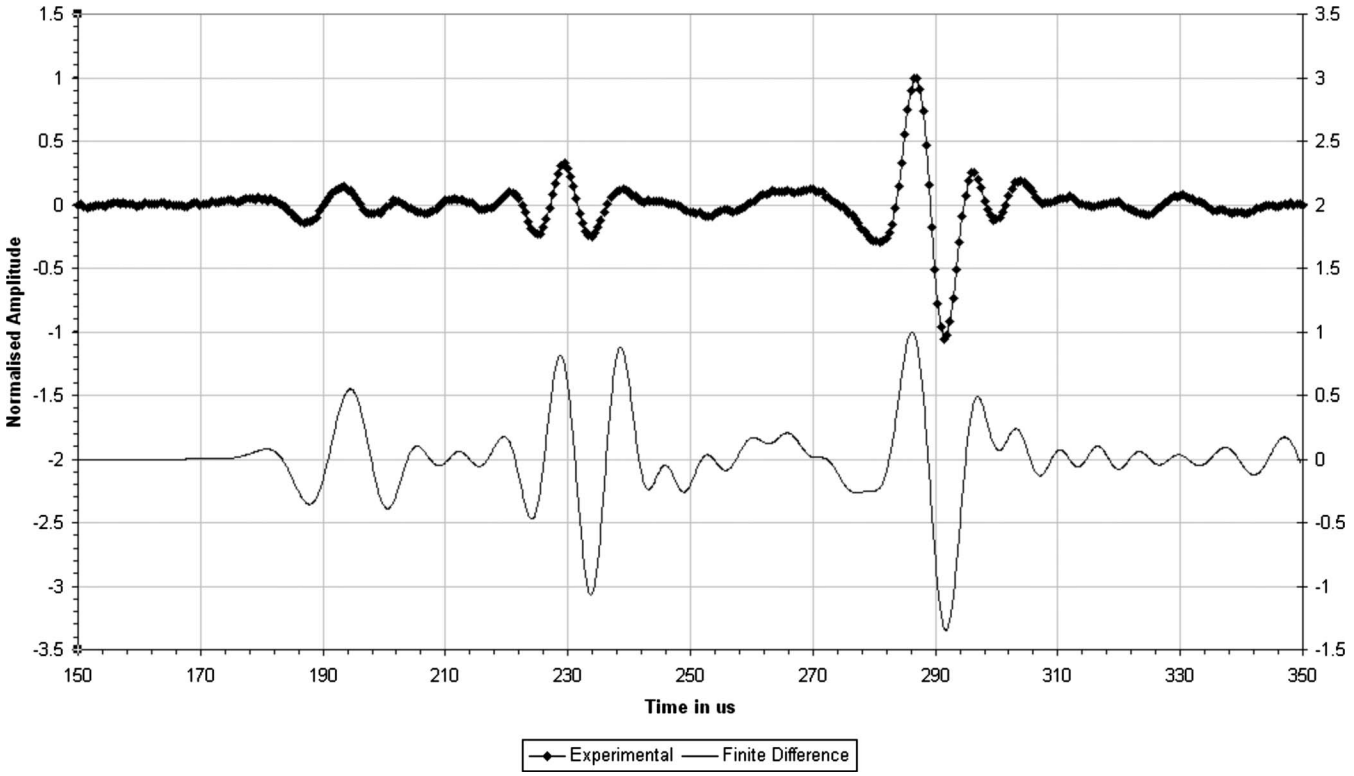


FIG. 11. Comparison of experimental and modeled acoustic waveforms in the shadow region of a two-layer barrier.

VI. CONCLUSIONS

The presence of wave components in the shadow region of an acoustically impenetrable two-layer barrier has been investigated. Initial investigation began with the analytical solution proposed by Kendig and Hayek for an idealized hard-soft barrier. This was then used to validate the results predicted by a FD model. An experimental investigation of the nonideal, but acoustically impenetrable two-layer barrier was then conducted. In addition to the main diffracted wave, two further wave artifacts were discovered in the shadow region of the barrier when an elastic layer was at the rear of the panel.

A numerical analysis of the same geometry also predicted the presence of these artifacts and was then used to identify them as being the first symmetric (S_0) and first antisymmetric (A_0) Lamb modes leaking back into the water. This identification process was confirmed by comparing the direction of propagation of the theoretical leaky Lamb waves to that obtained from the numerical model, wherein a very good agreement was found.

ACKNOWLEDGMENTS

The author is grateful to Professor V. F. Humphrey for many helpful discussions and to the directors of Precision Acoustics Ltd. for financial support during this work.

[1]J. C. Piquette, "An analytical technique for reducing the influence of edge diffraction in reflection measurements made on thin acoustical panels," J. Acoust. Soc. Am. **80**, 19–28 (1986).
[2]J. C. Piquette, "Direct measurements of edge diffraction from soft underwater acoustic panels," J. Acoust. Soc. Am. **95**, 3090–3099 (1994).
[3]A. D. Rawlins, "Diffraction of sound by a rigid screen with an absorbent edge," J. Sound Vib. **47**, 523–541 (1976).
[4]R. P. Kendig and S. I. Hayek, "Diffraction by a hard-soft barrier," J. Acoust. Soc. Am. **70**, 1156–1165 (1981).
[5]H. Kolsky, *Stress Waves in Solids* (Dover, New York 1963).
[6]A. M. Hurrell, "Finite difference modelling of acoustic propagation and its application in underwater acoustics," Ph.D. thesis, University of Bath, 2002.
[7]J. Virieux, "P-SV wave propagation in heterogeneous media: Velocity stress finite-difference method," Geophysics **51**, 889–901 (1986).
[8]A. R. Levander, "Fourth-order finite-difference p-sv seismograms," Geophysics **53**, 1425–1436 (1988).
[9]L. J. Bond, "A computer model of the interaction of acoustic surface waves with discontinuities," Ultrasonics **17**, 71–77 (1979).
[10]A. Harker, "Numerical modelling of the scattering of elastic waves in plates," J. Nondestruct. Eval. **4**, 89–106 (1984).
[11]G. A. Dablain, "The application of high-order differencing to the scalar wave equation," Geophysics **51**, 54–66 (1986).
[12]R. Madariaga, "Dynamics of an expanding circular fault," Bull. Seismol. Soc. Am. **66**, 639–666 (1976).
[13]J.-P. Berenger, "A perfectly matched layer for the absorption of electromagnetic waves," J. Comput. Phys. **114**, 185–200 (1994).
[14]F. Collino and C. Tsogka, "Application of the perfectly matched absorbing layer model to the linear elastodynamic problem in anisotropic heterogeneous media," Geophysics **66**, 294-–307 (2001).
[15]M. B. Moffett and R. H. Mellen, "Model for parametric acoustic sources," J. Acoust. Soc. Am. **61**, 325–337 (1977).
[16]I. A. Viktorov, *Lamb Waves* (Plenum, New York, 1967), Pt. 5, pp. 1–8.

Energy concentration at the center of large aspect ratio rectangular waveguides at high frequencies

F. B. Cegla
Department of Mechanical Engineering, Imperial College London, London SW7 2AZ, United Kingdom

(Received 24 October 2007; revised 21 February 2008; accepted 17 March 2008)

Waveguides in non-destructive evaluation (NDE) applications are commonly of a regular geometry (e.g., circular and ring cross section) for which analytical solutions exist. In this paper, wave propagation in infinitely long strips of large rectangular aspect ratio is discussed. Due to the finite width of strips, a large number of modes exist within the structure. This complicates the analysis and usually discourages the use of strip waveguides in NDE sensors. However, it is shown that among the many modes of a strip, there are some with very desirable properties. This is highlighted by the example of two guided wave modes of a large aspect ratio rectangular strip whose dispersion characteristics approach those of the fundamental modes of an infinitely wide plate at high frequencies. The energy of these modes concentrates in the central region of the strip and decays toward the edges so that the strip waveguide can easily be mechanically attached to other components without influencing the wave propagation. Dispersion curves and mode shapes were derived by using a semianalytical finite element technique and are presented over a range of frequencies. It is shown that selective excitation of both modes is possible in practice and the experimental setup is described. [DOI: 10.1121/1.2908273]

PACS number(s): 43.40.Le, 43.20.Mv [RLW] Pages: 4218–4226

I. INTRODUCTION

The analysis of guided wave propagation in infinitely wide plates, rods, and pipes is well established[1–3] and has been made use of in numerous NDE applications such as large area inspection (e.g., Refs. 4 and 5), long range pipe inspection (e.g., Ref. 6), and material property measurements (e.g., Refs. 7 and 8) to mention only a few. Applications often make use of the low frequency regime, where the modal density is low and selective mode excitation is easily possible. Areas of high modal density such as at high frequency are usually avoided because of the need for more sophisticated transducers and narrower bandwidth of operation. In plates of finite width (strips), the modal density is high due to additional modes in the width direction. The conventional wisdom would suggest to avoid these geometries since excitation of selected modes would be more difficult. However, in certain applications, a strip geometry can offer an advantage over other geometries.

The author was interested in the use of waveguides to convey ultrasonic signals from a remote transducer to a component that is to be investigated. Literature on source characteristics of sources of different geometry and surface loading on half-spaces[9,10] showed that antiplane shear line sources are potentially an attractive way to maximize transmission into half-spaces. Since line sources do not exist in real life, large aspect ratio rectangular sources (width $\gg$thickness) are the closest practically implementable approximation. To transmit surface loads that are applied over a rectangular cross section onto components, the author analyzed wave propagation in infinitely long strip waveguides of large rectangular aspect ratio. During the analysis, the attention was focused on two modes that resemble the SH and A0 modes of a plate of infinite width at high frequencies. Their properties and experimental excitation are described in the following sections.

II. THEORY: WAVE PROPAGATION IN RECTANGULAR STRIPS AND THEIR DISPERSION CHARACTERISTICS

The modeling of wave propagation remote from the edges in very wide plates is accurately achieved by modeling plane wave propagation under plane strain conditions. Software packages that accurately model the wave propagation in plates under these conditions [e.g., DISPERSE (Ref. 11)] exist. These models work well in cases where the width (W) of the plate and propagation distance (L) are comparable and much larger than the thickness (T). As the propagation distance increases to be larger than the plate width, reflections from the sides start to appear in the received signals and it becomes necessary to deal with the three dimensional problem. This can be considered as the propagation along a rectangular strip with a constant cross section.

Mindlin and Fox[12] were the first to describe the propagating modes of a bar of rectangular cross section. Their solution was made up of a superposition of the flexural, longitudinal, and shear modes that propagate in infinitely wide plates of two different thicknesses. The thickness of the plates corresponded to the width and thickness of the rectangular bar. The solutions for the infinitely wide plates were rotated by 90° relative to each other and superposed in order to fulfill the boundary conditions of zero stress all around the perimeter of the cross section. This method enabled them to determine the propagating modes of the bar at distinct frequencies and aspect ratios of the bar, but a solution for all frequencies and aspect ratios was not possible. Also, Fraser[13] presented an analytical method to calculate the dispersion

 0001-4966/2008/123(6)/4218/9/$23.00

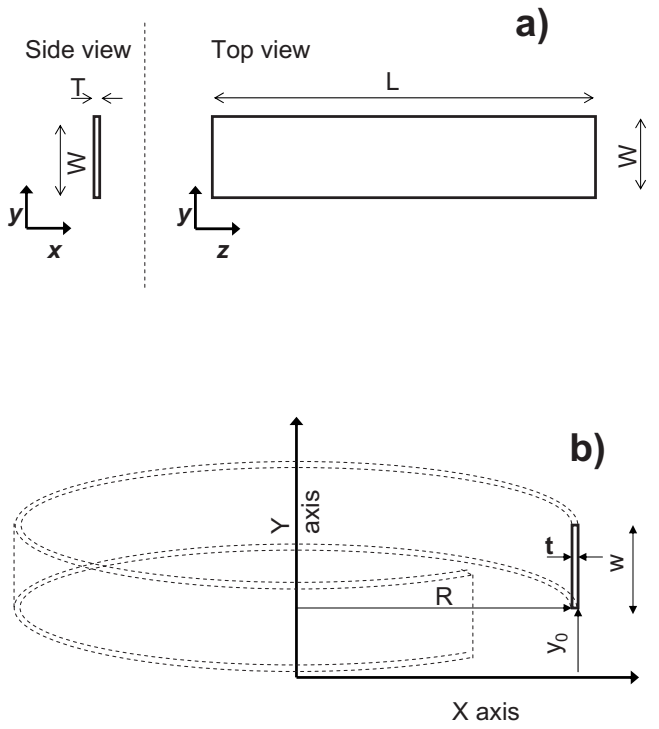


FIG. 1. Sketch (a) of the strip geometry considered in this paper and (b) of the geometry of the FE model used to obtain the dispersion curves of a strip of rectangular cross section (see text for dimensions).

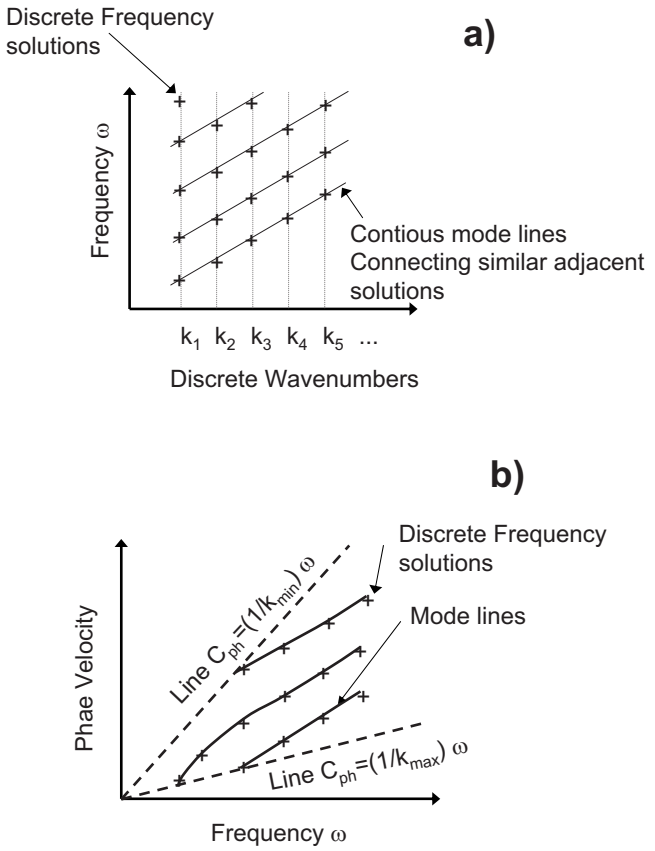


FIG. 2. (a) Sketch of the frequency–wave number results obtained from a FE eigensolver at different cyclic orders. (b) Sketch of the frequency–wave number results transformed into the phase velocity frequency domain.

curves for infinitely long rectangular bars accurately for a limited range of wave numbers. More recently, the continuous tracing of dispersion curves for wave propagation in structures of arbitrary cross section has become possible through the use of finite element (FE) eigensolvers. Wilcox *et al.*,[14] Mukdadi *et al.*,[15] Hayashi *et al.*,[16] Gavric,[17] Finnveden,[18] and others have reported methods of tracing dispersion curves for a range of different sections (L-shaped sections, rail heads, and rectangular strips). These techniques are now widely used and are often referred to as the semi-analytical finite element method. Predoi *et al.*[19] give an overview and a number of references about the development of finite element techniques for modeling of wave propagation in waveguides.

The method of Wilcox *et al.*[14] has been employed here to analyze the modes propagating in 1-mm-thick steel strips of a much larger width (>15 mm). The method works by defining an axisymmetric model with a very large radius compared to the dimensions of the cross section. The section of the axisymmetric body represents the cross section of the waveguide (see Fig. 1). Due to the very large radius, the structure approximates a straight waveguide. For the finite element eigensolver, a specific cyclic order can be specified. This specifies the number of wavelengths that exist around the circumference of the axisymmetric body. For example, a cyclic order of 1 corresponds to a wavelength equal to the circumference of the structure. For a cyclic order of 2, there are two wavelengths around the circumference and so on. Therefore, the wavelength of the solution is determined by the following equation:

$$\lambda = \frac{2\pi R}{C_{\text{order}}}, \tag{1}$$

where R is the radius of the model and C_{order} is the cyclic order of the FE eigensolver. At each cyclic order, the FE-eigensolver routine will determine several resonance frequencies, each frequency corresponding to a different mode. The wave number dispersion curves for the meshed cross section can now be determined by plotting the resonance frequencies against the wave number, which is determined by using Eq. (1) and the identity

$$k = \frac{2\pi}{\lambda}, \tag{2}$$

where k is the circular wave number and λ the wavelength. This method therefore yields a set of discrete frequency solutions at each wave number. A typical set of results is shown in Fig. 2(a). Wilcox *et al.*[14] developed a software that connects adjacent solution points to form a continuous solution line (mode) in the wave number frequency domain. The joining up of adjacent points to form a line is carried out by comparing the mode shapes of adjacent solutions and by using the slope of the curve of existing solutions to predict the continuation of the curve, which is similar to the method presented by Lowe.[3] The finite element software that determined the eigensolutions was the FINEL 77 code, which was developed by Hitchings[20] at Imperial College.

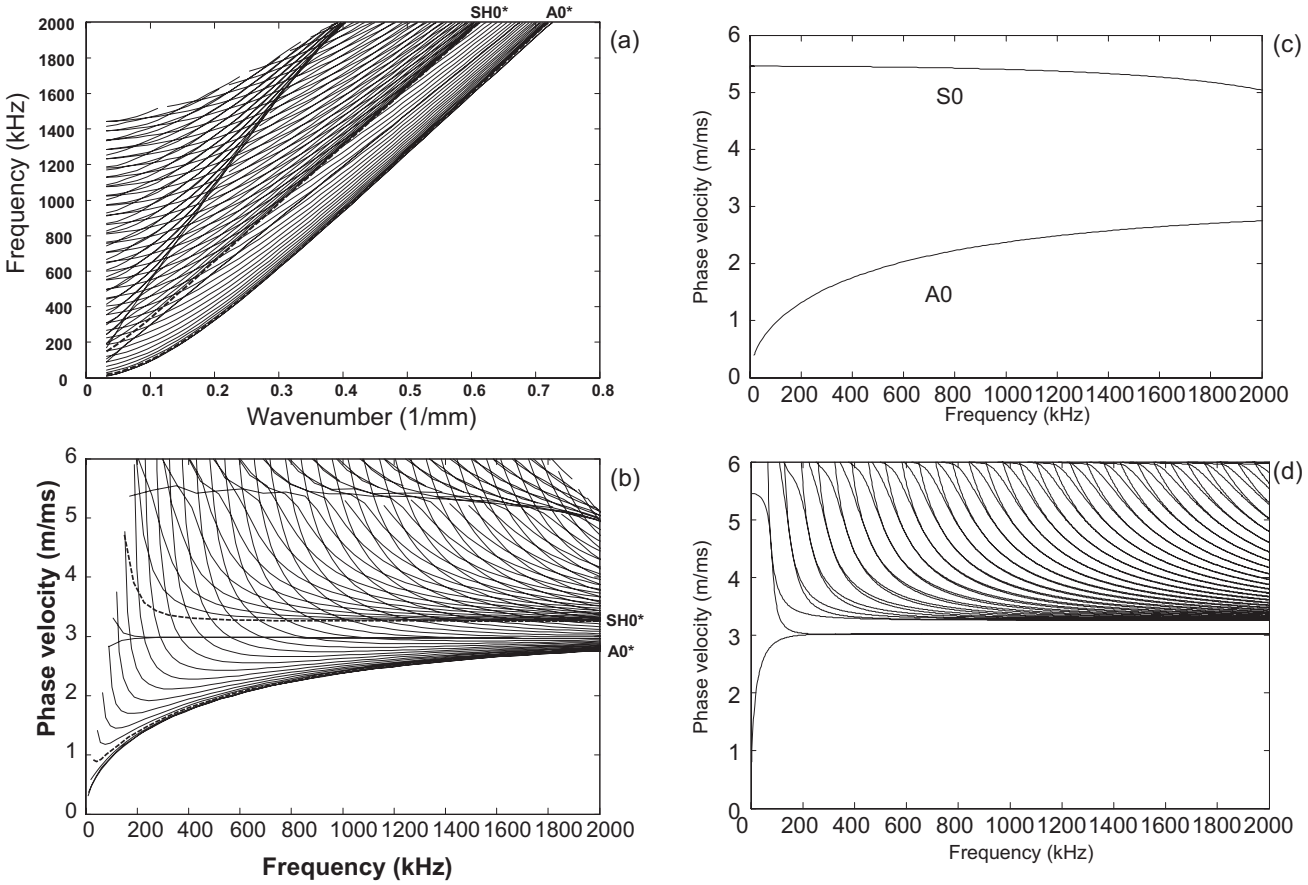


FIG. 3. Dispersion curves for a 1-mm thick and 30-mm wide rectangular steel strip determined by FEM (a) frequency–wave number and (b) phase velocity frequency. Two interesting modes that correspond to the lowest order shear horizontal mode (SH0*) and the lowest order flexural mode (A0*) that is symmetric with respect to its width are highlighted by the bold dashed (---) lines. For comparison, (c) and (d) show the phase velocity dispersion curve of a 1- and 30-mm-thick infinitely wide plate.

Phase velocity dispersion curves can be obtained from the wave number frequency plot by converting them using the following identity:

$$C_{\mathrm{ph}} = \frac{\omega}{k}. \tag{3}$$

Once the phase velocity is determined, the group velocity can be calculated by using

$$C_{\mathrm{gr}} = \frac{\partial \omega}{\partial k} = C_{\mathrm{ph}} + k\frac{\partial C_{\mathrm{ph}}}{\partial k}. \tag{4}$$

An interesting aspect to note about the technique is that the determined FE eigensolutions are confined to a rectangular domain in the frequency–wave number space. However, when this is transformed into a phase velocity–frequency space, the solutions will be bound in a space between the two lines $C_{\mathrm{ph}}=(1/k_{\mathrm{min}})\omega$ and $C_{\mathrm{ph}}=(1/k_{\mathrm{max}})\omega$. This is illustrated in Fig. 2(b). The mode shapes of each mode are a direct result of the FE analysis and can be extracted at each frequency.

The technique described was used to generate dispersion curves for rectangular strips. The results for a steel (ρ $=7932\ \mathrm{kg/m^3}$, $E=216.9$ GPa, $\nu=0.2865$, unless otherwise stated) strip geometry of 1 mm thickness and 30 mm width are presented here. The radius of curvature of the FE model was 2 m in order to ensure accurate results. Wilcox *et al.*[14] reported the appearance of discrepancies of the solution to the curved model and the analytical solution above frequencies of 4 MHz. Figures 3(a) and 3(b) show the frequency–wave number and phase velocity–frequency dispersion curves. For comparison, Figs. 3(c) and 3(d) show the phase velocity dispersion curves for a 1-mm-thick infinite plate (c) and a 30-mm-thick infinite plate (d) of the same material. The DISPERSE software[11] was used to trace the curves in Figs. 3(c) and 3(d). Many more modes with phase velocities below 3 m/ms exist at low frequencies in the strip than in the infinitely wide plate case; these extra modes are due to the finite width of the strip.

III. ANALYSIS OF THE SH0* AND A0* MODES

Two modes are highlighted in Fig. 3; These modes were named A0* and SH0* (the * here and for the remainder of the paper indicates a strip mode), after the well known fundamental A0 and SH0 plate modes because their high frequency dispersion characteristics tend toward those of these plate modes (assuming the smallest dimension to be the thickness). In addition to this, their mode shapes in the high frequency limit also are of the same polarization as the fundamental A0 and SH0 plate modes; however, they are concentrated at the center of the strip and their amplitudes decay toward the edges of the strip. (The naming of these two strip

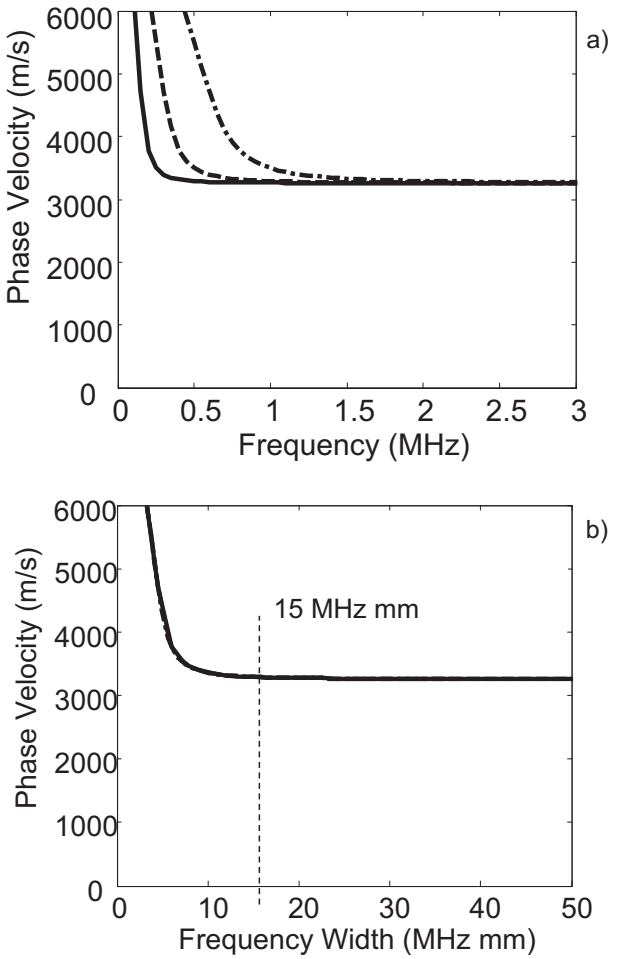


FIG. 4. (a) Phase velocity dispersion curves for a 1-mm-thick steel strip of widths of 30 mm (—), 15 mm (---), and 7.5 mm (-·····). (b) Phase velocity dispersion curves of the curves in (a) but plotted against the frequency-width product.

modes thus does not follow the same rules as the conventional naming of Lamb wave or Shear Horizontal (SH) wave modes but is done by comparing their properties and adding the * to indicate that it is a strip mode.) The A0* and SH0* modes are now investigated in detail.

A. The SH0* mode

Strictly speaking, the term "shear horizontal" does not make sense in a geometry other than an infinite plate. Therefore, it is stressed here again that the name SH0* mode was chosen due to the high frequency characteristics of the mode. At high frequencies, the SH0* mode travels with the shear velocity of the material and exhibits strong displacements in the y direction only that decay from the strip center toward the edges (the behavior on the center line being the same as that of the SH0 mode in an infinite plate). The same mode could also be described as a bending mode of the strip in the width-propagation direction (y-z) plane being similar to the A1 Lamb mode in an infinitely wide plate of thickness equal to the width of the strip. Just like the A1 Lamb mode, the SH0* modes possess a cutoff frequency below which it does not propagated. Due to the finite strip width, the SH0* mode

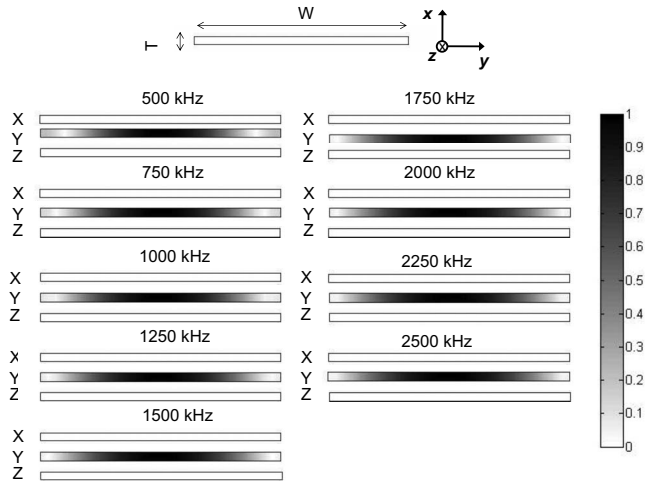


FIG. 5. Modulus of the displacement mode shapes of the SH0* mode of a 1-mm-thick and 30-mm-wide rectangular steel strip in the x, y, and z directions at different frequencies.

wave number in the propagation direction becomes imaginary below the cutoff frequency and the field exponentially decays along the waveguide axis.

The phase velocity dispersion curves for the SH0* mode of a 1-mm-thick steel strip of different widths (30, 15, and 7.5 mm) are shown in Fig. 4(a). The SH0* mode has a cutoff that depends on the width of the strip. At frequencies well above the cutoff, the phase velocity asymptotically approaches the bulk shear velocity of the strip material. Figure 4(b) shows that the dispersion behavior is a function of the frequency-width product, all three curves of Fig. 4(b) being coincident.

While strongly frequency (or width) dependent at low frequency-width products, at high frequency-width products, the SH0* mode phase velocity is constant and equal to the bulk shear velocity, allowing nondispersive wave propagation without signal distortion. The transition from highly dispersive to nondispersive is virtually complete at a frequency-width product of 15 MHz mm where the SH0* mode velocity differs by less than 1% from the bulk shear velocity. This point is marked by a vertical line in Fig. 4(b) and marks the frequency-width product at which the wave propagation

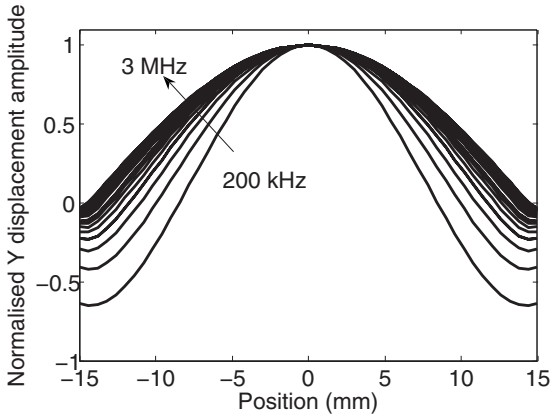


FIG. 6. Evolution of the SH0* y displacement mode shape of a 30-mm-wide and 1-mm-thick steel strip over a range of frequencies (200 kHz–3 MHz in steps of 100 kHz).

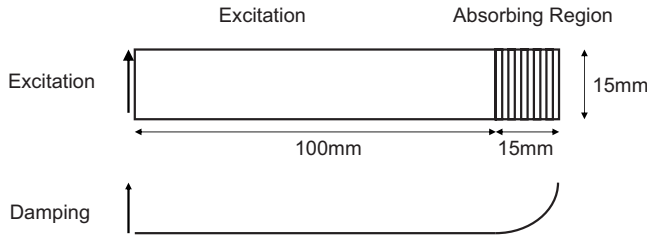


FIG. 7. Sketch of the 2D plane stress model that was defined in ABAQUS to analyze the effect of different excitation force profiles.

is becoming nondispersive for most practical purposes. The 15 MHz mm criterion above which SH0* propagation becomes nondispersive is material property dependent and is specific to steel that was used in the analysis here. To extend the criterion to other materials, it is useful to specify a minimum strip width in terms of shear bulk wavelengths. 15 mm equates to roughly five bulk shear wavelengths at 1 MHz in steel. As a rule of thumb, it can therefore be expected that the strip width has to be larger than five shear bulk wavelengths of the waveguide material in order to permit nondispersive wave propagation in form of the SH0* mode.

The mode shapes of the SH0* mode at different frequencies are shown in Fig. 5. The figure shows that at high frequencies, the y displacement component is dominant and concentrated in the center of the strip. The mode shape is constant across the thickness (x direction). Near cutoff (55 kHz for 30-mm-wide strip), there are displacements at the edges of the strip, which diminish as the frequency increases. This is better illustrated in Fig. 6, which shows the evolution of the dominant y displacement across the width of the strip over a range of frequencies. In the graph, each line represents a mode shape starting from 200 kHz and increasing in steps of 100 kHz up to 3 MHz. At a frequency-width product of 15 MHz mm (i.e., the fourth line in Fig. 6), the mode shape has started to concentrate in the center of the strip, displacements at the edge have decayed, and the shape has become similar to the final high frequency parabolic profile. A further increase in frequency will only slightly refine the mode shape toward its final shape.

If the mode is to be used in practice, it is important to be able to excite it in a reliable way. It was investigated how accurately an exciting transducer would have to reproduce the mode shape in order to selectively excite the SH0* mode. The influence of the distribution of the excitation force was investigated by using a finite element model. A finite element model for a 15-mm-wide steel strip was prepared in the ABAQUS finite element software.[21] The model was two dimensional with a plane stress condition in the thickness direction of the strip. Due to its special polarization, the SH0* mode only contains σ_{zy} stress components, which satisfy the plane stress condition ($\sigma_{xx}=\sigma_{xy}=\sigma_{xz}=0$). A frequency domain solver was used. At one end of the strip, a force was applied while an absorbing region at the other end of the strip prevented any reflections of the excited waves. This technique is commonly used to remove the influence of unwanted reflections from boundaries in FE models.[22,23]

A sketch of the FE model is displayed in Fig. 7. Square quadratic elements of size of 0.25 mm were used to mesh the strip and the absorbing region. The viscoelastic parameters of the absorbing region were increased in a cubic fashion from the interface with the strip. They were determined, as described by Drodz *et al.*[23] Different distributions of exciting force over the width of the steel strip ($\rho=7932$ kg/m^3, $E=216.9$ GPa, $\nu=0.2865$) were used to see the influence of the excitation force profile on the waves excited in the strip.

The results of the FE analysis are displayed in Fig. 8. In the figure, three different profiles of excitation force across the width are shown together with the y (width) direction displacement field that they produce; the excitation frequency was 2 MHz. In the picture, the displacement fields are normalized to show displacements on a scale between +100 and −100 so that the relative amplitude of the modes excited by the different force profiles could be assessed. For a uniformly applied force, the displacement field in the strip becomes relatively complicated. It can be concluded that many modes are excited and interfere. If the stress profile of the SH0* mode (from FE eigensolver) is applied at the strip end, a pure mode can be excited in the strip.[24] Displacements are concentrated at the center of the strip. For a triangular forcing profile, the SH0* mode is also preferentially excited. There is a marginal difference in the displacement field excited by the exact mode shape forcing compared to the triangular forcing profile; the difference is only visible at the edges of the strip where the amplitude of the SH0* mode is weakest. This shows that other modes are excited at much lower levels than the SH0* mode and their contribution to the overall displacement field is negligible. It was concluded that any excitation that is constant across the thickness of the strip and resembles the mode shape better than a triangular forcing in the width direction will allow the excitation of an almost pure SH0* mode in the strip.

B. The A0* mode

The A0* mode is very similar to the commonly known A0 mode in an infinitely wide plate. It is a flexural mode with respect to the thickness (x direction); however, due to the finite width, the mode also has a variation across the width (y direction). Just as the dominant displacements (y direction) of the SH0* mode are symmetric about the width, the dominant displacements (x direction) of the A0* mode are also symmetric about the width with maximum displacement at the center of the rectangular strip. The relationship of the mode to the A0 mode in an infinitely wide plate is underlined by their similarity in phase velocity. Figure 9 shows the phase velocity of both modes as a function of frequency, the difference being greatest at low frequency near the cutoff of the A0* mode and becoming negligible at high frequency. This trend is similar to the SH0* mode case where the strip mode asymptotically approaches the plate mode properties at high frequency.

The mode shape of the A0* mode also shows a similar behavior to the SH0* mode as the dominant displacement concentrates in the center of the strip with increasing frequency. Figure 10 shows the mode shape of the A0* mode of a 1-mm-thick, 30-mm-wide strip at different frequencies. The main displacement component is in the x direction and

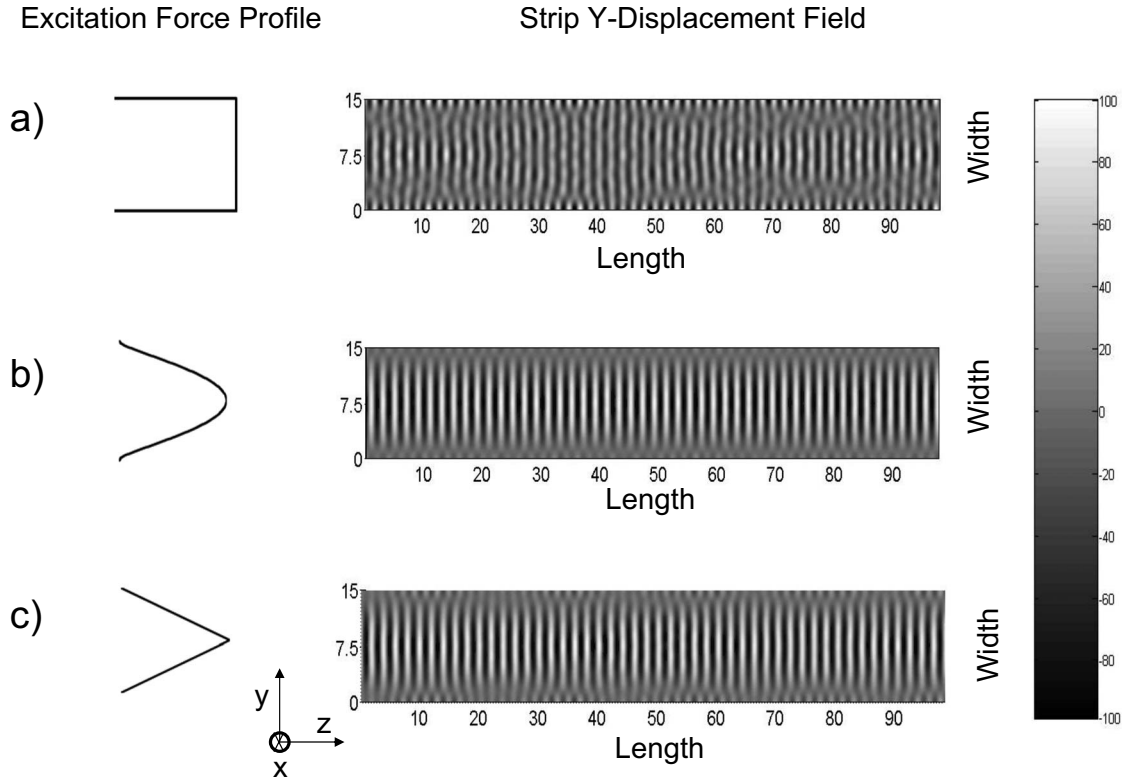


FIG. 8. *y* direction displacement field output of the plane stress steady state frequency domain finite element model of a steel strip (15 mm) under (a) rectangular excitation force profile, (b) exact mode shape excitation force profile and (c) triangular excitation force profile across the width of the strip at 2 MHz. The difference between cases (b) and (c) is marginal and only visible at the strip edge.

displacements are significantly concentrated at the center of the strip as the frequency increases. This is better seen in Fig. 11 where the *x* displacement component on the center line of the strip is shown over a range of frequencies.

So far, displacement components have been shown to illustrate the polarization of the mode as well as its concentration of energy at the center. Being proportional to the square of the displacement amplitude, the concentration of the mode energy is even more drastic.

IV. EXPERIMENTS

The experimental investigation of wave propagation in rectangular strips was focused on exciting A0* and SH0* modes. To achieve this, the exciting transducer has to mimic the mode shape as closely as possible and, for broadband signals, the mode shape should not change significantly over the range of excited frequencies. As shown in Fig. 8, it is insufficient to simply impose a uniform excitation across the waveguide width. However, as mode shapes are uniform across the thickness, the transducer output has to be varied across the width of the strip only (see Figs. 5 and 10). Displacements for both modes are strong at the center of the strip and decay toward the edges in approximately parabolic fashion. The main difference between the two modes is the

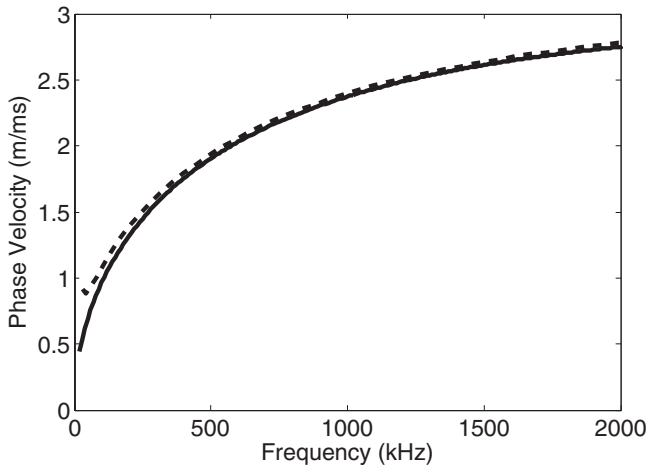


FIG. 9. Phase velocity dispersion curve of the A0 mode (—) of a 1-mm-thick steel plate and of the A0* mode (---) of a 1-mm-thick and 30-mm-wide steel strip.

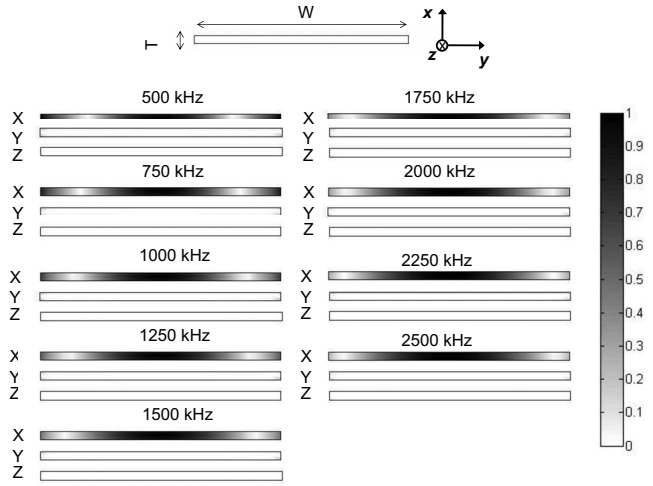


FIG. 10. Modulus of the displacement mode shape in the *x*, *y*, and *z* directions of the A0* mode in a 1-mm-thick and 30-mm-wide strip at different frequencies.

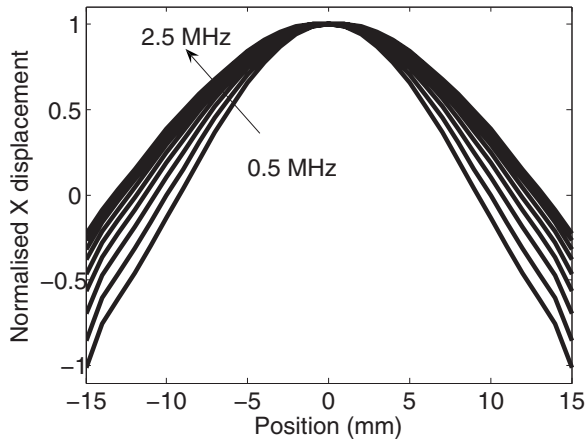


FIG. 11. Evolution of the A0* x displacement mode shape of a 30-mm-wide and 1-mm-thick steel strip over a range of frequencies (500 kHz–2.5 MHz in steps of 250 kHz).

polarization of the A0* mode in the thickness (x) direction and the polarization of the SH0* mode displacements in the width (y) direction.

It was found that good results could be achieved by simply coupling a standard circular ultrasonic shear transducer (Panametrics $V154$) to the end cross section of the strip. The circular shape of the piezoelectric element within the transducer was believed to transmit larger shear stresses at the center of the strip width than at the outside and thus lead to preferential excitation of the SH0* or A0* mode when rotated by 90°.

A. SH0* mode excitation

Figure 12 shows a 5 cycle, 2 MHz Hanning windowed toneburst that was sent and received in pulse echo mode from a 15-mm-wide, 1-mm-thick, and 300-mm-long stainless steel strip. The ⌀13 mm transducer (Panametrics $V154$) was clamped to the steel strip by a purpose made clamp using treacle (a very viscous fluid, similar to honey) as a shear couplant between the transducer face and the waveguide end section. The transducer was polarized in the direction of the width of the strip (y direction). Figure 12 shows that a very clean signal without significant dispersion can be excited and received in the strip. The presence of other modes about 30 dB weaker than the main signal can also be seen in Fig. 12. The presence of higher order modes that are notably slower can be explained by transducer misalignment and other imperfections within the strip and during reflection at the waveguide end.

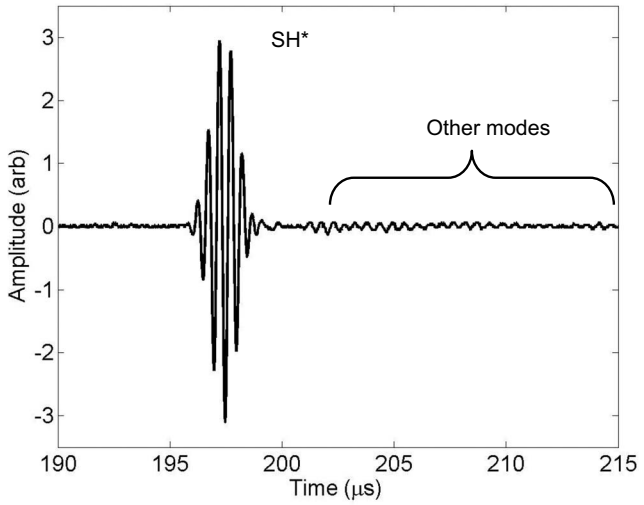


FIG. 12. 2 MHz center frequency SH0* signal received in pulse echo mode from a standard shear transducer coupled to the end of a 15-mm-wide, 1-mm-thick, and 300-mm-long stainless steel strip.

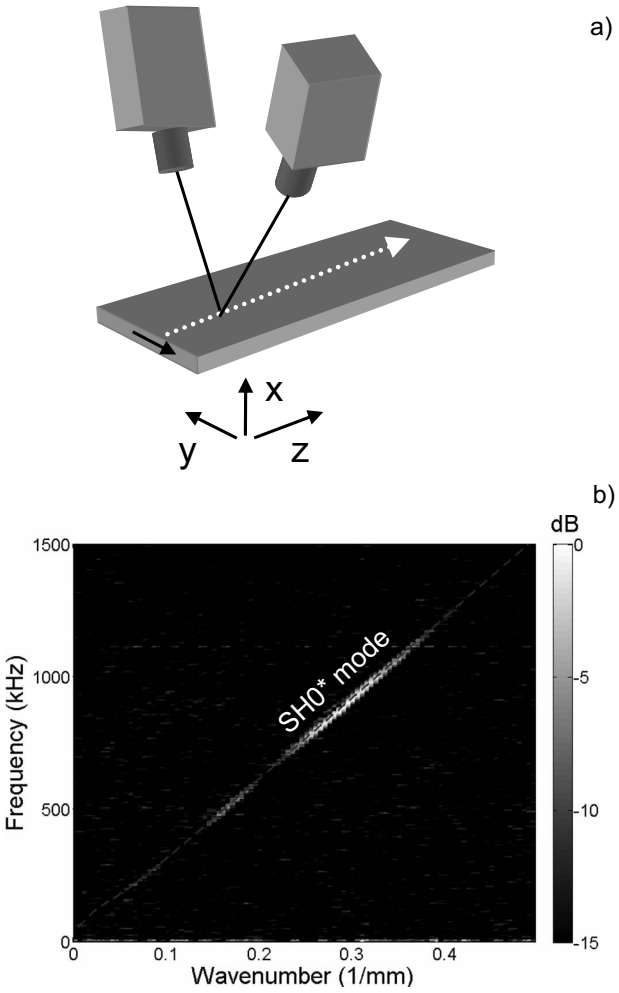


FIG. 13. (a) Sketch of the in-plane laser Doppler vibrometer scanning configuration along the strip. (b) Two dimensional Fourier transform of in-plane surface displacements (polarized in the width direction of the strip) along the center line of the 1-mm-thick and 30-mm-wide steel strip. The dashed line (---) shows the predicted dispersion relation for the SH0* mode of steel (ρ=7932 kg/m^3, C_l=6000 m/s, C_s=3060 m/s).

To be certain that the desired mode was excited in the experiment, an in-plane dual head laser Doppler vibrometer (Polytech OFV 512) was used to measure the in-plane surface displacement (y direction) of the strip along the center line of the strip. The signal was recorded every 0.5 mm over a distance of 200 mm at a sampling frequency of 10 MHz. This is schematically illustrated in Fig. 13(a). From the measurements, a two dimensional Fourier transform was computed.[25] The two-dimensional fast fourier transform (2D-FFT) displays the frequency–wave number relationship of the signals that have been measured in the waveguide. This plot can be directly compared to analytical frequency–wave

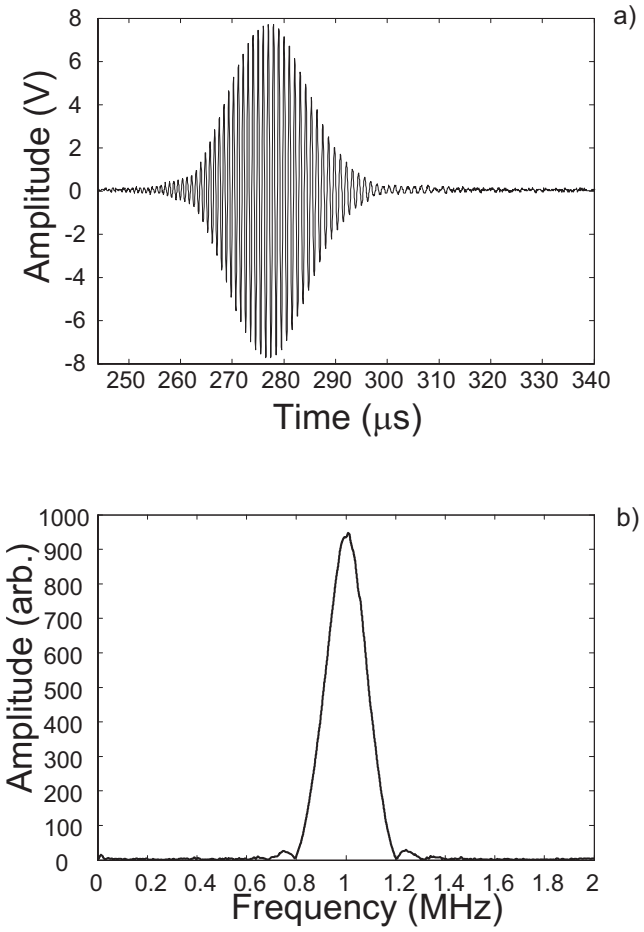


FIG. 14. (a) Pulse echo signal received from a 30-mm-wide and 0.2-mm-thick stainless steel strip using a 10 cycle Hanning windowed tone burst with 2 MHz center frequency. (b) Spectrum of the signal in (a).

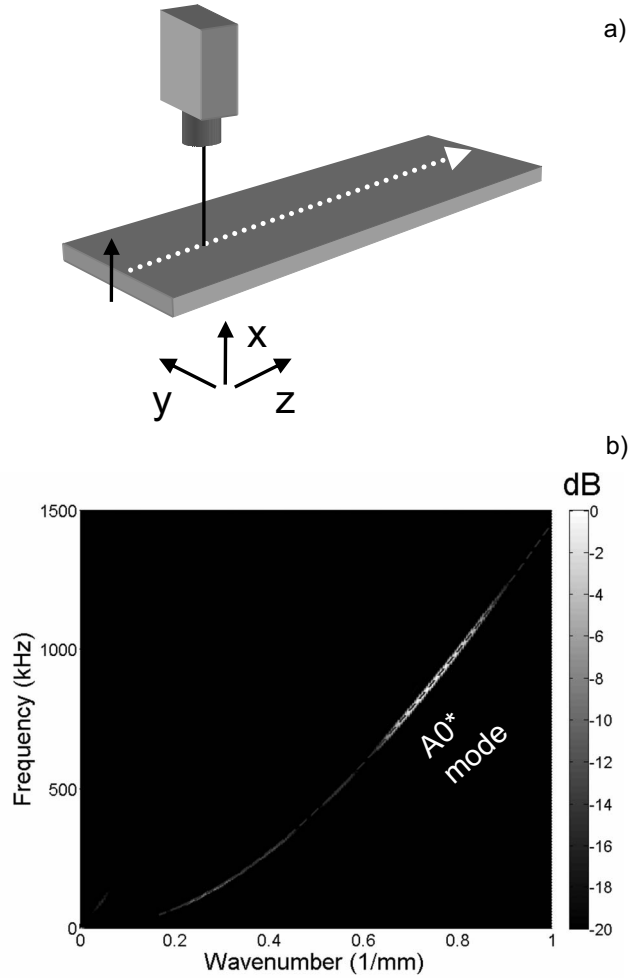


FIG. 15. (a) Sketch of the out-of-plane laser Doppler vibrometer measurements on a strip. (b) Two dimensional Fourier transform of the out-of-plane displacements of the center line of a 0.2-mm-thick and 30-mm-wide steel strip along the center line of the strip. The dashed line (---) shows the predicted dispersion relation for the A0* mode of steel (ρ=7932 kg/m^3, C_l=6000 m/s, C_s=2840 m/s).

number predictions. In Fig. 13(b), the 2D-FFT result for a 30-mm-wide and 1-mm-thick steel strip is plotted. A line indicating the theoretically predicted SH0* mode frequency–wave number relation for a 30-mm-wide and 1-mm-thick steel strip (ρ=7932 kg/m^3, C_l=6000 m/s, C_s=3060 m/s) is also displayed. There is very good agreement between the measured data and the predicted values for the SH0* mode. The noise floor in the 2D-FFT plot is relatively high; this is due to strong noise and drop outs in the laser vibrometer measurements.

B. A0* mode excitation

The setup was slightly changed to excite the A0* mode. A 30-mm-wide, 0.2-mm-thick, and 300-mm-long steel strip was used and the transducer was turned by 90° to excite displacements in the thickness (x) direction. Figure 14(a) shows the pulse echo signal received by the transducer following excitation by a 10 cycle Hanning windowed tone burst with 2 MHz center frequency. The A0* mode is very dispersive in this frequency range, which explains the very strong distortion of the signal. Despite the strong dispersion, the signal is due to a single dominant mode, as suggested by Fig. 14(b), which shows a smooth spectrum of the signal without dips and interference from other modes.

Again, to verify the excited modes in the strip, the laser vibrometer was scanned along the center line of the strip and a 2D-FFT was computed. Figure 15 shows the result. Out-of-plane (x direction) displacement measurements were carried out over a distance of 200 mm at increments of 0.5 mm with a temporal sampling frequency of 10 MHz. The 2D-FFT shows that the A0* mode is the dominant mode. Other modes are ~20 dB less strong than the A0* mode.

V. CONCLUSIONS

Two particular guided wave modes (SH0* and A0*) of a large aspect ratio rectangular cross section waveguide were investigated in detail. It is shown that in both modes, energy concentration at the center of the waveguide is observed at high frequencies. At the same time, the mode characteristics converge toward those of the fundamental modes (SH and A0) in the infinitely wide plate case. One of the modes also asymptotically approaches the shear velocity at high frequencies and therefore becomes almost nondispersive, which allows the propagation of undistorted signals over large distances.

It is experimentally demonstrated that these modes can easily be excited by standard techniques with signal to coherent noise ratios of 20 dB and better. Essentially, at large frequency-width products, the modes that concentrate at the strip center resemble the fundamental plane strain guided wave modes of a plate in the central region of the strip and decay toward the edges of the strip. For the SH0* mode, it was possible to define a minimum strip width above which wave propagation will be nondispersive. Any SH0* signal whose wavelength is smaller than a fifth of the strip width will propagate virtually nondispersively at the bulk shear velocity along the waveguide center.

Potential applications of these modes are in any field where guided plate waves are to be used but space is confined. An example from the author's experience is the development of nondispersive buffer waveguide strips for high temperature thickness gauging. The waveguide allows the separation of the transducer from the measurement zone, which can be useful for transducer longevity and accessibility reasons. Work on the development of the thickness gauge will be presented in a future publication.

ACKNOWLEDGMENTS

The author would like to thank Professor P. Cawley for encouragement and the continued discussions on the topic of this paper.

[1]K. F. Graff, *Wave Motion in Elastic Solids* (Dover, New York, 1973).
[2]J. L. Rose, *Ultrasonic Waves in Solid Media* (Cambridge University Press, Cambridge, 1999).
[3]M. Lowe, "Matrix techniques for modelling ultrasonic waves in multilayered media," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **42**, 525–542 (1995).
[4]J. C. P. McKeon and M. K. Hinders, "Parallel projection and cross-hole lamb wave contact scanning tomography," J. Acoust. Soc. Am. **106**, 2568–2577 (1999).
[5]P. Fromme, P. D. Wilcox, M. J. S. Lowe, and P. Canvley, "On the development and testing of a guided ultrasonic wave array for structural integrity monitoring," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 777–785 (2006).
[6]D. Alleyne and P. Cawley, "The long range detection of corrosion in pipes using lamb waves," *Review of Progress in Quantitative NDE*, edited by D. O. Thompson and D. E. Chimenti (Plenum, New York, 1994), Vol. 14.
[7]J. O. Kim and H. H. Bau, "On line, real-time densimeter—theory and optimsation," J. Acoust. Soc. Am. **85**, 432–439 (1989).
[8]P. B. Nagy and R. M. Kent, "Ultrasonic assessment of Poisson's ratio in thin rods," J. Acoust. Soc. Am. **98**, 269–2701 (1995).
[9]G. F. Miller and H. Pursey, "The field and radiation impedance of mechanical radiators on the free surface of a semi-infinite isotropic solid," Proc. R. Soc. London, Ser. A **223**, 521–541 (1954).
[10]J. D. Achenbach, *Wave Propagation in Elastic Solids* (North-Holland, Amsterdam, 1975).
[11]M. J. S. Lowe and B. N. Pavlakovic, *Disperse User Manual, Version 2.0.11d* (Imperial College of Science, Technology and Medicine, London, 2001).
[12]R. D. Mindlin and E. A. Fox, "Vibrations and waves in elastic bars of rectangular cross section," J. Appl. Mech. **27**, 152–158 (1960).
[13]W. B. Fraser, "Stress wave propagation in rectangular bars," Int. J. Solids Struct. **5**, 379–397 (1969).
[14]P. Wilcox, M. Evans, O. Diligent, M. Lowe, and P. Cawley, "Dispersion and excitability of guided acoustic waves in isotropic beams with arbitrary cross section," *Review of Progress in Quantitative NDE*, edited by D. O. Thompson and D. E. Chimenti (AIP, Melville, NY, 2002), Vol. 20.
[15]O. M. Mukdadi, Y. M. Desai, S. Datta, A. H. Shah, and A. J. Niklasson, "Elastic guided waves in a layered plate with rectangular cross section," J. Acoust. Soc. Am. **112**, 1766–1779 (2002).
[16]T. Hayashi, W. Song, and J. L. Rose, "Guided wave dispersion curves for a bar with an arbitrary cross-section, a rod and rail example," Ultrasonics **41**, 175–183 (2003).
[17]L. Gavric, "Computation of propagative waves in free rail using a finite element technique," J. Sound Vib. **185**, 531–543 (1995).
[18]S. Finnveden, "Evaluation of modal density and group velocity by a finite element method," J. Sound Vib. **273**, 51–75 (2004).
[19]M. V. Predoi, M. Castaings, B. Hosten, and C. Bacon, "Wave propagation along transversely periodic structures," J. Acoust. Soc. Am. **121**, 1935–1944 (2007).
[20]D. Hitchings, *Fe77 User Manual* (Imperial College of Science, Technology and Medicine, London, 1994).
[21]*ABAQUS Reference Manuals, Version 6.5* (ABAQUS Inc., Providence, RI, 2004).
[22]L. Moreau, M. Castaings, B. Hosten, and M. V. Predoi, "An orthogonality relation-based technique for post-processing finite element predictions of waves scattering in solid waveguides," J. Acoust. Soc. Am. **120**, 611–620 (2006).
[23]M. Drodz, L. Moreau, M. Castaings, M. J. S. Lowe, and P. Cawley, "Efficient finite element modelling of absorbing regions for boundaries of guided wave problems," *Review of Progress in Quantitative NDE*, edited by D. O. Thompson and D. E. Chimenti (AIP, Melville, NY, 2005).
[24]P. J. Torvik, "Reflection of wave trains in semi-infinite plates," J. Acoust. Soc. Am. **41**, 346–353 (1967).
[25]D. Alleyne and P. Cawley, "A two-dimensional fourier transform method for the measurement of propagating multimode signals," J. Acoust. Soc. Am. **89**, 1159–1168 (1991).

Evaluating the maximum playback sound levels from portable digital audio players

Stephen E. Keith,[a] David S. Michaud, and Vincent Chiu
Consumer and Clinical Radiation Protection Bureau, Health Canada, 775 Brookfield Road, Ottawa, Ontario, K1A 1C1 Canada

(Received 1 October 2007; revised 9 March 2008; accepted 11 March 2008)

To assess the maximum sound levels that may be experienced by young people in Canada from modern digital audio players, this study measured nine recent models of players and 20 earphones. Measurement methodology followed European standard BS EN 50332. Playback levels ranged from 101 to 107 dBA at maximum volume level. Estimated listener sound levels could vary from 79 to 125 dBA due to the following factors: (i) earphone seal against the ear, (ii) player output voltage, (iii) earphone sensitivity, and (iv) recorded music levels. There was a greater potential for high sound levels if intra-concha "earbud" earphones were used due to the effect of earphone seal. Simpler measurement techniques were explored as field test methods; the best results were obtained by sealing the microphone of a sound level meter to the earphone using a cupped hand and correcting for the free field response of the ear. Measurement of noise levels 0.25 m from the earphone showed that a bystander is unlikely to accurately judge listener sound levels.
 [DOI: 10.1121/1.2904465]

PACS number(s): 43.50.Hg, 43.50.Qp, 43.38.Lc, 43.50.Yw [BLM] Pages: 4227–4237

I. INTRODUCTION

Sound levels below approximately 70 dBA (Ward *et al.*, 1976; World Health Organization, 1999; NIH, 1990; ISO, 1990) pose no known risk of hearing loss, regardless of exposure duration. For higher sound levels, the duration of daily exposure becomes an important risk factor. For example, sounds with levels of 85 dBA pose no known risk of hearing loss when exposures are less than 45 min per day (Health Canada, 2006b). However, with sound levels of 85 dBA or higher, there is a risk of gradual permanent hearing loss, if a person is exposed for 8 h per day. As a result, several Canadian provinces have set the maximum permissible sound level for occupational noise exposure at 85 dBA energy averaged over an 8 h day (see references in Canadian Centre for Occupational Health and Safety, 2007). In the USA, OSHA has set this occupational exposure to be the threshold for the establishment of a hearing conservation program (OSHA, 2006); the maximum permissible sound level is 90 dBA with a 5 dB exchange rate.

To our knowledge, France is the only country with legislation governing both performance and labeling for personal stereo systems, including digital audio players (DAPs) (Legifrance, 2005). This legislation limits the sound levels from DAPs with headphones to 100 dBA as determined using the measurement methodology of two voluntary European standards (British Standards Institution, 2000; 2004). The French law also limits the output voltage of players to 150 mV and requires a list of compatible earphones that will meet the 100 dB limit when used with a personal stereo system. A warning is also required to be affixed to the device about possible hearing impairment. Devices imported from the European Economic Area and Turkey do not have to meet the law if the devices are made to provide the same level of safety and information to the consumer as is provided by the French law (Legifrance, 2005).

Personal stereo systems using earphones generate concern primarily because the source is close to the tympanic membrane and can potentially generate higher sound levels than produced by typical home or commercial sound systems. Furthermore, there continues to be an uncertainty with respect to user listening habits, type of earphone (including fit), type of music, influence of background sound levels, etc. These factors, coupled with the growth in popularity of modern portable DAPs such as MP3 players, can influence the sound levels to which individuals are exposing themselves.

Studies done at Health Canada, by Keith and coworkers, have assessed the potential for personal stereo systems with earphones to exceed sound levels that could potentially increase the risk of hearing impairment among users. These studies evaluated portable compact disc (CD) players at maximum volume settings with either the earphones included with the player at purchase, or with separately purchased earphones (Bly *et al.*, 1998; Keith *et al.*, 1999; 2001). It was found that, depending on the type of earphone used, type of music and the headphone seal at the ear, some of the devices were capable of emitting sound that exceeded safe levels at the user's ear. The equivalent free field levels covered a potential range from 65 to 122 dBA (Keith *et al.*, 2001). However, the physical size, limited playback capacity, and limited battery life that characterized typical portable CD players all served to discourage users, or at least make it less convenient, to utilize these devices for extended periods of time at high volume settings (TNO, 1998; Airo *et al.*, 1996; Aono, 1997; Felchlin *et al.*, 1998). Health Canada concluded from these studies that actual listening habits appeared to have kept the risk low. However, it

[a]Author to whom correspondence should be addressed. Electronic mail: skeith@hc-sc.gc.ca

was concluded that the risk should not be ignored, as so many young people use these devices (Health Canada, 2006b).

In contrast, DAPs that have become commonplace today are substantially smaller in size, have an increased storage capacity, and increased battery life. These factors make it very convenient to remain "plugged-in" for extended durations. Despite these factors and their widespread popularity, there is still a paucity of peer-reviewed publications dedicated to evaluating the risk of hearing impairment from DAPs.

Williams (2005) recently evaluated the risk to hearing associated with personal stereo use in environments characterized as having high background sound levels (mean value 73.2 dBA for Leq (5 min) morning, midday, and afternoon for two days) because it was hypothesized that high background levels should force users to increase volumes to "worst-case conditions." Williams concluded that while some individuals were exposing themselves to Leq (8 h) values that would pose a risk for hearing impairment, most individuals' Leq (8 h) exposures were considered safe. Males tended to listen to music at higher Leq (8 h) values. The study by Williams did not attempt to assess how the fit between the ear and headphone influenced exposure levels, nor did the author investigate how the types of headphone or music influenced sound levels. Furthermore, there was no attempt to specifically evaluate the newer generation DAPs.

A similar study, (Ahmed *et al.*, 2007) showed that while the majority of participants preferred playback volume settings around 67 dBA, there was a tendency to increase the volume depending on background sound levels. Measured sound levels were as high as 105 dBA and nine of the 24 participants selected a volume setting above 85 dBA. A recent survey of user habits among 150 university students revealed that the vast majority of participants owned a portable audio device (i.e., 82.7%), with nearly half of them using the device for 5–7 days per week at an average of 2 h per day. Out of the 150 students surveyed, 21 students reported listening to their device at volume settings between 80% and 100% levels, but the majority of users selected volumes that were between 25% and 75% of their device's maximum setting. This study also reported that as many as 31% of the participants indicated that their self-reported hearing health was worse at the time of the study than it was five years earlier, but this could not be associated with using a DAP (Ahmed *et al.*, 2007).

Portnuff and Fligor (2006) recently presented a paper that investigated sound levels from five popular DAPs. These authors found that at a given volume setting, the sound level output was dependent on the type of earphone used (i.e., sound levels from the earbud style were 5.5 dB higher compared to supra-aural style earphones). Levels at 100% volume ranged from approximately 96 to 105 dBA and were reduced by approximately 6 dB for each decrease in volume setting equal to 10% of the maximum volume. Based on volume setting and earphone type, the authors suggested corresponding maximum listening times based on conservative damage-risk criteria. The maximum listening time per day dropped considerably on all earphone types once volume settings exceeded 70%, but this quantity also depended on earphone type. For in-ear canal "isolator" earphones the maximum listening time was reported to be 3.4 h and 3 min at 70% and 100% volume settings, respectively. By comparison, the maximum listening times for supra-aural earphones at these volume settings were reported to be 20 h and 18 min, respectively. The authors indicated that music genre did not have any significant impact on sound level differences between the five DAPs, especially at higher volume settings.

The present study was undertaken to characterize the maximum playback sound levels for listeners using currently popular DAPs with different earphone types and fittings when listening to contemporary popular music samples. A second purpose was to compare the results with two relatively new standards (British Standards Institution, 2000; 2004), and to investigate the validity of a common "field test" procedure that uses a sound level meter to approximate the free field response when a head and torso simulator (HATS) is not available. This information will be used as a scientific basis for advice to manage the potential for hearing loss from DAPs sold in Canada.

To this end, a measurement survey was made of nine recent model DAPs and 20 separately purchased earphones. To compare to regulated limits in France, the European measurement standard was followed. The range of procedures adopted for earphone fitting was also intended to test the uncertainties in measurement using the standard. Also, for relevancy to the type of music listened to by youths, who are potentially most at risk, sound levels were obtained from the top ten hit singles for Canadian youth between 12 and 24 years identified from 2005 radio broadcast data in Canada (Nielsen Broadcast Data Systems, 2007). Tests were also made for headphone insertion loss because of evidence that background noise affects the volume setting that people use. Most testing used a HATS. However, since this is specialized equipment that is not available to many audiologists or acoustical consultants, additional tests were made to evaluate the validity of results obtained using only a sound level meter. Additional tests were made to determine: the usefulness of some procedures, and messages for educational purposes about the risks to hearing of DAPs.

II. METHODS

The measurement methodology followed BS EN 50332 part 1 (British Standards Institution, 2000) and part 2 (British Standards Institution, 2004) with the exception that measurements were repeated only three times before averaging. Parts of the methodology have been described previously (Bly *et al.*, 1998; Keith *et al.*, 1999; 2001) and more details are provided below. Measurements were made of the maximum sound levels from nine recent model DAPs with packaged earphones and 20 separately purchased earphones sold in Canada. Table II lists the DAPs that were selected based on informal surveys of the best sellers on various online internet sites, their availability and the sales experience of the local

TABLE I. Test audio files including 2005 top ten airplay for ages between 12 and 24 years (Nielsen Broadcast Data Systems, 2007).

Track[a]	Album	Artist	Measurement duration (s)	RMS level *re* full scale of the D/A converter (dBA)
IEC signal		Health Canada per IEC (1985)	128	−12.5
Caught up	Confessions	Usher	224	−16.1
Pon de replay	Music of the sun	Rihanna	246	−19.3
Hollaback girl	Love Angel Music Baby	Gwen Stefani	199	−12.2
Don't phunk with my heart	Monkey business	Black eyed peas	239	−14.3
1,2 step	Goodies	Ciara feat, Missy Elliot	202	−21.9
We belong together	The emancipation of mimi	Mariah Carey	201	−19.1
Rich girl	Love. Angel. Music. Baby.	Gwen Stefani	236	−14.5
Boulevard of broken dreams	American Idiot	Green Day	260	−13.4
Since u been gone	Breakaway	Kelly Clarkson	188	−14.4
Let me love you	Turing point	Mario	249	−18.5
Filler/I don't want to hear it	Undisputed attitude	Slayer	148	−12.3
A promise kept	Titanic music from the motion picture	James Horner	362	−42.3

[a]Filler/I don't want to hear it, A promise kept and IEC signal are not in the 2005 top ten airplay.

retail salesperson; the last 12 of the listed earphones were used in previous studies (Bly *et al.*, 1998; Keith *et al.*, 1999; 2001).

Five different types of earphones were used in the study. All DAPs were supplied with intra-concha earphones ("earbuds") that fit into the concha without entering the ear canal. Also tested were five insert earphones, ("canal" earphones), where the speaker driver was near the entrance to the ear canal and a flexible earpiece extended into and sealed the ear canal. There were three different types of headphones, i.e., earphones connected with a headband; supra-aural, circumaural and one with forward facing intra-concha earbuds.

A. Selected audio material

A filtered pink noise signal (IEC signal) as specified in IEC 60268-1 (IEC, 1985) for typical program material, was used in all testing. The RMS level was −10 dB, *re* full-scale of the digital-to-analog converter (D/A), and the crest factor was set at 6 dB as per IEC 60268-7 (IEC, 1996).

Test material also included the top ten hit singles for ages between 12 and 24 years, identified from 2005 radio broadcast data in Canada (Table I) (Nielsen Broadcast Data Systems, 2007). Two additional tests were made with the tracks that produced the highest and lowest measured levels in previous studies (Bly *et al.*, 1998; Keith *et al.*, 1999; 2001). In each case, the entire track was measured in the present study.

Table I lists the audio files used in testing, along with their duration and approximate RMS level *re* full scale of the D/A converter (in dBA). As found in a previous study, the average spectra of the top ten music closely matched the IEC signal, with the main exception that the low frequency limit in the top ten music extended one octave lower than the IEC signal (Keith *et al.*, 2001).

B. Earphone fitting

Sound level measurements were obtained using anatomically realistic, soft silicone rubber pinnas, Brüel & Kjær type DZ9752 on a Brüel & Kjær type 4128 HATS with type 2807 power supply. For "canal phones," the HATS was replaced by a Brüel & Kjær type 4157 ear simulator, which had a flared conical metal ear canal. According to the manufacturer, the internal acoustical passages in the HATS and ear simulator were the same.

Earphones with headbands were removed from the HATS, and the fitting force was estimated by placing the headphones on type DZ9752 pinnae mounted with appropriate spacing and geometry on a Phillips HR2385/A 5 kg digital electronic scale. The scale was checked before and after measurements using a Mettler 100 gm OIML class F1 calibration weight and a nominal 1000 gm laboratory weight to be within 2% of the correct reading. For the headphones with headbands: the fitting force was 0.2 N for earbud *e*23, 1.3–3 N for the supra aural, 2.7–4.5 N, for the circum aural, and approximately 9 N for the hearing protectors.

To simulate a normal fit, earphones were positioned as follows to obtain the maximum sound level using the IEC signal. Earbuds were inserted into the concha, gently moving the tragus forward, and then repositioned to obtain the maximum sound level. The procedure was slightly different due

TABLE II. Comparison of specifications of tested DAP with supplied earphones.

DAP/earphone ID	Model	player $V_{out,max}$, mV[a]	phone V_{WBCV}, mV	equivalent free fieldSPL, dBA[a]	high frequency insertion loss, dB[b]	bystander SPL for 94 dBA at ear, dBA[c]
*p*1/*e*1	ICE audio MP3 Personal Stereo 512 MB w/earbud	238	137	101.9	−0.5	50.1
*p*2/*e*2	Sony Portable 1C NW-E307, 1 GB w/earbud	201	84	101.1	−0.1	41.4
*p*3/*e*3	Iriver T10 1 GB w/earbud	416	93	102.9		45.5
*p*4/*e*4	Creative Zen Nano Plus 512 MB w/earbud	248	55	102.7	4.3	40.2
*p*5/*e*5	Apple Ipod nano MA004LL, 2 GB w/earbud	491	156	104.6	−0.3	49.5
*p*6/*e*6	Sony PSYC 1C NW-E103, 256 MB w/earbud	192	46	107.3	1.8	38.1
*p*7/*e*7	Apple Ipod MA147LL, 60 GB w/earbud	576	137	107.3	−0.5	51.0
*p*8/*e*8	Toshiba Gigabeat F20 MEGF20 w/earbud	469	109	104.9		53.8
*p*9/*e*9	Digitalway MYIO, FY500, 1 GB w/earbud	639	87	105.9	0.1	42.9
*e*10	Creative Zen DAP-MDD007, 8 GB w/earbud	150	73	100.2		45.5
*e*11	Apple Ipod shuffle earbud	150	146	94.2	−0.4	49.6
*e*12	Acoustic Authority ACM-800 noise canceling supra aural headphone	150	106	97.0	3.8	49.4
*e*13	Senheiser Noisegard PXC250 noise canceling supra aural headphone	150	229	90.3	8.5	42.1
*e*14	Apple Ipod In ear earphones canal	150	58	102.2	20.6	54.5
*e*15	Shure Sound isolating earphones EC3 canal	150	55	102.7	38.1	33.2
*e*16	Panasonic RP-HJE50PP-S canal	150	119	96.0	16.1	49.4
*e*17	Etymotic ER6i canal	150	77	99.8	40.5	<30
*e*18	Sony Fontopia MDR-EX70LP canal	150	173	92.7	18.1	52.4
*e*19	Gemini FAS7 earbud					57.6
*e*20	Gemini FAS5 earbud	150	114	96.4		44.8
*e*21	JVC HA-F65 earbud	150	66	101.1		48.1
*e*22	Koss HP/3 earbud	150	508	83.4		57.1
*e*23	Sony MDR-W20G headphone earbud	150	366	86.2		46.9
*e*24	Panasonic VMSS SL-SW515 headphone supra aural	150	114	96.4		55.6
*e*25	RCA RP7927-1 headphone supra aural	150	287	88.3		56.4
*e*26	Senheiser HD56 headphone supra aural	150	432	84.8		62.5
*e*27	Kenwood DPC-382 headphone supra aural	150				57.3
*e*28	Koss R/90 headphone circum aural	150	307	87.8	−1.7	66.1
*e*29	Optimus Pro XB 100 headphone circum aural	150	169	92.9	6.7	48.0
*e*30	Clark hearing protector muff headphone circum aural				24.9	
*e*31	Thunder 29 hearing protector muff headphone circum aural				30.9	

[a]Equivalent free field SPL is measured for player and earphone combination *p*1/*e*1 to *p*9/*e*9, and otherwise is based on measured headphone V_{WBCV} and an assumed player Vout,max of 150 mV which is the maximum output voltage specified in France (Legifrance, 2005) and BS EN 50332-2 (British Standards Institution, 2004).

[b]Some results for the insertion loss are negative, likely due to resonances.

[c]For all earphones, most of the sound energy at the bystander position (0.25 m) was in the 2–8 kHz octave bands. Notably, the tested earbuds had significantly reduced levels below 2 kHz.

to the shape of the left and right pinnae. Positioning was repeated on the left pinna until the earbud would stay in place by itself. In the right pinna, although the earbud would snap into position, repetitions were necessary to ensure the true maximum sound level was obtained.

For headphones (supra-aural, circum-aural and intra-concha), both transducers were simultaneously aligned for optimum coverage of the ear (without excessive deformation of the pinna), and then repositioned slightly to obtain simultaneous maximum levels. Canal phones were simply inserted into the conical opening of the ear simulator and a small object was placed behind the earphone to prevent it from slipping out. For all earphones, the labels "L" for left, or "R" for right, determined placement in the ears. Earphones *e*7 and *e*9 had an outward appearance suggesting that they could fit in either ear, but this was not attempted.

During all tests, the HATS output could be monitored by an external set of headphones, (ID *e*28 in Table II) worn by the experimenter. The response of these headphones was corrected using a TOA DP0204 signal processor so that the frequency response approximated that of the earphone under test. This was used to verify proper operation of the headphone without distortion, and assist in attaining the best fit, which was characterized by increased level, especially at lower frequencies.

For the earbuds and the supra-aural headphones, the effect on sound level of tighter and looser earphone fits to the ear was investigated.[1] This procedure was intended to explore plausible conditions under which earphones could be worn. For example, users may prefer a loose fit for ventilation or comfort. A tight fit could occur when earphones are worn under a hat or toque that covers the ears, or during exercise where accumulated sweat forms a liquid seal between the earphone and the ear. To simulate a tight fit, ear-

phones were pushed onto the soft pinnas of the HATS with a finger force of up to 20 N for the earbuds and 40 N for the larger headphones. For a loose fit, the earpieces were also oriented on the pinna so that, while still supported, visible gaps between the earpiece and pinna were allowed to occur.

C. Earphone wideband characteristic voltage, V_{WBCV}

Earphone sensitivity, measured as V_{WBCV} following BS EN 50332-2 (British Standards Institution, 2004), was obtained using a Bryston 8B amplifier (125 W per channel, <0.2 ohm output impedance) connected to a Sony XA20ES CD player. The amplifier output was adjusted to obtain approximately 94 dBA equivalent free field level at the HATS using the IEC signal. A Hewlett-Packard 3468A multimeter measured the unweighted RMS voltage into the earphones. The V_{WBCV} was the voltage calculated to produce a 94 dBA equivalent free field level at the ear.

D. Earphones packaged with DAP

For earphones packaged with a DAP, the volume and tone controls were set to obtain the maximum A-weighted equivalent free-field output on the HATS. A GW Model GPS-1850 dc power supply provided 1.5 V to players with replaceable batteries. Players with internal rechargeable batteries were operated using their battery charger, and if this was not possible, with the batteries fully charged between each test.

E. Earphone insertion loss in the presence of external noise

The insertion loss of the earphones in the presence of external noise was tested in the $13\times9\times7$ m^3 hemi-anechoic chamber at Health Canada. At the ear of the HATS, 88 dBA pink noise (145 Hz to 10 kHz varying between 66 dB and 71 dB in 1/12 octave bands), was produced by a Bruel & Kjaer Pulse® 10 analyzer, with the Bryston 8B amplifier driving a PSB Alpha LR1 speaker. The speaker was positioned 1m from the right side of the HATS and centered on the axis joining the two ears (at a height of 1.5 m). The insertion loss due to the earphones was the difference between measurements with the earphone, and the bare ear. For simplicity in reporting results, the insertion loss was reported using the formula for the single number noise reduction rating (NRR) (US EPA, 1979). A normal fit was used for all measurements. Preliminary measurements showed little to no attenuation for most earphones. The earphones were then exchanged for commercial hearing protectors ($e30$ and $e31$ in Table II). Using a normal fit with the hearing protectors yielded NRR values that did not exceed 3 dB. Repositioning the hearing protectors to eliminate gaps provided a tighter seal and increased the NRR to approximately 30. The difference was due to the low frequency attenuation. The improved fit caused the pinna to bend and would not originally have been classified as a normal fit in the context of this paper.

As a result of these preliminary findings, the insertion loss was evaluated with a normal fit, and estimated using the formula for NRR, but applied only to octave bands from 1 kHz to 8 kHz (i.e., assuming infinite attenuation of noise below 1 kHz). With the exception of the canal phones, the earphones tested would be expected to have negligible low frequency attenuation for a normal fit. Therefore, insertion loss values determined using the technique described here should be greater than corresponding NRR values. These fitting problems did not occur for the canal phones. Their insertion loss could also be higher than real world NRR values as NIOSH recommends subtracting 50% from the manufacturer's labeled NRR for similar design hearing protectors (NIOSH, 1998).

F. Comparison of bystander and user levels

During earphone testing, an additional external Bruel & Kjaer type 4165 free field microphone (bystander microphone) was positioned at a distance of 0.25 m from the right ear of the HATS. The reported bystander level was the sound level measured at the bystander microphone when the equivalent free field level measured at the HATS was 94 dBA.

G. Comparison of HATS and microphone

Sound levels from six earphones were also obtained using only a bare 1/2 in. microphone instead of the HATS. With the IEC signal set to produce an equivalent free field level of 94 dBA measured on the HATS, earphones were tested with the earphone lightly touching the grid of a Bruel & Kjaer type 4165 free field microphone. Despite the light contact, as there was no acoustical seal between the earphone and microphone, this was considered a simulation of a very loose fit. Then the microphone body was held in a fist by the little finger with the other fingers forming a tube (to simulate the ear canal), and the earphone was seated on top of the fist, being held in position with the other hand. A tight fit required up to 40 N of force to maintain what appeared to be the best seal possible. This force was estimated using the scale previously described in Sec. B. The normal fit used just enough force to keep the earphone stable. Both normal and tight fits were tested in this configuration. Finally, the microphone was slid fully into the fist so that it almost touched the earphone. Both normal and tight fits were tested. Results were obtained by applying either only A-weighting correction or both A-weighting and the free field correction for the HATS.

H. Equivalent sound level measurements

A Bruel & Kjaer Pulse® 10 analyzer with type 3110 front end provided real time 1/12 octave band sound pressure levels. Before and after each set of measurements, the microphone sensitivity was checked at 1 kHz using a Bruel & Kjaer type 4230 piezo calibrator. Occasionally, the microphone sensitivity was checked using a Brüel & Kjær type 4226 multifunction calibrator. A-weighted background noise was typically 50 dB lower than the measurements.

The sound levels of interest were the A-weighted, equivalent continuous sound pressure levels in a free field, incident normal to the forehead of the HATS, that would produce the same response as the earphone. To determine

these values, the 1/12 octave measurements from both the HATS and ear simulator were A-weighted and corrected using the free field frequency response supplied by the manufacturer for the HATS (i.e., Bruel & Kjaer listener response for the HATS). For the IEC signal, there is less than 1 dB difference between corrections using the manufacturer listener response, and either the diffuse-field response, or the free field 1/3 octave corrections in ISO 11904-2 (ISO, 2004).

The BS EN 50332 standard (British Standards Institution, 2000) specifies an arithmetic average of 10 Leq measurements, 5 on each ear (as noted above, each Leq was taken for an entire song). In this study, the real time display of the spectrum sound pressure level (SPL), aural monitoring, and the consistent fitting technique could occasionally lead to identical results when repeated sequentially. Nevertheless, for the earbuds, long term reproducibility variations could be as high as 8 dB in the right ear of the HATS. The effect was due to an unreliable tight fit that could occur in this ear. The other, left, ear of the HATS had a shape that prevented most earbuds from achieving a tight fit; this improved reproducibility. Due to these findings, it was judged that an arithmetic average of five nominally identical sequential measurements could lead to bias errors. To reduce this potential, three measurements with a normal fit were made over a time span of a few hours by being interspersed with the other measurements from this study. Three normal fit measurements were obtained for each ear. The three measurements were arithmetically averaged and the results for left and right ears reported separately, then all six measurements were arithmetically averaged and this result was also reported. Under the specified measurement conditions, the typical reproducibility of these Leq values was ±3 dB.

In addition, one measurement of each of the tight and loose fit conditions on both ears was also made and reported separately. In particular, the tight and loose fit conditions were assumed to characterize the range of levels that could be measured on the HATS.

I. DAP maximum output voltage, $V_{out,max}$

The IEC signal for typical program material was used in testing $V_{out,max}$, specified in BS EN 50332-2 (British Standards Institution, 2004). This voltage was obtained from each DAP by simulating the electrical load of the earphones using a precision 33 ohm resistor in parallel across the input. The un-weighted RMS voltage output was measured using the Bruel & Kjaer Pulse® 10 analyzer.

III. STATISTICS

A three-way analysis of variance was performed to analyze the data. The main independent variables were earphone Side (left versus right), Fit (normal, tight and loose) and Device (9 DAPs). The dependent variable was dBA SPL. All statistical analyses were performed using StatView® Version 5.0 and alpha levels less than 0.05 were considered statistically significant.

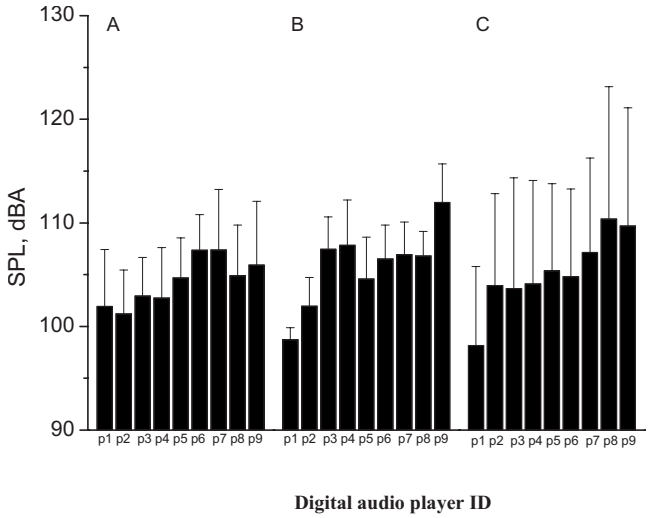


FIG. 1. Results for DAP with supplied earphones taken from averaged direct measurements for normal fits (panel A) and calculations (panel B) using $V_{out,max}$ and V_{WBCV} (normal fit). Panel C in Fig. 1 shows an average of all of the tight and loose fit results shown later in Fig. 3. Player ID taken from Table II, *p*1/*e*1; *p*2/*e*2; *p*3/*e*3; *p*4/*e*4; *p*5/*e*5; *p*6/*e*6; *p*7/*e*7; *p*8/*e*8; *p*9/*e*9.

IV. RESULTS

Table II provides a description of the DAP and earphones tested, including the: (i) DAP and earphone reference ID, (ii) DAP description, (iii) maximum DAP output voltage ($V_{out,max}$) as specified in BS EN 50332-2 (British Standards Institution, 2004) for separately sold components (iv) earphone sensitivity, V_{WBCV}, as specified in BS EN 50332-2 (British Standards Institution, 2004), (v) maximum equivalent free field SPL, (vi) high frequency insertion loss for external noise, and (vii) noise levels experienced by bystanders at a distance of 0.25 m when the earphone is producing 94 dBA at the ear.

For the earphones supplied with the DAP, the bars in Panel A in Fig. 1 are the arithmetic averages of the dB levels measured on the HATS (normal fits) with free field correction. Each data point is an average of six measurements, three on each ear. Measured values ranged from 101 to 107 dBA. A similar average is shown in Panel B, Fig. 1, which used separate measurements of the player $V_{out,max}$ and earphone V_{WBCV} (normal fits) to calculate the SPL. The calculated range of values shown in Panel B ranged from 99 to 112 dBA. Calculated and actual measurements were not statistically different ($F_{1,8}=2.166$, $p=0.18$) Panel C in Fig. 1 shows an average of all of the tight and loose fit results shown later in Fig. 3. The three sets of results are not statistically different ($p>0.05$). The measurements under normal fit conditions on the HATS were by chance almost exactly equal to the arithmetic averages of the tight and loose fits. Thus a tight fit can be estimated by adding 8 dB to the average values in Fig. 1, and a loose fit can be estimated by subtracting 8 dB. Calculated values of SPL from separate measurements of earphone and player voltages could be as much as 3 dB higher than directly measured SPL values since the output impedance on some DAP players was up to 7 ohms larger than the output impedance of the test amplifier

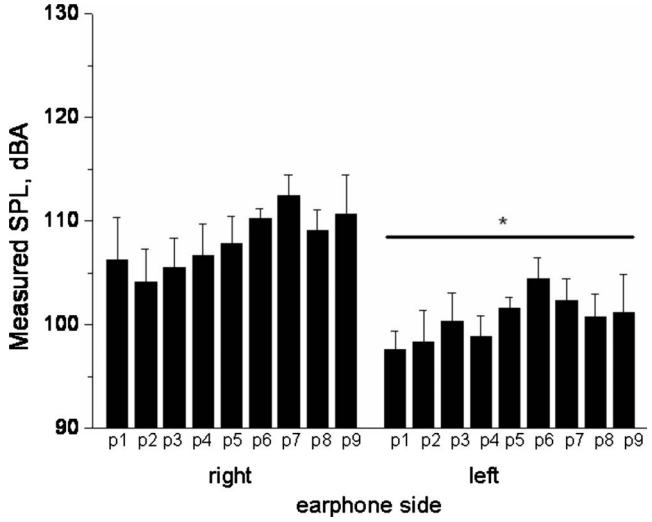


FIG. 2. Averaged measured results for DAP with supplied earphones, right ear; left ear. Error bars indicate one standard deviation. Player and earphone ID taken from Table II, $p1/e1$; $p2/e2$; $p3/e3$; $p4/e4$; $p5/e5$; $p6/e6$; $p7/e7$; $p8/e8$; $p9/e9$. * Statistically significant from right, $p<0.01$.

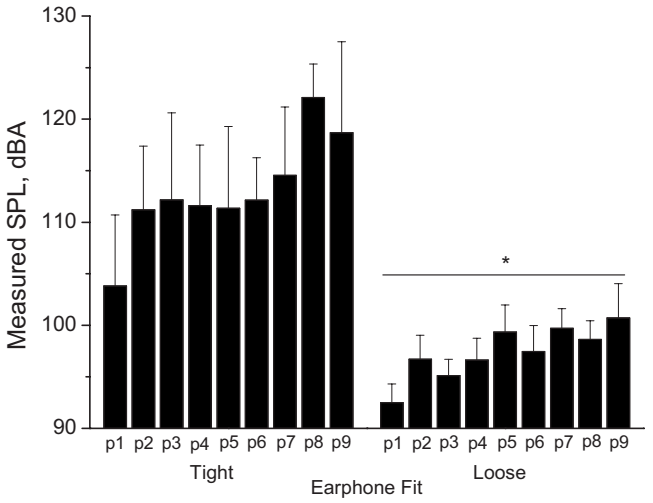


FIG. 3. Range of measured levels averaged over both ears for tight fit and loose fit. Error bars indicate one standard deviation. Player and earphone ID taken from Table II, $p1/e1$; $p2/e2$; $p3/e3$; $p4/e4$; $p5/e5$; $p6/e6$; $p7/e7$; $p8/e8$; $p9/e9$. * Statistically significant from tight fit, $p<0.0001$.

used to determine V_{WBCV}, and the earphone impedances could be as low as 16 ohms, half the value used in determining $V_{out,max}$.

Figure 2 shows the measured SPL results for each DAP with the supplied earphones showing the contribution of the left and right earphones separately. The error bars show one standard deviation, which was typically 3 dB. There was a main effect of Earphone Side with an approximate difference of 8 dB between the left and right ears (respective means, 108.07 dB and 100.6 dB), ($F_{1,8}=24.259$, $p<0.01$). There was also a significant main effect of DAP, ($F_{8,32}=7.726$, $p<0.0001$), but no interaction between Earphone Side and DAP ($p=0.3366$). Similar results were obtained when levels were calculated using separate measurements of the player $V_{out,max}$ and earphone V_{WBCV} to calculate SPL, and were also found in our earlier studies of compact disc players (Keith *et al.*, 1999). The maximum measured level in these trials was 115 dBA using DAP ID $p9$ with the earbud in the right ear and the minimum measured level was 99.6 dBA in the left ear using the DAP ID $p1$.

During the fitting procedure with the earbuds, a tighter fit was obtained in the right ear of the HATS. There was difficulty in maintaining the earbud position in the cavum of the left ear, and decreased low frequency output was noted compared to the right ear (see, for example, Keith *et al.*, 1999). Due to the right left asymmetry of both the pinnae and earphones, any confounding effect of potential level differences between right and left earphones could not be quantified.

Figure 3 shows the range of measured levels found for a tight and loose fit. The error bars are one standard deviation, which were typically 6.5 dB for a tight fit and 2.2 dB for a loose fit. There was a main effect of fit, owing to the observation that when data from all the players was pooled, the tight fit was approximately 16 dB higher in level than the loose fit ($F_{1,8}=169.6$, $p<0.0001$). The tight fit not only produced higher sound pressures, but compared to the loose fit could increase the bass response (below 125 Hz) anywhere from 2 to 12 dB, depending on the earphone. The maximum measured level with a tight fit was 120.4 dBA using DAP ID $p8$ with the earbud in the right ear and the minimum measured level was 93.4 dBA in the left ear using the DAP ID $p1$.

Figure 4 shows a comparison of measurements using free field corrected data from the HATS or ear simulator and uncorrected levels using a bare microphone or microphone held in the hand, as described above. Placing the earphone on top of the bare microphone resulted in a close match to the loose fit on the HATS in the tests of earbud and supra aural earphones. However, this type of measurement was lower than the sound level of a normal fit in the right ear by about 7 dB (results were slightly improved if right and left

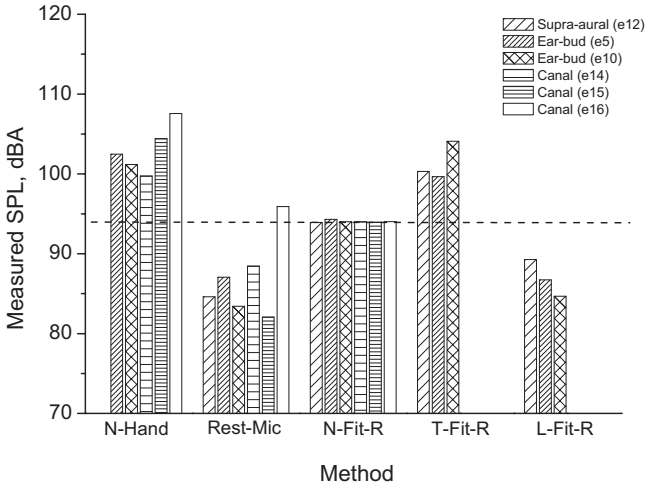


FIG. 4. Comparison of earphone measurements of free field response corrected values using a fixture (ear simulator or HATS) and uncorrected hand held microphone measurements. The dashed line shows the value for a normal fit on the fixture of 94 dBA. An ear simulator was used for the canal phones, and a HATS was used for the other earphones. Normal fit sealed against cupped hand (N-Hand); earphone resting on bare microphone (Rest-Mic); Normal fit fixture-right (N-Fit-R); Tight fit fixture-Right (T-Fit-R); Loose fit fixture right (L-Fit-R). Earphone ID taken from Table II.

ear were averaged together). Applying the HATS correction for the free field response to the bare microphone results increases the difference from the HATS normal fit measurements to approximately 16 dB.

The figure also shows that for these earphones, a "normal fit" using a cupped hand to create a seal (see Sec. G above) was higher than obtained in the HATS by 9 dB on average. Loose and tight fits could also be simulated using the hand, the same 16 dB range obtained in loose and tight fitting on the HATS was produced using a cupped hand. Correction for the free field response improved results so that the loose, normal and tight fit in the hand (data not shown) roughly matched the corresponding fit on the HATS. Best results were obtained with the microphone close to the earphone. Attempts to simulate an ear canal using the fingers to form a tube produced weak resonances which had little effect other than to make results less repeatable. When earphone levels measured on the HATS were set to 94 dBA, variations in measurements with either a bare microphone (Rest-Mic) or a cupped hand (N-Hand) yielded standard deviations of approximately 4 dB when calculated across the different earphone types. The close agreement between the measurements on the HATS and the earphone fitting in the cupped hand may be due to the experience of the operator and his ability to both hear the earphone output using headphones and monitor the measured spectra in real time.

V. DISCUSSION

In this study, following the European standards (British Standards Institution, 2000; British Standards Institution, 2004), the level produced by the tested units ranged from 101 to 107 dBA, with 104 dBA as the arithmetic average for the tested units. With a tight fit, the highest measured value was 120.4 dBA so that instantaneous traumatic damage to the ears appears to be unlikely under our test conditions. A permanent threshold shift could potentially occur if this sound level was maintained continuously for about 10 min (Ward, 1991). It should, however, be emphasized that a sound level of 107 dBA for about 5 min exceeds the common occupational limit, of 85 dBA for an 8 h day with a 3 dB exchange rate (Canadian Centre for Occupational Health and Safety, 2007).

All nine of the one package units (including earphones) exceeded the 100 dB SPL limit specified in French regulations (Legifrance, 2005). All nine DAPs failed the $V_{out,max}$ <150 mV criterion of the French regulation for separate components. It was found that $V_{out,max}$ varied from 201 to 639 mV, up to 12.6 dB higher than the regulated level. Among the earphones tested, the average sensitivity met the BS EN Standard (British Standards Institution, 2004) recommendation that $V_{WBCV}>75$ mV for 13 of 17 newly purchased earphones and for all but one of the previously tested earphones (Bly *et al.*, 1998).

These results show approximately a 5 dB increase in arithmetically averaged levels when compared to a study of portable CD players made 7 years earlier (Keith *et al.*, 1999; 2001). In that study, sound levels obtained using the IEC signal with portable CD players ranged from 91 to 107 dBA. The average level in the previous study was lower because six of the eight CD players were packaged with supra aural headphones which produced lower average levels than the earbud earphones packaged with the two loudest CD players. The average level of the top-ten pop music increased by only 1 dB compared to the previous study. However, the loudest current popular song in the present study was 4 dB louder than any of the previously tested top-ten music tracks and comparable to the loudest supplemental heavy metal track used in the previous studies (Keith *et al.*, 1999; 2001).

For the recently purchased earphones and players, this study showed the range of Leq values could vary due to differences in each of the following factors: DAP player (10 dB), earphone sensitivity, V_{WBCV} (14 dB), top-ten music track (10 dB), and fit of the earphone (16 dB). Taken together, this represents possible Leq values between 79 and 125 dBA. The wide variation in sound levels found in this study suggests that, to manage this risk, one option may be education of users about subjective criteria for limiting noise exposure.

Discussions regarding the insertion loss testing, ear phone fit, free field response correction, and safe volume settings are below.

A. Insertion loss testing in the presence of background noise

In high levels of environmental background noise, users may increase the playback levels of DAP players to levels that are hazardous to hearing. In such an environment, canal earphones could allow listening levels that pose no risk to hearing as this study showed measured high frequency insertion loss between 13 and 38 dB above 700 Hz. None of the other types of earphones were very effective above 700 Hz in reducing external noise. This was also demonstrated by Fligor and Ives at a variety of background noise levels (Fligor and Ives, 2006). At a background level of 80 dBA, that study found the average listening level was 7 dB below the background noise using a canal phone similar to *e*17, while the listening level was 7 dB higher than the background noise with an earbud similar to the *e*5 earbud.

However, for safety, the combination of listening levels and background noise reduction must not reduce the awareness of warning signals. With the exception of the canal earphones, most tested earphones would not significantly attenuate warning signals.

It was assumed that noise reducing headphones or earphones would allow listening at lower volume settings when background noise is dominated by low frequencies. Unfortunately the tests in this study were inadequate for this purpose, and formal subjective testing would be required to make such a conclusion.

B. Measurement method, earphone fit

The largest uncontrolled source of variation in earphone measurements appears to be related to the fit, i.e., seal, of the earphone on the ear. For headphones, the force on the pinna is measured to evaluate this parameter; however, for other earphone types there is no comparable parameter. The effect

of the fitting procedure was demonstrated by the variation in earphone sound levels and the high variation found in NRR values for muff-type hearing protectors (i.e., *e*30 and *e*31). The importance of the tightness of fit was noted even with supra-aural earphones, which merely sit on top of the pinna. This suggests that the need to describe the seal would be advised when measuring any device used close to the ear (Health Canada, 2006a; ISO, 2000).

Regarding headphone/earphone fit, Part I of the BS EN standard only states that intra-concha earphones shall be positioned to fit normally. For supra-aural and circumaural headphones, fitting is so that the measured sound level is maximized. In Part II, for all headphones/earphones, positioning is to be done correctly, taking into account manufacturer's instructions, while maximizing the sound level. Depending on the pinna morphology, hardness (Keith *et al.*, 2001) (not specified in the BS EN standards) and the user's interpretation of the standard's instructions, the results of this study suggest that a 16 dB or more variation could occur in reported levels.

Reasons for selection of conditions that could lead to a tight or loose fit are indicated in Sec. II above. In use, the exposure from an earphone will depend on the geometry of the ear and earphone, as well as user preferences. Figure 2 shows that a natural loose fit is usually obtained for earbud earphones in the left ear of our HATS, apparently due to the shape of the left concha. Conversely, a tight fit is associated with the highest possible levels and improved bass response. These factors could bias user's earphone preference to those giving tightest fit, either at the time of purchase or through manual pressure (during a favored track).

C. Free-field response correction

The next largest effect on the results was due to the correction for free field response (related to the ear canal resonance around 3 kHz). This correction, an approximate 7 dB reduction, when applied to the hand-held measurements brought them into closer agreement with the HATS results.

For the canal phones, fitting was not a large source of variation, but it was necessary to use the free field response correction. Neglecting this correction could overestimate the levels by up to 15 dB using an ear simulator or hand-held measurement.

D. Safe volume settings

Despite the variability found in this study, all of the players had options through supplied software or external programs to allow volume equalization so that all music would play at comparable levels. This can reduce the possibility of unintentional overexposure by eliminating the need to change volume settings (e.g., between pop music and orchestral soundtracks). This software did not appear to use an A-weighted Leq as would be preferable for the protection of hearing.

Other investigations (Portnuff and Fligor, 2006) have suggested a 50%–60% volume setting is safe and according to one study (Ahmed *et al.*, 2007) this would appear to be the level at which most people listen to their DAP. The range of levels measured by Portnuff and Fligor was approximately 5 dB between players and 10 dB due to earphones, suggesting a specific safe volume setting is feasible. The present study included a wider range of DAP's and earphones, increasing the possible range of levels to 40 dB. This much wider range brings into question the usefulness of a specific safe volume setting. The range arises from the following contributions; 16 dB due to fit, 14 dB due to recent model earphone V_{WBCV} ($e1-e17$) and 10 dB due to player, (this could increase a further 30 dB if the quietest measured music in Table I was included). Potentially, there could be three ways to minimize these effects on the determination of a safe volume setting: (i) the industry standardizes a safe volume setting, (ii) the user has the output measured, or (iii) the user relies on subjective criteria (described below). In all cases, if a new set of earphones are used with a different V_{WBCV} then a new safe volume setting would have to be determined. The fit during testing should match that on the user as closely as possible since Leq values were observed to vary in a range of ± 8 dB. This range could be narrowed with a more reproducible fit, such as a loose fit with earbuds, or the tight fit with the canal phones in this study.

The third method using subjective criterion to determine a safe volume setting has been used in the past on its own. In two *It's Your Health* public information documents published by Health Canada (Health Canada, 2005; 2006b), the suggested subjective approach to gauging safe sound levels included the following advice:

"If someone standing a metre away from you has to shout to be understood, the sound levels around you probably exceed 85 dBA. You face a significant risk of permanent hearing loss if you are exposed to these sound levels for eight hours or more per day."

"If someone standing 30 cm away has to shout to be understood, the levels probably exceed 95 dBA. This means a significant risk of permanent hearing loss if you are exposed for about 45 minutes or more per day."

Using the subjective criterion also helps to ensure that warnings can be heard. In any potentially hazardous environment, earphone levels must always be adjusted so that the user is aware of external noise and can hear warning signals.

However, it seems reasonable to expect an improvement in the determination of a safe volume setting if methods (i) and (ii) above are feasible and the subjective criterion could be relegated to acting only as a check. This stems from the fact that there are additional uncertainties of the subjective definition of "shout," as well as the time and frequency variation in music from a steady noise, for which the result was determined, as well as variation from person to person in their ability to understand spoken communication. The subjective method would also be more difficult to implement with tight fitting canal phones, which were found to significantly attenuate external noise.

The wide range of bystander sound levels in Table II measured at 0.25 m from the earphone showed that an external observer is unlikely to be able to judge what represents a safe volume level. Assuming a DAP user is listening at 85 dBA, the bystander level at 0.25 m for the e8 earbud

would be 45 dBA, which would be clearly audible in many private indoor settings. Perception of intermittent musical notes would also be possible in a variety of noisy environments. For example, on an idling city bus with an interior sound level of 60 dBA with a 5 dB per octave slope, the level in the 4 kHz octave band was measured at about 43 dB (data not shown). This is roughly the same as the bystander level for earbuds *e*8 at an 85 dBA listener level. Therefore, it is plausible that the closest bystanders would be aware of intermittent high frequency sounds from these earphones. However, when background environmental noise is high (e.g., 85 dB) and the DAPs are readily audible to an external observer, it is reasonable to expect that the DAP user is experiencing potentially unsafe sound levels.

VI. CONCLUSIONS

In the past seven years, the sound levels at maximum volume of top-ten music from personal music players have increased by as much as 5 dB. For the normal fit, playback levels ranged from 101 dBA to 107 dBA at maximum volume level. This means that Canadian occupational noise limits could be exceeded after listening to only a single track per day. Also, for comparison, the World Health Organization (1999) guideline levels state that patrons at entertainment events should not be exposed to more than 100 dBA for 4 h, more than four times per year, to avoid any known risk of hearing loss. The highest level measured with a tight fit was 120.4 dBA. A permanent threshold shift could occur if this sound level was maintained continuously for about 10 min (Ward, 1991).

Despite the high measured levels found in this study, usage patterns reported elsewhere suggest that hearing loss is unlikely to occur for the majority of users. However, a minority of users appears to be at risk (Ahmed *et al.*, 2007; Williams, 2005; Hodgetts *et al.*, 2007; Airo *et al.*, 1996). Given the large population of users and the increasing potential for longer listening times, further research on listening habits may be warranted.

The testing reported here extends the earlier work on CD players (Keith *et al.*, 1999; 2001) and continues to find large variations in sound level occurring as a result of variations in earphone fit, program material, DAP $V_{out,max}$, and earphone sensitivity. Also, as noted previously, for a given DAP, there was a tendency for sound levels to be significantly greater if intra-concha earbud earphones were used.

The study shows that there are potential benefits of further development of methods to standardize safe volume settings (i.e., with reference to one or a few benchmark listening durations). The determination of safe volume setting ultimately also depends on acceptable level of risk.

Of the earphones in this study, proper fitting canal phones were the only earphones that were effective at reducing external noise. Reduction of external noise could help reduce the potential to listen at levels that are hazardous to hearing when background noise is high. However, such large reductions in external noise could also reduce the ability to hear and be aware of warning sounds, thus increasing the risk of potentially serious accidents. In any potentially hazardous environment, earphone levels must always be adjusted so that the user is aware of external noise and can hear warning signals.

This study suggests an interpretation for levels obtained in field testing using a sound level meter as compared to HATS measurements. A bare microphone underestimates sound levels by about 8 dB. Simulation of the fit using the hand overestimates levels by approximately 9 dB, with up to ±8 dB variation due to the tightness of fit. The most accurate field test methodology requires measurement in 1/3 octave or narrower bands, correction for the frequency response of the ear, and simulation of the fit using the hand. An experienced operator might be able to estimate results on a HATS with a standard deviation as low as 4 dB. For an inexperienced operator, the added difficulty in defining a loose and tight fit on a cupped hand could make the standard deviation twice as large. It is not clear how to quantify the fitting on the HATS, and it is even more difficult to quantify the fitting with a hand held measurement.

Nomenclature

CD = compact disc
DAP = digital audio player
dB = decibel
dBA = decibel A-weighted
HATS = head and torso simulator system
IEC = International Electrotechnical Commission
ISO = International Standardization for Organization
Leq = energy average equivalent sound level
mV = millivolts
MP3 = MPEG-1 audio layer 3
N = newtons
NRR = noise reduction rating
NIOSH = National Institute for Occupational Health and Safety
RMS = root mean square
SPL = sound pressure level
V = volts
$V_{out,max}$ = maximum output voltage
V_{WBCV} = earphone wideband characteristic voltage

[1]Although a loose fit also appeared to be plausible for canal phones, attempts to define the method in a reproducible way were unsuccessful.

Ahmed, S., Fallah, S., Garrido, B., Gross, A., King, M., Morrish, T., Pereira, D., Sharma, S., Zaszewska, E., and Pichora-Fuller, K. (**2007**). "Use of portable audio devices by university students," Can. Acoust. **35**, 35–52.

Airo, E., Pekkarinen, J., and Olkinuora, P. (**1996**). "Listening to music with earphones: An assessment of noise exposure," Acta Acust. **82**, 885–894.

Aono, S., Ohta, T., Yama-naka, N., Kudo, T., and Takagi, K. (**1997**). "Listening level of music from the personal stereo player and the effect on hearing in terms of TTS," J. Acoust. Soc. Jpn. **53**, 440–447.

Bly, S., Keith, S., and Hussey, R. (**1998**). "Sound levels from headphone/portable compact disc player systems," Proceedings of Canadian Acoustical Association **26**, 74–75 (London, Ontario).

British Standards Institution (**2000**). "Sound System Equipment-Headphones and earphones associated with portable audio equipment-Maximum sound pressure level measurement methodology and limit considerations-Part 1: General method for one package equipment," Standard No. BS EN 50332-1:2000.

British Standards Institution (**2004**). "Sound system equipment: Headphones and earphones associated with portable audio equipment. Maximum sound pressure level measurement methodology and limit considerations Part 2: Matching of sets with headphones if either or both are offered separately," Standard No. BS EN 50332-2:2003.

Canadian Centre for Occupational Health and Safety (**2007**). "What are the occupational exposure limits for workplace noise?," (accessed on 07/06/2007) http://www.ccohs.ca/oshanswers/phys_agents/exposure_can.html

Felchlin, I., Hohmann, B. W., and Matefi, L. (**1998**). "Personal cassette players: A hazard to hearing?," In *Protection Against Noise*, Vol. 2, edited by D. Prasher, L. Luxon, and I. Pyykko (Whurr, London), pp. 95–100.

Fligor, B. J., and Ives, T. E. (**2006**). "Does earphone type affect risk for recreational noise-induced hearing loss?," NIHL in Children Meeting, Cincinnati, OH http://www.hearingconservation.org/docs/virtualPressRoom/FligorIves.pdf (accessed 01/23/2008).

Health Canada (**2006a**). "Industry Guide to Canadian Safety Requirements for Children's Toys and Related Products, 2006," http://www.hc-sc.gc.ca/cps-spc/pubs/indust/toys-jouets/mechanical-mecaniques_html (accessed on 09/10/2007).

Health Canada (**2006b**). "Personal Stereo Systems and the Risk of Hearing Loss," http://www.hc-sc.gc.ca/iyh-vsv/life-vie/stereo-baladeur_e.html (accessed on 09/10/2007).

Health Canada (**2005**). "Hearing Loss and Leisure Noise," http://www.hc-sc.gc.ca/iyh-vsv/environ/leisure-loisirs_e.html (accessed on 09/10/2007).

Hodgetts, W. E., Rieger, J. M., and Szarko, R. A. (**2007**). "The effects of listening environment and earphone style on preferred listening levels of normal hearing adults using an MP3 player," Ear Hear. **28**(3), 290–297.

IEC (**1985**). "Sound system equipment. Part 1: General. Report No. IEC 60268-1."

IEC (**1996**). "Sound system equipment-Part 7: Headphones and earphones. Report No. IEC 60268-7."

ISO (**1990**). "Acoustics-determination of occupational noise exposure and estimation of noise-induced hearing impairment," ISO 1999.

ISO (**2000**). "Safety of toys-Part 1: Safety aspects related to mechanical and physical properties," ISO 8124-1.

ISO (**2004**). "Acoustics-Determination of sound immission from sound sources placed close to the ear-Part 2: Technique using a manikin," ISO 11904-2.

Keith, S., Bly, S., Chiu, V., and Hussey, R. (**1999**). "Sound levels from headphone/portable compact disc player," *Inter-Noise Proceedings*, Fort Lauderdale.

Keith, S., Bly, S., Chiu, V., and Hussey, R. G. (**2001**). "Sound levels from headphone/portable compact disc player systems III," *Inter-Noise Proceedings*, Haia, Holanda.

Legifrance (**2005**). "Arréte du 8 novembre 2005 portant application de l'article L. 5232-1 du code de la santé publique relatif aux baladeurs musicaux," [Order of November 8, 2005 on the application of article L. 5232-1 of the Public Health Code on portable music] L. 5232-1.

Nielsen Broadcast Data Systems (**2007**). "Canada Top 40," http://www.bdsonline.com/about.html (accessed on 06/27/07)

NIH (**1990**). "Noise and Hearing Loss. NIH Consens Statement," Jan. 22–24; **8**(1), 1–24.

NIOSH (**1998**). "Occupational noise exposure, revised criteria," U.S. Department of Health and Human Services.

OSHA (**2006**). Noise Standard Code of Federal Regulations, Title 29, Part 1910.95 (c)(1), Subpart G.

Portnuff, C. D. F., and Fligor, B. J. (**2006**). "Sound output levels of the iPod and other mP3 players: Is there a potential risk to hearing?," NIHL in Children Meeting, Cincinnati, OH http://www.hearingconservation.org/docs/virtualPressRoom/portnuff.htm (accessed 07/30/2007).

TNO (**1998**). "Pop music through headphones and hearing loss," TNO Report No. 98. 036.

US EPA (**1979**). "Noise Labeling Requirements for Hearing Protectors," United States Environmental Protection Agency, Fed. Regist. 44(190), 40CFR Part 211, 56130-56147.

Ward, E. D., Cushing, E. M., and Burns, E. M. (**1976**). "Effective quiet and moderate TTS: Implications for noise exposure standards," J. Acoust. Soc. Am. **59**, 160–165.

Ward, W. D. (**1991**). "Hearing loss from noise and music," Proceedings Audio Eng. Soc., 1–8, New York.

Williams, W. (**2005**). "Noise exposure levels from personal stereo use," Int. J. Audiol. **44**, 231–236.

World Health Organization (**1999**). "Guidelines for Community Noise," WHO, Geneva.

Eigenvalue equalization filtered-x algorithm for the multichannel active noise control of stationary and nonstationary signals

Jared K. Thomas
Department of Mechanical Engineering, Brigham Young University, 435 CTB, Provo, Utah 84602

Stephan P. Lovstedt
Department of Physics and Astronomy, Brigham Young University, N283 ESC, Provo, Utah 84602

Jonathan D. Blotter
Department of Mechanical Engineering, Brigham Young University, 435 CTB, Provo, Utah 84602

Scott D. Sommerfeldt
Department of Physics and Astronomy, Brigham Young University, N283 ESC, Provo, Utah 84602

(Received 9 August 2007; revised 8 February 2008; accepted 6 March 2008)

The FXLMS algorithm, which is extensively used in active noise control, exhibits frequency dependent convergence behavior. This leads to degraded performance for time-varying and multiple frequency signals. A new algorithm called the eigenvalue equalization filtered-*x* least mean squares (EE-FXLMS) has been developed to overcome this limitation without increasing the computational burden of the controller. The algorithm is easily implemented for either single or multichannel control. The magnitude coefficients of the secondary path transfer function estimate are altered while preserving the phase. For a reference signal that has the same magnitude at all frequencies, the secondary path estimate is given a flat response over frequency. For a reference signal that contains tonal components of unequal magnitudes, the magnitude coefficients of the secondary path are adjusted to be the inverse magnitude of the reference tones. Both modifications reduce the variation in the eigenvalues of the filtered-*x* autocorrelation matrix and lead to increased performance. Experimental results show that the EE-FXLMS algorithm provides 3.5–4.4 dB additional attenuation at the error sensor compared to normal FXLMS control. The EE-FXLMS algorithm's convergence rate at individual frequencies is faster and more uniform than the normal FXLMS algorithm with several second improvement being seen in some cases.
[DOI: 10.1121/1.2903857]

PACS number(s): 43.50.Ki, 43.40.Vn [KAC] Pages: 4238–4249

I. INTRODUCTION

An active noise control (ANC) system relies on the theory of superposition of sound waves—propagating waves can constructively and destructively interfere to either increase or decrease the sound, respectively. Applications of ANC are widespread but can, in general, be categorized into two types of controllable signals: signals which are stationary in time and signals which are nonstationary or time varying. Signals of both types may be single frequency, multiple frequency, broadband, or some combination of these three. The most common control approach for the ANC of these signals is based on some version of the filtered-*x* least mean squares (FXLMS) algorithm.[1,2] The FXLMS algorithm has proven successful for applications such as single frequency noise in a duct,[2] broadband noise in an enclosure,[3] multiple frequency noise in a helicopter,[4] and time-varying frequency noise in a tractor.[5]

One of the limitations of the FXLMS algorithm is that it exhibits frequency dependent convergence behavior that can lead to a significant degradation in the overall performance of the control system. The performance degradation is evident for the case of noise that is time varying, such as that of a tractor engine, where the frequency changes as the speed of the engine, in rpm, changes during operation. If the frequency associated with the engine speed changes faster than the algorithm can converge and attenuate that particular frequency, then performance of the ANC system will be degraded. The degradation is also evident for the case of stationary multiple frequency noise, such as that in a helicopter, where multiple harmonics of the engine, tail rotor, and main rotor can be controlled. Poor performance is expected at frequencies where the convergence of the algorithm is slow. The frequency dependent problem is not manifested for stationary single frequency noise, as optimal performance is still possible by the correct selection of the convergence parameter μ. Since μ also exhibits frequency dependence, optimal performance by selection of the correct μ is not guaranteed for the case of multiple stationary frequencies and time-varying frequencies.

Solutions to the frequency dependent problem for multiple stationary frequency noise have been proposed such as the higher harmonic filtered-*x* (HLMS) algorithm by Clark and Gibbs,[6] similar work by Lee *et al.*,[7] and the modified FXLMS algorithm.[8] The drawback of these approaches is that they add complexity and computational burden to the algorithm. The work of Kuo *et al.*[9,10] suggested a relatively simple solution for the case of internally generated sinusoids. More of their work will be discussed at a later point. For the

 0001-4966/2008/123(6)/4238/12/$23.00

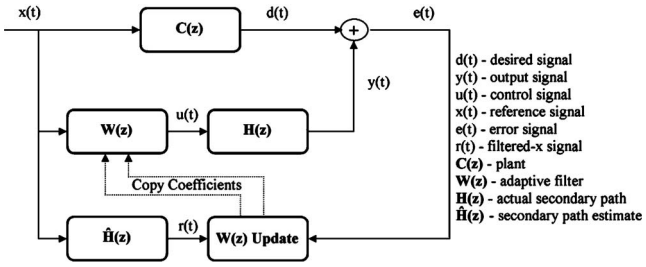


FIG. 1. Block diagram of the FXLMS algorithm.

case of time-varying frequencies, the filtered-x gradient adaptive lattice (FXGAL) algorithm by Vicente and Masgrau[11] improves the convergence behavior when an acoustic reference signal is used at the expense of computational complexity. For the case of a single time-varying frequency, the normalized FXLMS can be an effective solution.

This paper will discuss two simple approaches for dealing with noise characterized as multiple stationary or time-varying frequencies, which largely overcomes this frequency dependent performance and improves the overall performance of the ANC system. These approaches are appropriate for both single and multiple channel controls, are relatively simple to implement, and do not increase the computational burden of the algorithm. The effectiveness of these approaches will be experimentally demonstrated.

II. BACKGROUND

For this research, a feedforward multiple channel implementation of the FXLMS algorithm is used, which relies on a reference signal being "fed" forward to the control algorithm so that it can predict in advance the control signal needed to attenuate the unwanted noise. A feedforward implementation of the FXLMS algorithm involves adaptive signal processing to filter the reference signal in such a way that the measured residual noise is minimized. The measured residual is called the error signal and for this research it will be measured as an energy density (ED) quantity. The advantages of an ED based FXLMS algorithm[12] for noise in an enclosure[3,13] and for the application of tractor engine noise[5,14] are well documented. For simplicity in developing the control approaches, a brief derivation of the general FXLMS algorithm for a single channel is given. The extension of the approaches for multiple channel control[15] is straightforward. The use of an ED based FXLMS is also straightforward and well documented in Ref. 12.

A. Single channel FXLMS

The goal of the FXLMS algorithm is to reduce the mean-squared error of the error signal at a location where the sound is to be minimized. Boucher *et al.*[16] provided a good reference for the derivation of the single channel FXLMS algorithm, which is shown in block diagram form in Fig. 1. In the figure and in all equations presented, the variable t is used as a discrete time index and the variable z is used as a discrete frequency domain index. Signals in the time domain are represented as lower case letters, while capital letters are used in the frequency domain. Vectors in each domain are represented as bold letters.

The mean-squared error is a quadratic function (a "bowl") with a unique global minimum. For each iteration of the algorithm, $\mathbf{W}(z)$, which is represented as an adaptive finite impulse response (FIR) control filter, takes a step of size μ, the convergence coefficient, times the gradient in search of a single global minimum that represents the smallest attainable mean-squared error. The control filter update equation for $\mathbf{W}(z)$ can be expressed in vector notation as

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \mu e(t)\mathbf{r}(t), \tag{1}$$

where $e(t)$ is the error signal and $\mathbf{r}(t)$ and $\mathbf{w}(t)$ are defined as

$$\mathbf{r}^T(t) = [r(t), r(t-1), \ldots, r(t-I+1)], \tag{2}$$

$$\mathbf{w}^T(t) = [w_0, w_1, \ldots, w_{t-1}]. \tag{3}$$

The filtered-x signal $r(t)$ is the convolution of $\hat{\mathbf{h}}(t)$, which is the estimate of the impulse response of the secondary path transfer function, and $x(t)$, which is the reference signal. The secondary path transfer function includes the effects of digital-to-analog converters, reconstruction filters, audio power amplifiers, loudspeakers, the acoustical transmission path, error sensors, signal conditioning, antialias filters, and analog-to-digital converters. The reference signal contains information correlated with the unwanted noise that the ANC system will target when control is enabled.

B. Secondary path transfer function

One difficulty in implementing the FXLMS algorithm is that the secondary path, which is represented as $\mathbf{H}(z)$ in Fig. 1, is unknown. An estimate, $\hat{\mathbf{H}}(z)$, of the secondary path must be used. The estimate is obtained through a process called system identification (SysID).

The SysID process is performed either online (while ANC is running) or offline (before ANC is started). For the fastest convergence of the algorithm, an offline approach is used. The offline SysID process is performed before ANC is started and consists of playing white noise through the control speaker(s) and measuring the output at the error sensor. The measured impulse response is obtained as a FIR filter $\hat{\mathbf{h}}(t)$ that represents $\hat{\mathbf{H}}(z)$. The coefficients of $\hat{\mathbf{h}}(t)$ are stored and used to run control. For multiple channel control, there is an $\hat{\mathbf{h}}(t)$ estimate for every error sensor and control speaker combination. Each is obtained in turn through the SysID process.

C. Reference signal

The reference signal may be an acoustic signal (e.g., from a microphone) or a nonacoustic signal (e.g., a tachometer signal from an engine) depending on the control application. Generally, the reference signal will be either stationary or time varying. Signals of either type may be single frequency, multiple frequency, broadband, or some combination of these three.

Significant signal conditioning may be required to get the reference signal in a form suitable for control. For ex-

ample, for control of engine noise, a tachometer signal related to the engine speed (in rpm) is typically used as the reference signal. The tachometer signal is usually a multiple or some fraction of the engine firing frequency and must be filtered and passed through a frequency multiplier to be directly used as the reference signal. If harmonics are also targeted for control, they too are usually generated either in hardware or software from the fundamental frequency. Where multiple noise sources are present, a reference signal may be obtained for each and combined into a single reference. The resulting signal will, in general, have varying magnitude at the various tonal components.

III. FREQUENCY DEPENDENT CONVERGENCE BEHAVIOR

The inclusion of $\hat{\mathbf{H}}(z)$, while necessary for algorithm stability, degrades performance by slowing the algorithm's convergence. One reason for the decreased performance is the delay associated with $\hat{\mathbf{H}}(z)$. For many ANC applications, such as enclosures of less than a few meters, the delay is on the order of 10 ms or less and convergence is still rapid.[17] A more significant problem is that the inclusion of $\hat{\mathbf{H}}(z)$ causes a frequency dependent convergence behavior. The frequency dependent behavior can be better understood by looking at the eigenvalues of the autocorrelation matrix of the filtered-x signal, which is a function of $\hat{\mathbf{H}}(z)$ and $\mathbf{X}(z)$.

A. Eigenvalues

The eigenvalues of the autocorrelation matrix of the filtered-x signal relate to the dynamics or time constants of the modes of the system. Typically, a large spread is observed in the eigenvalues of this matrix, which corresponds to fast and slow modes of convergence. The slowest modes limit the performance of the algorithm because they determine the overall convergence of the algorithm to the optimum. The fastest modes have the fastest convergence and the greatest reduction potential but limit how large of a convergence parameter μ can be used.[18] For stability, μ is set based on the slowest converging mode (the maximum eigenvalue), which leads to degraded performance. If μ is increased, the slower modes will converge faster, but the faster modes will drive the system unstable.

The autocorrelation matrix of the filtered-x signal is defined as

$$E[\mathbf{r}(t)^{*}\mathbf{r}^{T}(t)], \tag{4}$$

where $E[]$ denotes the expected value of the operand, which is the filtered-x signal vector $\mathbf{r}(t)$ multiplied by the filtered-x signal vector transposed $\mathbf{r}^{T}(t)$. In general, it has been shown that the algorithm will converge and remain stable as long as the chosen μ satisfies the following equation:[16]

$$0 < \mu < \frac{2}{\lambda_{\max}}, \tag{5}$$

where $\lambda_{\max}$ is the maximum eigenvalue of the autocorrelation matrix in the range of frequencies targeted for control.

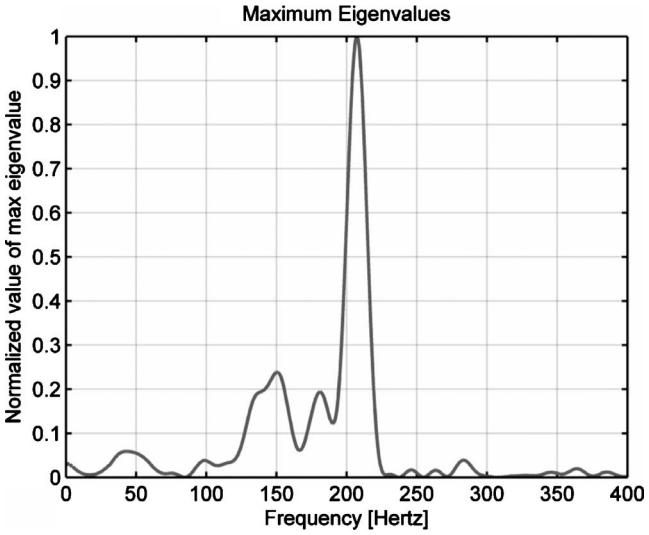


FIG. 2. Plot of normalized maximum eigenvalues over frequency—equally weighted reference signal.

In practice, it is computationally demanding to obtain a real-time estimate of the autocorrelation matrix so the optimal μ is often selected through experimentation. An offline estimate of the autocorrelation matrix is made by taking an actual $\hat{\mathbf{H}}(z)$ model from a mock cabin and importing it into a numerical computer package. If a single frequency reference signal is used, $\lambda_{\max}$ can be computed for that frequency. If the simulation is repeated over a range of frequencies, $\lambda_{\max}$ for each frequency can be found. Figure 2 shows an offline simulation using an actual $\hat{\mathbf{H}}(z)$ from a mock cabin and equal amplitude tonal inputs from 0 to 400 Hz. The eigenvalues in the figure have been normalized to the largest eigenvalue in the range. In this eigenvalue simulation, and all others in this paper, the results are shown for only a single $\hat{\mathbf{H}}(z)$ for a single channel. The results for multiple channel and/or ED control [where the filtered-x signal is a combination of each $\hat{\mathbf{H}}(z)$ component of each channel] follow a similar trend as the single channel case and so, to facilitate explanation of the concepts presented in this paper, only a single $\hat{\mathbf{H}}(z)$ for a single channel is shown.

Figure 2 illustrates the frequency dependent behavior. The largest eigenvalue occurs at about 208 Hz. This location corresponds to the smallest stable μ in the frequency range from 0 to 400 Hz, as given by Eq. (5). Most other frequencies have a smaller eigenvalue and could use a larger μ and still be stable, if just that particular frequency was targeted for control. Frequencies at the valleys of the figure have the smallest eigenvalues and could use the largest μ's and still be stable, again if they were the only frequencies targeted for control. The larger μ's are desirable as they lead to faster convergence and increased attenuation.

If the frequency range for control is 0–400 Hz, the μ associated with 208 Hz (the smallest in the range) must be used for stability. If, for example, 100 Hz was the only tone targeted for control, a μ larger than the 1 used at 208 Hz could be used and convergence would be faster. If both 100 and 208 Hz were targeted for control, the smaller μ associated with 208 Hz must be used for stability and degraded

performance at 100 Hz is expected. In summary, because the μ associated with the largest eigenvalue in the range of frequencies targeted for control must be used for stability, degraded performance is expected at the other frequencies in the range that would benefit from the use of a larger μ.

IV. EIGENVALUE EQUALIZATION

If the variance in the eigenvalues of the autocorrelation matrix was minimized, a single convergence parameter could then be chosen that would converge at nearly the same rate over all frequencies. As previously stated, the autocorrelation matrix is directly dependent on the filtered-x signal, which is computed by filtering the input reference signal $\mathbf{X}(z)$ with $\hat{\mathbf{H}}(z)$. Thus, to adjust the eigenvalues, changes can either be made to $\mathbf{X}(z)$ or to $\hat{\mathbf{H}}(z)$. For either $\mathbf{X}(z)$ or $\hat{\mathbf{H}}(z)$ changes must be carefully done so that control is not only still possible, but at worst, it is as good as if they were left unmodified.

The choice of whether to adjust $\mathbf{X}(z)$ or $\hat{\mathbf{H}}(z)$ largely depends on the control application being investigated. For simplicity, it can be said that there are two possible cases: (1) applications where changes can easily be made to $\mathbf{X}(z)$ [leave $\hat{\mathbf{H}}(z)$ unmodified] and (2) applications where changes cannot be easily made to $\mathbf{X}(z)$ [modify $\hat{\mathbf{H}}(z)$]. The adverb "easily" is included in the previous sentence to emphasize that in some control cases, it may be a simple procedure to make adjustments to $\mathbf{X}(z)$, and for other control cases, although it may be possible to make changes to $\mathbf{X}(z)$, it may be a difficult or undesirable procedure. An example of the first case would be if the fundamental and higher harmonics of the reference signal were computer generated. If such was the case, it would be a straightforward process to digitally adjust the weightings of each tone in the signal. An example of the latter case would be if an acoustic reference signal was used that included several tones, each with a different amplitude. Adjustments to the weightings of individual tones could require significant signal processing, which may add an undesired complexity to the system.

A solution for case 1 was proposed by Kuo *et al.*[9,10] The solution is most applicable for stationary multiple frequency single channel control where the fundamental and harmonics are internally generated. A general solution for case 2 is presented in this paper as the eigenvalue equalization filtered-x least mean squares (EE-FXLMS) algorithm. The algorithm has been developed to handle both the case of multiple stationary frequency control and the case of time-varying frequency control for either single or multiple channel control. The EE-FXLMS algorithm has two unique implementations to handle the two possible conditions of the reference signal. These are the following: (1) the frequencies of interest in the reference signal are equally weighted or (2) the frequencies of interest in the reference signal are unequally weighted.

A. Case 1 solution: The method of Kuo *et al.*

Kuo *et al.* observed that for multiple frequency control, the convergence parameter μ must be chosen to ensure that the system is stable at the frequency where the magnitude response of $\hat{\mathbf{H}}(z)$ is largest and that this causes the convergence at frequencies where the magnitude response of $\hat{\mathbf{H}}(z)$ is small to be slow.[9] They showed that if the amplitude of each frequency in the reference signal is optimized as the inverse of the magnitude response of $\hat{\mathbf{H}}(z)$ at that frequency, then the performance of the algorithm is greatly improved: the biggest improvement being seen at the frequencies where the magnitude response of $\hat{\mathbf{H}}(z)$ is small and convergence was originally slow. In terms of the filtered-x autocorrelation matrix eigenvalues, they show that the eigenvalue spread is close to 1 for the frequencies of interest, which should result in better convergence properties. This method was developed for single channel control.

A strength of their method is that it can be performed offline (before control is enabled) and thus does not increase the computational burden on the algorithm. One drawback is that, as they suggest, it is applicable for cases where frequency information is first obtained through a source such as a tachometer or accelerometer and is then used to digitally synthesize a reference signal that contains the fundamental frequency and appropriate harmonics.[9] Because the reference signal is digitally synthesized, it is a simple process to adjust the amplitude of each frequency in the reference signal to its optimal value. In many control applications, however, it is desirable to directly use the reference signal from its source, which makes the adjustment of the individual frequency amplitudes a more difficult task requiring extensive filtering and signal conditioning. Such might be the case when an acoustic reference signal is used. Directly using the reference signal from its source is especially important when time-varying frequencies are involved. For example, when controlling engine noise, it is desirable to directly use the tachometer signal so that the engine firing frequency and harmonics can be tracked and controlled as the speed of the engine changes during operation.

B. Case 2 solution: EE-FXLMS

Often times, it is either difficult or undesirable to alter the reference signal. Assuming that the reference signal is left unchanged, changes to the autocorrelation matrix must stem from changes to $\hat{\mathbf{H}}(z)$ but must be done carefully as any errors in its estimation already contribute to lower convergence rates and instability. Estimation errors can be considered in two parts: errors in the amplitude estimation and errors in the phase estimation.[19] It has been shown that phase estimation errors greater than $\pm 90°$ cause algorithm instability,[16] but errors as high as 40° have little effect on the performance.[16] Magnitude estimation errors can be compensated for by the choice of μ (Ref. 20), and consequently do not affect stability. Ideally, changes would be made to the magnitude information of $\hat{\mathbf{H}}(z)$, while the phase information is preserved. The method to equalize the eigenvalues of the autocorrelation matrix by changing the magnitude information of $\hat{\mathbf{H}}(z)$ while preserving the phase information will be referred to as the EE-FXLMS algorithm.

Two implementations of the EE-FXLMS algorithm have been developed to handle the two possible conditions of the

reference signal. These are the following: (1) the frequencies of interest in the reference signal are equally weighted or (2) the frequencies of interest in the reference signal are unequally weighted. If the frequencies of interest in the reference signal are equally weighted, it is proposed that the magnitude coefficients of $\hat{\mathbf{H}}(z)$ be optimized to also have an equal (flat) weighting over frequency. One example of a signal that would have an equal weighting over frequency would be the use of a tachometer signal to control the engine firing frequency of an engine. As the rpm of the engine changes during operation, the engine firing frequency will change, but the voltage level of the tachometer signal will remain constant. In other words, the amplitude of the reference signal will not change as the frequency changes. If the frequencies of interest in the reference signal are unequally weighted, it is proposed that the magnitude coefficients of $\hat{\mathbf{H}}(z)$ be optimized to have the inverse of the magnitude response of $\mathbf{X}(z)$. An example would be multiple frequency noise where the signal is nominally stationary, such as helicopter noise. With helicopter noise, each of the major noise sources (engine, main rotor, and tail rotor) will require a different reference signal obtained through a tachometer, photocell, or other method. Each of these signals will contain harmonics, each with a unique amplitude (generally, each successive higher harmonic will have a lower amplitude), resulting in an unequally weighted multiple frequency reference signal.

1. EE-FXLMS Flat $\hat{\mathbf{H}}(z)$ implementation

If all frequencies in $\mathbf{X}(z)$ are equally weighted, then Kuo *et al.*[9] suggested that μ must be chosen based on the frequency where the magnitude of $\hat{\mathbf{H}}(z)$ is the largest. This slows down the convergence at frequencies where the magnitude of $\hat{\mathbf{H}}(z)$ is small. This agrees with the eigenvalue simulation shown in Fig. 2. Note that in that simulation, each frequency was given an equal weighting. As previously mentioned, μ must be chosen based on the maximum eigenvalue in the frequency range of interest, and performance is degraded at frequencies where the eigenvalues are small. For the equally weighted $\mathbf{X}(z)$, the eigenvalue spread is mostly a function of the magnitude response of $\hat{\mathbf{H}}(z)$. This can be seen in Fig. 3. In Fig. 3, the magnitude coefficients of $\hat{\mathbf{H}}(z)$ are overlaid on the same plot of the maximum eigenvalues shown in Fig. 2. It can be seen that the magnitude coefficients are highly correlated with the eigenvalues. The maximum eigenvalue occurs where the response of $\hat{\mathbf{H}}(z)$ is large.

The data in Fig. 3 suggest that manipulating the magnitude coefficients of $\hat{\mathbf{H}}(z)$ should modify the eigenvalue spread. If the magnitude coefficients were "flat" over frequency, the eigenvalue spread should also be more flat over frequency. A method for flattening the magnitude coefficients has been developed, which is simple to implement and does not increase the computational burden of the algorithm.

The basic procedure for implementing the EE-FXLMS is to adjust the coefficients of $\hat{\mathbf{h}}(t)$ before ANC control is started as follows.

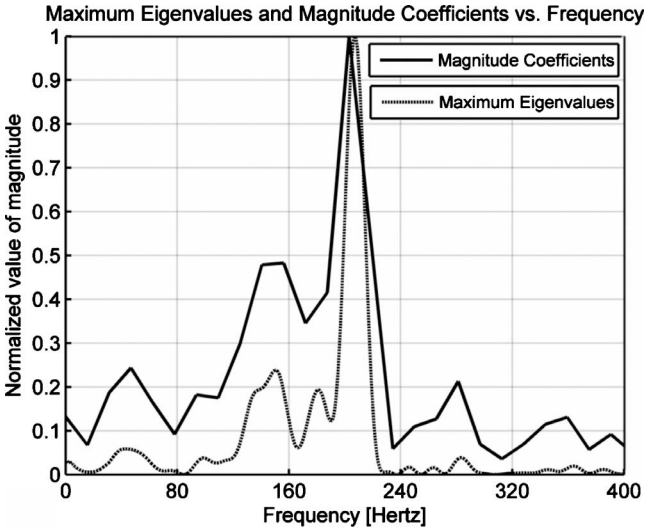


FIG. 3. Maximum eigenvalues and magnitude coefficients vs frequency for a mock cab—equally weighted reference signal.

(i) Obtain the time domain impulse response $\hat{\mathbf{h}}(t)$ for each $\hat{\mathbf{H}}(z)$ through an offline SysID process.
(ii) Take the fast fourier transform (FFT) to obtain $\hat{\mathbf{H}}(z)$.
(iii) Divide each value in the FFT by its magnitude and then multiply by the mean value of the FFT.
(iv) Compute the inverse FFT to obtain a new $\hat{\mathbf{h}}(t)$ and use the new modified $\hat{\mathbf{h}}(t)$ in the FXLMS algorithm as normal.

If using multiple channel and/or ED control, the process is repeated for each $\hat{\mathbf{H}}(z)$ estimate. This procedure flattens the magnitude coefficients of $\hat{\mathbf{H}}(z)$ while preserving the phase. It is an offline process directly done by following SysID and can be incorporated into any existing algorithm with only a few lines of code. As an offline process, it adds no computational burden to the algorithm while control is running. The results of the flattening process can be seen in Fig. 4. Figure 4 shows the original and modified $\hat{\mathbf{H}}(z)$ magnitude coefficients in the top plot and shows that the phase information of $\hat{\mathbf{H}}(z)$ has been preserved in the bottom plot. Note that the two lines representing the original and modified phase information of $\hat{\mathbf{H}}(z)$ are directly on top of each other in the bottom plot. The plots for the other $\hat{\mathbf{H}}(z)$ models for the other channels and ED components are similar.

An attempt to quantify any improvement in the eigenvalue spread has been made by using the following metrics:

(1) *Span*—λ_{max} divided by λ_{min}. Ideally 1.
(2) *rms value*—root mean square. Ideally 1.
(3) *Crest factor*—λ_{max} divided by rms value (how close the rms value is to the peak value) Ideally 1.

The effect of the flattening process on the eigenvalues can be seen in Fig. 5. The data for the figure were computed as before by an offline estimate of the autocorrelation matrix by using an actual $\hat{\mathbf{H}}(z)$ model from a mock cabin and finding the λ_{max} for each frequency from 0 to 400 Hz. The curve

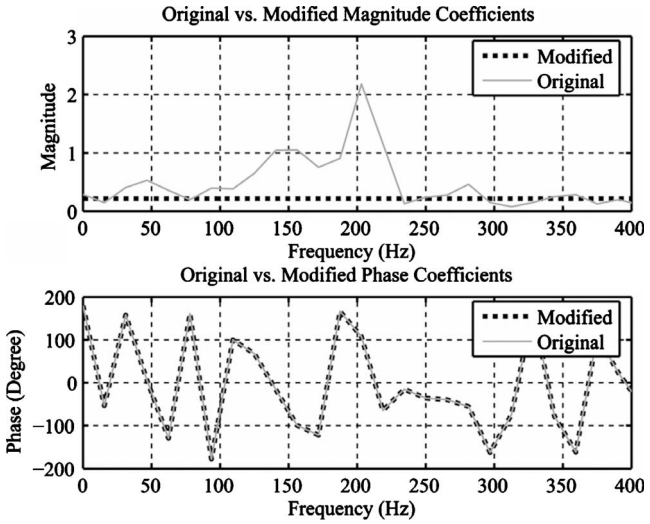


FIG. 4. Original and modified magnitude coefficients of $\hat{\mathbf{H}}(z)$–EE-FXLMS flat $\hat{\mathbf{H}}(z)$ implementation and the original and modified phase coefficients of $\hat{\mathbf{H}}(z)$–EE-FXLMS flat $\hat{\mathbf{H}}(z)$ implementation.

labeled "original" represents the same data shown in Fig. 2, and the curve labeled "modified" is an estimate of the eigenvalues by using the modified $\hat{\mathbf{H}}(z)$ model. In Fig. 5, the eigenvalues in both the original and modified cases have been normalized by the largest of the original eigenvalues. It is seen that the modified eigenvalues are more uniform ("equalized") over all frequencies. While the variation in the modified eigenvalues would ideally be zero, the decreased variation compared to the original eigenvalues should produce an observable performance advantage. The algorithm should converge at near the same rate over all frequencies and should not exhibit the frequency dependent behavior of the standard FXLMS.

Table I shows the improvement of the modified eigenvalues according to the defined metrics over the range from 0 to 400 Hz. The range from 0 to 400 Hz was selected because the experimental hardware has a cutoff frequency at 400 Hz.

TABLE I. Comparison of original and modified eigenvalues by using defined metrics—EE-FXLMS flat $\hat{\mathbf{H}}(z)$ implementation.

Metric	Original	Modified	% improvement
Span	2.37×10^5	2920	99
rms	0.19	0.3	58
Crest factor	5.283	3.413	35

In Table I it can be seen that the modified case has a lower span, a higher rms value, and a lower crest factor. In all three metrics, the values for the modified case are closer to the optimum values that would be present if the eigenvalues across all frequencies were exactly the same. These modifications to $\hat{\mathbf{H}}(z)$ should make a noticeable improvement in the performance of the algorithm.

2. EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(z)|$ implementation

In this case, the reference signal is unequally weighted, and so the eigenvalue simulation shown in Fig. 2 must be redone by using an unequally weighted reference signal. For this simulation, the reference signal was chosen so that there was a linear descending trend in the amplitude of each successive frequency. Figure 6 shows the offline simulation by using an actual $\hat{\mathbf{H}}(z)$ from a mock cabin and tonal inputs from 0 to 400 Hz. The eigenvalues in the figure have been normalized to the largest eigenvalue in the range.

The same ideas behind flattening the magnitude coefficients of $\hat{\mathbf{H}}(z)$ when $\mathbf{X}(z)$ is equally weighted apply for the case when the frequencies in $\mathbf{X}(z)$ have unequal weighting. The magnitude coefficients of $\hat{\mathbf{H}}(z)$ at each frequency bin must be made to compensate for the unequal weighting of each tone in $\mathbf{X}(z)$ so that the eigenvalue spread becomes essentially flat over frequency. As with the flat $\hat{\mathbf{H}}(z)$ implementation, the phase information must be preserved.

When $\mathbf{X}(z)$ has unequally weighted frequencies, the magnitude coefficients of $\hat{\mathbf{H}}(z)$ are set to be $1/|\mathbf{X}(z)|$. This

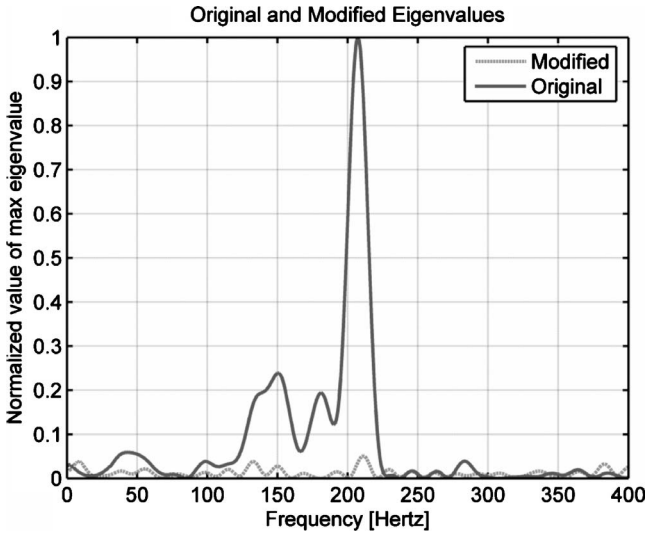


FIG. 5. Normalized original and modified eigenvalues—EE-FXLMS flat $\hat{\mathbf{H}}(z)$ implementation.

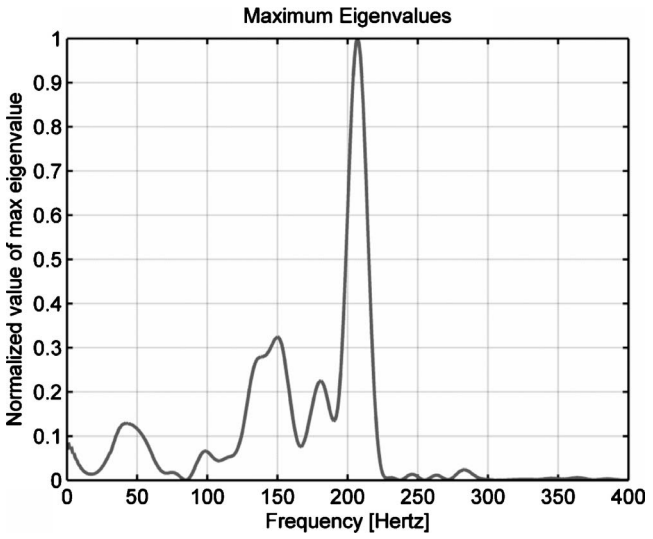


FIG. 6. Plot of normalized maximum eigenvalues over frequency—unequally weighted reference signal.

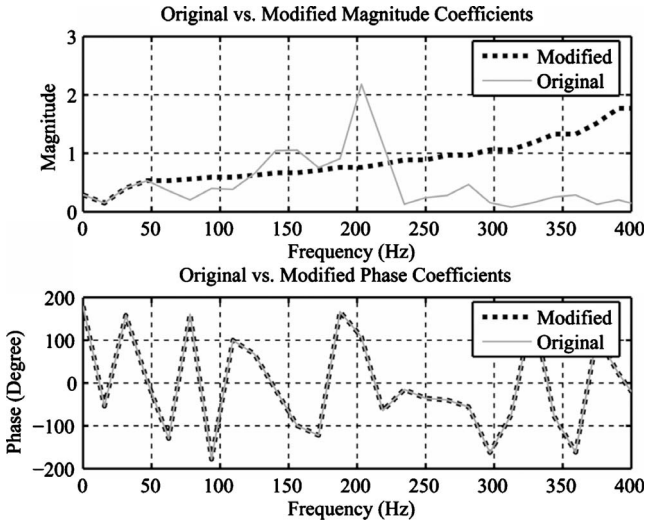


FIG. 7. Original and modified magnitude coefficients of $\hat{\mathbf{H}}(z)$–EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(t)|$ implementation and the original and modified phase coefficients of $\hat{\mathbf{H}}(z)$–EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(t)|$ implementation.

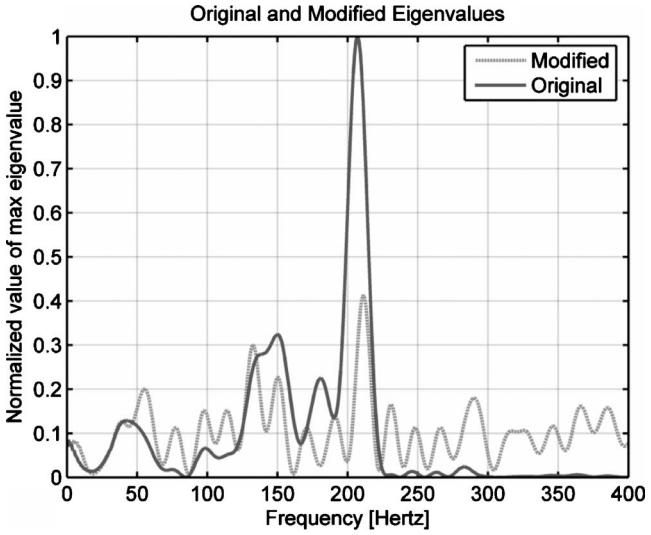


FIG. 8. Normalized original and modified eigenvalues—EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(t)|$ implementation.

has the same effect as flattening the magnitude coefficients of $\hat{\mathbf{H}}(z)$ when $\mathbf{X}(z)$ has equally weighted frequencies. The method is simple to implement and does not increase the computational burden of the algorithm. The basic procedure is run before ANC control and is as follows:

(1) Obtain the time domain impulse response $\hat{\mathbf{h}}(t)$ for each $\hat{\mathbf{H}}(z)$ through an offline SysID process.
(2) Take the FFT to obtain $\hat{\mathbf{H}}(z)$.
(3) Compute the phase coefficients from the FFT of $\hat{\mathbf{H}}(z)$ and save them in a vector.
(4) Take a time sample of $\mathbf{x}(t)$ and compute its FFT, $\mathbf{X}(z)$.
(5) Find the magnitude coefficients of $\mathbf{X}(z)$ at the frequencies of interest.
(6) Create a vector of magnitude coefficients that is equal to $1/|\mathbf{X}(z)|$ from the coefficients computed in step (5).
(7) Take the vector of phase coefficients from step (3) and combine them with the vector of magnitude coefficients in step (6) to create a new single vector of complex coefficients having the original phase and the $1/|\mathbf{X}(z)|$ magnitudes.
(8) Take the inverse FFT of the new vector and use it as the new modified $\hat{\mathbf{h}}(t)$ in the FXLMS algorithm as normal.

If using multiple channel and/or ED control, the process is repeated for each $\hat{\mathbf{H}}(z)$ estimate. This procedure adjusts the magnitude coefficients of $\hat{\mathbf{H}}(z)$ while preserving the phase. As an offline process, it adds no computational burden to the algorithm when control is running. The results of the process can be seen in Fig. 7. The top plot in Fig. 7 shows the original and modified $\hat{\mathbf{H}}(z)$ magnitude coefficients. The bottom plot shows that the phase information of the same $\hat{\mathbf{H}}(z)$ has been preserved. Again, note that the two lines representing the original and modified phase information of $\hat{\mathbf{H}}(z)$ are directly on top of each other. The plots for the other $\hat{\mathbf{H}}(z)$ models for the other channels and ED components are similar.

The effect of the modification process on the eigenvalues can be seen in Fig. 8. The data for the figure were computed as before by an offline estimate of the autocorrelation matrix by using an actual $\hat{\mathbf{H}}(z)$ model from a mock cabin and finding the λ_{max} for each frequency from 0 to 400 Hz. The curve labeled "original" represents the same data shown in Fig. 7, and the curve labeled "modified" is an estimate of the eigenvalues using the modified $\hat{\mathbf{H}}(z)$ model. In Fig. 8, the eigenvalues in both the original and modified cases have been normalized by the largest of the original eigenvalues. It is seen that the modified eigenvalues are more uniform (equalized) over all frequencies, though not perfectly flat. While the variation in the modified eigenvalues would ideally be zero, the decreased variation compared to the original eigenvalues should produce an observable performance advantage.

Table II shows the improvement of the modified eigenvalues according to the same metrics defined for the flat $\hat{\mathbf{H}}(z)$ implementation over the range from 0 to 400 Hz. In Table II it can be seen that the modified case has a lower span a higher rms value, and a lower crest factor. In all three metrics, the values for the modified case are closer to the optimum values that would be present if the eigenvalues across all frequencies were exactly the same. It is also of note that setting the magnitude coefficients of $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(z)|$ with an unequally weighted reference signal offered nearly the same percentage improvement as flattening the magnitude coefficients of $\hat{\mathbf{H}}(z)$ with an equally weighted reference signal (compare Tables I and II).

TABLE II. Comparison of original and modified eigenvalues using defined metrics—EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(t)|$ implementation.

Metric	Original	Modified	% improvement
Span	4.77×10^5	99.24	100
rms	0.2	0.3	50
Crest factor	4.983	3.33	33

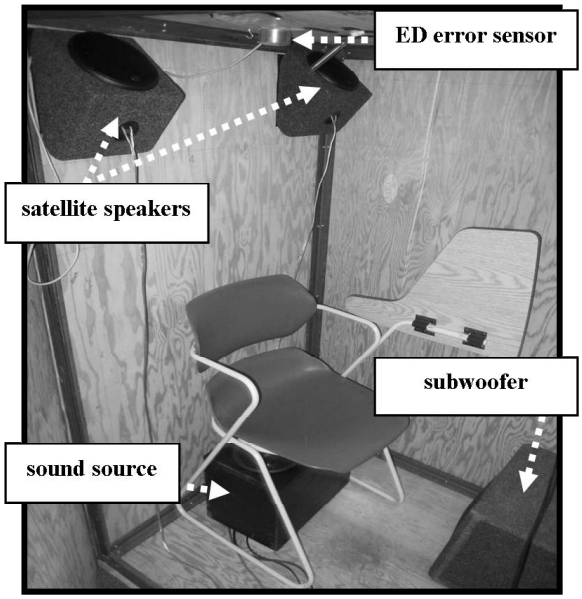


FIG. 9. Photo of inside of mock cab.

V. EXPERIMENTAL RESULTS

The performance advantages of the EE-FXLMS algorithm were verified for both single time-varying frequency and stationary multiple frequency test cases. First, the experimental setup will be explained, second the results for the EE-FXLMS flat $\hat{\mathbf{H}}(z)$ implementation will be shown, and, lastly, the results for the EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(z)|$ implementation will be shown.

A. Experimental setup

The experiments were conducted inside a mock cabin enclosure with nominal dimensions of $1.0\times1.5\times1.1$ m^3. The cabin has a steel frame, 0.01-m-thick plywood sides, and a 0.003-m-thick Plexiglass® front panel. A speaker placed under a chair served as the sound source and three loudspeakers were set up in a two channel control configuration. The control signals were routed through a crossover circuit to route the low-frequency content (below 90 Hz) of both channels to a subwoofer on the floor of the cab and to route the high-frequency content (above 90 Hz) of each control channel to one of two smaller satellite speakers mounted in the top corners of the cab, near the back. An ED error sensor consisting of four equally spaced microphones around a small disk was placed on the ceiling near where an operator's head would be. The performance of the algorithms will be reported at the error sensor. A photo of the cab, error sensor, and speakers is seen in Fig. 9.

The control algorithms were implemented on a Texas Instruments TMS320C6713 DSP processor, capable of 1.350×10^6 floating point operations/s. Both adaptive control filters consisted of 32 taps for control of single tones and 100 taps for multiple tones, and all secondary path transfer functions were modeled with 128 taps. All input channels were simultaneously sampled at 2 kHz, and all input and output signals had 16 bits of resolution. Fourth-order Butterworth low pass filters (400 Hz cutoff) provided antialiasing and reconstruction of input and output signals, respectively.

TABLE III. Comparison of EE-FXLMS flat $\hat{\mathbf{H}}(z)$ implementation and normal FXLMS control for time-varying frequency experimentation. A positive number indicates that EE-FXLMS control performed better.

Sweep rate (Hz)	Control type	Average reduction at error mic (dB)	Difference[a] (dB)
2	Normal	6.5	3.5
	EE	10.0	
4	Normal	5.2	2.1
	EE	7.3	
8	Normal	4.3	1.4
	EE	5.7	
16	Normal	4.4	1.1
	EE	5.5	
32	Normal	4.0	0.3
	EE	4.3	
64	Normal	3.9	0.2
	EE	4.1	
128	Normal	3.9	0.0
	EE	3.9	
256	Normal	3.9	0.0
	EE	3.9	
TOTAL AVERAGE			1.1

[a]Positive number indicates that EE-FXLMS performed better.

B. Results for EE-FXLMS Flat $\hat{\mathbf{H}}(z)$ implementation

The EE-FXLMS flat implementation was tested for a time-varying frequency. For the time-varying frequency, a swept sine signal with different sweep rates was used. The signal maintained the same amplitude at each frequency in the sweep.

1. Time-varying frequency results

Several swept sine test signals with different sweep rates were created. Each test signal consisted of a swept sine from 50 to 200 Hz and the rates ranged from 2 to 256 Hz/s. The time-averaged sound pressure level over the entire duration of the test signal was measured with and without control running. The convergence coefficient μ was experimentally determined by finding the largest stable value for the entire frequency range and then scaling it back by a factor of 10 to ensure stability. The μ for EE-FXLMS control was found to be 1×10^{-7} and the μ for standard FXLMS control was found to be 1×10^{-8}. Each measurement was repeated three times and the average and standard deviation were computed. The actual attenuation for both control types at the error sensor is shown in Table III. The difference in attenuation between EE-FXLMS and FXLMS controls is also shown in Table III. A positive number indicates the EE-FXLMS performed better. The standard deviation for each test case was small (usually less than 0.02 dB) and is not reported in the table.

The data show that when averaged over all of the data, EE-FXLMS performs 1.1 dB better than normal FXLMS at the error sensor. The data also show that the slower the sweep rate, the more advantage the EE-FXLMS control provides. For

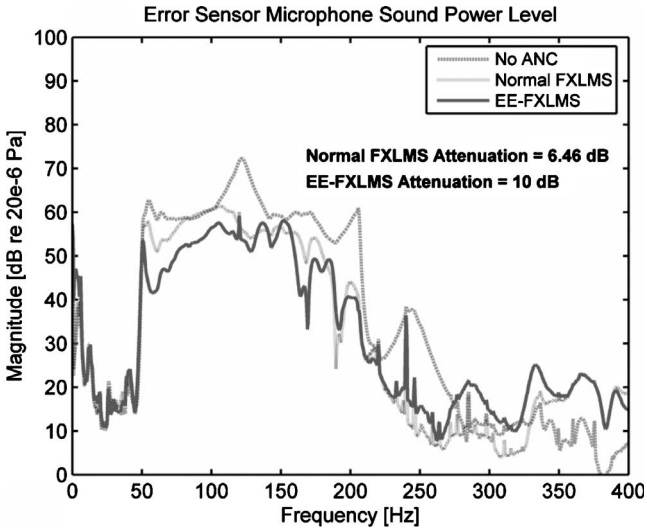


FIG. 10. SPL at the error sensor for normal FXLMS and EE-FXLMS control.

the 2 Hz sweep rate, EE-FXLMS control provides 3.5 dB more reduction at the error sensor. Figure 10 shows a plot of the control results for both normal FXLMS and EE-FXLMS for the 2 Hz sweep rate. For this run, the sound pressure level (SPL) at the error sensor before control was enabled was 87.9 dB (calculated over the entire frequency range). The SPL dropped to 81.5 dB for normal FXLMS control and 77.9 dB for EE-FXLMS control.

At the fastest sweep rates, the differences were almost negligible. An explanation of this can be found by looking at the fastest convergence times for the single frequency case. For this case, the fastest convergence times were seen to be on the order of 0.10 s. At the faster sweep rates, such as 128 Hz/s, the algorithm has 0.0078 s (1/128 Hz/s =0.0078 s/Hz) to converge at each frequency. At the slower sweep rates, such as 2 Hz/s, the algorithm has 0.5 s (1/2 Hz/s=0.5 s/Hz) to converge at each frequency. When the sweep rates are faster, the convergence times are several orders of magnitude larger than the time the algorithm has to converge on each frequency before it shifts, which leads to poor performance and little gain from the faster convergence times of the EE-FXLMS. When the sweep rates are slower, the convergence times are on the same order of time that the algorithm has to converge on each frequency before it shifts, which leads to better performance and noticeable gains from the faster convergence times of the EE-FXLMS.

C. Results for EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(z)|$ implementation

The EE-FXLMS was compared again to the FXLMS algorithm for a multiple frequency test signal, this time with the magnitude of the tones in the reference signal decreasing with increasing frequency. Since the frequency resolution in $\hat{\mathbf{H}}(z)$ was not high, the magnitude coefficients in $\hat{\mathbf{H}}(z)$ bracketing the frequencies in the reference were adjusted to be the inverse magnitude of the tones in the reference. Five tone (50, 125, 200, 250, and 300 Hz) and 11 tone (50, 75, 100, 125, 150, 175, 200, 225, 250, 275, and 300 Hz) noise and reference signals were created for this comparison. Many of the tones in the signal were intentionally chosen to match frequencies where the magnitude response of $\hat{\mathbf{H}}(z)$ is large; frequencies where the advantages of the EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(z)|$ implementation should be the most observable. Additionally, they were chosen far enough apart that the magnitude of $\hat{\mathbf{H}}(z)$ could be individually adjusted for each tone. Control was run with both the normal FXLMS algorithm and the EE-FXLMS algorithm with the $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(z)|$ implementation. The number of control taps was increased to 100 for these test cases.

1. Multiple frequency results

The convergence coefficient μ was determined as before. The scaled μ for EE-FXLMS and normal FXLMS controls for the noise signal containing five tones were found to be 8×10^{-9} and 1×10^{-9}, respectively. The scaled μ for EE-FXLMS and normal FXLMS controls for the noise signal containing 11 tones were found to be 2×10^{-8} and 4×10^{-9}, respectively. The measured performance for each configuration was the amount of attenuation in decibels and the convergence time in seconds at each frequency, as well as for the total error signal. The convergence time was defined as the time it takes the signal to converge to $1/e$ (natural log e) of its initial value. A convergence time of 9 s means that the signal did not converge at that frequency in the time period of the measurement, which was 9 s. Each measurement was performed three times for computation of an average and standard deviation.

The average for the three test runs and the difference between normal FXLMS and EE-FXLMS controls are shown for the 5 tone test case in Table IV and the 11 tone case in Table V. In both tables, a linear average of the reduction at each frequency is merely given to give a sense of the performance of the algorithms at the frequencies of interest. The actual overall reduction of the entire error signal is also given. Again, a positive number for the difference indicates that EE-FXLMS performed better.

In the tables, it can be seen that EE-FXLMS control performed about 4 dB better overall at the error sensor for the 5 tone case and about 2 dB better overall for the 11 tone case. Observing the convergence time and attenuation at each frequency shows the more uniform performance over frequency of the EE-FXLMS approach. At some frequencies, the EE-FXLMS algorithm provides as much as 16 dB additional attenuation and converged several seconds faster. At higher frequencies, where the weighting of the tones in the reference was smaller, the FXLMS algorithm had very long convergence times and poor attenuation. In many cases, those frequencies did not appreciably converge during the measurement. The EE-FXLMS algorithm outperformed the normal FXLMS algorithm in both attenuation and convergence speed at all frequencies except 150 Hz, which did not converge well for any test case. 150 Hz corresponds to a large resonance mode of the mock cabin. Further investigation found that the error sensor was at a nodal position for this frequency, which leads to reduced performance. The frequency spectrum for the error signal with no control and FXLMS and

TABLE IV. Comparison of EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(t)|$ implementation and normal FXLMS control for multiple stationary frequency experimentation of five tones. A positive number indicates that EE-FXLMS control performed better. The overall attenuation and convergence times reported at the bottom of the table are for the entire error signal and not the average of all the values at each tone.

	Control of 5 tones					
	Normal FXLMS		EE-FXLMS		Difference[a]	
Frequency (Hz)	Average reduction at error mic (dB)	Convergence time (s)	Average reduction at error mic (dB)	Convergence time (s)	Average reduction at error mic (dB)	Convergence time (s)
50	25.5	1.84	30.2	0.58	4.7	1.26
125	25.9	0.44	28.3	0.33	2.4	0.11
200	30.7	0.89	30.1	0.44	−0.6	0.45
250	5.2	9	19.8	0.99	14.6	8.01
300	0.2	9	16.6	3.04	16.4	5.96
Linear Average of reduction at 5 tones	18.2	3.61	25.2	0.95	7.0	2.66
Overall reduction for entire error signal	21.6	0.5	26.0	0.34	4.4	0.16

[a]Positive number indicates that EE-FXLMS performed better.

TABLE V. Comparison of EE-FXLMS $\hat{\mathbf{H}}(z)=1/|\mathbf{X}(t)|$ implementation and normal FXLMS control for multiple stationary frequency experimentation of 11 tones. A positive number indicates that EE-FXLMS control performed better. The overall attenuation and convergence times reported at the bottom of the table are for the entire error signal and not the average of all the values at each tone.

	Control of 11 Tones					
	Normal FXLMS		EE-FXLMS		Difference[a]	
Frequency (Hz)	Average reduction at error mic (dB)	Convergence time (s)	Average reduction at error mic (dB)	Convergence time (s)	Average reduction at error mic (dB)	Convergence time (s)
50	29	1.4	31	0.763	2	0.637
75	12.3	3.95	21.1	1.05	8.8	2.9
100	7.2	7.27	11	2.15	3.8	5.12
125	25.6	0.43	28.5	0.39	2.9	0.04
150	1.7	9	2.2	9	0.5	0
175	8.5	7.42	13.3	0.82	4.8	6.6
200	27.7	0.68	28.9	0.52	1.2	0.16
225	5.2	9	14.7	3.4	9.5	5.6
250	4.9	9	21.2	1.34	16.3	7.66
275	1.5	9	9.8	4.18	8.3	4.82
300	0.1	9	12.5	3.54	12.4	5.46
Linear Average of reduction at 11 Tones	11.4	5.56	17.5	2.30	6.1	3.26
Overall reduction for entire error signal	12.8	0.61	15.2	0.49	2.4	0.12

[a]Positive number indicates that EE-FXLMS performed better.

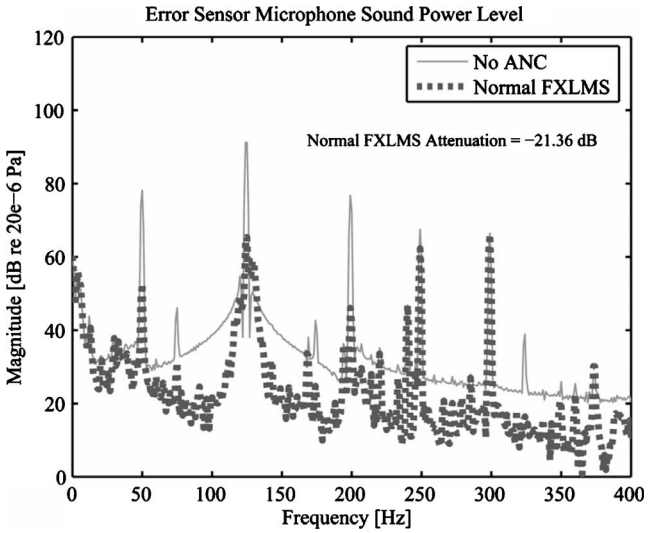


FIG. 11. SPL at the error sensor for a normal FXLMS.

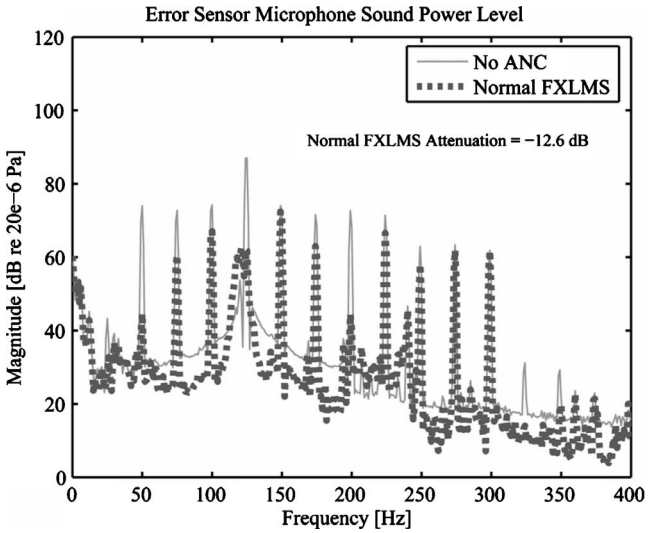


FIG. 13. SPL at the error sensor for a normal FXLMS.

EEFXLMS controls for the 5 tone case is shown in Figs. 11 and 12 and that for the 11 tone case is shown in Figs. 13 and 14.

VI. CONCLUSIONS

A new eigenvalue equalization approach (EE-FXLMS) has been demonstrated for time-varying and multiple frequency noise. It has been shown that adjustments to the magnitude coefficients of $\hat{\mathbf{H}}(z)$ while preserving the phase leads to a smaller eigenvalue spread, faster convergence times, and increased attenuation. Two offline methods for adjusting the magnitude coefficients of $\hat{\mathbf{H}}(z)$ to complement the magnitude of the reference signal have been demonstrated.

The EE-FXLMS implementation to flatten the magnitude coefficients of $\hat{\mathbf{H}}(z)$, when the magnitude of the reference signal is the same for all frequencies, led to as much as 3.5 dB additional attenuation at the error sensor for the slower sweep rates. An additional attenuation of 1.0 dB at the error sensor was seen at sweep rates of up to 16 Hz/s, with a slight increase still being seen at rates as high as 64 Hz/s. When averaged over all of the sweep rates tested, EE-FXLMS provided 1.1 dB additional attenuation at the error sensor.

The EE-FXLMS implementation of adjusting the magnitude coefficients of $\hat{\mathbf{H}}(z)$ to be the inverse magnitude of the reference signal, when the magnitude of the reference signal is different for all frequencies, led to as much as 4.4 dB additional overall attenuation at the error sensor and as much as 16 dB additional attenuation at an individual tone. The EE-FXLMS algorithm's convergence rate at individual frequencies was faster and more uniform than the normal FXLMS with several second improvement being seen in some cases.

The performance advantages of the EE-FXLMS become more meaningful when considering the simplicity of its implementation. It can be incorporated into any FXLMS algorithm with only a few lines of code. As an offline process, it does not increase the computational burden of the algorithm.

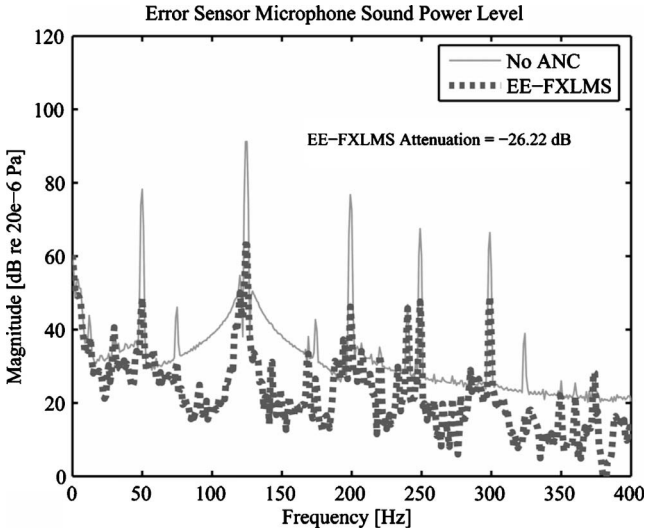


FIG. 12. EE-FXLMS control for reference signal with five tones.

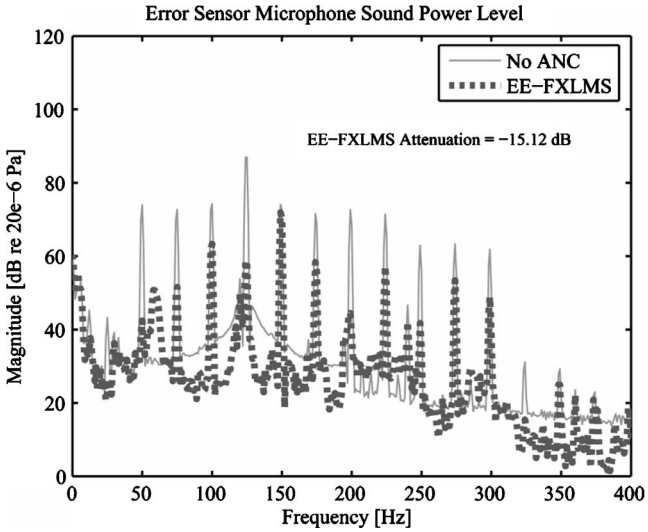


FIG. 14. EE-FXLMS control for reference signal with 11 tones.

Additionally, it does not require that the reference be internally generated or extensively modified.

These two methods of adjusting the magnitude coefficients of $\hat{\mathbf{H}}(z)$ provide a way to reduce the frequency dependent convergence of the FXLMS algorithm. As noted, the eigenvalue span resulting from these modifications is still not perfectly flat. Other alteration schemes may be developed that can further reduce this variation.

[1]D. R. Morgan, "An analysis of multiple correlation cancellation loops with a filter in the auxiliary path," IEEE Trans. Acoust., Speech, Signal Process. **28**, 454–467 (1980).

[2]J. C. Burgess, "Active adaptive sound control in a duct: A computer simulation," J. Acoust. Soc. Am. **70**, 715–726 (1981).

[3]Y. C. Park and S. D. Sommerfeldt, "Global attenuation of broadband noise fields using energy density control," J. Acoust. Soc. Am. **101**, 350–359 (1997).

[4]C. C. Boucher, S. J. Elliot, and K.-H. Baek, "Active control of helicopter rotor tones," *Proceedings of the INTER-NOISE 96*, pp. 1179–1182 (1996).

[5]B. Faber and S. D. Sommerfeldt, "Global control in a mock tractor cabin using energy density," in *Proceedings of the ACTIVE '04 (1994)*, edited by R. H. Cabell and George C. Maling, Jr..

[6]R. L. Clark and G. P. Gibbs, "A novel approach to feedforward higher-harmonic control," J. Acoust. Soc. Am. **96**, 926–936 (1994).

[7]S. M. Lee, H. J. Lee, C. H. Yoo, D. H. Youn, and I. W. Cha, "An active noise control algorithm for controlling multiple sinusoids," J. Acoust. Soc. Am. **104**, 248–254 (1998).

[8]M. Rupp and A. H. Sayed, "Modified FXLMS algorithms with improved convergence performance," *IEEE Proceedings of the ASILOMAR-29* (1995).

[9]S. M. Kuo, X. Kong, S. Chen, and W. Hao, "Analysis and design of narrowband active noise control systems," IEEE Trans. Acoust., Speech, Signal Process. **6**, 3557–3560 (1998).

[10]S. M. Kuo, M. Tahernezhadi, and W. Hao, "Convergence analysis of narrow-band active noise control systems," IEEE Trans. Circuits Syst., II: Analog Digital Signal Process. **46**, 220–223 (1999).

[11]L. Vicente and E. Masgrau, "Performance comparison of two fast algorithms for active control," *Proceedings of the ACTIVE 99*, pp. 1089–1100 (1999).

[12]S. D. Sommerfeldt and P. J. Nashif, "An adaptive filtered-x algorithm for energy-based active control," J. Acoust. Soc. Am. **96**, 300–306 (1994).

[13]J. W. Parkins, S. D. Sommerfeldt, and J. Tichy, "Narrowband and broadband active noise control in an enclosure using acoustic energy density," J. Acoust. Soc. Am. **108**, 192–203 (2000).

[14]D. C. Copley, B. Faber, and S. D. Sommerfeldt, "Energy density active noise control in an earthmoving machine cab," in *Proceedings of the NOISE-CON '05*, edited by J. Stuart Bolton, P. Davies, and G. C. Maling, Jr.. (1995)

[15]S. M. Kuo and D. R. Morgan, in *Active Noise Control Systems: Algorithms and DSP Implementations*, edited by J. G. Proakis (Wiley, New York, 1996), Chap. 5, pp. 147–186.

[16]C. C. Boucher, S. J. Elliot, and P. A. Nelson, "Effects of errors in the plant model on the performance of algorithms for adaptive feedforward control," IEE Proc. F, Commun. Radar Signal Process. **138**, 313–319 (1991).

[17]S. J. Elliott and P. A. Nelson, "Active Noise Control," IEEE Signal Process. Mag. **10**, 12–35 (1993).

[18]L. A. Sievers and A. H. von Flotow, "Comparison and extensions of control methods for narrow-band disturbance rejection," IEEE Trans. Signal Process. **40**, 459–461 (1992).

[19]S. D. Snyder and C. H. Hansen, "The effect of transfer function estimation errors of the filtered-x LMS algorithm," IEEE Trans. Signal Process. **42**, 950–953 (1994).

[20]S. D. Snyder and C. H. Hansen, "The influence of transducer transfer functions and acoustic time delays on the implementations of the LMS algorithm in active noise control systems," J. Sound Vib. **141**, 409–424 (1990).

Testing a theory of aircraft noise annoyance: A structural equation analysis

Maarten Kroesen,[a] Eric J. E. Molin, and Bert van Wee
Faculty of Technology, Policy and Management, Delft University of Technology, P.O. Box 5015, 2600 GA Delft, The Netherlands

(Received 2 August 2007; revised 18 March 2008; accepted 6 April 2008)

Previous research has stressed the relevance of nonacoustical factors in the perception of aircraft noise. However, it is largely empirically driven and lacks a sound theoretical basis. In this paper, a theoretical model which explains noise annoyance based on the psychological stress theory is empirically tested. The model is estimated by applying structural equation modeling based on data from residents living in the vicinity of Amsterdam Airport Schiphol in The Netherlands. The model provides a good model fit and indicates that concern about the negative health effects of noise and pollution, perceived disturbance, and perceived control and coping capacity are the most important variables that explain noise annoyance. Furthermore, the model provides evidence for the existence of two reciprocal relationships between (1) perceived disturbance and noise annoyance and (2) perceived control and coping capacity and noise annoyance. Lastly, the model yielded two unexpected results. Firstly, the variables noise sensitivity and fear related to the noise source were unable to explain additional variance in the endogenous variables of the model and were therefore excluded from the model. And secondly, the size of the total effect of noise exposure on noise annoyance was relatively small. The paper concludes with some recommended directions for further research. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2916589]

PACS number(s): 43.50.Rq, 43.50.Qp, 43.50.Lj [BSF] Pages: 4250–4260

I. INTRODUCTION

The global aviation sector has rapidly developed since the beginning of the 1960s. Air travel has grown due to numerous factors such as economic and demographic growth, decreasing market prices, globalization, increasing quality, the introduction of the hub-and-spoke concept, and liberalizing measures. With this growth, the negative externalities of the aviation market have also become more evident: noise, local and global air pollution, and decreasing external safety.

The noise policies adopted by national governments in relation to major airports mainly focus on reducing the level of noise exposure and the number of people who are exposed. However, there is no one-on-one relationship between noise exposure and noise annoyance. Based on 39 empirical studies, Job (1988) concluded that the correlation coefficient for group data (aggregate models) is 0.82 (standard deviation of 0.14) and for individual data 0.42 (standard deviation of 0.12). This means that in the latter case, only 18% of the variance in noise annoyance is explained by noise exposure.[1] One explanation for this weak relationship is that factors other than the level of noise exposure, the so-called nonacoustical factors, influence noise annoyance. Guski (1999) concluded that approximately one-third of the variation in noise annoyance can be explained by acoustical factors (e.g., the sound level, peak levels, sound spectrum, and number of noise events) and a second third by nonacoustical factors. The last third can either be attributed to measurement errors (which decreases the proportion of explained variance in the dependent variables), the presence of yet unknown factors which influence noise annoyance, or stochastic variation related to idiosyncrasies of individuals.

Past studies that investigated relevant nonacoustical factors, however, have two major shortcomings. Firstly, the research can be characterized as highly inductive, which generally means that it lacks a sound theoretical basis (Taylor, 1984). As Taylor noted (1984) (p. 245), "many of the models which are tested by using path analysis are exploratory. As such, they probably do not adequately represent the processes leading to the outcome in question (e.g., noise annoyance). In such cases, causal claims stand on weak ground indeed and sensibly are best avoided." In addition, although not mentioned by Taylor, the lack of elementary understanding related to the topic of noise annoyance can result in misspecification of the statistical model and hence even lead to false inferences related to the effect sizes of relevant variables.

Secondly, the practical relevance and significance of nonacoustical factors in relation to noise annoyance are often based on correlational analysis or multiple regression analysis. Both these methods have severe deficiencies in modeling noise annoyance. As Alexandre (1976) has shown, the results of correlational analysis can be misinterpreted since the effect of the factor under investigation is not controlled for noise exposure or other factors. In addition, the direction of causation remains uncertain. Of the three commonly accepted conditions needed to qualify something as a causal relationship, i.e., time precedence, nonspuriousness, and simple association, only the last one is satisfied. The result is that the relative importance of different factors may be under—or overestimated. With multiple regression analysis,

[a]Electronic mail: m.kroesen@tudelft.nl.

 0001-4966/2008/123(6)/4250/11/$23.00

the effects of different nonacoustical factors *can* be controlled for noise exposure and other factors. However, this method is not suited to model indirect and reciprocal effects. Without being able to include these relationships, the model may contain serious misspecifications and hence lead to false inferences about the parameter estimates associated with different causes of annoyance.

This paper aims to overcome these shortcomings by developing and estimating a causal model of aircraft noise annoyance based on theory which includes nonacoustical and acoustical variables. The model is based on a conceptualization of noise annoyance by Stallen (1999) which is rooted in the psychological stress theory of Lazarus (1966). To the authors' knowledge, this is, as of yet, the only theory that gives an explanation for noise annoyance. Since the conceptual model, besides direct relationships, includes indirect and reciprocal relationships between variables, structural equation modeling (SEM) is applied to estimate the model. This method is especially suitable to model these complex causal relationships (Bollen, 1989). An additional advantage of SEM is that it can take measurement errors into account, which results in less bias in the estimated coefficients and potentially larger portions of explained variance. Data to estimate the model are gathered through a survey among residents living inside the 45 $DENL^2$ contour around Amsterdam Airport Schiphol (AAS), the largest airport in The Netherlands.

The structure of this paper is as follows. Section II discusses the causal model to explain aircraft noise annoyance which is based on the psychological stress theory. The third section presents the research approach and data gathering procedure. Section IV discusses the model results. The last section presents the main conclusions and concludes with some reflective remarks and related recommended directions for further research.

II. TOWARD A CAUSAL MODEL OF NOISE ANNOYANCE

This section first discusses the definition of noise annoyance, after which the model of Stallen (1999) is presented, which forms the core of the noise annoyance model to be tested in this paper. Following this, relevant acoustical and nonacoustical factors are identified and the constructed causal model is elaborated.

Based on a survey among experts, Guski *et al.* (1999) concluded that noise annoyance is a multifaceted concept, which covers immediate behavioral noise aspects, such as disturbance and interference with activities, and long-term evaluative aspects such as nuisance, unpleasantness, and getting on one's nerves. Although the two components of noise annoyance, i.e., disturbance and nuisance, can be theoretically distinguished, Guski (1999) noted that it is unknown how the integration of short-term experiences and long-term evaluation related to the acoustic environment takes place. It is unknown whether, for example, the most severe disturbances are remembered or whether a respondent averages all the disturbances he or she can remember. Guski *et al.* (1999 p. 525) also emphasised that noise annoyance is not just reflecting acoustic characteristics: "noise annoyance describes a situation between an acoustic situation and a person who is forced by noise to do things he or she does not want to do, who cognitively and emotionally evaluates this situation and feels partly helpless." This statement is in line with Stallen's (1999) definition of noise annoyance as a form of psychological stress, which constitutes the fundamental idea behind his conceptual model of noise annoyance and is discussed below.

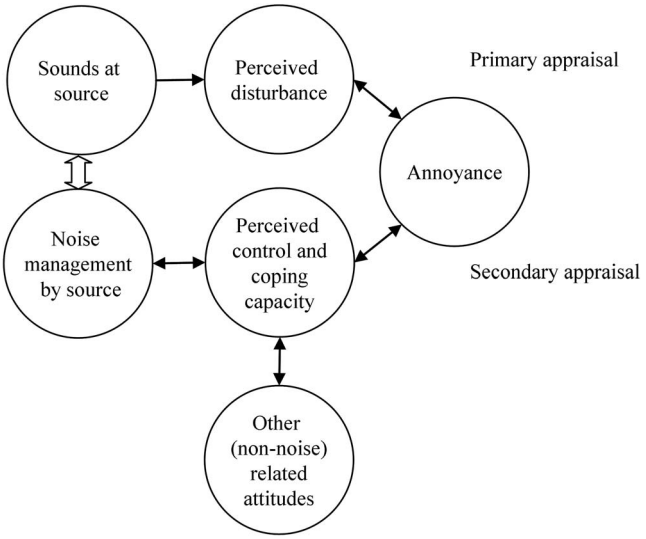


FIG. 1. The conceptual model of Stallen (1999) used to explain noise annoyance. Noise annoyance is defined as a form of psychological stress, which is determined by the perceived impact of a stressor and the perceived resources to cope with this stressor.

Different models have been developed that aim to provide insight into the processes that result in noise annoyance (Taylor, 1984; Job, 1996; Guski, 1999). However, all these models are developed based on empirical evidence related to previously found correlations between noise annoyance and other variables. Since these associations between noise annoyance and nonacoustical factors have been found in an exploratory manner, these models are based on implicit theory rather than on a predefined theory of noise annoyance. In his application of the psychological stress theory of Lazarus (1966) on the phenomenon noise annoyance, Stallen (1999) developed an explicit theoretical framework for describing the process of noise annoyance. Empirical research by Lazarus (1966) and others has revealed two major determinants of stress: perceived threat and perceived control. Stallen (1999) argued that the perceived disturbance (i.e., short-term or immediate annoyance) and the perceived threat basically form equal concepts. Subsequently, noise annoyance as a form of psychological stress is determined by the extent to which a person perceives a threat (i.e., perceived disturbance) and the possibilities or resources that a person has with which to face this threat (i.e., perceived control) (Stallen, 1999). Stallen's conceptual model is presented in Fig. 1. The presented model is a simplified and slightly adapted version of the original model (i.e., perceived control and coping capacity are treated as one factor).

The level of perceived disturbance, also called the primary appraisal, is a person's evaluation of the impact of the threat or harm in relation to his or her well being. The acous-

tic situation to which one is exposed is considered the main determinant of this evaluation. After a threat or harm is recognized, a process of secondary appraisal is triggered. Within this process, the resources to face the threat are evaluated. One potential resource results from the relationship one has with the noise source. If this relationship is good, one is better able to handle the impact of the stressor. However, in the words of Maris *et al.* (2007) (p. 2001): "if the exposed has little control over the source, or little trust in the source, the perceived coping resources will be reduced and psychological stress will arise." Next to the noise management by the source, other non-noise related attitudes can be considered as potential coping resources. In this respect, Stallen (1999) mentioned nonacoustical factors such as beliefs about the importance of the noise source and annoyance with non-noise impacts of the noise source, which were identified by Fields' extensive review as supported by sufficient evidence (Fields, 1993).

Based on his model, Stallen (1999) argued that if the perceived threat (i.e., noise) is larger than the perceived resources to face the threat (i.e., perceived control and coping capacity), psychological stress (i.e., noise annoyance) will arise. In addition, even though the perceived disturbance may be very high, no noise annoyance will arise if there are sufficient coping resources. Lastly, since the process of coping is in a constant flux, the theoretical framework includes multiple reciprocal relationships between variables.

To further extend the conceptual model of aircraft noise annoyance, relevant acoustical and nonacoustical factors that play a significant role in the noise-reaction relationship are supplemented. These variables are identified based on the results of past studies. In Table I, the variables found by Lercher (1996) and Guski (1999), based on reviews of studies that investigated the effects of personal, social, and contextual variables on annoyance, are enumerated. The current overview is complemented with studies by Miedema and Vos (1999) and Fields (1993) who assessed the influence of (non-)acoustical factors on annoyance via metaanalyses.

To limit the number of variables in order to avoid problems in the data collection phase only the variables of which the evidence is sufficiently present as indicated by the cited authors are included in the extended model of noise annoyance. An additional criterion for inclusion is that a theoretical notion must exist that explains how each variable influences one or more dependent variables (i.e., the "mechanism of causation") in the conceptual model of Stallen (Fig. 1). Such theoretical notions could not be given for neighborhood satisfaction (for which it is more likely to be a dependent variable itself), education, occupational/social status, and household size. In addition, since these latter three variables have only a small effect size on noise annoyance, causing an estimated extra annoyance equivalent to 2 dB day-night level or less (Miedema and Vos, 1999), their exclusion will not substantially affect the model. In addition, the variable "change in noise environment" is omitted. The reason for this is that the structural equation modeling approach assumes that an estimated model and hence the "process of noise annoyance" are in a stable state. The fact that the dose-response function, which predicts the percentage of highly annoyed people for varying levels of noise exposure, has not significantly changed for nearly a decade for residents living around AAS (RIVM and RIGO, 2006), suggests that this assumption holds in our study.[3] The exclusion of noise insulation will be explained in the next section.

The relevant acoustical and nonacoustical variables (Table I) and the conceptual model of Stallen (1999) (Fig. 1) are combined in an extended model of noise annoyance, which is constructed as follows. In line with Stallen's framework, noise annoyance is assumed to have two determinants, the perceived level of disturbance and the perceived level of control and coping capacity, which have a positive and a negative effect on noise annoyance, respectively.

The level of perceived disturbance is assumed to be positively influenced by the level of noise exposure and noise sensitivity. In turn, since noise sensitivity has been shown to be significantly associated with age and length of residence in noisy areas (for a brief review, see Van Kamp *et al.*, 2004), this variable is assumed to be influenced by these variables. Although on the balance of existing evidence, it is concluded that this length of residence in noisy areas has no significant relationship with annoyance (Table I; Fields, 1993), it is plausible that length of stay indirectly influences annoyance through the noise sensitivity of a person.

The perceived level of control and coping capacity are assumed to be directly influenced by the negative attitude toward noise source authorities and the noise policy (i.e., the noise management by the source) and by other nonacoustical variables (i.e., non-noise related attitudes). Dependent on whether the respective variables "add" or "extract" coping potential, the sign of the hypothesised relationship is either positive or negative. In addition, the effects of nonacoustical variables on the perceived level of control and coping capacity are assumed to be mediated by the negative attitude toward noise source authorities and the noise policy. For all included nonacoustical variables, the assumption that these variables can deteriorate or improve the relationship between residents around the airport and the noise source authorities (i.e., the government and airport operators) is plausible. Therefore, the hypothesis that these factors influence the attitude toward the source authorities and the policies they adopt to control the noise will also be tested. Hence, it is hypothesized that the nonacoustical variables directly influence the perceived level of control and coping capacity as well as indirectly via the negative attitude toward noise source authorities and the noise management. For example, a strong belief that noise can be prevented can directly lead to a perceived loss of coping potential (i.e., a lack of control over the situation) as well as increase distrust in the authorities and the adopted noise policy through which the coping potential also decreases.

Since, in the words of Stallen (1999), (p. 77), coping is a process with information flowing back and forth (i.e., the process of coping can be seen as a constant reappraisal of the person-environment relationship), Stallen's framework included several reciprocal relationships. In relation to the extended model described here, it is assumed that the perceived level of disturbance not only influences noise annoyance but also noise annoyance in turn, influences the degree of per-

TABLE I. Overview of acoustical and nonacoustical variables.

Nonacoustical variables	Sufficient evidence	Reference[c]	Included in the extended model
Critical tendencies[a]		1	
Negative affectivity		1	
Neuroticism/extraversion		1	
Locus of control		1	
Type A/B[b]		1	
Noncomplaining attitude		1	
Noise sensitivity	···	1, 2, 3, 4	···
Misfeasance in relation to source authorities	···	1, 2	···
Preventability beliefs	···	1, 4	···
Fear of noise source	···	1, 2, 3, 4	···
Concern about negative health effects of noise	···	1, 2	···
Social evaluation of the source/attitude towards the source	···	1, 2, 4	···
Interference with activities (i.e. activity disturbances)	···	1, 2	···
Controllability/predictability/adaptability in relation to noise situation	···	1	···
Annoyance in relation to non-noise effects (odour, vibrations)	···	1, 4	···
Neighborhood satisfaction	···	1	
Home ownership/concern about property devaluation	···	1, 3, 4	···
Aesthetic appearance of site		1	
Negative expectations related to future development of noise	···	2	···
Coping capacity	···	1, 2	···
Gender		3, 4	
Age	···	3	···
Education	···	3	
Income		4	
Occupational/social status	···	3	
Household size	···	3	
Personal evaluation of the source/dependency on the noise source	···	2, 3, 4	···
Length of residence/length of residence in noisy areas		4	···
Acoustical variables			
Noise exposure (e.g., DENL)	···	5	···
History of noise exposure levels/exposure time		1, 2, 4	
Change in noise environment/time since change	···	1, 4	
Home type and design/rooms facing noise source		1, 4	
Noise insulation	···	4	
Background noise level		4	

[a]The general tendency of individuals to express critical or negative judgments.
[b]Type A personality is a set of characteristics that includes being impatient, excessively time conscious, insecure about one's status, highly competitive, hostile and aggressive, and incapable of relaxation (Friedman and Rosenman, 1974).
[c]References: 1=Lercher (1996) and references presented in this paper; 2=Guski (1999) and references presented in this paper; 3=Miedema and Vos (1999); 4=Fields (1993); 5=Job (1988).

ceived disturbance. The hypothesis is that an annoyed person is more prone to be frequently disturbed by the acoustic environment. A second reciprocal relationship is assumed to be present between noise annoyance and the perceived level of control and coping capacity. It is hypothesized that more stress (annoyance) increases the incentive for people to find direct or indirect ways to cope with the stressor. In other words, it is assumed that being in a state of "high annoyance" leads people to adopt cognitive or direct coping strategies to reduce their level of stress. Glass and Singer (1972) used the term adaptation to characterize this process. They argue that since humans can rely on cognitive processes to achieve adaptation, they have a large variety of adaptive mechanisms at their disposal to protect themselves (Glass and Singer, 1972).[4]

It needs to be noted that the reciprocity assumed between noise annoyance and perceived disturbance is purely cognitive, while the path from perceived control and coping capacity toward noise annoyance is both cognitive (i.e., including latent mental processes such as emotional regulation) and behavioral (i.e., including direct coping strategies such as closing a window). Hence, to correctly model this process would require inclusion of such behavioral strategies in a feedback loop from noise annoyance to perceived control and coping capacity, in addition to the direct feedback loop

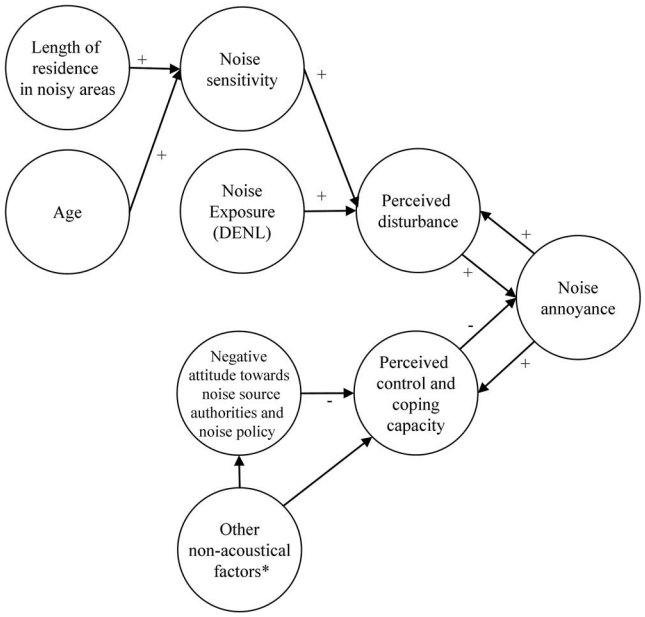


FIG. 2. The developed causal model of aircraft noise annoyance. Included nonacoustical factors are the following. (1). Belief noise can be prevented (−). (2). Positive social evaluation of the noise source (+). (3) Negative expectations related to noise development (−). (4). Personal dependency on noise source (+). (5) Concern about negative health effects of noise and pollution (−). (6) Annoyance by non-noise effects (i.e., vibrations, dust, and odor) (−). (7) Fear related to noise source (−) and (8) Concern about property devaluation (−). Note that the sign in the parentheses relates to the hypothesized relationship of the respective variable with perceived control and coping capacity (the sign of the assumed relationship with the negative attitude toward noise source authorities and the noise policy is the opposite of this sign).

which represents mental coping. However, by considering the range of different behavioral coping strategies, the fact that such strategies can have both positive and negative outcomes, and the fact that such behavioral responses have other antecedents next to noise annoyance (which would also needed to be taken into account), inclusion of this behavior in the present model would be too complex to achieve. Therefore, this additional indirect feedback loop is not explicitly modeled but assumed to be sufficiently captured by the direct feedback loop. Hence, it is assumed that these behavioral coping strategies have a net positive effect.

The extended causal model is depicted in Fig. 2.

III. RESEARCH APPROACH

A. Sample

The extended model depicted in Fig. 2 is parametrized in the form of structural equation model. Data to estimate this model were gathered via a survey among residents living inside 45 DENL contour around Amsterdam Airport Schiphol (AAS) in The Netherlands. Approximately 1.5×10^6 people live within this area. The lower limit of 45 DENL is chosen to physically constrain the size of the geographic survey area. Approximately 85% of all people around AAS who are being highly annoyed by aircraft noise live within this contour (RIVM and RIGO, 2006). Highly annoyed in this respect is defined according to the convention definition of a score of 72 or higher on a scale from 0 (no annoyance at all) to 100 (very high annoyance) (e.g., Miedema and Vos, 1998). The level of noise exposure in the dataset[5] ranges from 45 DENL through 58 DENL (only 0.8% of the people around AAS exposed to 45 DENL or more are exposed to higher levels than 58 DENL). Since only residents who are exposed to 60 DENL or more are eligible to receive noise insulation, and because the upper limit of the level of noise exposure in the sample is 58 DENL, the effects of this variable could not be estimated and it was therefore excluded from further analysis.

From the chosen geographical survey area, a random sample of dwellings was selected. Per selected dwelling, one resident was approached via a letter (delivered at the home address) that invited him or her to fill in an online questionnaire. The letter contained the URL of the website where this questionnaire could be reached. The survey was conducted at the beginning of April 2006. Although issues surrounding Airport Schiphol are highly controversial (i.e., expansion, noise pollution, and emissions) (Van Eeten, 2001), there was no public debate or explicit media attention at the time of data collection or in the preceding months.

Considering the large amount of variables and to avoid problems with multicollinearity and/or deviations from normality, the sample size had to be sufficiently large (at least more than 400). Based on an expected response ratio of 10%, 7000 residents were approached. With 646 useable responses, the actual response ratio was 9.2%. The completion ratio was 91.8%, which indicates that there was no serious matter of questionnaire fatigue.

The choice for an internet questionnaire was based on the advantages this method brings in term of speed and costs. Based on a comparison of a large internet sample and 500 traditional samples, Gosling *et al.* (2004) concluded that internet findings are consistent with findings from traditional methods and that these methods can contribute to many areas of psychology. However, the use of this method has been criticized due to (1) problems of internet coverage of the general population (Couper, 2000), (2) the difficulty of drawing probability samples (Couper, 2000), and (3) high nonresponse rates (Braunsberger *et al.*, 2007).

In relation to the first, it can be noted that internet access in The Netherlands is among the highest in the world. In 2005 83% of the Dutch population had access to the internet (CBS, 2006), which suggests that the internet population accurately reflects the general population. However, usual differences found between the general population and the internet population, i.e., people with internet access are generally better educated, have higher incomes and are generally younger, have also been found in our sample (although this also might be due to the fact that, in general, these people are more motivated to participate in surveys). More specifically, a small overrepresentation exists of well-educated respondents and respondents with high incomes. However, the mean age of the respondents in the sample (mean standard deviation=49.8(14.5)) is not much different from (even higher than) the average age of the Dutch population of 18 years and older (mean=46.7).[6] In addition, since these

variables are not strongly related to the main variable of interest, i.e., noise annoyance (see Table I), the bias present in the sample is considered to be negligible.

With respect to the second point of critique in relation to internet research, it can be noted that the usual problem of self-selection in web-based surveys, which prohibits generalizations in relation to a larger population, has been limited through the use of traditional methods for the sampling and recruitment of respondents. As mentioned earlier, a random sample was drawn from the survey area and respondents were approached via a letter that was delivered at their home address. In addition, the use of cookies prevented multiple entries from the same respondents.

This leaves the issue of nonresponse, which is of course also present in traditional postal or telephone surveys, unaddressed. Nonresponse is undesirable insofar there are main differences between the respondents and nonrespondents on the variables of interest. In this study, it is likely that annoyed people are more (than less annoyed people) inclined to participate. Based on the positive correlation found between the difference in the actual and expected response per municipality and the average noise annoyance score per municipality ($r=0.235$, $p=0.000$), it can be concluded that the sample has indeed a small bias toward people who experience more noise annoyance than the average person living in the 45 DENL contour. However, since, to the authors' knowledge, previous empirical research has never indicated that the relation between nonacoustical factors and noise annoyance is different for varying degrees of noise annoyance, it is assumed that this small overrepresentation did not bias the estimated relationships between noise annoyance and other factors. Yet, this remains an issue of empirical investigation.

B. Measurements

Except for age, length of residence in noisy areas, concern about property devaluation, and noise exposure, the variables presented in Fig. 2 represent complex concepts that are considered to be latent variables. Latent variables are not measured by a single question in the questionnaire, but these are measured with multiple indicators. Noise annoyance and noise sensitivity were measured by previously validated scales. For noise annoyance, two standardized noise reaction items were used (Fields *et al.*, 2001). Since three items per scale form a preferable minimum the scale was expanded with one item, which relates to annoyance due to disturbances. To measure noise sensitivity, the 21-item scale of Weinstein is used (Weinstein, 1978). Because of limited space in the questionnaire, a selection of ten items was included. It has been previously shown that this selection provides a reliable scale for noise sensitivity (Breugelmans *et al.*, 2004). In addition, to increase the reliability of this scale, it is expanded with one general noise sensitivity question measured with an 11-point scale. All other scales are composed of newly formulated indicator variables, which are measured on seven-point Likert-type scales.

Normal procedure in structural equation modeling is to include all indicators of each latent variable into the structural equation model and thereby taking measurement error into account. However, to reduce the overall complexity of the model (i.e., the number of free parameters to be estimated) and since our interest lies in testing the structural part of the model, we constructed the different latent constructs *a priori* by calculating sum scores of the multiple indicators and including only these summated scales as the indicators of the latent variables in the structural equation model. Following this procedure, the measurement error can still be taken into account (thereby retaining the benefits of a measurement model) if the measurement error of the summated scale is specified in the structural equation model. This is done by fixing the measurement error of the summated scales (the single indicator variable) at a value of 1 minus the Cronbach's alpha of the summated scale (Kelloway, 1998).

To that effect, the Cronbach's alpha of each summated scale was calculated in the statistical software package SPSS. By calculating the Cronbach's alpha, one assumes that the items represent a unidimensional scale, but the measure itself does not reveal whether this is the case or not. Therefore, factor analysis was conducted prior to calculating the Cronbach's alpha to check the unidimensionality of the each intended scale. Except for the construct "belief noise can be prevented," a single factor was found for each construct, implying that the summated scales are unidimensional. For the construct belief noise can be prevented, the item that has the highest correlation with the central variable, i.e., noise annoyance, is chosen as a single indicator to represent that latent variable. Furthermore, to ensure that each item sufficiently contributed to the measurement of the complex construct, only those items remained in the scale that had a factor loading larger than 0.50. Table II presents an overview of the included scales and their respective items. Since no reliability value can be derived for the single item constructs belief noise can be prevented and "concern about property devaluation," these variables are assumed to be measured with the average reliability of all scales ($\alpha=0.83$). All other constructs were represented by summated scale scores computed as the sum of the individual item scores.

IV. RESULTS

The model tests and parameters estimates are based on the covariance matrix and used maximum likelihood estimation as implemented in Lisrel 8 (Jöreskog and Sörbom, 1992).[7] After estimation of the full model in Fig. 2, the insignificant paths are deleted and the model is re-estimated. Insignificant paths can be considered irrelevant to the model and should, based on the parsimony criterion, be deleted from the model (Byrne, 1998). Hence all insignificant paths are fixed to zero. Variables that are left with no path are deleted from the model, after checking the modification indices to assess whether paths should be drawn that were not theoretically expected. These indices indicate the decrease in the chi-square value (i.e., improved fit) if an extra path between two factors is added. After this step, the following five factors are removed from the model as these have no significant relationships with other variables left: noise sensitivity, fear of noise source, personal dependency on the noise source, length of residence in noisy areas, and age. Hence,

TABLE II. Overview of scales, Cronbach alpha's, items, and item ranges. Items with factors loadings smaller than 0.50 were removed from the solutions.

Scale/latent variable	Alpha	Item	Range
Noise annoyance(past 12 months)	0.89	Level of annoyance due to air traffic 1	0=not annoyed at all–10=very high annoyance
		Level of annoyance due to air traffic 2	1=not annoyed at all–5=extremely annoyed
		Level of annoyance due to disturbances	1=not annoyed at all–7=very high annoyance
Perceived disturbance (past 12 months)	0.88	Disturbances by aviation traffic during daytime	1=never–5=daily
		Disturbances by aviation traffic in sleep	1=never–5=daily
		Disturbances by aviation traffic during conversations	1=never–5=often
		Disturbances by aviation traffic during activities that demand concentration	1=never–5=often
		Disturbances by aviation traffic during resting	1=never–5=often
Negative expectations toward noise development	0.83	Belief personal noise situation will worsen	1=noise situation will improve–5=noise situation will deteriorate
		General belief noise exposure will increase	1=noise level will decrease–7=noise level will increase
Noise sensitivity	0.86	General noise sensitivity	0=not at all noise sensitive–10=highly noise sensitive
		I get used to most noises without much difficulty	1=completely agree–5=completely disagree
		I am good at concentrating no matter what is going on around me	1=completely agree–5=completely disagree
		I am easily awakened by noise	1=completely agree–5=completely disagree
		I find it hard to relax in a place that is noisy	1=completely agree–5=completely disagree
		I am sensitive to noise	1=completely agree–5=completely disagree
		Sometimes noises get on my nerves and get me irritated	1=completely agree–5=completely disagree
		I get angry with people making noise	1=completely agree–5=completely disagree
Fear of noise source	0.76	Fear of aircraft crash in neighborhood	1=no fear at all–7=extremely fearful
		Frightened when aircrafts fly over	1=not frightened at all–7=extremely frightened
Positive social evaluation of noise source	0.79	I believe Schiphol is valuable for the region	1=completely agree–7=completely disagree
		I believe Schiphol is important for the Dutch economy	1=completely agree–7=completely disagree
		I believe flying is a sustainable transportation mode	1=completely agree–7=completely disagree
Negative attitude toward noise source authorities and the noise policy	0.92	I believe Schiphol must be able to grow at its current location	1=completely agree–7=completely disagree
		General attitude toward Schiphol	1=very negative–7=very positive
		Satisfaction with Schiphol policy, in general	1=not satisfied at all–7=extremely satisfied
		Belief Schiphol abuses its power	1=no abuse at all–7=a lot of abuse
		I trust the government to maintain a good balance betweenenvironmental and economic factors	1=completely agree–7=completely disagree
		I trust the government to uphold the environmental norms	1=completely agree–7=completely disagree
		I believe the government acknowledges the noise problem	1=completely agree–7=completely disagree
		Satisfaction with government policy on noise	1=not satisfied at all–7=extremely satisfied
Concern about the negative health effects of noise and pollution	0.91	Concern that pollution leads to negative health effects	1=not concerned at all–7=very much concerned
		Concern that noise leads to negative health effects	1=not concerned at all–7=very much concerned
		Concern that noise leads to sleep loss	1=not concerned at all–7=very much concerned
		Concern that noise leads to more stress	1=not concerned at all–7=very much concerned
Annoyance related to non-noise effects	0.85	Annoyed by odour due to aircrafts	1=not annoyed at all–7=very much annoyed
		Annoyed by vibrations due to aircrafts	1=not annoyed at all–7=very much annoyed
		Annoyed by particles, dust or smoke due to aircrafts	1=not annoyed at all–7=very much annoyed
Personal dependency on the noise source	0.65	Importance of Schiphol in relation to job	1=not important at all–7=very important
		Dependency on Schiphol due to travel needs	1=not dependent at all–7=very dependent
		Financial dependency on Schiphol	1=not dependent at all–7=very dependent
Perceived control and coping capacity	0.77	Feeling of direct control (via physical measures) over the experienced level of noise exposure	1=no control at all–7=very much control
		Feeling of being powerless in relation to the noise situation	1=very powerless–7=not powerless at all
		Capacity to deal with aircraft noise	1=very low capacity–7=very high capacity

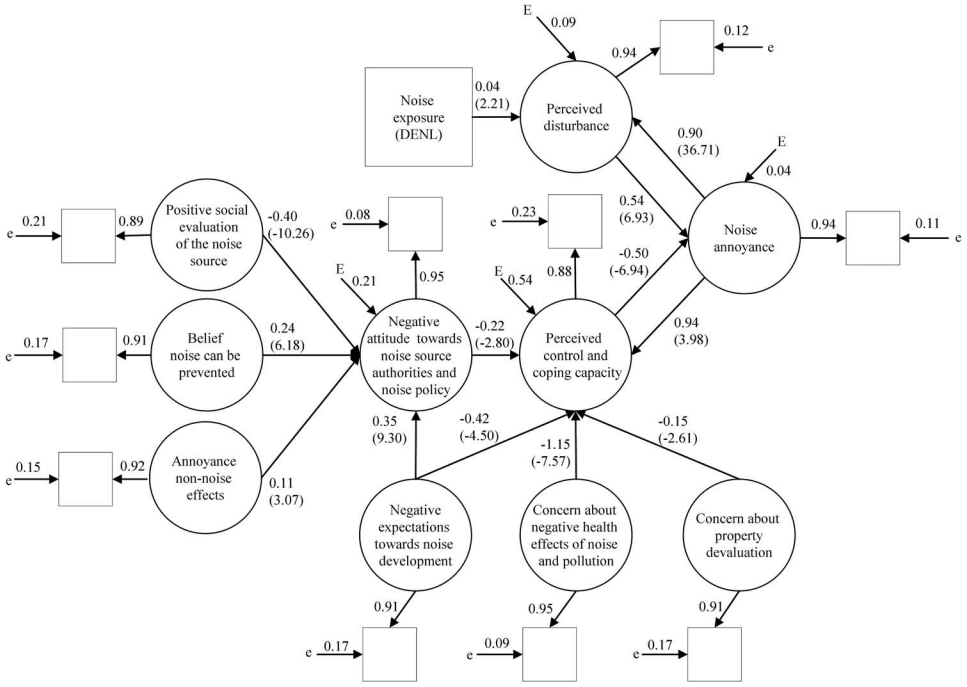


FIG. 3. The estimated aircraft noise annoyance model. $n=646$, $\chi^2=54.45$, p-value$=0.000\,08$, $Df=21$, GFI$=0.99$, CFI$=1.00$, and RMSEA$=0.044$. The standardized path estimates are shown. The values in the parentheses represent the t-values of the structural parameter estimates. All parameter estimates are significant ($p<0.05$). (○) Latent variable; (E) error/unexplained variance of latent variable; (□) observed variable (based on single-item composite scale); (e) error/unexplained variance of observed variable.

taking into account the other variables that are still in the model, these variables are unable to explain additional variance in the endogenous variables. Figure 3 presents the final model.

The chi-square value is statistically significant ($\chi^2=54.45$, $p=0.000\,08$), which means that the model implied covariance matrix is significantly different from the observed covariance matrix. However, since this statistic is very sensitive for large sample sizes ($N>500$), the review of other fit indices is recommended (Browne and Cudeck, 1993; Hu and Bentler, 1995; Schermelleh-Engel *et al.*, 2003). The values for the goodness-of-fit index (GFI) and the comparative fit index (CFI) are well above the recommended lower limit of 0.90, which suggests a good model fit. The root mean square error of approximation (RMSEA), a badness-of-fit index, has a value below the recommended upper limit of 0.05, which again suggests a good model fit. Overall, it can be concluded that the model fit is good. In addition, all the signs of the hypothesised relationships between the variables are as expected.

The values related to each path in Fig. 3 represent the standardized parameter estimates. Standardization of the estimates makes comparisons in terms of the relative importance of each path possible. It can be concluded that the effect sizes of perceived disturbance and perceived control and coping capacity on noise annoyance, 0.54 and −0.50, respectively, are quite similar. The effects of noise annoyance on perceived disturbance and perceived control and coping capacity, 0.90 and 0.94, respectively, are also of the same magnitude. It can be concluded that to a large extent, the reciprocal effects between noise annoyance and perceived disturbance and noise annoyance and perceived control and coping capacity cancel each other out.

The only significant determinant of perceived disturbance is the level of noise exposure. However, the effect size of 0.04 can be qualified as small. The significant determinants of the perceived level of control and coping capacity are the negative attitude toward noise source authorities and the noise policy (−0.22), the negative expectations related to noise development (−0.42), the concern about negative health effects of noise and pollution (−1.15), and the concern about property devaluation (−0.15). Especially, the concern about negative health effects has a large effect on the capacity of people to handle the noise situation.

The positive social evaluation of noise source (−0.40), the belief that noise can be prevented (0.24), and annoyance related to non-noise effects (0.11) influence the negative attitude toward noise source authorities and the noise policy. The most important determinant of this factor is the positive social evaluation.

Only the negative expectation related to the future noise development has both a direct (−0.42) and an indirect effect ($0.35*-0.22=-0.077$) on the perceived level of control and coping capacity. The presence of both effects is theoretically plausible. The indirect effect, via the negative attitude toward source authorities, can be explained by the mechanism that, if the belief exists that the noise situation will worsen, the noise source authorities are to blame for the expected increase in noise, which negatively influences the attitude toward the authorities. The direct effect, on the other hand, can be explained by the mechanism that a negative expectation related to the future noise development creates an immediate sense of despair (i.e., expecting that the situation will be become worse makes the appreciation of the current situation worse).

In order to assess the total effect of each variable on the central variable noise annoyance, the standardized total effects need to be assessed. These are presented in Table III.

The total effect of a variable is the combination of the indirect and direct effects. It can be concluded that the concern about negative health effects of noise and pollution, the

TABLE III. Standardized total effects of each variable on noise annoyance.

Variable	Effect
Concern about negative health effects of noise and pollution	0.59
Perceived disturbance	0.56
Perceived control and coping capacity	–0.51
Negative expectations toward noise development	0.26
Negative attitude toward source authorities	0.11
Concern about property devaluation	0.08
Positive social evaluation of the noise source	–0.05
Belief noise can be prevented	0.03
Noise annoyance	0.02
Noise exposure (DENL)	0.02
Annoyance non-noise effects	0.01

perceived disturbance, and the perceived control and coping capacity are the most important determinants of noise annoyance. Noise annoyance (via the reciprocal relationships), noise exposure, and annoyance non-noise effects have the lowest total effects on noise annoyance.

The error terms of the latent constructs (i.e., the E's in Fig. 3) indicate the proportions of unexplained variance of the endogenous variables. Since the model is nonrecursive (i.e., it includes feedback loops), the interpretation of the proportion of explained variance ($1-E$ or R^2) is not the same as it would be in traditional regression analysis (Jöreskog, 2000). This interpretation only holds for the negative attitude toward source authorities and the noise policy. Jöreskog (2000), therefore, advised assessing the R^2's calculated from the reduced form equations, which indicate the proportions of variance in the endogenous variables solely explained by the exogenous variables. For each endogenous variable, the R^2 of the reduced form is presented in Table IV.

Even though the unique portions of variance explained by the endogenous variables in each other are not included, the R^2 values are still high. In addition, the explained variance in noise annoyance, 78%, is considerably higher than in a path model previously estimated on this topic, which was able to explain 42% in noise annoyance (Taylor, 1984).

V. CONCLUSION AND DISCUSSION

In this study, a structural equation model is developed and estimated to explain aircraft noise annoyance. In contrast to existing models that largely lack a sound theoretical basis, the model presented in this paper is theoretically well founded. As a result, the model provides a better insight into the factors and causal processes that precede and result in aircraft noise annoyance. In addition, the use of SEM to model and explain noise annoyance has proven itself to be a suitable method in overcoming the shortcomings of previously used methods such as correlational analysis and multiple regression analysis. The final model provides a good model fit and supports the presence of indirect and reciprocal effects, which empirically have not previously been identified. It can be concluded that the concern about the negative health effects of noise and pollution, the level of perceived disturbance, and the level of perceived control and coping capacity have the highest total effects on noise annoyance. Finally, the proportion of explained variance in noise annoyance is higher than in previous models.

TABLE IV. The proportions of explained variance in the endogenous variables (R^2's) based on the reduced form equations.

Endogenous variable	R^2
Noise annoyance	0.78
Perceived disturbance	0.65
Perceived control and coping capacity	0.79
Negative attitude toward source authorities	0.79

Controlled for other variables still in the model, the variables noise sensitivity, fear related to the noise source, personal dependency on the noise source, length of residence in noisy areas, and age have no significant relationships with endogenous variables in the model and were therefore excluded from the model. The exclusion of the variables noise sensitivity and fear in relation to the noise source is especially remarkable, since many studies emphasise the importance of these factors (e.g., see Fields, 1993; Van de Kamp *et al.*, 2004; Miedema and Vos, 1999). Although these variables show significant correlations with noise annoyance, 0.51 and 0.50, respectively, they are unable to explain additional variance given the other variables still in the model. With respect to the exclusion of fear related to the noise source, a probable explanation is that the concern about the negative health effects of noise and pollution explains the same variance in the perceived control and coping capacity variable. This explanation seems reasonable since a fairly strong correlation between fear related to the noise source and concern about negative health effects of noise and pollution exists ($r=0.54$, $p=0.000$). An explanation of a similar form can be found for the exclusion of noise sensitivity. The variable perceived control and coping capacity show a significant correlation with noise sensitivity ($r=-0.48$, $p=0.000$), and its influence in the model is the probable cause for the suppression of the effect of noise sensitivity. However, as opposed to the relation between fear and the concern about negative health effects, we cannot identify a theoretical explanation why noise sensitivity and perceived control and coping capacity are empirically associated. Based on this finding, we recommend future research to address this particular relationship and the theoretical mechanism that underlies it, as well as, from a more general perspective, the relationship between noise sensitivity and other nonacoustical factors.

In relation to this study, some reflective remarks and related recommended directions for further research can be made. The first remark and associated research direction is related to the theoretical framework, developed by Stallen (1999), on which our causal model is based. Based on this theoretical framework, the specified model structure presented in this paper is deemed the most plausible one. The fact that the model is not falsified, however, does not exclude the validity of other theoretical frameworks. With respect to the apparent lack of theoretical insights in the phenomenon noise annoyance, we stress that future research related to the acoustical and nonacoustical antecedents of noise annoyance

should focus on the fundamental causal mechanisms that exist between variables, in addition to finding statistically significant associations between them. Rich qualitative descriptions related to the causal mechanisms "at work" between variables can be used to verify or falsify the used model structure or can be used to develop a new theoretical framework and related model structure(s). Since these descriptions cannot be derived from current theoretical insights or from traditional quantitative approaches, other means to derive these will have to be explored. A qualitative research approach (e.g., using in-depth interview techniques) is considered to be suitable in this respect.

A second remark and associated research opportunity, which partly overlaps with the previous one, relates to the assumed temporal causal order between variables. The estimated relationships in the model depicted in Fig. 3 are based on the assumption that the identified causes (independent variables) precede the effects (dependent variables) in time. However, as opposed to other causal models, the assumption of time precedence in our aircraft noise annoyance model is questionable. All the variables in the model, except noise exposure, constitute concepts such as beliefs, attitudes, perceptions, expectations, and evaluations. These types of variables are by nature very abstract. Although a causal ordering can be assumed based on theoretical notions (e.g., a general belief precedes a specific attitude), this assumption cannot be empirically investigated. The reason for this is that the model is based on cross-sectional data. Inferences based on this model about the temporal order between variables and the directions of causation are therefore inherently less strong. This is especially true for the estimated reciprocal relationships (i.e., does perceived control and coping capacity cause noise annoyance, vice versa, or does indeed a reciprocal relationship exist?). Hence, with respect to our developed aircraft noise annoyance model as well as future models to explain noise annoyance, special attention to the tenability of the assumption of time precedence is justified. A suitable approach to empirically investigate the tenability of the time-precedence criterion is through the use of panel data. More specifically, a SEM panel design can yield empirical evidence of a specific causal ordering between two variables (Finkel, 1995).[8]

A third direction for further research is to apply the model to residents around other airports in varying countries and explore similarities and differences between them. It should be taken into account that country or airport specific variables can play a role. These variables can be related to cultural characteristics of the country or to the specific policy context of the airport. For example, the qualitative research Bröer (2006) shows that the policy discourse at an airport influences the meaning people attribute to the sound of aircrafts. This, in turn, influences their experienced level of annoyance. In addition, through cross-national comparative research "best practices" of (nonacoustical) sound management can be identified.

The last research direction is related to the inclusion of acoustical and situational factors (e.g., frequency, tone, impulsiveness, time of day, the presence of noise insulation, arrangement of rooms and home type, and background noise level). The model in this study included only a year's mean noise exposure metric (DENL). The limited range of this metric (i.e., 45–58 DENL) has likely contributed to the relatively low effect of this variable in the estimated model (see also Job, 1988). The assessment of the influence of noise exposure can be improved by taking into account a larger geographical area for sample selection (to include levels below 45 DENL) and by oversampling (and subsequent weighting) of residents with high exposure levels (above 58 DENL). Especially in the case of AAS, oversampling is necessary since a relatively small proportion of the total population is exposed to these high levels of noise exposure. In addition, the assessment can possibly be improved through the inclusion of noise descriptors based on other weighting filters (e.g., C-weighting) or a dynamic filter (Schomer, 2001). Lastly, to estimate the relative importance of other acoustical and situational variables, we recommend inclusion of these factors in future models of aircraft noise annoyance.

To conclude, we believe that insights into the preceding factors and causal processes of aircraft noise annoyance open the door for revision of existing policies and the design of new policies to reduce this adverse effect. Treating aircraft noise annoyance around airports as a mere technical problem, involving exposure levels and dose-response functions is only one side of addressing the noise problem.

ACKNOWLEDGMENTS

The authors wish to thank two anonymous reviewers for their useful comments and suggestions on an earlier draft of this paper.

[1]Different metrics exist to indicate the level of noise exposure (e.g., energy-based indices and number of events). A study of Vincent *et al.* (2000) revealed that correlations between these different noise exposure metrics and noise annoyance are both low ($r \sim 0.30$) and very similar. Hence, it can be concluded that noise level descriptors, in general, are unable to explain individual levels of noise annoyance.

[2]DENL (day-evening-night level) is an equivalent sound level of 24 h expressed in decibels on the "A" weighted scale dB(A), which, in this study, is calculated for the period of a year. Sound levels during the evening (7 pm–11 pm) and during the night (11 pm–7 am) are increased by penalties of 5 and 10 dB(A), respectively. This metric is selected by the European Council to monitor and assess noise problems in its member states. It needs to be noted, however, that this metric has been criticized for its use to assess environmental noise because A-weighting approximates the response characteristics of the human ear only for narrow band sounds at low levels. It has been shown to underestimate the effects of low-frequency noise on pleasantness and annoyance ratings (e.g., Schomer *et al.*, 2001). In addition, broad band sounds such as aircraft noises are underestimated by A-weighting with respect to their loudness and annoyance by typically 15 dB.

[3]However, this also means that in situations where drastic changes in exogenous factors take place, our model cannot be used.

[4]Glass and Singer (1972) (p. 10) also note that "continued exposure to a stressor may produce cumulative effects that appear only after stimulation is terminated; it is as though the organism does not experience maximal stress until he is no longer required to cope with the stressor." Since this effect only incurs after the stressful situation has passed, it is not included in our causal model.

[5]For every respondent in the sample, the level of noise exposure (a year mean DENL) was calculated by the National Aerospace Laboratory (NLR). This was done by transforming the four-digit two-letter postal code of each respondent's residence, which includes on average an area of 50 m^2 (approximately 15 households) (Batty *et al.*, 2004), into *XY*-coordinates, which are subsequently used to determine the level of

noise exposure at the particular location. Calculations for the level of noise exposure are based on the 12 month period before the execution of the survey (the period from May 2005 to April 2006).

[6]Ideally, the sample should be compared to the chosen population (i.e., all residents within the 45 DENL contour). However, since no demographic information was available for this population, the sample was compared to the Dutch population of 18 years and above.

[7]With respect to the developed causal model in Fig. 2, two theoretical uncertainties were identified. Since the ultimate objective was to develop a model that is both theoretically meaningful and statistically well fitting, these two theoretical uncertainties were combined in four alternative model specifications, which all represented plausible views on reality. To find the model that was "most plausible," all four models were estimated and compared. In comparison to the other models, the model discussed in this paper, which is presented in Fig. 2, provided the best fit to the data and was therefore assumed to reflect the most plausible view on reality. To be able to present a concise paper, the choice was made not to include the discussion of these alternative models nor their results. These can, however, be requested in correspondence with the authors.

[8]Within a SEM panel design, the effect of an independent variable X_0 (read: X at time point 0) on a dependent variable Y_1 is controlled for Y's own stability. Hence, if X_0 is able to explain variation in Y_1, over and above the variation Y_0 can explain in Y_1 (the stability of Y), it can be empirically inferred that X is a causal predictor of Y.

Alexandre, A. (**1976**). "An assessment of certain causal models used in surveys on aircraft noise annoyance," J. Sound Vib. **44**, 119–125.

Batty, M., Besussi, E., Maat, K., and Harts, J. J. (**2004**). "Representing multifunctional cities: density and diversity in space and time," Build. Environ. **30**, 324–337.

Bollen, K. A. (**1989**). *Structural Equations with Latent Variables* (Wiley, New York).

Braunsberger, K., Wybenga, H., and Gates, R. (**2007**). "A comparison of reliability between telephone and web-based surveys," J. Bus. Res. **60**, 758–764.

Breugelmans, O. R. P., Van Wiechen, C. M. A. G., Van Kamp, I., Heisterkamp, S. H., and Houthuijs, D. J. M. (**2004**). "Gezondheid en beleving van de omgevingskwaliteit in de regio Schiphol: 2002 ("Health and the perception of environmental quality in the Schiphol area")," Bilthoven, The Netherlands.

Bröer, C. (**2006**). *Beleid vormt overlast (Policy frames annoyance)* (Aksant, Amsterdam).

Browne, M. W., and Cudeck, R. (**1993**). in *Testing Structural Equations Models*, edited by K. A. Bollen and J. S. Long (Sage, Newbury Park), pp. 136–162.

Byrne, B. M. (**1998**). *Structural Equations Modeling with Lisrel, Prelis and Simplis: Basic Concepts, Applications and Programming* (LEA, London).

Central Bureau of Statistics (CBS) (**2006**). "E-mailen en chatten populairste internetactiviteiten ("E-mail and chatting are the most popular internet activities")," by G. Linden, CBS Webmagazine, Voorburg, The Netherlands.

Couper, M. P. (**2000**). "Web surveys: a review of issues and approaches," Public Opin. Q. **64**, 464–494.

Fields, J. M. (**1993**). "Effect of personal and situational variables on noise annoyance in residential areas," J. Acoust. Soc. Am. **46**, 2753–2763.

Fields, J. M., de Jong, R. G., Gjestland, T., Flindell, I. H., Job, R. F. S., Kurra, S., Lercher, P., Vallet, M., Yano, T., Guski, R., and Schumer, R. (**2001**). "Standardised general-purpose noise reaction questions for community noise surveys: research and a recommendation community response to noise team of ICBEN (the International Commission on the Biological Effects of Noise)," J. Sound Vib. **242**, 641–679.

Finkel, S. E. (**1995**). *Causal Analysis with Panel Data* (Thousand Oaks, London).

Friedman, M., and Rosenman, R. H. (**1974**). *Type A behavior and Your Heart* (Knopf, New York).

Glass, D. C., and Singer, J. E. (**1972**). *Urban Stress. Experiments on Noise and Social Stressors* (Academic, New York).

Gosling, S. D., Vazire, S., Srivastava, S., and John, O. P. (**2004**). "Should we trust web-based studies?: A comparative analysis of six preconceptions about internet questionnaires," Am. Psychol. **59**, 93–104.

Guski, R. (**1999**). "Personal and social variables as co-determinants of noise annoyance," Noise Health **1**, 45–56.

Guski, R., Felscher-Suhr, U., and Scheuemer, R. (**1999**). "The concept of noise annoyance: how international experts see it," J. Sound Vib. **223**, 513–527.

Hu, L., and Bentler, P. M. (**1995**). in *Structural Equation Modeling: Concepts, Issues, and Applications*, edited by R. H. Hoyle (Sage, Thousand Oaks, CA), pp. 76–99.

Job, R. F. S. (**1988**). "Community response to noise: A review of factors influencing the relationship between noise exposure and reaction," J. Acoust. Soc. Am. **83**, 991–1001.

Job, R. F. S. (**1996**). "The influence of subjective reactions to noise on health effects of the noise," Environ. Int. **22**, 93–104.

Jöreskog, K. G. (**2000**). The interpretation of R^2 revisited. Retrieved 12 July 2007 from http://www.ssicentral.com/lisrel/techdocs/r2rev.pdf.

Jöreskog, K. G., and Sörbom, D. (**1992**). *LISREL VIII: Analysis of Linear Structural Relations* (Scientific Software, Mooresville).

Kelloway, E. K. (**1998**). *Using Lisrel for Structural Equations Modeling, A Researcher's Guide* (Sage, California).

Lazarus, R. S. (**1966**). *Psychological Stress and the Coping Process* (McGraw-Hill, New York).

Lercher, P. (**1996**). "Environmental noise and health: an integrated research perspective," Environ. Int. **22**, 117–129.

Maris, E., Stallen, P. J., Vermunt, R., and Steensma, H. (**2007**). "Noise within the social context: annoyance reduction through fair procedures," J. Acoust. Soc. Am. **121**, 2000–2010.

Miedema, H. M. E., and Vos, H. (**1998**). "Exposure-response relationships for transportation noise," J. Acoust. Soc. Am. **104**, 3432–3445.

Miedema, H. M. E., and Vos, H. (**1999**). "Demographic and attitudinal factors that modify annoyance from transportation noise," J. Acoust. Soc. Am. **105**, 3336–3344.

RIVM and RIGO (**2006**). "Evaluatie Schipholbeleid. Schiphol beleefd door omwonenden ("Policy evaluation Schiphol: Schiphol experienced by residents")," Bilthoven, The Netherlands.

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (**2003**). "Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures," MPR-online 8, 23–74.

Schomer, P. D., Suzuki, Y., and Saito, F. (**2001**). "Evaluation of loudness-level weightings for assessing the annoyance of environmental noise," J. Acoust. Soc. Am. **110**, 2390–2397.

Stallen, P. J. M. (**1999**). "A theoretical framework for environmental noise annoyance," Noise Health **1**, 69–80.

Taylor, S. M. (**1984**). "A path model of aircraft noise annoyance?" J. Sound Vib. **96**, 243–260.

Van Eeten, M. J. G. (**2001**). "Recasting intractable policy issues: The wider implications of The Netherlands civil aviation controversy," J. Policy Anal. Manage. **20**, 391–414.

Van Kamp, I., Job, R. F. S., Hatfield, J., Haines, M., Stellato, R. K., and Stansfeld, S. A. (**2004**). "The role of noise sensitivity in the noise-response relation: A comparison of three international airport studies," J. Acoust. Soc. Am. **116**, 3471–3479.

Vincent, B., Vallet, M., Olivier, D., and Paque, G. (**2000**). "Evaluation of variations of the annoyance due to aircraft noise," Internoise 2000, Nice, France, 27–30 August.

Weinstein, N. D. (**1978**). "Individual differences in reactions to noise: A longitudinal study in a college dormitory," J. Appl. Psychol. **63**, 458–566.

Modeling the sound transmission between rooms coupled through partition walls by using a diffusion model

Alexis Billon
University of Liège, Sart-Tilman B28, B-4000 Liège, Belgium

Cédric Foy
CEBTP-SOLEN, 12 Avenue Gay Lussac, ZAC La Clef Saint Pierre, 78990 Elancourt, France

Judicaël Picaut[a)]
Laboratoire Central des Ponts et Chaussées, Section Acoustique Routière et Urbaine, Route de Bouaye, B.P. 4129, 44341 Bouguenais Cedex, France

Vincent Valeau
Université de Poitiers, LEA UMR CNRS 6609, 40 Avenue du Recteur Pineau, 86022 Poitiers Cedex, France

Anas Sakout
Université de La Rochelle, LEPTIAB, Avenue Michel Crépeau, 17042 La Rochelle Cedex 01, France

(Received 28 September 2007; revised 14 February 2008; accepted 13 March 2008)

In this paper, a modification of the diffusion model for room acoustics is proposed to account for sound transmission between two rooms, a source room and an adjacent room, which are coupled through a partition wall. A system of two diffusion equations, one for each room, together with a set of two boundary conditions, one for the partition wall and one for the other walls of a room, is obtained and numerically solved. The modified diffusion model is validated by numerical comparisons with the statistical theory for several coupled-room configurations by varying the coupling area surface, the absorption coefficient of each room, and the volume of the adjacent room. An experimental comparison is also carried out for two coupled classrooms. The modified diffusion model results agree very well with both the statistical theory and the experimental data. The diffusion model can then be used as an alternative to the statistical theory, especially when the statistical theory is not applicable, that is, when the reverberant sound field is not diffuse. Moreover, the diffusion model allows the prediction of the spatial distribution of sound energy within each coupled room, while the statistical theory gives only one sound level for each room.
 [DOI: 10.1121/1.2905242]

PACS number(s): 43.55.Br, 43.55.Ka [NX] Pages: 4261–4271

I. INTRODUCTION

In two coupled rooms, separated by a partition wall (Fig. 1), exchanges of acoustical energy occur. This phenomenon is of great significance for building comfort. The ISO 140-4 standard,[1] which permits the evaluation of the transmission loss of a partition wall through measurements and, the EN 12354-1 standard,[2] which is used for predicting the sound transmission through partition walls, are based on statistical theory. However, this theory becomes invalid when the reverberant sound field departs from the diffuse field assumptions,[3] like in long or flat rooms. Valeau *et al.*[4] showed that the transmission through a partition wall can be evaluated by using a model based on a diffusion equation. The so-called diffusion model can be viewed as an extension of the statistical theory with the main interest in that the predicted reverberant sound field is no longer related to the diffused field assumptions. This model, which was introduced by Ollendorff,[5] has been validated experimentally and numerically in configurations such as rooms with proportionate dimensions,[4,6–8] long rooms,[4,6–9] flat rooms[4,6,8,10] and a system of rooms coupled through an aperture.[11] Nevertheless, in the model of Valeau *et al.*,[4] the transmission through a partition wall between a source room and an adjacent room is treated simply by considering that the partition wall acts like an equivalent sound source (a surface source) in the adjacent room. The energy of the equivalent sound source is then obtained by applying the statistical theory in the source room without coupling with the adjacent room. The model is then not accurate since it does not take the sound transmission on both sides of the partition wall into account. Moreover, if the room with the sound source is not a diffuse one, the statistical theory should not be applied.

In this paper, an extension of the diffusion model is proposed to deal with the complete phenomenon of the reverberant sound field transmission between two rooms through a partition wall in order to predict the acoustic energy density on both sides of the partition wall for both diffuse and nondiffuse configurations (long enclosures, rooms with nonuniform and high absorption). In Sec. II, the statistical theory and the modified diffusion model are presented. The extended diffusion model is then validated by comparison with

[a)]Author to whom correspondence should be addressed: electronic mail: judicael.picaut@lcpc.fr

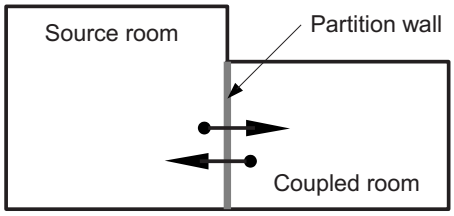


FIG. 1. Sketch of two rooms coupled through a partition wall: source room (V_1, S_1, α_s); adjacent room (V_2, S_2, α_a). The partition wall is defined by its transmission loss R and surface S_{12}.

the statistical theory in Sec. III. A comparison with experimental data is also carried out in Sec. IV. Section V concludes the paper.

II. MODEL PRESENTATION

The problem under consideration is presented in Fig. 1: the first room (the source room) of volume V_1 and surface area S_1 (without the coupling area) contains a sound source of power P. This room is separated from the second room (the adjacent room) of volume V_2 and surface area S_2 (without the coupling area) by a partition wall. The part of the wall, that allows sound transmission is the coupling area: it is defined by its surface S_{12} and its absorption coefficient α_{12}. By considering internal losses in the partition wall,[12] the absorption coefficient α_{12} can be written as a function of the transmission loss R (or the transmission coefficient $\tau = 10^{-R/10}$) and the dissipation coefficient δ, such as $\alpha_{12} = \tau + \delta$. The absorption coefficients in the source room and in the adjacent room are noted down as α_s and α_a, respectively, and are the mean absorption coefficient of each room.

A. Statistical theory

In the statistical theory, two constant energy densities, w_1 in the source room and w_2 in the adjacent room, are considered. The energy balance in the adjacent room can be written as[12]

$$-A_{20}w_2\frac{c}{4} - \alpha_{12}S_{12}w_2\frac{c}{4} + \tau S_{12}w_1\frac{c}{4} = 0, \tag{1}$$

where c is the sound velocity. The first term represents the absorption of sound energy by the walls in the adjacent room, $A_{20} = \alpha_a S_2$ being its equivalent absorption area, including all surfaces except S_{12}. The second term is the absorption of sound energy due to internal loss and transmission to the source room at the coupling area S_{12}. The last term describes the energy transfer from the source room to the adjacent room through the coupling area S_{12}. If there is no dissipation in the partition wall (i.e., $\delta = 0$), Eq. (1) can be simplified by considering that $\alpha_{12} = \tau$. The last equation can be rewritten using the total equivalent absorption area A_{22}, including the coupling area S_{12}, instead of A_{20}:

$$A_{22} = A_{20} + \alpha_{12}S_{12}, \tag{2}$$

leading to the simplified equation

$$-A_{22}w_2\frac{c}{4} + \tau S_{12}w_1\frac{c}{4} = 0, \tag{3}$$

and thus to the following room energy density ratio:

$$\frac{w_2}{w_1} = \frac{\tau S_{12}}{A_{22}}. \tag{4}$$

Considering now the sound levels, L_1 in the source room and L_2 in the adjacent room, we obtain[12]

$$L_1 - L_2 = R - 10\log\frac{S_{12}}{A_{22}}. \tag{5}$$

The ISO 140-4 and EN 12354-1 standards[1,2] use Eq. (5) to relate the sound pressure level (SPL) difference $L_1 - L_2$ (dB) to the partition wall transmission loss R. In these standards, the equivalent absorption area is evaluated using the Sabine absorption coefficient. In this paper, the Eyring absorption coefficient, which is more accurate for high absorptions, is used instead.[8]

B. Diffusion model

Based on the diffusion of particles in a scattering medium,[13] the acoustic energy density distribution in an enclosure of volume V and surface S can be described by the following diffusion equation:[4]

$$\frac{\partial w(\boldsymbol{r},t)}{\partial t} - D\nabla^2 w(\boldsymbol{r},t) = P(\boldsymbol{r},t) \quad \text{in } V, \tag{6}$$

where $w(\boldsymbol{r},t)$ is the acoustic energy density, ∇^2 is the Laplace operator, $P(\boldsymbol{r},t)$ is the source term, and $D = \lambda c/3$, with $\lambda = 4V/S$ the room mean free path, is the so-called diffusion constant. The energy exchanges at the boundaries (wall absorption) can be expressed as a mixed boundary condition in the following form:[4]

$$\boldsymbol{J}\cdot\boldsymbol{n} = -D\frac{\partial w(\boldsymbol{r},t)}{\partial \boldsymbol{n}} = hw(\boldsymbol{r},t) \quad \text{on } S, \tag{7}$$

where h is the local exchange coefficient, $\boldsymbol{n}$ is the outgoing normal vector, and $\partial/\partial\boldsymbol{n}$ is the normal derivative. The exchange coefficient can be expressed using the absorption coefficient α as

$$h = \frac{c\alpha}{4} \tag{8}$$

for low absorption (i.e., using Sabine's formula)[4,9] or

$$h = -\frac{c\ln(1-\alpha)}{4} \tag{9}$$

for high absorption (i.e., using Eyring's formula).[6,8] An improved mixed boundary condition was also recently proposed as[7]

$$h = \frac{c\alpha}{2(2-\alpha)} \tag{10}$$

in order to overcome the singularity problem of Eq. (9) when the absorption coefficient becomes 1. In the numerical simu-

lations presented in this paper, the exchange coefficient based on Eyring's formula is considered.

The SPL can be expressed as[14]

$$\mathrm{SPL}(\boldsymbol{r},t) = 10\log[w(\boldsymbol{r},t)\rho c^2/P_{\mathrm{ref}}^2], \tag{11}$$

where P_{ref} is equal to 2×10^{-5} Pa. The diffusion equation (6) associated with the boundary conditions of Eq. (7) is solved numerically by means of the finite element method.[4]

To extend this model to two enclosures coupled through a partition wall, two diffusion equations must be considered, one for each enclosure:

$$-D_1\nabla^2 w_1(\boldsymbol{r},t) = P(\boldsymbol{r},t) \quad \text{in } V_1, \tag{12}$$

$$-D_2\nabla^2 w_2(\boldsymbol{r},t) = 0 \quad \text{in } V_2, \tag{13}$$

where D_1 and D_2 are the diffusion coefficients of the source and adjacent rooms, respectively. Note that only the stationary problem is considered here, the time-related terms being discarded. It must be emphasized that the proposed model can also be applied to time-dependent problems by keeping the time-dependent terms in the equations. The sound decay due to both sound propagation and transmission could then be evaluated in the same manner as the stationary SPL. However, only stationary problems are considered in this paper as they are the most commonly treated in acoustic transmission problems. The boundary conditions can be expressed as

$$D_1\frac{\partial w_1(\boldsymbol{r},t)}{\partial \boldsymbol{n}_1} + h_1 w_1(\boldsymbol{r},t) = 0 \quad \text{on } S_1, \tag{14}$$

$$D_2\frac{\partial w_2(\boldsymbol{r},t)}{\partial \boldsymbol{n}_2} + h_2 w_2(\boldsymbol{r},t) = 0 \quad \text{on } S_2, \tag{15}$$

where $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are the exterior normal vectors.

At the coupling area, the energy exchanges between the enclosures must be expressed. To model the energy transmission from the source room to the adjacent one, Valeau *et al.*[4] assumed the partition wall as a sound source in the adjacent room. The ingoing energy flux J_{in} can be written as[4]

$$\boldsymbol{J}\cdot\boldsymbol{n} = -D\frac{\partial w}{\partial \boldsymbol{n}} = -J_{\mathrm{in}}, \tag{16}$$

the negative sign accounting for the ingoing nature of the energy in the adjacent room. The value of J_{in} is related to the mean density w_0 in the source room and to the transmission coefficient through[4]

$$J_{\mathrm{in}} = -\frac{\tau c w_0}{4}. \tag{17}$$

In this last model, w_0 was evaluated using the statistical theory. The SPL, which was obtained in the adjacent room by using this diffusion model, matched satisfactorily with the statistical theory. Here, this boundary condition is extended on both sides of the coupling area. If we also consider the energy absorption occurring at the wall surfaces, we now obtain

$$D_1\frac{\partial w_1}{\partial \boldsymbol{n}_1} + h_{12} w_1 = \frac{\tau c}{4} w_2 \quad \text{on } S_{12} \tag{18}$$

on the source room side and

$$D_2\frac{\partial w_2}{\partial \boldsymbol{n}_2} + h_{12} w_2 = \frac{\tau c}{4} w_1 \quad \text{on } S_{12} \tag{19}$$

on the adjacent room side, where h_{12} is the exchange coefficient of the wall partition.

Equation (18) accounts for the energy transfer from the adjacent room to the source room and Eq. (19) accounts for the one from the adjacent room to the source room. Similarly to the statistical theory, Eq. (18), describing the energy transfer from the adjacent room to the source room, could be neglected. However, in this paper, all the energy density exchanges are considered. Then, Eqs. (12) and (13), together with the boundary conditions (14), (15), (18), and (19), set up a system that must be solved by a numerical method. This model can be easily extended to an arbitrary number of coupled rooms, as well as the number of sound sources. In this case, a diffusion equation, the appropriate boundary, and coupling conditions must be considered for each room. The consistency of the proposed model is checked in the following section by comparison with the statistical theory.

C. Coupled-room energy balance

Equation (1) evaluates the coupled-room energy balance by using the statistical theory. For the diffusion model, this energy balance can be obtained by integrating Eq. (13) over the volume V_2:

$$-\int_{V_2} D_2\nabla^2 w_2(\boldsymbol{r})dV_2 = 0. \tag{20}$$

Using Gauss' theorem, this equation can be rewritten as

$$D_2\int_{S_2}\frac{\partial w_2(\boldsymbol{r})}{\partial \boldsymbol{n}_2}dS_2 + D_2\int_{S_{12}}\frac{\partial w_2(\boldsymbol{r})}{\partial \boldsymbol{n}_2}dS_{12} = 0. \tag{21}$$

By using the boundary conditions of Eqs. (15) and (19), it follows that

$$-\int_{S_2} h_2 w_2(\boldsymbol{r})dS_2 - \int_{S_{12}} h_{12} w_2(\boldsymbol{r})dS_{12} + \int_{S_{12}}\frac{\tau c}{4} w_1(\boldsymbol{r})dS_{12} = 0. \tag{22}$$

By considering the exchange coefficient defined in Eq. (8) for rooms with low absorption, the last equation gives

$$-\frac{c\alpha_a}{4}\int_{S_2} w_2(\boldsymbol{r})dS_2 - \frac{c\alpha_{12}}{4}\int_{S_{12}} w_2(\boldsymbol{r})dS_{12} + \int_{S_{12}}\frac{\tau c}{4} w_1(\boldsymbol{r})dS_{12} = 0. \tag{23}$$

Let us consider now the case of a diffuse sound field in the source and adjacent rooms, meaning that $w_1(\boldsymbol{r})$ and $w_2(\boldsymbol{r})$ are constant. Finally, Eq. (23) can be simplified as

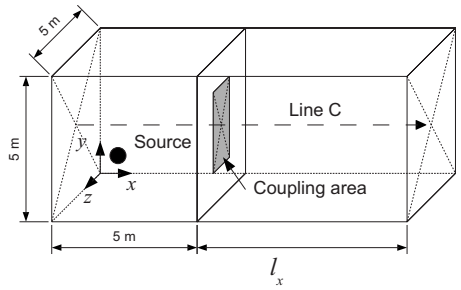


FIG. 2. Sketch of the simulated configuration (dimensions in m).

$$-\frac{c\alpha_a}{4}w_2S_2-\frac{c\alpha_{12}}{4}w_2S_{12}+\frac{\tau c}{4}w_1S_{12}=0. \tag{24}$$

According to Eq. (2), this last equations gives

$$-A_{22}w_2\frac{c}{4}+\tau S_{12}w_1\frac{c}{4}=0, \tag{25}$$

which is identical to Eq. (3). By using the stationary diffusion equation (13), together with the proposed boundary conditions of Eqs. (15), (18), and (19), and by considering the diffuse field assumptions, the diffusion model and the statistical theory are equivalent. Then, if the diffusion sound field assumptions are not considered, the diffusion model can be seen as a direct extension of the classical theory of sound transmission to spatially varying sound fields.

III. NUMERICAL VALIDATIONS

In this section, the diffusion model is compared to the statistical theory. The studied geometry is a $5\times5\times5$ m^3 room containing the sound source (Fig. 2). In the diffusion model, the source is modeled as a 17 cm sphere located at point (2,2,2) m with a sound power level of 100 dB. The source room, with a varying length l_x, is linked to the adjacent room through a coupling area of variable size. In each room, the wall absorption coefficient is uniform. The SPL difference between the source and the adjacent rooms (L_1-L_2 in dB) and the SPL (dB) along line C are used for the comparison between the diffusion model and the statistical theory. For the diffusion model, the SPLs L_1 and L_2 are evaluated by averaging the mean SPL over the whole volume of each room (without the volume source). To perform accurate comparisons with the statistical theory, only simulations of the reverberant field are carried out (the direct sound field is not computed).

Several parameters affecting the SPL difference between the rooms are investigated: the transmission loss R and the surface S_{12} of the coupling area, the absorption coefficients of the source and the adjacent rooms, α_s and α_a, respectively, as well as the length l_x of the adjacent room.

To perform the numerical simulations for the diffusion model, a finite element solver is used.[4] A maximum number of 6000 linear Lagrange elements is used, giving a computational time lower than 30 s on a personal computer. The numerical system of Eqs. (12) and (13), coupled with the boundary conditions (14), (15), (18), and (19), must be solved numerically as two coupled equations. An iterative solving of this set of equations is then required.

A. Effect of the transmission loss

The coupled-room dimensions are $5\times5\times5$ m^3 and the absorption coefficients of both rooms are set to 0.1. The coupling area surface is equal to 25 m^2 and its transmission loss is varied between 10 and 50 dB.

Figure 3 presents the SPL difference and shows good agreement between the diffusion and statistical models, with a maximum discrepancy lower than 0.5 dB. Figure 4 shows the SPL along line C (crossing the partition wall at $x=5$ m): very small variations (about 1 dB) in the SPL can be observed along the line for each room, particularly in the vicinity of the sound source; overall, the reverberant sound field is very closed to a diffuse sound field in each room. Indeed, it is well known that rooms with proportionate dimensions and

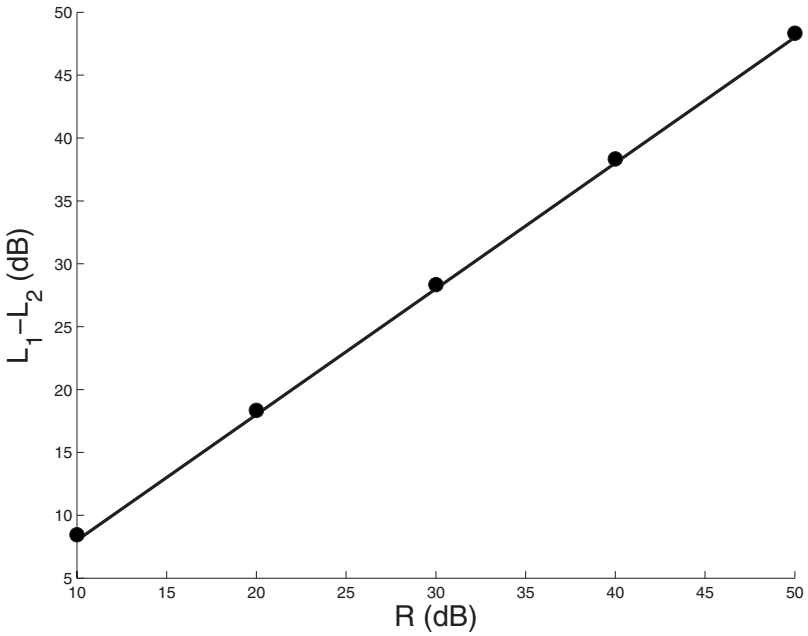


FIG. 3. SPL difference as a function of the transmission loss: statistical theory (—) and diffusion model (●).

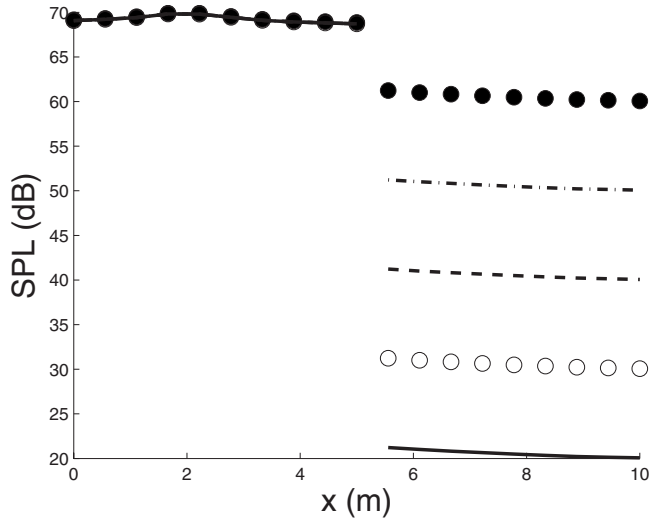


FIG. 4. Distribution of the sound pressure along line C calculated using the diffusion model: $R=10$ dB (●), $R=20$ dB (–·–), $R=30$ dB (– – –), $R=40$ dB (○), and $R=50$ dB (—).

uniform absorption lead to a diffuse sound field. In accordance with the statistical theory, the diffusion model shows that the transmission loss values have no significant effect on the sound level in the source room.

B. Effect of the coupling area surface

The coupled-room dimensions are $5\times5\times5$ m^3 and the absorption coefficients of both rooms are set to 0.1. The coupling area transmission loss is set to 20 dB and its area S_{12} is varied between 1 and 25 m^2. Figure 5 shows again very good agreement between the models, with a maximal discrepancy of about 0.3 dB.

C. Effect of the absorption coefficients of the source and the adjacent rooms

The coupled-room dimensions are $5\times5\times5$ m^3, and the coupling area transmission loss is set to 20 dB and its surface is set to 25 m^2. The absorption coefficient of the source room is equal to 0.1, while the adjacent room absorption is varied between 0.05 and 0.5. Figure 6 shows again very good agreement between the models, with a maximal discrepancy of lower than 0.7 dB. As expected, the more absorbent the adjacent room is, the greater the SPL difference between the rooms is.

The absorption coefficient of the adjacent room is now set to 0.1, while the source room absorption coefficient is varied between 0.05 and 0.5. Results of the source room alone (uncoupled to the adjacent room) are also presented. The agreement between the models is not as good as for the previously tested parameters, with a maximal discrepancy of 1.6 dB (Fig. 7). While the results given by the statistical theory is independent of the absorption coefficient of the source room [see Eq. (5)], those obtained using the diffusion show an increase of the SPL difference of about 1.5 dB from $\alpha_s=0.05$ to $\alpha_s=0.5$.

Simulations carried out with the source room alone (uncoupled to the adjacent room) show similar results as for the coupled system (Fig. 8): as widely discussed in the literature,[3] for high room absorptions, and even for a cubic room, the reverberant sound field is no longer diffuse, leading to an increase of the SPL at the source vicinity.

D. Volume of the adjacent room

The length l_x of the adjacent room is now varied from 2 to 30 m. The coupling area transmission loss is set to 20 dB and its area is equal to 25 m^2. The absorption coefficient of both rooms is equal to 0.1. Again, the agreement between the models is good, with a maximal discrepancy of

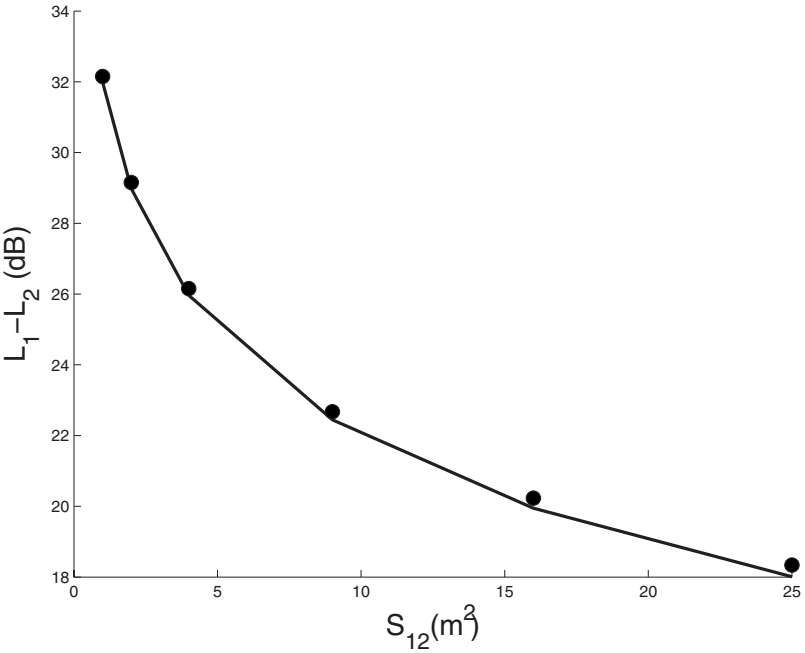


FIG. 5. SPL difference as a function of the surface of the coupling area: statistical theory (—) and diffusion model (●).

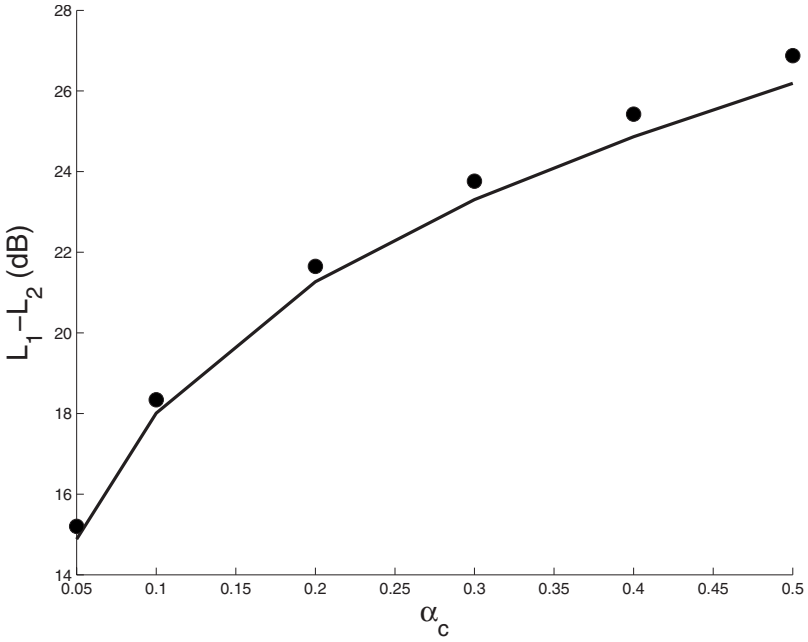


FIG. 6. SPL difference as a function of the absorption coefficient α_a of the adjacent room: statistical theory (—) and diffusion model (●).

lower than 2.3 dB (Fig. 9). However, the difference between the models increases with the adjacent room length. As expected, when the adjacent room length increases, the reverberant sound field departs from a diffuse one (Fig. 10). For example, for l_x=30 m, the adjacent room length is six times greater than the other dimensions: it can then be assimilated to a long room for which it is known that the diffuse field assumptions are not valid.[15] Thus, Fig. 9 shows one of the main interests of the diffusion model since it can account for nondiffuse sound fields in the adjacent room as well as in the source room. In the present configuration, the SPL attenuation varies by about 14 dB along the length of the adjacent room (Fig. 10). For a more absorbent room or a more elongated room, this variation will be greater.

IV. EXPERIMENTAL VALIDATION

A. Experimental setup

Measurements were carried out in the Orbigny teaching building of La Rochelle University. Two empty coupled classrooms with nearly identical geometries were considered (Fig. 11): room C27 ($9.37\times6.67\times3.03$ m^3) and room C28 ($9.37\times6.75\times3.03$ m^3). The partition wall (noted as 3/5 in Fig. 11) between the rooms is made of a 10.5-cm-thick multilayer material (plaster/wood/glass wool/wood/plaster) and a 4-cm-thick wooden door (0.83×2.06 m^2). Along the corridor, walls 2, 6, and 7 are also made with the same multilayer material. Both walls 4 and 8 are entirely made of window glasses. Wall 1 of room C27 is a concrete wall.

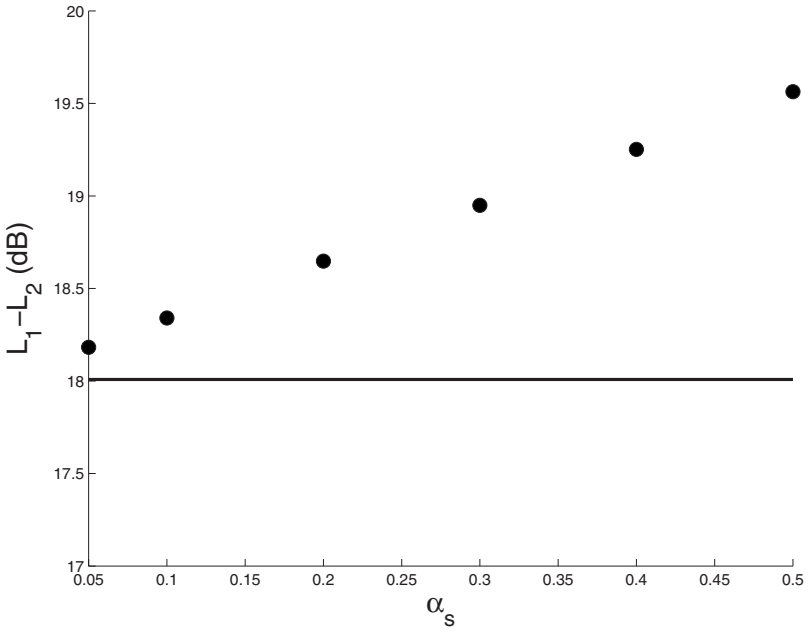


FIG. 7. SPL as a function of the absorption coefficient α_s of the source source: statistical theory (—) and diffusion model (●).

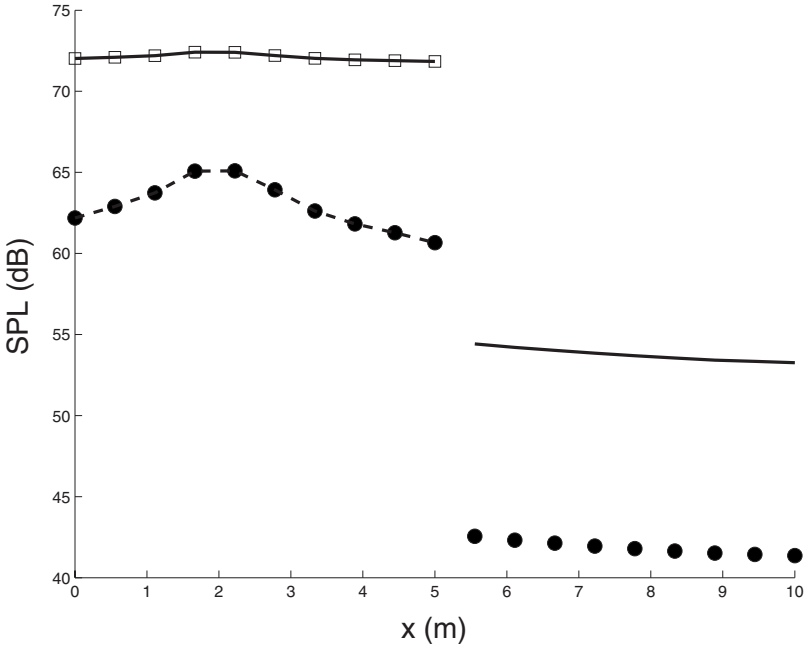


FIG. 8. Distribution of the SPL along line C, which is calculated using the diffusion model: α_s=0.1 (—) and α_s=0.5 (●) for the source room coupled with the adjacent room; α_s=0.1 (□) and α_s=0.5 (– – –) for the source room alone.

Ceilings are made with 2-cm-thick wood fiber plates set up on a 20 cm plenum.

In accordance with the ISO 140-4 standard,[1] four measurement configurations (Fig. 12) were considered: two positions of the sound source in room C27 (1 and 2) and two in room C28 (3 and 4). For each source position, the SPL and the temporal sound decay were evaluated at five positions 1.2 m high within each room, numbered from S1 to S5 in room C27 and R1 to R5 in room C28. For all measurements, a reference receiver was located 2 m from the sound source at a height of 1.2 m.

Reverberation times (RTs) for each source and receiver location were calculated from the impulse response following the ISO 3382 standard.[16] Due to the low signal-to-noise ratio (SNR), the sound decays from −5 to −25 dB were used instead of the conventional ones from −5 to −35 dB. The mean RT of each room was obtained by averaging the RT over the five receiver locations for both source locations in each room. Measurements were carried out with an omnidirectional loudspeaker (type B&K 4296) connected to a power amplifier (type B&K 2716) and by using two $\frac{1}{4}$ in. microphones (type B&K 4135) connected to 2619 preamplifiers, all manufactured by Brüel & Kjær. Both microphones were connected to a NEXUS conditioning amplifier (B&K 2690). The NEXUS and the source power amplifier were connected to a personal computer by using a high-quality sound card. Impulse response measurements were realized with the DSSF3 acoustic analysis software by using the time-stretched pulse (TSP) method. A TSP signal is sent to the sound source while data acquisition is carried out on both microphones with a sample frequency of 48 kHz and a mea-

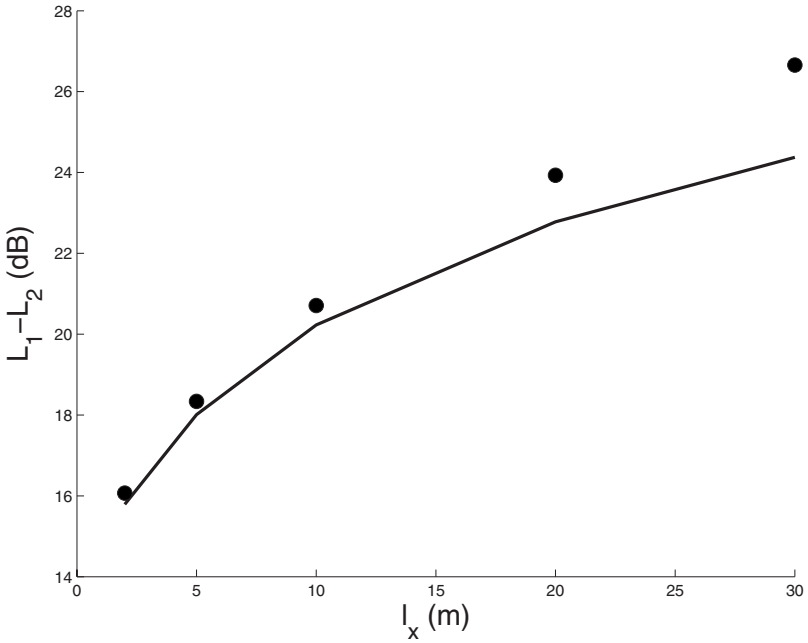


FIG. 9. SPL difference as a function of the length l_x of the adjacent room: statistical theory (—) and diffusion model (●).

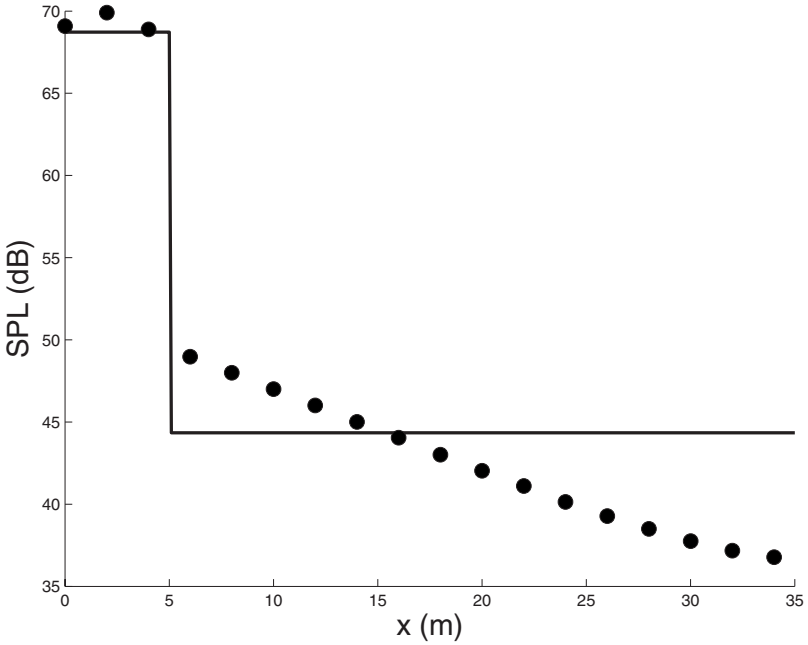


FIG. 10. Distribution of the sound pressure along line C with l_x=30 m: statistical theory (—) and diffusion model (●).

suring time of 2.731 s. In order to avoid the small fluctuations due to the background noise and to increase the SNR, ten impulse responses were averaged at each receiver location. The sound source level is automatically adjusted by the DSSF3 software in order to achieve an optimal SNR at each receiver location. The SPL at each receiver location is then normalized by considering the SPL at the reference microphone.

The RTs were evaluated in each room by averaging the values obtained at the two sound source locations and over the five microphone positions (Table I). Likewise, the SPL in each room, L_{C27} and L_{C28}, were calculated by averaging over the five microphone positions for each source location. To obtain the SPL difference between the rooms by third octave bands, $L_{C27}-L_{C28}$ is evaluated from the average of the SPL differences of configurations 1 and 2 (source in C27) and $L_{C28}-L_{C27}$ is evaluated from the average of configurations 3 and 4 (source in C28). This entire procedure is in accordance with the ISO 140-4 standard. The equivalent absorption areas A of each room were evaluated from the RT by using the statistical theory. The transmission loss R of the partition wall was obtained (Table I) again in accordance with the ISO 140-4 standard.

B. Numerical parameters

In order to achieve accurate simulations with the diffusion model, the mean absorption coefficients of both rooms

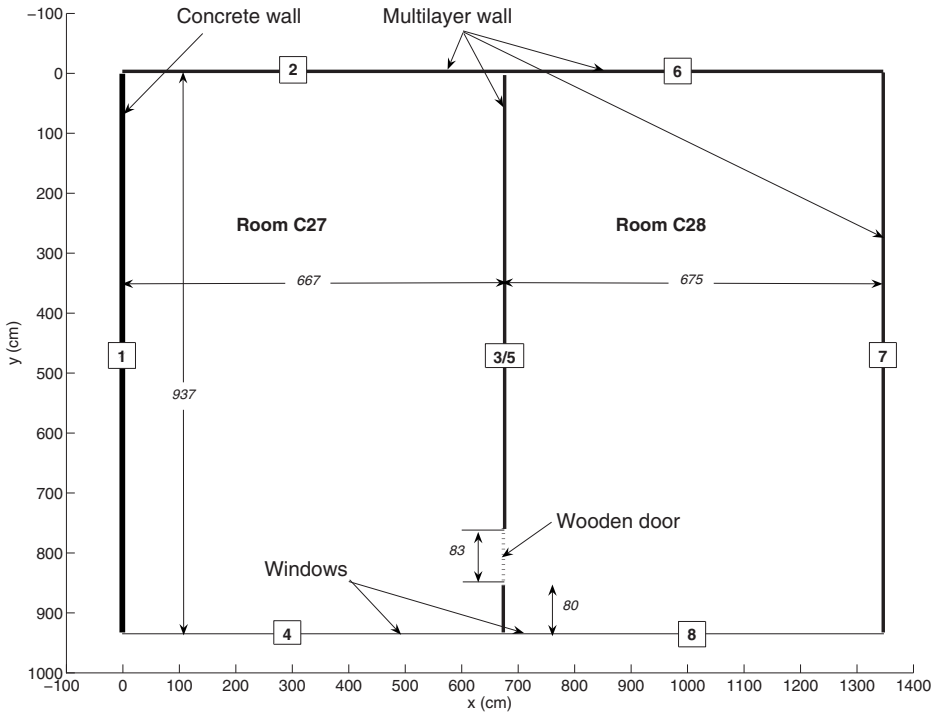


FIG. 11. Room geometry.

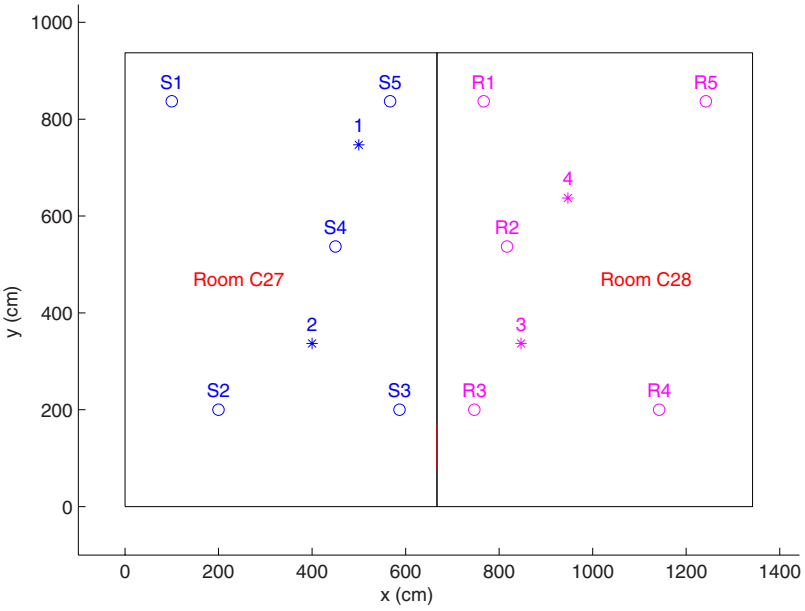


FIG. 12. (Color online) Measurement configurations: four locations of the sound source (∗, noted 1 to 4); eight locations of the microphones (○, noted S1 to S4 and R1 to R4 in the source and in the adjacent rooms, respectively).

C27 and C28 have to be determined with accuracy. Since the depth of the plenum behind ceilings has a great impact over the *in situ* absorption coefficient of ceilings, the use of absorption coefficient databases may be a source of errors. As RTs were available, the mean absorption coefficients were evaluated using the Eyring formula (Table I). It may be noted that the estimated absorption coefficients probably include the modal effects of the coupled-room system for frequencies below Schroeder's frequency of the uncoupled rooms (about 130 Hz).

For the diffusion model, the simulated geometry is discretized into 2000 elements and the stationary responses are obtained in a few seconds. All the four configurations were simulated and the SPLs were evaluated at the same locations as those in the experiments. Similarly with the experiments, two SPL differences are calculated: $L_{C27}-L_{C28}$ and $L_{C28}-L_{C27}$.

C. Results and discussion

As shown in Fig. 13, the results obtained using the diffusion model is in very good agreement with the experimental data. The maximal discrepancy is about 1.9 dB when the sound source is located in room C27 and 0.7 dB when the sound source is located in room C28, while the mean discrepancies are 0.7 and 0.4 dB, respectively. Although it is not presented in Fig. 13, the statistical theory results match perfectly the experimental results, too. This result was expected since the boundary conditions (absorption coefficients and transmission loss) are derived from the experimental data by using the statistical theory.

One can remark very good agreement of the diffusion model with experiments even at the lowest third octave bands which are situated below the Schroeder frequency, which is probably due to the method that is used for estimating the absorption coefficients. At these frequencies, modal effects are expected to occur and predictions of the diffusion model, which is a mid- to high-frequency model, are expected to be no longer accurate. The good agreement with the experimental results comes from the fact that the statistical theory, which was used to obtain the acoustical characteristics of the coupled-room system, also ignores the modal effect acoustic characteristics of the system. At the lowest frequencies, different receiver locations would give different

TABLE I. RT, equivalent absorption area A, Eyring absorption coefficient α, SPL difference, and transmission loss R of rooms C27 and C28 by the third octave band.

Frequency (Hz)	100	125	160	200	250	315	400	500	630	800	1000	1250	1600	2000	2500	3150	4000	5000	6300	8000	10 000
RT_{C27} (s)	2.31	0.83	0.68	0.79	0.74	0.81	0.79	0.80	0.91	0.90	1.04	1.07	1.11	1.06	0.96	0.98	0.95	0.84	0.77	0.69	0.61
A_{C27} (m^2)	13.5	37.5	45.8	39.4	42.1	38.5	39.5	39.0	34.3	34.7	30.0	29.2	28.1	29.4	32.5	31.8	32.8	37.1	40.5	45.2	51.1
α_{C27}	0.06	0.15	0.18	0.16	0.17	0.16	0.16	0.16	0.14	0.14	0.13	0.12	0.12	0.12	0.13	0.13	0.14	0.15	0.17	0.18	0.20
$L_{C27}-L_{C28}$ (dB)	9.20	16.6	20.5	27.8	29.8	30.3	30.2	33.8	36.6	36.8	35.8	35.4	35.9	34.8	31.3	32.3	33.7	32.8	34.1	35.5	37.4
R_{C27} (dB)	12.7	15.4	19.5	26.5	28.8	29.0	28.9	32.1	34.9	34.8	33.7	33.7	34.7	33.4	30.0	31.4	32.8	31.6	32.5	33.5	34.9
RT_{C28} (s)	2.47	0.84	0.88	0.83	0.88	0.82	0.81	0.76	0.74	0.68	0.68	0.74	0.83	0.80	0.81	0.90	0.89	0.84	0.76	0.69	0.61
A_{C28} (m^2)	2.63	37.1	35.4	37.6	35.4	38.0	38.5	41.0	42.2	45.9	45.9	42.2	37.6	39.0	38.5	34.7	35.0	37.1	41.0	45.2	51.1
α_{C28}	0.06	0.16	0.19	0.16	0.17	0.16	0.16	0.16	0.14	0.14	0.13	0.12	0.12	0.12	0.14	0.13	0.14	0.15	0.17	0.18	0.20
$L_{C28}-L_{C27}$ (dB)	8.80	18.4	21.8	25.9	28.7	29.3	28.9	33.4	35.9	36.4	34.9	34.4	34.7	34.2	31.8	32.2	33.2	32.2	34.5	34.3	37.6
R_{C28} (dB)	12.0	17.2	19.7	24.5	27.0	28.0	27.5	32.0	35.0	35.6	34.6	34.3	34.8	34.0	31.2	31.7	32.5	31.0	33.0	32.3	35.1

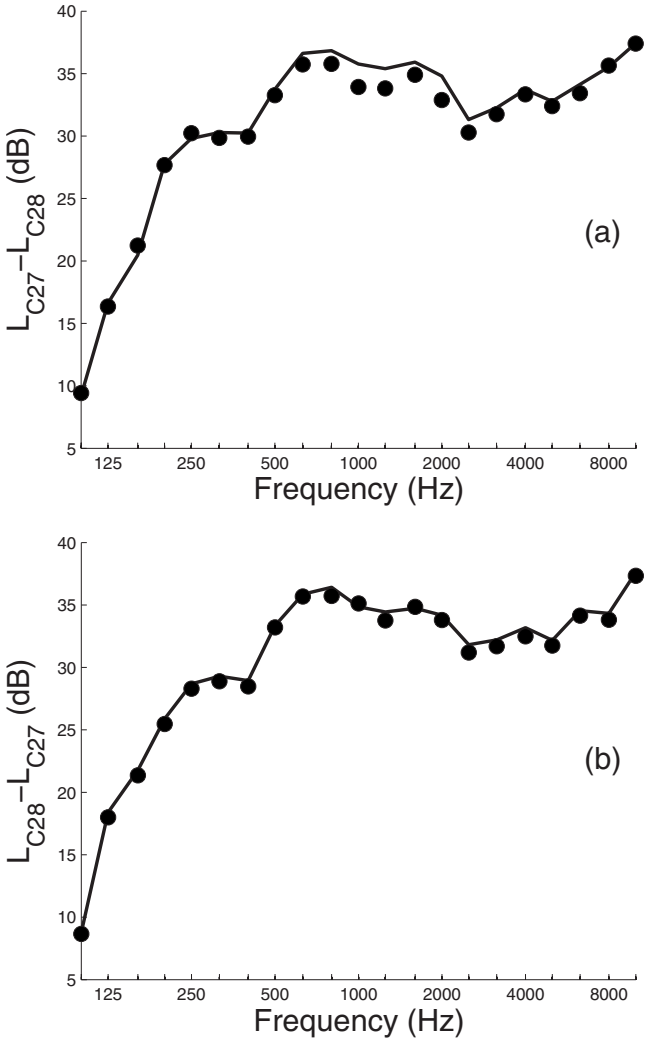


FIG. 13. SPL differences by third octave bands when the sound source is located in rooms C27 (a) and C28 (b): experimental data (—) and diffusion model (●).

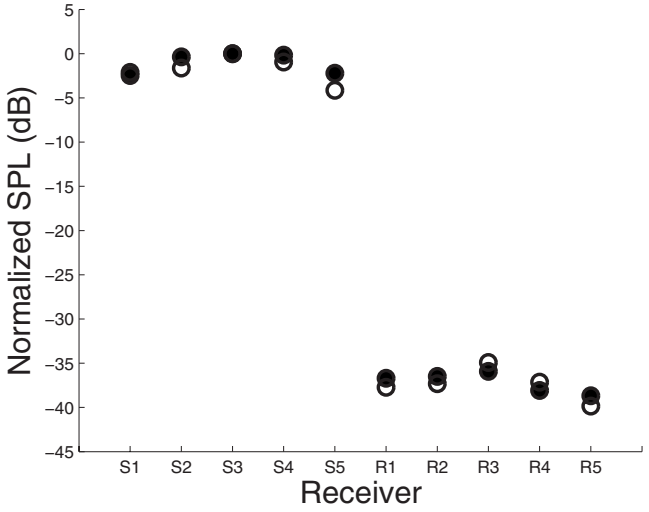


FIG. 14. Normalized SPL for the 800 Hz third octave band at receivers S1 to S5 in room C27 and receivers R1 to R5 in room C28. The sound source is located at position 2 in room C27 (see Fig. 12). Experimental data (○) and diffusion model (●).

acoustic characteristics. The diffusion model does not handle the modal behavior of the rooms at the lowest frequencies.

Figure 14 shows the SPL for the 800 Hz third octave band at receivers S1 to S5 in room C27 and receivers R1 to R5 in room C28 when the sound source is located at location 2 in room C27. Since the power of the experimental sound source is unknown, the SPL has been normalized with respect to receiver S3 to compare the experimental data to the diffusion model results. The agreement of the diffusion model with the experimental data is very good. The maximal discrepancy is less than 2 dB, occurring at receiver S5. Moreover, in the adjacent room, the spatial energy distribution obtained using the diffusion model matches closely the one obtained experimentally, with a maximum discrepancy of 1 dB. One can also observe that the reverberated sound field is not perfectly diffuse in both rooms, with variations in the SPL of 6 dB. At receiver R3, one can remark that the predicted SPL is slightly lower than the measured one: the door situated in the vicinity of the receiver is not as soundproofing as the wall.

The observed variations of SPL could not be predicted using the statistical theory, which considers only a diffuse sound field. So, even in configurations where a diffuse sound field is expected, the diffusion model can lead to more accurate predictions than those of the statistical theory.

V. CONCLUSION

A modification of the diffusion model for room acoustics has been proposed to account for sound transmission between two rooms coupled through a partition wall. In this model, the reverberant sound fields are evaluated with two diffusion equations, one for each room. A coupling condition at the partition wall is introduced and the considered system is then numerically solved. This modified diffusion model has been validated with numerical simulations by comparison with the statistical theory. The different parameters governing the sound transmission, the transmission loss and the surface of the coupling area, the absorption coefficients of the source and the adjacent rooms, as well as the length of the adjacent room, have been varied. Very good agreement was found with the statistical theory: discrepancies arise when one of the reverberant sound fields departs from a diffuse sound field. Then, a comparison with experimental data in a configuration composed of two coupled classrooms has been carried out. Again, the diffusion model results agree very well with the experimental data, with a maximal discrepancy of lower than 1.9 dB. The diffusion model can then be used as an alternative method to the statistical theory as a prediction tool in building acoustics. While the statistical theory is limited to configurations where the diffuse field assumptions hold, the diffusion model can predict the transmission in configurations where the reverberant sound field is not diffuse at the expense of a longer computation time. Moreover, this approach can be generalized to an arbitrary number n of coupled rooms by considering n coupled equations in the diffusion model.

ACKNOWLEDGMENTS

The authors wish to thank the Agence de l'Environnement et de la Maîtrise de l'Énergie (ADEME) for providing financial support of this work.

[1]"Acoustics—measurement of sound insulation in buildings and of building elements—Part 4: Field measurements of airborne sound insulation between rooms," International Organization for Standardization Document No. ISO 140-4, 1998.
[2]"Building acoustics—Estimation of acoustic performance of buildings from the performance of elements—Part 1: Airborne sound insulation between rooms," European Committee for Standardization Document No. EN-12354-1, 2000.
[3]M. Hodgson, "When is diffuse-field theory applicable?," Appl. Acoust. **49**, 197–207 (1996).
[4]V. Valeau, J. Picaut, and M. Hodgson, "On the use of a diffusion equation for room-acoustic prediction," J. Acoust. Soc. Am. **119**, 1504–1513 (2006).
[5]F. Ollendorff, "Statistical room-acoustics as a problem of diffusion—A proposal," Acustica **21**, 236–245 (1969).
[6]Y. Jing and N. Xiang, "A modified diffusion equation for room-acoustic predication," J. Acoust. Soc. Am. **121**, 3284–3287 (2007).
[7]Y. Jing and N. Xiang, "On boundary conditions for the diffusion equation in room-acoustic prediction: Theory, simulations, and experiments," J. Acoust. Soc. Am. **123**, 145–153 (2008).
[8]A. Billon, J. Picaut, and A. Sakout, "Prediction of the reverberation time in high absorbent room using a modified-diffusion model," Appl. Acoust. **69**, 68–74 (2008).
[9]J. Picaut, L. Simon, and J. D. Polack, "Sound field in long rooms with diffusely reflecting boundaries," Appl. Acoust. **56**, 217–240 (1999).
[10]V. Valeau, M. Hodgson, and J. Picaut, "A diffusion-based analogy for the prediction of sound fields in fitted rooms," Acust. Acta Acust. **93**, 94–105 (2007).
[11]A. Billon, V. Valeau, A. Sakout, and J. Picaut, "On the use of a diffusion model for acoustically coupled rooms," J. Acoust. Soc. Am. **120**, 2043–2054 (2006).
[12]L. Cremer and H. A. Müller, *Principle and Applications of Room Acoustics* (Applied Science, London, 1982), Vol. **1**.
[13]P. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953).
[14]A. Pierce, *Acoustics: An Introduction to its Physical Principles and Applications* (Acoustical Society of America, New York, 1981).
[15]J. Kang, *Acoustics of Long Spaces* (Thomas Telford, London, 2002).
[16]"Acoustics—Measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization Document No. ISO 3382, 1997.

The effect of visual and auditory cues on seat preference in an opera theater

Jin Yong Jeon[a] and Yong Hee Kim
School of Architectural Engineering, Hanyang University, Seoul 133-791, Korea

Densil Cabrera and John Bassett
Faculty of Architecture, Design and Planning, University of Sydney, New South Wales 2006, Australia

(Received 18 November 2006; revised 9 March 2008; accepted 28 March 2008)

Opera performance conveys both visual and auditory information to an audience, and so opera theaters should be evaluated in both domains. This study investigates the effect of static visual and auditory cues on seat preference in an opera theater. Acoustical parameters were measured and visibility was analyzed for nine seats. Subjective assessments for visual-only, auditory-only, and auditory-visual preferences for these seat positions were made through paired-comparison tests. In the cases of visual-only and auditory-only subjective evaluations, preference judgment tests on a rating scale were also employed. Visual stimuli were based on still photographs, and auditory stimuli were based on binaural impulse responses convolved with a solo tenor recording. For the visual-only experiment, preference is predicted well by measurements taken related to the angle of seats from the theater midline at the center of the stage, the size of the photographed stage view, the visual obstruction, and the distance from the stage. Sound pressure level was the dominant predictor of auditory preference in the auditory-only experiment. In the cross-modal experiments, both auditory and visual preferences were shown to contribute to overall impression, but auditory cues were more influential than the static visual cues. The results show that both a positive visual-only or a positive auditory-only evaluations positively contribute to the assessments of seat quality.
 [DOI: 10.1121/1.2912435]

PACS number(s): 43.55.Fw, 43.55.Gx, 43.55.Hy, 43.55.Mc [NX] Pages: 4272–4282

I. INTRODUCTION

Multisensory processes are now well understood as being deeply involved in the perception of everyday events. Auditory and visual cues generated from single events are simultaneously perceived through different sensory systems, and these signals lead us to acoustically and visually recognize the event.[1] However, multisensory integration is not a simple combination of unimodal perceptions because different sensory cues are not independently perceived and combinations of multisensory stimuli can yield quite different results to unimodal perception due to intersensory interaction.[2,3] Multimodal perception can either sensitize or desensitize a person to particular characteristics of a stimulus.[3] Generally, the spatial and temporal coincidences of the visual and auditory cues in stimuli maximize the effects of multisensory interaction.[4]

A number of studies on auditory-visual cross modality in diverse multisensory research fields confirm the interaction between auditory and visual cues and often show dominance of one sense over the other.[5–9] Video quality has been found to be more influential on the overall perceived audio-visual quality[5] when viewing television commercials. Similarly, in a more static video speech clip, the effect of video quality on perceived audio quality is slightly greater than the effect of audio quality on perceived video quality. The potential for the visual modality to influence sound perception is exemplified by both the McGurk effect[6] and ventriloquist effect.[7] In the McGurk effect, a video of a face speaking one phoneme but with the sound of another phoneme yields the perception of a third intermediate phoneme. In the ventriloquist effect, a visual stimulus dominates in the multimodal localization of a sound source. Generally, the dominance of vision over audition can be found in spatial tasks with spatial attention.[4] Nevertheless, sound can also influence visual perception, for example, it has been shown to alter visual resolution in the temporal domain,[8] and the perceived intensity of a visual stimulus is enhanced by the presence of sound.[9] Hence, dominance of audition can be found in temporal tasks with temporal attention.[4]

Combined auditory and visual cues for a single event in a musical performance space are not independently perceived.[2] Some evidence for this comes from a survey of prominent acousticians, almost 70% of whom indicated (using the expression "definitely") that acoustical quality in concert halls is affected by vision.[10,11] However, very few studies are found in the auditorium acoustics field regarding cross-modal interactions. Hyde[10] has conjectured that visual elements within a concert hall affect acoustical intimacy and the perception of the loudness of sound, which are related to the distance between the stage and the position within the audience area. Nathanail *et al.*[12] reported that auditory distance was perceived to be less than the visually perceived distance when the concert hall stage was close to the listener, but greater when it was distant. Larsson *et al.*[13] carried out

[a]Author to whom correspondence should be addressed. Tel.: 82 2 2220 1795. FAX: 82 2 2220 4794. Electronic mail: jyjeon@hanyang.ac.kr

an auditory experiment that analyzed perceptual room parameters, when music sources were presented with and without images. They found that the "distance to sound source" is perceived shorter with visual information concerning the room than without. Barron[14] pointed out that in making loudness judgments, listeners in auditoria compensate for the actual source-receiver distance and make judgments relative to expectations: the same sound level is judged louder in more remote seats. Cabrera *et al.*[15] examined auditory and visual space perceptions, separately, by using recordings and images from concert halls, finding that while similar terms, such as "intimacy" or "envelopment" can be used in these two perceptual modes, the results for these two modes are not necessarily related. These studies in multimodal perception of auditoria are mainly concerned with concert halls, but the present study is concerned with an opera hall.

A theatrical performance in an opera hall involves numerous and complex visual and auditory stimuli: musical instruments are played in the orchestra pit; songs are sung by the soloists and chorus; there are changing stage sets; and variable lighting conveys a desired mood. The need for the singers and the orchestra to see the conductor makes a proscenium stage essential, and the orchestra pit distances the stage from the stall seating area. The situation means that of all auditorium forms, the opera house is the most constrained in terms of design.[16] Because of this, in opera halls, visual information is as important as acoustical conditions in performances. Audiences prefer seats that are close to the stage, with unobstructed stage views, because the stage performance rather than the orchestra is of primary interest.[16,17] Opera theaters may have boxes in which the quality of sound is affected by the enclosures, and the different plan forms of opera houses are more due to sightlines than acoustics.[16,18] Therefore, the aim of the present study is to determine contributive elements to overall seat preference in terms of auditory-visual perception in the context of an opera theater because of the potential for such information to contribute to theater design. This study also investigates how auditory-visual interactions apply to opera performances. Does a seat close to the stage with an unobstructed view of it enhance the auditory sensation of the viewer? Would a viewer prefer a seat with good visual and poor acoustical conditions or might they prefer a seat with poor visual but good acoustical conditions? These questions were the impetus for the present study. A range of acoustical and visual conditions in an opera house was collected through field measurements. The objective characteristics of a given seat position were evaluated from a measured impulse response and a photographed image of the stage view. Three subjective investigations were performed by using paired-comparison and preference judgment on a rating scale. The tests examine the subjective impression of visual-only stimuli (stage view images), of auditory-only stimuli (auralized music excerpts), and of combined auditory-visual stimuli. The hypothesis of this series of experiments is that the factors that influence visual-only perception (related to the view of the stage) and factors that influence auditory-only perception (related to the sound from the performer) combine to influence auditory-visual perception.

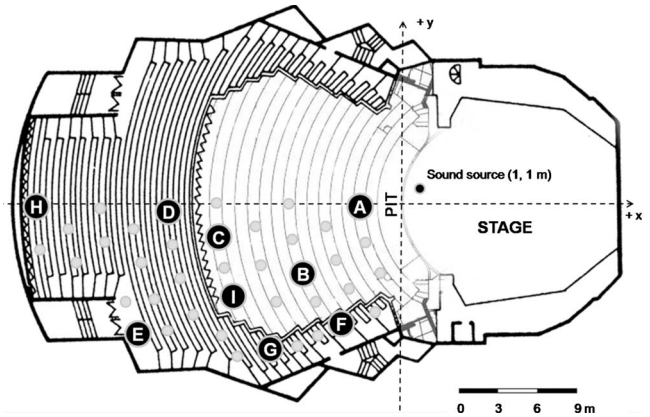


FIG. 1. The source position on the stage and the selected nine receiver positions (A–I) as well as all other measured seats (gray circles).

II. ACOUSTICAL MEASUREMENTS

A. Opera theater

Acoustical measurements were conducted in an opera theater that is a primary venue for a national opera company. This opera theater has a volume of about 15 000 m^3, with a seating capacity of 1547 seats (Fig. 1). The proscenium arch is 11.5 m wide and 7.0 m tall. The pit has space for 75 orchestra members. The theater is regularly used for opera, ballet, and contemporary dance performances.

B. Source and receiver positions

Acoustical measurements of the opera theater were made with a sound source on the stage and receivers at various positions in the audience area. The sound source, a B&K type 4296 Omni-Power loudspeaker, was on the stage and positioned 1 m from both the front of the stage and the centerline of the stage, at a height of 1.5 m.[19,20] The dummy head, a B&K type 4128C head and torso simulator, for binaural recording of swept-sinusoid signals, was pointed toward the sound source on the stage. The acoustical and visual measurements were carried out at 40 seats, as shown in Fig. 1. Following these measurements, nine seats (A–I) were chosen in consideration of the distance from the stage, the size of the visual image, and the diverse distribution of the acoustical characteristics. The acoustical parameters measured at the dummy positions were reverberation time (T30), sound pressure level (SPL), clarity index and definition (C80 and D50), early decay time (EDT), interaural cross-correlation coefficient (IACC), and lateral energy fraction (LF). In calculating the IACC value, the early component of a binaural impulse response was considered (being 0–80 ms from the direct sound) and was averaged at the octave band frequencies of 500, 1000, and 2000 Hz (referred to as 1-$IACC_{E3}$ or binaural quality index, BQI).[21]

C. Measurement results

Table I shows the measured acoustical parameters at each of the nine selected seats. Reverberation time (T30) ranged from 1.07 to 1.23 s in the unoccupied hall. (The mean T30 in the unoccupied condition was 1.17 s.) SPL (i.e., the unweighted L_{eq} of our anechoic singing recording con-

TABLE I. Acoustical parameters of the opera theater. Δt_1: initial-time delay gap between direct sound and the first reflection; SPL: A-weighted equivalent sound pressure level of the stimulus representing each seat; EDT: measured time in decay curve between 0 and −10 dB extrapolated to a range of 60 dB (averaged from 125 to 4000 Hz); T30: measured time in decay curve between −5 and −35 dB extrapolated to a range of 60 dB (averaged from 125 to 4000 Hz); C80: logarithmic energy ratio between 0 to 80 ms and 80 to infinite (averaged from 125 to 4000 Hz); D50: energy ratio between 0 and 50 ms and 0 and infinite (averaged from 125 to 4000 Hz); LF_{E4}: energy ratio between 0 and 80 ms from omnidirectional reception and 5 to 80 ms for figure-of-8 reception directed to the sides of the auditorium (averaged from 125 to 1000 Hz); BQI $=1\text{-}IACC_{E3}$: interaural cross correlation with integration interval of 0–80 ms (averaged from 500 to 2000 Hz).

Seat No.	Δt_1 (ms)	SPL [dB(A)]	EDT (s)	T30 (s)	C80 (dB)	D50 (%)	LF_{E4}	$1\text{-}IACC_{E3}$
A	24	68.5	0.94	1.02	6.1	64	0.15	0.39
B	16	64.2	1.13	1.06	3.0	45	0.17	0.52
C	19	65.8	0.75	1.01	5.4	52	0.29	0.52
D	16	62.7	1.38	1.15	−0.3	34	0.27	0.50
E	11	66.3	1.22	1.18	3.2	60	0.13	0.71
F	18	65.1	1.22	1.13	5.0	67	0.18	0.48
G	12	64.6	1.20	1.08	2.7	50	0.15	0.70
H	13	65.8	1.26	1.14	2.2	44	0.24	0.50
I	15	67.0	0.84	1.02	5.2	58	0.19	0.48

volved with the impulse response for a seat) varied from 62.7 to 68.5 dB. C80 values (i.e., the ratio of impulse response energies before and after 80 ms from the direct sound, expressed in decibels) ranged from −0.3 to 6.1 dB. The $1\text{-}IACC_{E3}$ value was around 0.7 at positions G and E, which are close to lateral walls and was around 0.5 at the other seats, except for seat A, which was located near the front center of the seating area. These objective parameters were later used for correlation with the results of a subjective evaluation.

III. LABORATORY TEST PROCEDURE

Three experimental conditions were designed for investigating seat preference in terms of visual cues only, auditory cues only, and auditory-visual combined cues.

In the vision-only experiment, visual scene and room geometric parameters were derived from the photographed stage view images and the location of each seat in the hall. Values for overall visual impression and for specific visual attributes were obtained in a subjective experiment. Then relationships between objective and subjective values were examined.

In the sound-only experiment, acoustical parameters were derived from the measured impulse responses at each seat. Values for overall impression of sound for specific questions on auditory impression were obtained in a listening test by using auralized music excerpts. Then, relationships between objective and subjective values were examined.

In the auditory-visual combined experiment, seat preference as an overall impression was obtained in two subjective tests. In the first test, the images and auralizations were those for each of the nine seats. In the second test, three images and three auralizations were selected, based on the auditory and visual preference results of previous experiments, and all combinations of the cross-matched auditory-visual stimuli were examined so that the interaction between auditory and visual cues could be better understood.

A. Subjects and stimuli

A total of 50 subjects (35 males and 15 female) with self-reported normal hearing and vision, and at ages from 20 to 46 years, participated in the experiments. All subjects participated in all experiments in order to compare between-condition results.

Visual stimuli. A digital still photograph of the stage view was taken at the each of the nine selected seats. The focal length of the camera was constant. The photographs were then combined with a two-dimensional view of a three-dimensional computer model of a stage scene, taking into account the angle and distance of each measurement position, with the area outside the stage scene dark, as would be the case during opera performances. The processed images are shown in Fig. 2.

Auditory stimuli. Auralization was conducted with a music source recorded in an anechoic chamber convolved with binaural impulse responses for each audience position. An 8 s anechoic recording of an operatic tenor singing solo in Italian was used as a music excerpt. This recording was chosen because the aim of this study was to examine opera performance. Constant gain was used in the convolution, so that the SPL would vary between the stimuli, as would be the case in the real auditorium.

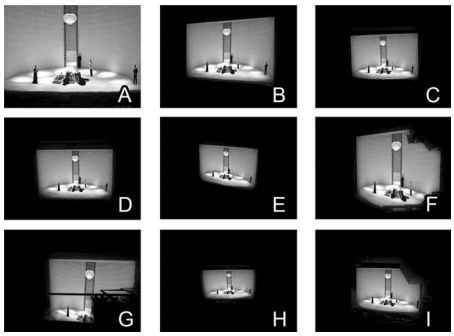


FIG. 2. Images of the stage view used to represent each seat.

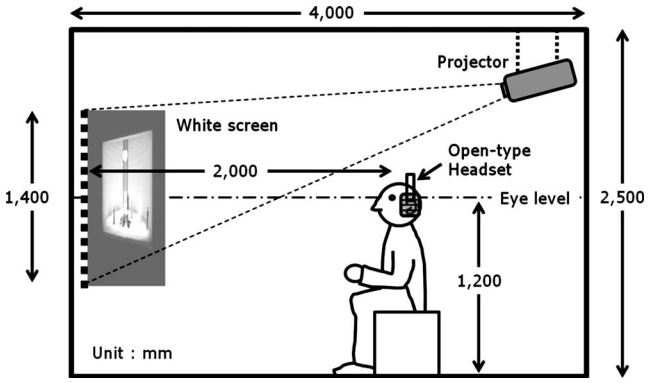


FIG. 3. Subjective test facility setup (unit: mm).

Auditory-visual combined stimuli. Two types of stimuli were prepared for parallel and cross matching of sound and vision. For parallel sound-vision matching, the respective auditory and visual stimuli from each of the nine positions were combined. Cross-matched stimuli consisted of all nine combinations of three visual stimuli (D, E, and H) and three auditory stimuli (A, D, and H), selected based on the results of the respective single-modal preference experiments.

B. Test facilities

Figure 3 shows the experiment setup in a dark room with low background noise (L_{eq} less than 20 dB A). The room dimensions were 4.5×4.0 m^2 with a ceiling height of 2.5 m. All lighting, except for the projector was turned off. The reverberation time of the test room was less than 0.5 s for octave bands centered on 125 Hz–4 kHz. The projector was located at the back of the room and above the participant in order to reproduce a real-scale stage view. Visual stimuli were projected onto a white screen (1.9 m wide and 1.4 m high) with a 1:1 angular scale. Position C had the same altitude as the center of the proscenium arch; this position was taken as a reference image for adjusting the screen height. Walls around the screen were covered with black curtains to prevent disturbance of the visual scene.

The auditory stimuli were presented to subjects through an open-type headset (STAX SR-303+SRM-313). Reproducibility of the headset sound was checked by comparison of virtual impulse responses with real impulse responses that were used for making the auditory stimuli. Virtual impulse responses of the headset were recorded through a dummy head (B&K type 4100) by using convolved swept-sine signals from the real impulse response. The reproduced nine sounds have no significant difference in terms of acoustical parameters to the respective original measurements; the maximum differences were 1 ms for Δt_1, 0.02 s for EDT, 0.03 s for T30, 0.2 dB for C80, 1% for D50, and 0.03 for BQI. These differences are less than the known perceptual thresholds for each parameter.

C. Subjective evaluation methods

Two test methods were employed to quantify subjective impression for auditory and visual qualities: paired comparisons and preference judgment on a rating scale.

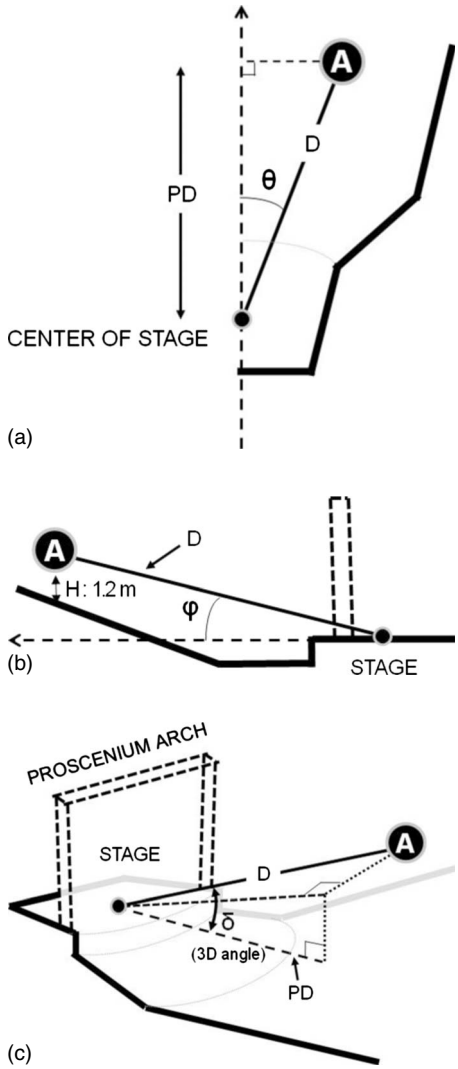


FIG. 4. Definition of room dimensional parameters. Spatial relationship between stage center and a seat position can be simply measured by using distance and angular factors. (a) Floor plan; (b) section; (c) perspective.

Paired-comparison method. In each test condition of visual-only, auditory-only, and auditory-visual, the paired combination of nine stimuli leads to a total of 36 pairs without reversal. Within a pair, a sound or image was successively presented, and the subject's task was to judge which of the pair was preferred. A pair consisted of "sound A–sound B" in the auditory preference test, "image A–image B" in the visual preference test, and "sound A with image A–sound B with image B" in the auditory-visual preference. The significance of intrasubject difference was verified through a consistency test, and the significance of intersubject difference was examined through an agreement test.[22,23] The consistency test detected subjects who answered inconsistently, by calculating the number of circular triads. The agreement test indicated the representativeness of the intersubject averaged results. By using subjects who passed the consistency test, a scale value of preference was calculated by applying Thurstone's law of comparative judgment (case V)[24] which linearizes the probability of stimulus selection by using an inverse function of the normal distribution. A scale value of zero

TABLE II. Room dimensional parameters of the opera theater. D: distance from the stage center to each seat; θ: horizontal angle of each seat to longitudinal center plane. φ: vertical angle of each seat to ground plane; δ: three-dimensional angle of each seat to ground and longitudinal center plane; PD: projected distance on the longitudinal center plane; SR: relative ratio between projected stage area and all visible area in a viewed image at each seat; LC: luminosity contrast between stage area and outer stage area in a viewed image at each seat; VP: product of PD and SR for expressing "visual potential" of each seat.

Seat No.	D (m)	θ (rad)	φ (rad)	δ (rad)	PD (m)	SR (%)	LC (%)	VP
A	7.9	0.19	0.11	0.22	7.7	83.4	1.18	6.42
B	17.4	0.36	0.20	0.41	16.0	37.5	2.63	6.00
C	24.3	0.19	0.21	0.29	23.3	20.7	4.68	4.82
D	24.6	0.05	0.35	0.35	21.1	26.0	3.74	5.49
E	27.9	0.39	0.37	0.54	24.0	15.7	6.26	3.77
F	15.1	0.82	0.31	0.87	9.8	38.2	2.54	3.74
G	15.0	0.88	0.49	1.00	8.5	29.4	3.29	2.50
H	34.0	0.03	0.37	0.37	31.6	10.4	9.34	3.29
I	24.4	0.37	0.21	0.43	22.2	17.8	5.22	3.95

means that the number of preferred and not-preferred responses was equal ("neutral").

Preference judgment method on a five-point rating scale. The paired-comparison method is designed to provide a simple preference score based on the probability of stimulus selection. However, the paired-comparison method usually employs only one question due to the number of stimuli pairs required, so a preference judgment method on a rating scale can supplement this and facilitate interpretation through the subjective evaluation of a range of potentially relevant concepts. In this study, the five-point ordinal scale for the rating tests was "disagree (1 pt)—(2 pt)—agree (3 pt)—(4 pt)—strongly agree (5 pt)."[25] Different questions were used for the visual and auditory experiments (see Tables III and VI).

IV. VISUAL PERCEPTION

A. Visual perspective

Figure 4 shows that any position in the audience seating has a horizontal angle (θ), a vertical angle (φ), and a distance (D) from the stage center position. These parameters are classified as room dimensional factors. Table II shows the room dimensional parameters measured at each seat. Stage ratio (SR) is defined as the ratio between the projected stage area and the entire visible area in a stage view. In this paper, SR values were calculated from photographs taken at each seat. The focal length of digital camera was fixed at 38 mm in relation to the 35 mm format. The rationale for SR is that a large stage view is likely to be desirable because the performance occurs on the stage. Within a given theater, SR should be proportional to the inverse of the squared distance from the stage, notwithstanding the effect of partial occlusion of the stage view and angle-of-view effects. For this theater, the linear relationship between SR and $1/D$ is a little stronger than that between SR and $1/D^2$ ($R^2=0.94$, $p<0.01$) and is given in Eq. (1), which has a correlation of $R^2=0.96$ ($p<0.01$),

$$\mathrm{SR} \approx 7.16 \times \frac{1}{D} - 0.09. \tag{1}$$

Projected distance (PD), which can be thought of as how far back a seat is from the stage projected onto the center line of the auditorium, is determined by multiplying the distance and cosine values of horizontal and vertical angles, as shown in the following equation:

$$\mathrm{PD} = D \cos\theta \cos\varphi = D \cos\delta \ (\mathrm{m}). \tag{2}$$

In Eq. (2), PD increases when either the horizontal or vertical angle decreases. The rationale for this measure of PD could be a supposition that being close to the stage center contributes positively to visual preference ratings, since these ratings provide a large stage image with an undistorted perspective. However, in this study, we also use PD (a simple geometric parameter) for comparison and in combination with SR (which is directly derived from images).

Equation (3) shows the product of SR and PD and is defined as visual potential (VP) (Table II). Derived in this

TABLE III. Questions for preference judgment on a five-point rating scale of visual cues and related subjective terms.

No.	Question	Subjective terms
1	The stage is distant	Perceived distance
2	The stage is bright	Brightness
3	I feel intimacy with the performance	Visual intimacy
4	I can see the whole stage without obstruction	Unobstructed stage view
5	The stage appears to be large	Apparent stage size
6	It is visually comfortable to enjoy opera	Visual comfort
7	I like the stage image from this seat	Overall impression

TABLE IV. Scale values of visual preference (S.V.v), auditory preference (S.V.a), and multimodal seat preference (S.V.seat).

	A	B	C	D	E	F	G	H	I
S.V.v	1.15	0.74	0.30	0.91	−0.53	−0.34	−1.76	−0.08	−0.38
S.V.a	1.14	−0.56	0.47	−1.28	0.35	−0.26	−0.37	−0.03	0.55
S.V.seat	1.28	−0.24	0.36	−0.99	0.09	−0.22	−0.82	0.08	0.47

way, VP cancels out some of the effect of distance, and in so doing emphasizes the deviation of SR from a simple distance function (primarily due to visual obstructions of the stage and perspective-related distortion). For example, a given obstructed area of the stage image will reduce the VP by a greater proportion for a distant than a close seat. The resulting values of VP still have a small negative correlation with distance ($R=-0.38$), but a stronger negative correlation with three-dimensional angle δ ($R=-0.71$). In summary, VP provides a visual measure that combines angle of view, distance, and obstruction, but it emphasizes the nondistance attributes of the image.

$$\mathrm{VP} = \mathrm{PD} \times \mathrm{SR}. \tag{3}$$

B. Evaluation procedure

Visual preference was subjectively measured for the selected seats by the method of paired comparisons. A total of 36 pairs of stage views (consisting of all pair combinations of nine distinct stage images) were presented to subjects to derive scale values as a function of subjective preference. Fifty subjects participated in this experiment, and the results of every subject passed the consistency test ($p<0.05$). The average result of these 50 subjects' scores showed significant agreement ($p<0.01$). The subjects were also asked to answer the questions on the visual qualities of each image on a five-point ordinal scale, as shown in Table III. All subjective terms were explained in advance. The intersubject averaged scores for each question were investigated in relation to the room dimensional parameters for stage image and position.

C. Results and discussion

Table IV shows the calculated visual preference values for each seat from the paired-comparison test. Correlation analysis with room dimensional factors and visual preference was performed, and Table V shows the calculated correlation coefficients (Pearson coefficient, two-tailed test). As shown in Table V, the visual preference of stage view was correlated with angular components (θ, φ, and δ), as well as the compound value VP. Of the angular measurements, the three-dimensional angle δ showed the highest correlation to visual preference. Interestingly, distance components (D, PD, and SR) had no significant correlations with visual preference or VP, even though VP is based on D and SR.

The relationships between visual preference and the room dimensional parameters were examined by using regression analysis, and the best regression for the scale value of preference is given by the following equation;

$$\mathrm{S.V.v} \approx a_1(\mathrm{VP}) + C_1. \tag{4}$$

The model shown in Eq. (4) accounts for the effect of VP (VP=PD×SR) ($a_1=0.629$, $C_1=-2.796$, $R^2=0.86$, and $p<0.01$). This relationship indicates that visual preference is affected not only by the distance element (PD), but also by the size and distortion of stage view (SR). The relationship between the observed and predicted scale values of visual preference is shown in Fig. 5(a).

From the subjective evaluations using a five-point ordinal scale, the averaged scores of visual impressions were calculated to investigate various factors affecting visual preference. The averaged five-point score of "overall impression" is highly correlated with the paired-comparison test scale value of visual preference ($R=0.98$, $p<0.01$). The significant factors in terms of correlation coefficients with the visual preference were "visual comfort" ($R=0.98$, $p<0.01$), "perceived stage size" ($R=0.83$, $p<0.01$), "unobstructed stage view" ($R=0.80$, $p<0.01$), "visual intimacy" ($R=0.76$, $p<0.05$), and "brightness" ($R=0.73$, $p<0.05$). All subjective factors can be grouped into two categories; distance-related and comfort-related factors. "Perceived distance," "brightness," "visual intimacy," and "apparent stage size" were highly correlated to each other so as to be grouped as

TABLE V. Correlation coefficients (R) between room dimensional parameters and visual preference.

	D	Θ	φ	δ	PD	SR	LC	VP
S.V.v	⋯	−0.73[a]	−0.74[a]	−0.83[b]	⋯	⋯	⋯	0.93[b]
Θ	⋯							
φ	⋯	⋯						
δ	⋯	0.93[b]	0.69[a]					
PD	0.97[b]	−0.67[a]	⋯	⋯				
SR	−0.87[b]	⋯	⋯	⋯	−0.76[a]			
LC	⋯	⋯	⋯	⋯	0.88[b]	−0.79[a]		
VP	⋯	⋯	−0.79[a]	−0.72[a]	⋯	⋯	⋯	

[a] $p<0.05$.
[b] $p<0.01$.

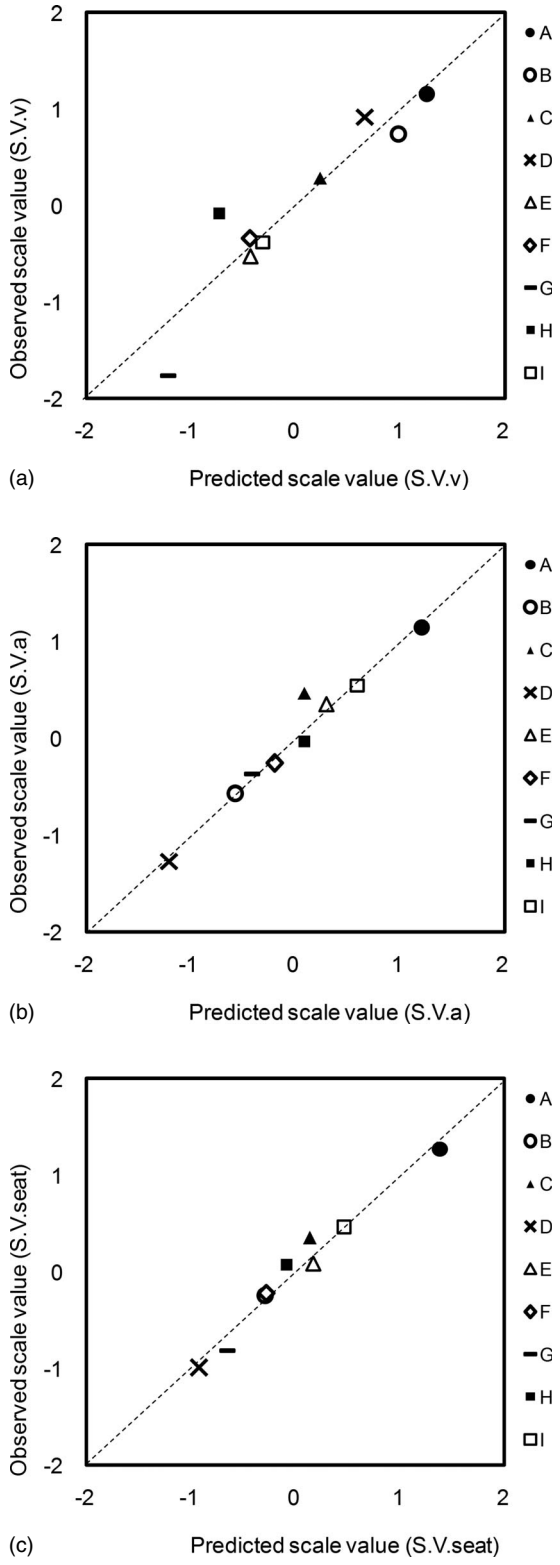


FIG. 5. Relationship between the observed and predicted scale values. These graphs show good agreement between subjective responses and objective measures. Most of response in the seat preference (c) follow auditory preference (b), but some responses were clearly affected by visual preference. For example, the multimodal seat preference score of *G* position was still poor but was increased relative to the auditory preference score due to its good visual preference score. (a) visual preference; (b) auditory preference; (c) seat preference.

distance-related factor ($R>0.9$, $p<0.01$). Other factors, unobstructed stage view and visual comfort, are highly correlated to each other ($R=0.86$, $p<0.01$) so as to be grouped into comfort-related factor. The relationships between the visual preference and results of the other scales were examined by using a multiple regression analysis. The regression for the predicted scale value is given by the following equation:

$$\text{S.V.v} \approx b_1(\text{visual comfort}) + b_2(\text{visual intimacy}) + C_2, \tag{5}$$

where coefficients b_1, b_2, and constant C_2 are 0.757, 0.211, and −2.773, respectively. These coefficients were statistically significant ($R^2=0.98$, $p<0.01$). The contribution of visual comfort to visual preference was much higher than visual intimacy. The VP parameter also showed a high correlation with visual comfort ($R=0.91$, $p<0.01$) and visual intimacy ($R=0.74$, $p<0.05$).

TABLE VI. Questions for preference judgment on a five-point rating scale of auditory cues and related subjective terms.

No.	Question	Subjective terms
1	The sound is heard clearly	Clarity
2	The sound is reverberant	Reverberance
3	The sound is loud	Loudness
4	The apparent size of sound source is large	Apparent source width
5	The sound is enveloping	Envelopment
6	I like the sound of this seat	Overall impression

V. AUDITORY PERCEPTION

A. Test procedure

The subjective evaluation procedure for auditory preference tests was the same as that of visual preference tests. A total of 36 pairs of sounds (consisting of all pair combinations of nine distinct auralized music excerpts) were presented to subjects to derive scale values as a function of subjective preference. Fifty subjects participated in this experiment, and the results of 42 subjects passed the consistency test ($p<0.05$) for the paired-comparison method. The test results of these 42 subjects showed significant agreement ($p<0.01$). The subjects were also asked to answer questions on auditory preference by using a series of five-point ordinal scales, as shown in Table VI. All subjective terms were explained in advance for understanding. The intersubjects' averaged scores for each question were investigated in relation to the measured acoustical parameters such as T30 and SPL.

B. Results and discussion

Table IV shows the calculated auditory preference values for each seat from the paired-comparison test. Correlation analysis between the acoustical parameters and auditory preference was performed. Table VII indicates the calculated correlation coefficients (Pearson coefficient, two-tailed test).

Table VII shows that high SPL, C80, and D50 and low EDT appear to be favored in this opera theater. As for frequency characteristics, it was found that SPL in the high frequency band (2 and 4 kHz octave bands), D50 in the mid-frequency band (500 Hz and 1 kHz octave bands), and C80 in the low frequency band (125 and 250 Hz octave bands)

TABLE VII. Correlation coefficients between acoustical parameters and auditory preference.

		D	Δt_1	SPL	EDT	T30	C80	D50	LF_{E4}	BQI (1-$IACC_{E3}$)
	S.V.a	⋯	⋯	0.98[a]	−0.74[a]	⋯	0.85[b]	0.72[a]	⋯	⋯
	SPL	⋯	⋯							
	EDT	⋯	⋯	⋯						
	T30	⋯	⋯	⋯	0.88[b]					
	C80	⋯	⋯	0.80[b]	−0.81[b]	⋯				
	D50	⋯	⋯	0.73[a]	⋯	⋯	0.84[b]			
	LF_{E4}	⋯	⋯	⋯	⋯	⋯	⋯	⋯		
	BQI (1-$IACC_{E3}$)	⋯	−0.80[b]	⋯	⋯	⋯	⋯	⋯	⋯	
S.V.a	Low freq. (125–250 Hz)	⋯	⋯	0.80[a]	−0.75[a]	⋯	0.86[b]	⋯	⋯	⋯
	Middle freq. (500–1 kHz)	⋯	⋯	0.85[b]	−0.72[a]	⋯	0.78[a]	0.84[b]	⋯	⋯
	High freq. (2 k–4 kHz)	⋯	⋯	0.97[b]	−0.67[a]	−0.71[a]	0.69[a]	⋯	⋯	⋯

[a] $p<0.05$.
[b] $p<0.01$.

had high correlations with the auditory preference. Clarity indicators (C80 and D50) are highly correlated with SPL. Reverberation time (T30) and spatial factors (LF_{E4}, 1-$IACC_{E3}$) are not significantly correlated with auditory preference. The relationships between auditory preference and acoustical characteristics were examined by using a multiple regression analysis. The regression for the predicted scale value is given by the following equation:

$$\text{S.V.a} \approx c_1(\text{SPL}) + C_3. \tag{6}$$

This model, as shown in Eq. (6) expresses auditory preference as a function of SPL only ($c_1=0.417$, $C_3=-27.351$, $R^2=0.98$, and $p<0.01$). C80 was the second predictor for auditory preference, but adding C80 to the prediction model did not increase the R squared value due to collinearity with SPL. The relationships between the observed and predicted scale values of auditory preference are shown in Fig. 5(b). For these stimuli, there is only a very weak relationship between distance and SPL ($R=-0.20$), and distance does not significantly correlate with the results of the auditory experiment.

From the subjective evaluations using a five-point ordinal scale, the averaged scores of auditory impressions were calculated to investigate various factors affecting auditory preference. The averaged five-point score of the overall impression has high correlation with the scale value of auditory preference ($R=0.96$, $p<0.01$), so the results of paired-comparison test are explained by subjective terms evaluated from the preference judgment test on a rating scale. The significant factors in terms of correlation coefficients with the auditory preference were "clarity" ($R=0.91$, $p<0.01$), "loudness" ($R=0.89$, $p<0.01$), "apparent source width" ($R=0.80$, $p<0.05$), and envelopment ($R=0.70$, $p<0.05$). Among these parameters, "clarity", "loudness", "apparent source width", and "envelopment" (but not "reverberance") were highly correlated to each other ($R>0.8$, $p<0.01$). The relationships between the auditory preference and evaluation results of each factor were examined by using a multiple regression analysis. The regression for the predicted scale value is given by the following equation:

$$\text{S.V.a} \approx d_1(\text{loudness}) + d_2(\text{reverberance}) + C_4, \tag{7}$$

where coefficients d_1, d_2, and constant C_4 are 0.904, −0.503, and −1.419, respectively. These coefficients were statistically significant ($R^2=0.95$, $p<0.01$). Reverberance itself was not significantly correlated with auditory preference, but adding of reverberance to the regression model augments the significance of the model. This regression in Eq. (7) directly corresponds to the regression result of the paired-comparison test in Eq. (6). The contribution of loudness to the auditory preference was much greater than reverberance. Loudness showed a high correlation with SPL ($R=0.89$, $p<0.01$). Reverberance was not significantly correlated with EDT, but did have a negative correlation with C80 ($R=-0.73$, $p<0.05$).

VI. AUDITORY-VISUAL PREFERENCE

A. Test procedure

Two experiments on audio-visual cross modality were conducted. The first experiment is a paired-comparison test of seat preference by using nine auditory-visual stimuli from seat positions A-I. The purpose of this test was to derive scale values in order to determine the contribution of visual and auditory preferences on seat preference. Stimulus pairs were prepared in the same manner as the previous paired-comparison tests. Fifty subjects participated in this experiment, and the results of 46 subjects passed the consistency test ($p<0.05$) by using the paired-comparison method. The average test results of these 46 subjects showed significant agreement ($p<0.01$).

The second experiment was another paired-comparison test similar to the first experiment. Some stimuli that were used in the first experiment may have "preferred" auditory

TABLE VIII. Scale values of seat preference by auditory and visual cues. nine auditory-visual combinations were evaluated by using three images and three sounds which were selected from the previous visual-only and auditory-only experiments.

		Visual cues			
		Less preferred	Neutral	Preferred	Mean
Auditory cues	Less preferred	−1.16	−1.11	−0.69	−0.98
	Neutral	−0.20	−0.36	0.87	0.10
	Preferred	0.97	0.28	1.39	0.88
	Mean	−0.13	−0.40	0.53	0.00

cues but "less preferred" visual cues or vice versa. Thus, cross-matched, nine combinations of three auditory and three visual stimuli were used for investigation of audio-visual interaction. Three visual stage views at E, H, and D positions were selected according to their scale value of visual preference (−0.53/−0.08/0.91). Three auralized sounds from D, H, and A positions were selected according to their scale value of auditory preference (−1.28/−0.03/1.14). For example, a combination of D image and A sound is both visually and acoustically preferred. A stimulus with a scale value of near zero was selected for "neutral preference."

B. Results and discussion

The first experiment. Table IV shows the calculated seat preference values for each seat from the paired-comparison test. The correlation coefficients of seat preference were $R=0.36$ ($p>0.05$) for visual preference and $R=0.94$ ($p<0.01$) for auditory preference (Pearson's coefficient, two-tailed test). The following equation shows the prediction model of seat preference in terms of visual and auditory preferences. The standardized partial regression coefficients of variables e_1 and e_2 in Eq. (8) were 0.324 and 0.922, respectively. These coefficients and the model were statistically significant ($p<0.01$ for e_1 and e_2; $R^2=0.98$, $p<0.01$ for the model). The "(S.V.v)⊗(S.V.a)" term indicating auditory-visual interaction was found not to be significant.

$$\text{S.V.seat} \approx e_1(\text{S.V.v}) + e_2(\text{S.V.a}). \tag{8}$$

From these standardized coefficients, the contribution of each component to seat preference is estimated. The calculated contribution reveals that the effects of auditory preference are more contributive to seat preference (85%) than those of visual preference (10%) in the natural sound-vision compositions. Preference is also expressed in terms of objective parameters in Eq. (9),

$$\text{S.V.seat} \approx f_1(\text{VP}) + f_2(\text{SPL}) + C_5, \tag{9}$$

where coefficients f_1, f_2, and constant C_5 are 0.147, 0.374, and −25.18, respectively ($R^2=0.96$, $p<0.01$). The relationship between the observed and predicted scale values of seat preference are shown in Fig. 5(c).

From the existing literature, auditory dominance is mainly found in the temporal tasks. The experience of opera in the testing condition is based on two dimensional and static stage image, but operatic singing voice was quite realistically reproduced by headset. Though the audible and visible ranges for the subject in the testing room were taken from the actual hall, this difference of reproduction quality emphasized the temporal variation in terms of SPL. However, it turned out to be clear that both visual and auditory cues subjectively and objectively affected seat preference from Eqs. (8) and (9). In addition, there was no significant correlation for seat preference with distance, and as a simple variable, distance does not make a significant contribution to any multiple regression model.

The second experiment. Table VIII and Fig. 6 show the calculated seat preference values by auditory and visual preferences from the paired-comparison test. As shown in both Table VIII and Fig. 6, auditory cues were found to affect seat preference linearly, but the visual cues showed nonlinear effects on seat preference. This nonlinearity can be considered as an outcome of auditory-visual interaction. To evaluate the effect of cross-modal interaction, two-way analysis of variance was employed. As shown in Table IX, significant interaction was found between auditory cues and visual cues when cross-matched stimuli were evaluated. The results for seat preference have significant differences for both auditory and visual cues. Like the first experiment, the contribution of auditory cues (73.8%) was much larger than that of visual cues (17.3%).

The nature of the significant interaction effect was investigated by using a simple main effect (SME) analysis (Table IX). With regard to the SME for visual cues, all three cases of sound showed a significant difference between vi-

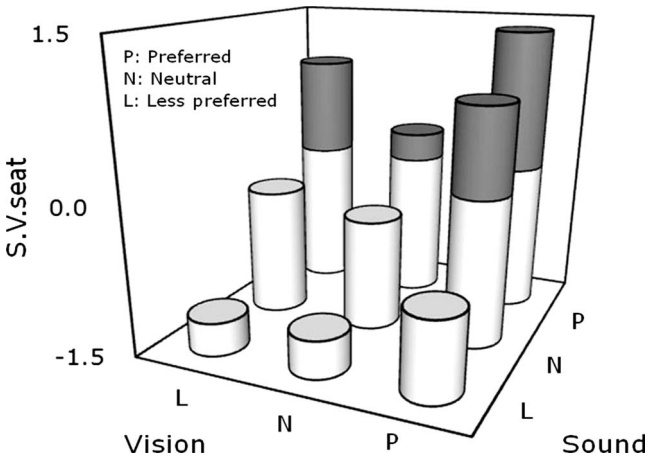


FIG. 6. Scale values of seat preference for the second auditory-visual test. In the case where the auditory stimulus is fixed as preferred (P), nonlinear response of seat preference was observed by changing visual cues. More generally, it can be seen that auditory preference has a stronger influence than visual preference on the results.

TABLE IX. Summary table of the two-way analysis of variance for multimodal seat preference including results of simple main effect (SME) analysis. The contribution of auditory cues on seat preference was much larger than that of visual cues.

Variance source	Sum of squares (SS)	*df*	Mean square (MS)	*F* ratio	*p*	Contribution (%)
Auditory cues	38.58	2	19.29	252.14	<0.01	73.8
Preferred vision	17.46	2	8.73	108.04	<0.01	
Neutral vision	6.92	2	3.46	69.25	<0.01	
Less preferred vision	16.85	2	8.42	85.31	<0.01	
Visual cues	11.42	2	5.71	74.63	<0.01	17.3
Preferred sound	5.01	2	2.50	33.31	<0.01	
Neutral sound	7.87	2	3.94	33.77	<0.01	
Less preferred sound	1.18	2	0.59	16.70	<0.01	
Interaction (Auditory × Visual)	2.65	4	0.66	8.65	<0.01	
Error	10.33	135	0.08			
Total	62.97	144				

sual stimuli. As shown in Table IX, it was found that visual difference clearly affected seat preference when the auditory cues were preferred and "neutral." This means that a good acoustical condition already contributed sufficiently to seat preference due to auditory dominance, hence sensitivity to visual cues was less than that under poor acoustical conditions, resulting in the nonlinear effect of visual cues. With regard to the SME for auditory cues, all three cases of visual image showed significant differences between auditory stimuli as expected from the results of the first experiment. Most notably, auditory difference strongly affected seat preference when the visual cue was preferred. The interaction between auditory and visual cues which were found in the previous studies on concert halls[12,13,16] has been confirmed in this result for an opera theater. However, while the previous studies[12,13] on the cross-modal interaction were only concerned with particular perceptual attributes such as reverberance, apparent source width, or apparent room size, this study deals with the auditory and visual impressions that contribute to overall seat preference. In the previous findings on general multisensory processes, the dominance of visual cues has been emphasized.[5–7] However, the present study has found that the auditory cues were more influential in the overall impression than the visual cues in assessing stimuli derived from an opera theater. Because the range of auditory and visual cues in this study was collected through the field measurement, this dominance of auditory cues is indicative, at least to some extent, of the balance of cue influence that would occur in the context of a real opera theater with a static visual scene.

VII. CONCLUSIONS

This study of auditory-visual interaction undertakes objective and subjective assessments of visual and auditory cues that contribute to seat preference in an opera theater. Seat preference was assumed to be derived from visual and auditory aspects, so those visual-only and auditory-only conditions were evaluated in advance, followed by an evaluation of bimodal stimuli.

From the visual-only experiment, visual preference is primarily affected by angle from the centerline and can be degraded by obstruction of stage view and factors related to distance. In this study, influence of distance alone is not as strong as might be supposed from general observations about auditoria.[16,17] For the visual preference test, VP is introduced, which in this instance is a strong predictor of visual preference. SR, which is derived from photographic images and approximated from an inverse function of distance, is useful in determining VP because SR includes the influence of visual obstructions. The preference results of the present study indicate that both visual comfort and visual intimacy contribute to visual preference and are highly related to VP. Therefore, visual intimacy does contribute to seat preference, which supports the previously mentioned survey findings of Hyde[10] on the importance of vision in judgments of intimacy in auditoria. While SR is easily defined in a proscenium theater, it would be more difficult to define and use it in a theater if there were no frame around the stage. More generally, the findings on visual preference support the concept that geometric and visual parameters reliably predict visual preference in a given auditorium and suggest that some refinement of these measures could be a fruitful area for future research directed toward improved theater design.

In the auditory experiments, it was found that SPL was the strongest predictor of auditory preference. Acoustic clarity was also found to be a contributive factor, confirmed by the results of the preference judgment test on a rating scale. However, considering that this auditorium has quite high clarity for an opera theater (relative to Beranek's survey of major opera theaters[14]), the results hint that greater clarity is a more general way of improving opera theater acoustics, at least for the sound of an unaccompanied opera soloist.

Seat preference was augmented by VP in addition to SPL when a visual image accompanied the auditory stimulus.

From the results of the auditory-visual cross-modality experiments, it can be concluded that both the perceived acoustical quality and the perceived visual quality in opera houses correlate with positive evaluation of the auditory-visual quality. Thinking back to the question raised in the Introduction, auditory cues were more influential in evaluating seat preference than visual cues in the opera theater. From a theater design perspective, an important observation from this and similar studies of concert halls is that auditory and visual qualities may be used to reinforce or compensate for each other. This concept supports the idea that both must be carefully considered in theater design.

The balance between auditory and visual influences in subjective assessments would vary with the amount of salient information conveyed through each mode. In this study, the visual stimuli were static. However, one could hypothesize that dynamic visual stimuli would have greater impact on multimodal perception and evaluation than static visual stimuli. It must also be borne in mind that the simulations in the present experiments have the subject as an audience member without any possibility of interaction with the performers. In a real theater, communication between performers and the audience occurs in both directions, and auditory and visual factors are likely to influence the effectiveness of this interaction. Audience members also interact with each other, and it may be that a view of the audience from a seat could be beneficial. These are areas for further research.

ACKNOWLEDGMENTS

The authors thank Pontus Larsson and the anonymous reviewers for their detailed advice toward the preparation of this paper.

[1]C. Spence and J. Driver, *Crossmodal space and crossmodal attention* (Oxford University Press, New York, 2004).
[2]R. B. Welch and D. H. Warren, in *Handbook of Perception and Human Performance*, edited by K. R. Boff, L. Kaufman, and J. P. Thomas (Wiley, New York, 1986).
[3]B. E. Stein and M. A. Meredith, *The Merging of the Senses* (MIT Press, Cambridge, 1993).
[4]G. Calvert, C. Spence, and B. E. Stein, *The Handbook of Multisensory Processes* (MIT Press, Cambridge, 2004).
[5]J. Beerends and F. Caluwe, "The influence of video quality on perceived audio quality and vice versa," J. Audio Eng. Soc. **47**, 355–362 (1999).
[6]H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature (London) **264**, 746–748 (1976).
[7]I. P. Howard and W. B. Templeton, *Human Spatial Orientation* (Wiley, London, 1966).
[8]C. R. Scheier, R. Nijwahan, and S. Shimojo, "Sound alters visual temporal resolution," Invest. Ophthalmol. Visual Sci. **40**, 4169 (1999).
[9]B. E. Stein, N. London, L. K. Wilkinson, and D. D. Price, "Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis," J. Cogn Neurosci. **8**, 497–506 (1996).
[10]J. R. Hyde, "Acoustical intimacy in concert halls: Does visual input affect the aural experience?," Paul S. Veneklasen Research Foundation, Santa Monica, Report No. 010301 (2003).
[11]J. R. Hyde, "Multisensory integration and the concert experience: An overview of how visual stimuli can affect what we hear," J. Acoust. Soc. Am. **115**, 2402 (2004).
[12]C. Nathanail, C. Lavandier, J. D. Polack, and O. Warusfel, "Influence of the visual information on auditory perception. Consequences on the subjective and objective characterization of room acoustic quality," Proceedings of the International symposium on simulation, visualization and auralization for acoustic research and education, Tokyo, Japan, 1997, pp. 285–290.
[13]P. Larsson, D. Västfjäll, and M. Kleiner, "Multimodal interaction in real and virtual concert halls," J. Acoust. Soc. Am. **115**, 2403 (2004).
[14]M. Barron, "Subjective study of British symphony concert halls," Acta Acust. **66**, 1–14 (1988).
[15]D. Cabrera, A. Nguyen, and Y. J. Choi, "Auditory versus visual spatial impression: A study of two auditoria," Proceedings of 10th International Conference on Auditory Display, Sydney, Australia, 2004.
[16]M. Barron, *Auditorium Acoustics and Architectural Design* (E & FN Spon, London, 1993), pp. 297–337.
[17]I. Mackintosh, *Architecture, Actor and Audience* (Routledge, London, 2002), pp. 126–141.
[18]L. L. Beranek, *Concert Halls and Opera Houses* (Springer-Verlag, New York, 2004), pp. 536–537.
[19]R. Pompoli and N. Prodi, "Guidelines for acoustical measurements inside historical opera houses: Procedures and validation," J. Sound Vib. **232**, 281–301 (2000).
[20]L. Tronchin and A. Farina, "The acoustics of the former Teatro 'La Fenice' in Venice," J. Audio Eng. Soc. **45**, 1051–1062 (1997).
[21]T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," J. Acoust. Soc. Am. **98**, 988–1007 (1995).
[22]F. Mosteller, "Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," Psychometrika **16**, 207–218 (1951).
[23]E. Parizet, "Paired Comparison Listening Tests and Circular Error Rates," Acta. Acust. Acust. **88**, 594–598 (2002).
[24]L. L. Thurstone, "A Law of Comparative Judgment," Psychol. Rev. **34**, 273–286 (1927).
[25]F. Rumsey, S. Zielinski, and R. Kassier, "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences," J. Acoust. Soc. Am. **117**, 3832–3840 (2005).

Cross Ψ_B-energy operator-based signal detection[a)]

Abdel-Ouahab Boudraa[b)]
IRENav, Ecole Navale, Lanvéoc Poulmic BP600 29200 Brest–Armées, France, and E^3I^2 (EA3876), ENSIETA, 2 rue Francois Verny, 29806 Brest Cedex 9, France

Jean-Christophe Cexus
E^3I^2 (EA3876), ENSIETA, 2 rue Francois Verny, 29806 Brest Cedex 9, France

Karim Abed-Meraim
TSI department, ENST-Paris, 46 rue Barrault 75634, Paris Cedex 13, France

(Received 6 April 2007; revised 7 March 2008; accepted 4 April 2008)

In this paper, two methods for signal detection and time-delay estimation based on the cross Ψ_B-energy operator are proposed. These methods are well suited for mono-component AM-FM signals. The Ψ_B energy operator measures how much one signal is present in another one. The peak of the Ψ_B operator corresponds to the maximum of interaction between the two signals. Compared to the cross-correlation function, the Ψ_B operator includes temporal information and relative changes of the signal which are reflected in its first and second derivatives. The discrete version of the continuous-time form of the Ψ_B operator, which is used in its implementation, is presented. The methods are illustrated on synthetic and real signals and the results compared to those of the matched filter and the cross correlation. The real signals correspond to impulse responses of buried objects obtained by active sonar in iso-speed single path environments.
 [DOI: 10.1121/1.2916583]

PACS number(s): 43.60.Bf, 43.60.Kx [WMCs] Pages: 4283–4289

I. INTRODUCTION

Signal detection and time-delay (TD) estimation between signals are important problems in communications and signal processing. Typical applications involve radar, sonar, machine fault diagnostics, biomedicine, and geophysics.[1] For example, in sonar signal processing the difference in arrival time of a signal at two or more spatially separated receivers is used to estimate range and bearing of the source. A common method of signal detection based on TD involves cross correlation of the receiver output, whereby an estimate of the TD is given by the argument that maximizes the cross-correlation (CC) function. Although the CC method is the simplest similarity measure used in signal processing, it is not sensitive to nonlinear dependence of the signals. In practice, the interaction between signals may be nonlinear. Furthermore, sensor characteristics may also cause nonlinear distortions. Consequently, the maximum CC may not correspond to the maximum signal interaction. In this case, the resulting TD may be erroneous. To tackle this problem a new similarity function that measures the nonlinear interaction between signals is proposed.[2–4] The method is based on the cross Ψ_B-energy operator, recently introduced by the authors.[2] The Ψ_B-energy operator is derived from a second energy-like function, called the Cross Teager-Kaiser Energy Operator (CTKEO)[5,6] which measures the interaction between two real time functions. Compared to the CC, Ψ_B includes temporal information and accounts for relative changes of the signal by using the first and the second derivatives of the signal. Since it is based on a nonlinear operator, Ψ_B, the proposed method can be viewed as a nonlinear matched filter and we note it MBF (for Matched Ψ_B Filter).[7] Ψ_B can also be used to "detect" the presence of known signals as components of more complicated signals by measuring how much one signal is present in another one.[3] Thus, the Ψ_B operator can be used as a strategy for signal detection. As shown in Ref. 3, Ψ_B or methods based on Ψ_B are well suited for nonstationary signals such as mono-component AM-FM signals. Based on the CTKEO, both the Ψ_B operator and the associated methods for TD estimation or signal detection are limited to narrowband signals.

The paper is organized as follows. In the following section the Ψ_B operator is presented. Section III deals with the discrete form of Ψ_B which is the basis of its numerical implementation. In Sec. IV a Ψ_B-based signal detection framework is introduced followed by the Ψ_B-based estimation approach in Sec. V. A pseudo code of the MBF is presented in Sec. VI. Simulation and experimental results are presented in Sec. VII and concluding remarks are stated in Sec. VIII. We illustrate the method with an underwater acoustics application, where each signal is the impulse response of a buried object obtained by active sonar in a single path environment.

II. THE CROSS Ψ_B-ENERGY OPERATOR

Recently the CTKEO has been extended to complex-valued signals and an operator called Ψ_B introduced.[2] Given two complex signals $x(t)$ and $y(t)$, Ψ_B is denned as follows:[2]

[a)]Preliminary results of this work were presented at IEEE ISSPA, Sydney, Australia, 2005.

[b)]Author to whom correspondence should be addressed. Electronic mail: boudra@ecole-naval.fr; URL: http://www.ecole-navale.fr/fr/irenav/cv/boudra/index.html

$$\Psi_B(x,y) = 0.5[\Psi_C(x,y) + \Psi_C(y,x)] \tag{1}$$

$$= 0.5[\dot{x}^*\dot{y} + \dot{x}\dot{y}^*] - 0.25[x\ddot{y}^* + x^*\ddot{y} + y\ddot{x}^* + y^*\ddot{x}], \tag{2}$$

where $\Psi_C(x,y) = 0.5[\dot{x}^*\dot{y} + \dot{x}\dot{y}^*] - 0.5[x\ddot{y}^* + x^*\ddot{y}]$. It has been shown that the cross Ψ_B-energy operator of $x(t)$ and $y(t)$ is equal to the cross-Teager energies of their real and imaginary parts,[2]

$$\Psi_B(x,y) = \Psi_B(x_r,y_r) + \Psi_B(x_i,y_i) \tag{3}$$

$$\Psi_B(x_l,y_l) = \dot{x}_l\dot{y}_l - 0.5[x_l\ddot{y}_l + \ddot{x}_l y_l], \quad l \in \{r,i\}, \tag{4}$$

where $x(t) = x_r(t) + jx_i(t)$ and $y(t) = y_r(t) + jy_i(t)$. The complex form of the signals is obtained using the Hilbert transform. For Ψ_B implementation, different derivative approximations can be used.

III. DISCRETIZING THE CONTINUOUS-TIME Ψ_B OPERATOR

Discretized derivatives of the continuous Ψ_B operator are combined to obtain an expression closely related to the discrete form of the operator Ψ_{B_d} and operating on discrete-time signals $x(n)$ and $y(n)$. Different sample differences can be used, but only the two-sample backward difference is detailed herein. For simplicity, we replace t by nT_s where T_s is the sampling period, and $x(t)$ with $x(nT_s)$ or simply $x(n)$.

$$\dot{x}(t) \to [x_k(n) - x_k(n-1)]/T_s$$

$$\ddot{x}(t) \to [x_k(n) - 2x_k(n-1) + x_k(n-2)]/T_s^2$$

$$\begin{aligned}\Psi_B(x_k(t),y_k(t)) \to\ & x_k(n-1)y_k(n-1)/T_s^2 \\ & - 0.5[x_k(n)y_k(n-2) \\ & + y_k(n)x_k(n-2)]/T_s^2\end{aligned}$$

$$\begin{aligned}\Psi_B(x_k(t),y_k(t)) \to\ & \Psi_{B_d}(x_k(n-1),y_k(n-1))/T_s^2 \\ & k \in \{i,r\}.\end{aligned} \tag{5}$$

The discrete form of $\Psi_B(x(t),y(t))$ is given by

$$\begin{aligned}\Psi_B(x(t),y(t)) \to\ & [\Psi_{B_d}(x_r(n-1),y_r(n-1)) \\ & + \Psi_{B_d}(x_i(n-1),y_i(n-1))]/T_s^2,\end{aligned} \tag{6}$$

where $\mapsto$ denotes the mapping from continuous to discrete. Thus, from Ψ_B we obtain Ψ_{B_d} shifted by one sample to the left and scaled by T_s^{-2}. It is easy to show that using the two-sample forward difference from Ψ_B, we obtain Ψ_{B_d} shifted by one sample to the right and scaled by T_s^{-2}. For both asymmetric two-sample differences, Ψ_B is shifted by one sample and scaled by T_s^{-2}. If we ignore the one-sample shift and the scaling parameter, we transform $\Psi_B(x(t),y(t))$ into $\Psi_{B_d}(x(n),y(n))$ as follows:

$$\Psi_B(x(t),y(t)) \to \Psi_{B_d}(x_r(n),y_r(n)) + \Psi_{B_d}(x_i(n),y_i(n)), \tag{7}$$

$$\begin{aligned}\Psi_{B_d}(x_k(n),y_k(n)) = {} & x_k(n)y_k(n) - 0.5[x_k(n+1)y_k(n-1) \\ & + y_k(n+1)x_k(n-1)] \quad k \in \{i,r\}.\end{aligned} \tag{8}$$

The three-sample symmetric difference can also be used but it leads to a more complicated expression compared to asymmetric two-sample differences. Indeed, the asymmetric approximation is less complicated for implementation and faster than the symmetric one because it requires fewer operations.

IV. Ψ_B-BASED SIGNAL DETECTION

We motivate Ψ_B-based detection by considering the classical binary hypothesis testing problem encountered in radar or in sonar.[8] Let $s(t)$ denote a baseband transmitted signal. The received signal $R(t)$ is processed over an interval $[T_i, T_f]$ to detect the presence of a target. The hypotheses on $R(t)$ are

$$\begin{cases} H_0: R_0(t) = n(t), t \in [T_i,T_f], \\ H_1: R_1(t) = \alpha s(t-t_0) + n(t), t \in [T_i,T_f]. \end{cases}$$

Under the null hypothesis, H_0, the received signal contains only an additive noise, $n(t)$. Under the alternative hypothesis, H_1, a received time-shifted version of the transmitted signal $\alpha s(t-t_0)$ is received in the presence of noise where α denotes an unknown gain parameter. The unknown time, t_0, represents the delay of the received signal and corresponds to the unknown distance of the target. Let $T = [T_{\min}, T_{\max}]$ denote the possible range of values for t_0. The required observation interval is $[T_i, T_f] = [T_{\min}, t_0 + T_{\max}]$ in this case. For any given value of $t_0 \in T$, the decision whether to reject H_0 is given by computing

$$T_B = \arg\max_{t \in T} \left[\int_T \Psi_B(R_0(t),R_1(t))dt \right], \tag{9}$$

where T_B corresponds to the time of maximum of interaction between $R_0(t)$ and $R_1(t)$.[3] T_B is then compared to a threshold to determine the presence (H_1) or absence (H_0) of a target. Thus, the best detector calculates the interaction between the received signal and all possible time-shifted versions of the transmitted signal and picks the largest energy interaction as the basis for the detection decision. The location of the peak is the estimate of the unknown parameter $t_0 = T_B$.

V. Ψ_B-BASED TIME-DELAY ESTIMATION

We have shown that the TD estimation problem measures the interaction between two FM signals by using the Ψ_B operator.[3] Consider a signal from a remote source being received in the presence of noise at two spatially separated receivers. The time histories of the receiver outputs, denoted by $r_m(t)$ and $r_k(t)$, are given by

$$\begin{cases} r_m(t) = s(t) + n_m(t), \\ r_k(t) = \beta s(t-(k-m)r) + n_k(t), \end{cases}$$

where $s(t)$ is the signal waveform, $n_m(t)$ and $n_k(t)$ are the noise waveforms at the respective receivers, β is an attenuation factor, and τ is the difference in wavefront arrival times at two consecutive receivers $(k=2, m=1)$. We assume that

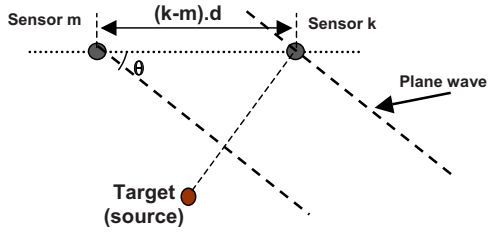


FIG. 1. (Color online) Geometry used to estimate the time delay associated with plane waves.

$n_m(t)$, $n_k(t)$ and $s(t)$ are mutually uncorrelated.[3] Propagation TD between receivers k and m is given by $\Delta d/c=(k-m)\tau$, where $\Delta d=d_k-d_m$ is the path length difference, and c is the sound speed in the medium. When the target is sufficiently distant from the receivers the wavefronts can be approximated by plane waves and the theoretical TD, T_{Theor}, is given by

$$T_{\text{Theor}} = (k-m)d \sin\theta/c, \tag{10}$$

where d is the distance between two consecutive receivers and θ is bearing angle (See Fig. 1).

VI. PSEUDO CODE OF MBF

The complex form of the signal is obtained using the Hilbert transform. The MBF implementation involves the following steps:

Inputs:

Emitted signal: $s(p)=s_a(p)+js_b(p)$, $p\in\{1,2,\ldots,w\}$,

Received signal: $r(n)=r_a(n)+jr_b(n)$, $n\in\{1,2,\ldots,N\}$,
where a and b indicate the real and imaginary parts of the complex signal respectively. w and N are time durations of $s(t)$ and $r(t)$, respectively, and $z=1$ to M denotes the index of the sliding window.

Outputs: T_{B}.

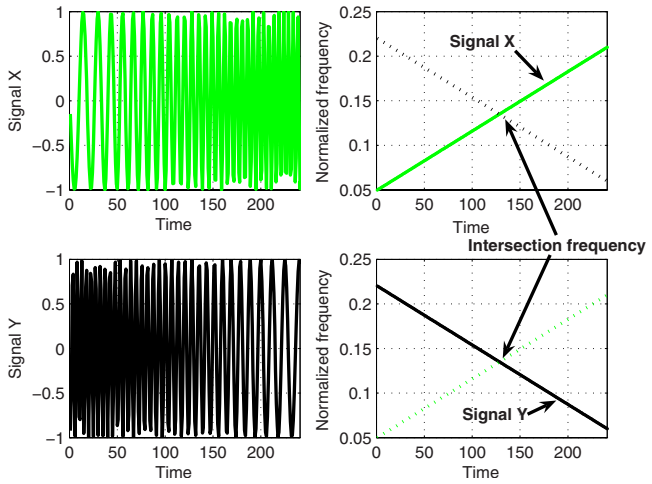


FIG. 2. (Color online) Linear chirp test signals (left) and their respective IFs (right).

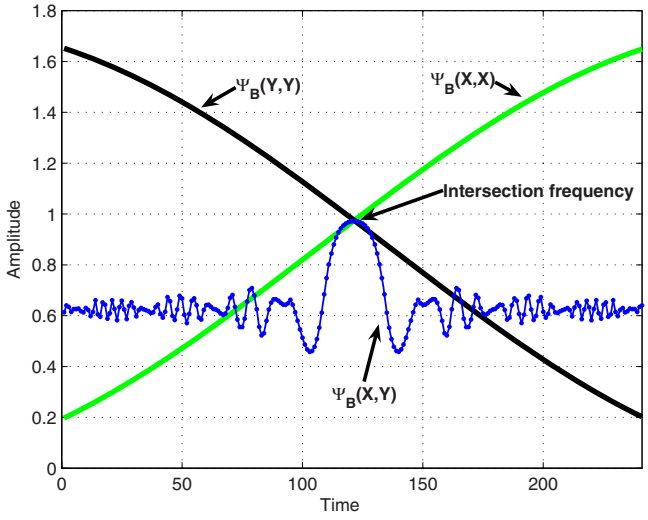


FIG. 3. (Color online) Interaction measure between $x(t)$ and $y(t)$.

Pseudo Code:
$M=N-w+1$

For $z=1$ to M
$g^z(n-z+1)\leftarrow r(n), n\in F=\{z,\ldots,w+z-1\}$

Compute $\Psi_a^l=\Psi_{\text{B}_d}(s_a(l),g_a^z(l))$ and
$\Psi_b^l=\Psi_{\text{B}_d}(s_b(l),g_b^z(l))$ using Eq. (8)
$\Psi_{\text{B}_d}(s(l),g^z(l))\leftarrow\Psi_a^l+\Psi_b^l,\quad l\in\{1,\ldots,w\}$

Compute the sum $\mathsf{I}(\mathsf{z})$ of Ψ_{B_d} values over F:

$$\mathsf{I}(\mathsf{z})=\sum_{l=z}^{w+z-1}\Psi_{\text{B}_d}(s(l),g^z(l))$$

EndFor

Compute $\text{T}_{\text{B}}=\arg\max_{1\le z\le M}[\mathsf{I}(\mathsf{z})]$

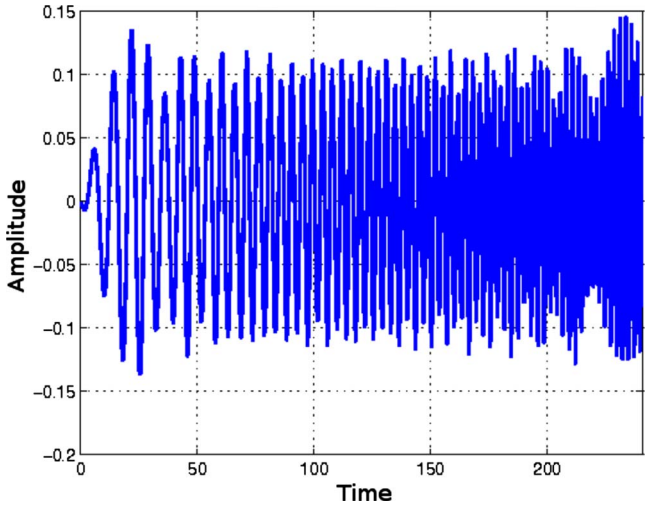


FIG. 4. (Color online) Interaction measure between $x(t)$ and $y(t)$ using CC.

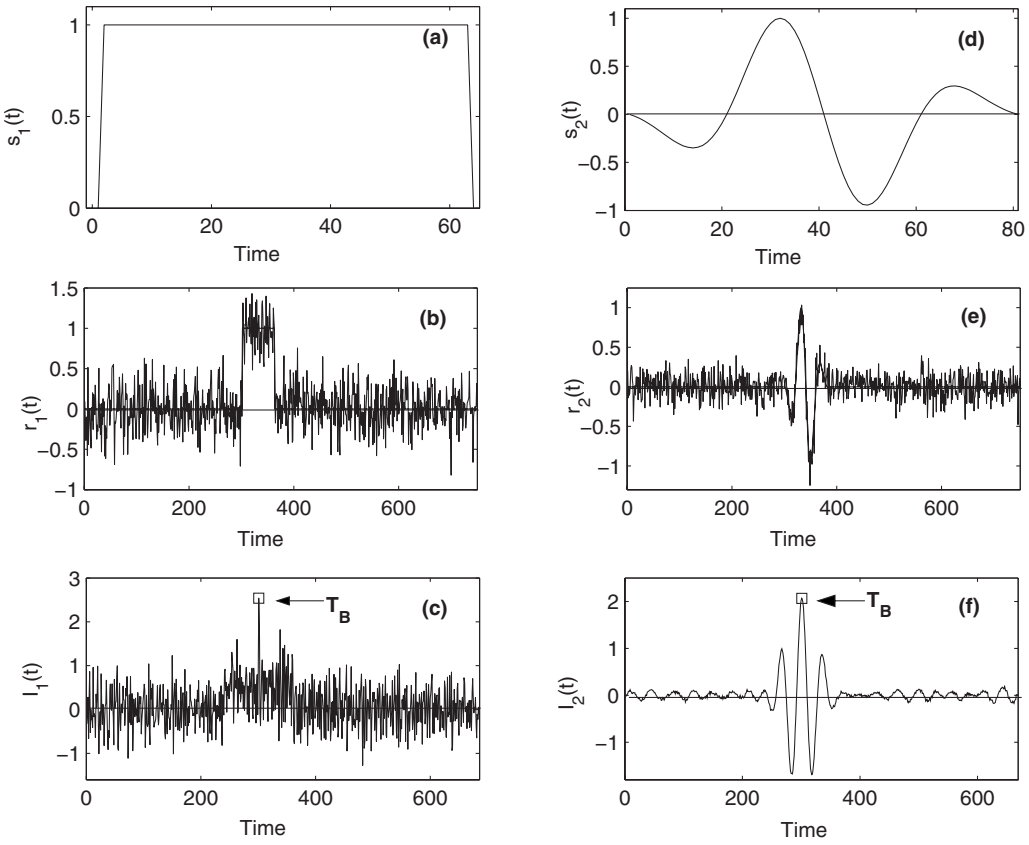


FIG. 5. Signal detection results with $a=0.95$: (a) First emitted signal. (b) Received echo of the first signal, (c) Integral, I, of Ψ_{B_d} calculated between the emitted signal and the echo of the first signal, (d) Second emitted signal. (e) Received echo of the second signal, and (f) Integral, I, of Ψ_{B_d} calculated between the emitted signal and the echo of the second signal.

VII. RESULTS

The results presented below highlight the contributions of the paper. The proposed method is illustrated on synthetic and real signals and the results are compared to those of the CC method and the Matched Filter (MF). The real signals correspond to impulse responses of buried objects. The present study is limited to signals obtained by active sonar in iso-speed single path environments. Both Ψ_B and MBF are implemented using asymmetric two-sample differences [Eqs. (7) and (8)].

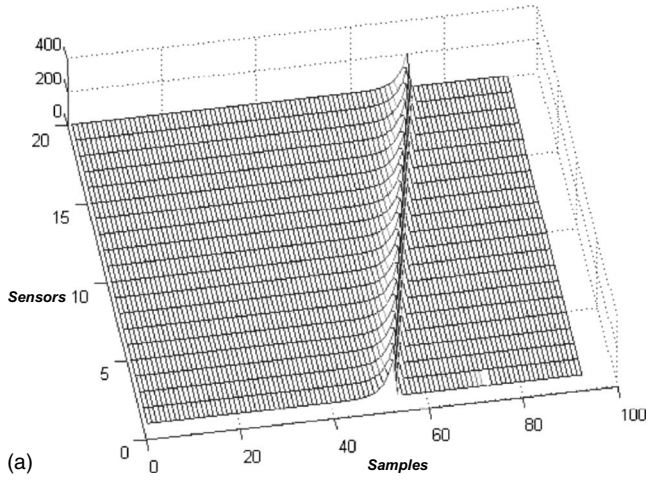
A. Synthetic signals

We first show an example of interaction between two nonstationary signals measured by Ψ_B operator and compare the result to the CC approach. Figure 2 shows an example of two linear FM signals $x(t)$ and $y(t)$ with the corresponding Instantaneous Frequencies (IFs). The IF of $x(t)$ increases linearly with time while that of $y(t)$ decreases with time. The inter-action between $x(t)$ and $y(t)$ is calculated using Eq. (2). Figure 3 shows the energy of each signal and the energy of their interaction. The maximum of interaction corresponds to the instant when the two IFs intersect (Figs. 2 and 3) and also where the energy of $x(t)(\Psi_B(x,x))$ and that of $y(t)(\Psi_B(y,y))$ are equal (Fig. 3). The point where the IFs intersect is located at $t=125$. Maximums of interaction between $x(t)$ and $y(t)$ occur at $t=125$ and $t=240$ for Ψ_B and CC, respectively, as shown in Figs. 3 and 4. The CC fails to point out, as expected, the point of maximum of interaction. This result shows that the CC measure is insensitive to nonlinear dependency between $x(t)$ and $y(t)$. Away from the point where the IFs cross, the amplitude of the interaction decreases because there is less similarity between the two signals. As the IFs converge from the time origin to their intersection, the interaction intensity of the signals increases and the maximum of interaction is achieved at the intersection. This example shows that Ψ_B is more effective with nonstationary signals than the CC. This is due to the fact that Ψ_B is a nonlinear operator while CC is linear.

In this simulation an example of delay estimation performed between two signals is presented. Two synthetic ref-

TABLE I. Estimated T_B versus SNR signals $s_1(t)$ and $s_2(t)$ using MF and MBF methods.

	SNR=−6 dB		SNR=−2 dB		SNR=1 dB		SNR=3 dB		SNR=5 dB		SNR=9 dB	
Signals	MBF	MF	MBF	MF	MBF	MF	MBF	MF	MBF	MF	MBF	MF
$s_1(t)$	300 ± 1	300 ± 1	300 ± 1	300	300	300	300	300	300	300	300	300
$s_2(t)$	300 ± 2	300 ± 2	300 ± 1	300 ± 1	300 ± 1	300 ± 1	300 ± 1	300	300 ± 1	300	300	300



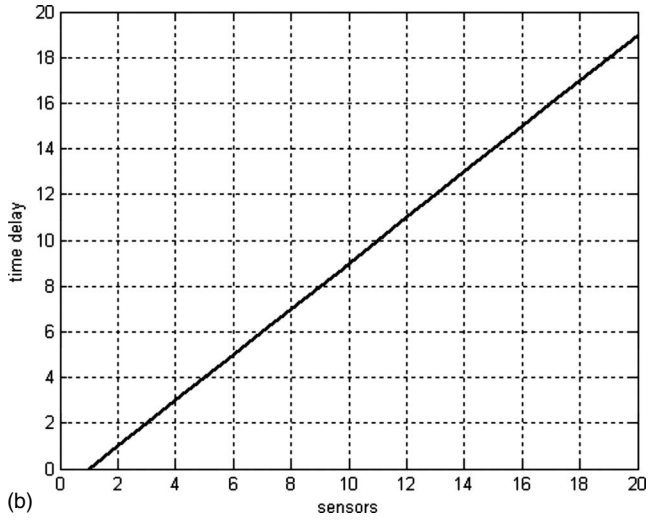


FIG. 6. TD estimation results (solid line: MBF method, dotted line: Theoretical method). (a) Time-sensors representation in synthetic case. (b) Estimated TD versus the sensor position indexes along the array in synthetic case.

erence signals, $s_1(t)$ and $s_2(t)$, with size window (in samples) w set to 65 and 81, respectively, are shown in Figs. 5(a) and 5(d). The received signals, $r_1(t)$ and $r_2(t)$ shown in Figs. 5(b) and 5(e), are obtained by time-shifting 300 samples, adding noise, and attenuating the reference signals. In this example, α is set to 0.95. Outputs of the MBF are shown in Figs. 5(c) and 5(f) indicating a net maximum at $t=T_B$. As expected, the peak of the function $I(t)$ is located at $T_B=300$. Table I lists the T_B values calculated for $s_1(t)$ and $s_2(t)$ with different signal-to-noise-ratios (SNRs) ranging from −6 to 9 dB, with α set to 0.7. Each value of Table I corresponds to the average of an ensemble of 25 trial TD estimates. These results show that the performance of the MBF is very close to that of the MF. Both methods point to the same theoretical value for T_B.

In this third example we consider the TD estimation in the case of a linear array composed of 20 uniformly spaced sensors. Each received signal sensor corresponds to the backscattered echo of a punctual target (acoustic source). Observed sensor output signals are shown in Fig. 6(a) as two-dimensional plots (time sensors). The delay estimation is performed between the first sensor of the array, taken as the

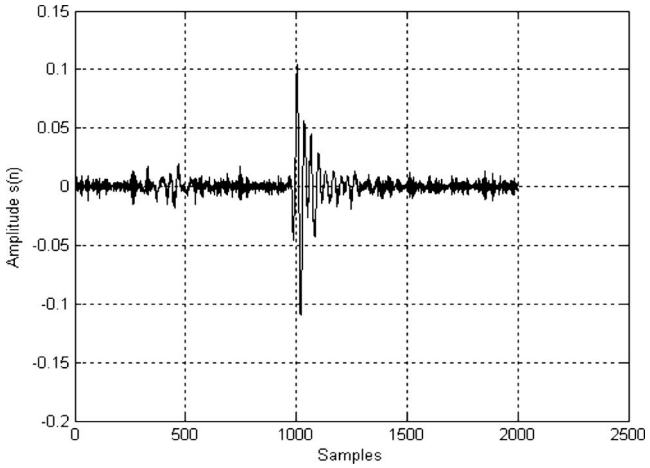


FIG. 7. Example of one buried target. The selected reference signal of Data2.

reference sensor, and each sensor of the array. The MBF delay estimation, of the received signal on two sensors, versus the sensor indexes along the linear array is shown in Fig. 6(b). Estimated delays, for all array sensors, are plotted as a function of the position indexes of the sensors along the array. Figure 6(b) shows a perfect agreement between theoretical estimation using Eq. (10) and that of the MBF.

B. Real signals

We demonstrated the performance of the Ψ_B operator on real data. These data are acoustic measurements conducted in a tank with a linear array of 20 sensors ($n=20$) where air-filled cylindrical objects are slightly buried under the sand bottom. Two cylinders were used for Data set 1 and one cylinder was used for Data set 2. The remaining parameter settings were $c=1485$ ms^{-1}, $d=2$ mm, $\theta_{Data1}=22°$, $\theta_{Data2}=64.15°$, frequency band [150 kHz,250 kHz] and sampling rate=2 MHz. Sensor outputs of signals (backscattered echoes) arriving from one (Data2) and two (Data1) cylinders are shown in Figs. 7 and 8, respectively. These signals correspond to the output of the first sensor of the array and are selected as reference signals. Both the MBF and the MF

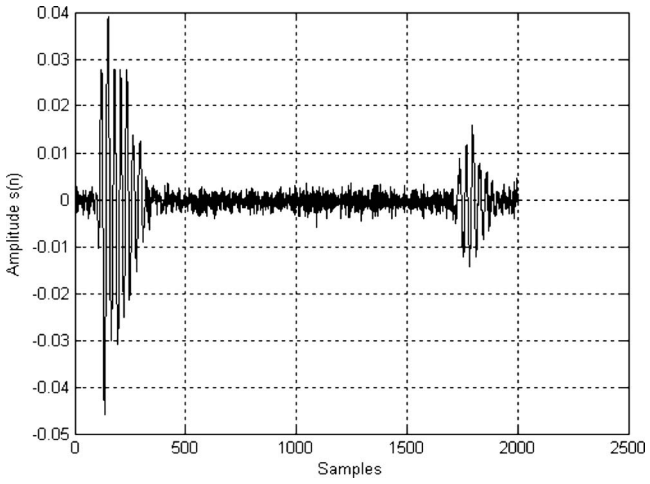


FIG. 8. Example of two buried targets. The selected reference signal of Data1.

TABLE II. RMSE between MBF, MF, and theoretical TD values for Datal and Data2.

Datal		Data2	
$\mathrm{RMSE}_{(\mathrm{MBF-Theor})}$	$\mathrm{RMSE}_{(\mathrm{MF-Theor})}$	$\mathrm{RMSE}_{(\mathrm{MBF-Theor})}$	$\mathrm{RMSE}_{(\mathrm{MF-Theor})}$
0.217	0.250	0.0526	0.180

methods, applied to the filtered signals, are implemented in the time domain. Signals are smoothed using a third-order Savitzky–Golay filter over a moving window of width set to 51.[9] TDs are estimated using the Theoretical [Eq. (10)], MBF, and the MF methods. In each case delay estimation is performed between the first sensor of the array and the remaining ones. Root mean square error (RMSE) between pair of sensors for Data1 and Data2 is reported in Table II:

$$\mathrm{RMSE}_{(\mathrm{A-Theor})} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{TD}_{\mathrm{A}}(i) - T_{\mathrm{Theor}}(i))^2}{n-1}} \tag{11}$$

C. Analysis

As shown in Fig. 9(b), for Data set 2 there is a perfect agreement (except for sensor 2 where the error is of one sample) between the MBF and the Theoretical method [Eq. (10)]. This is confirmed by the RMSE value (5.26%). The RMSE of the MF is 3.42 times higher than that of the MBF. Figure 9(a) shows that, for Datal, the TD estimated by the MBF and the MF deviate moderately from TD values given by Eq. (10). Note that the MBF performs slightly better than the MF with a RMSE of 21.7%. For both Data1 and Data2, globally, the MBF performs better than the MF. This may be due, even partially, to the nonlinear relationship between the signals that linear method such as the MF cannot account for. The mismatch between expected TD and MBF TD values may be due to the estimation error of the bearing angles θ_{Data1} and θ_{Data2} and to c, the sound speed in the water, which depends on the temperature. It is important to keep in mind that there is no odd way for estimating the TD value. The method based on Eq. (10) can be used, in the far field case, as a reference method if we have good measures of both θ and d.

VIII. CONCLUSION

In this paper, two methods, based on the Ψ_{B} operator, for signal detection and time-delay estimation [called Matched Ψ_{B} Filter (MBF)] are introduced. MBF measures how much one signal is present in another one. The discrete version of the continuous-time operator Ψ_{B}, which is used in its implementation, is presented. Preliminary results of signal detection and TD estimation and their comparison with matched filter and a reference method are presented. These signals show that Ψ_{B} is sensitive to nonlinear dependencies between signals compared to the classical CC method. For signal detection the MBF gives the same results as the MF for different SNRs. For TD estimation, the MBF results are very close to those of the reference method. The result for real acoustic signals shows that the MBF globally outperforms the MF. The processed signals are either noiseless or moderately noisy. For very noisy signals, the robustness of the MBF must be studied. As future work, we plan to use smooth splines to give more robustness to the MBF. To confirm the effectiveness of the MBF, the method must be evaluated with a large class of signals and in different experimental conditions including high noise levels and sampling rates and sample sizes. As future work, we plan to estimate TD values in the case of a nonuniform (distorted) array. We also plan to modify the proposed scheme to analyze situations in which signals and noises are mutually correlated.

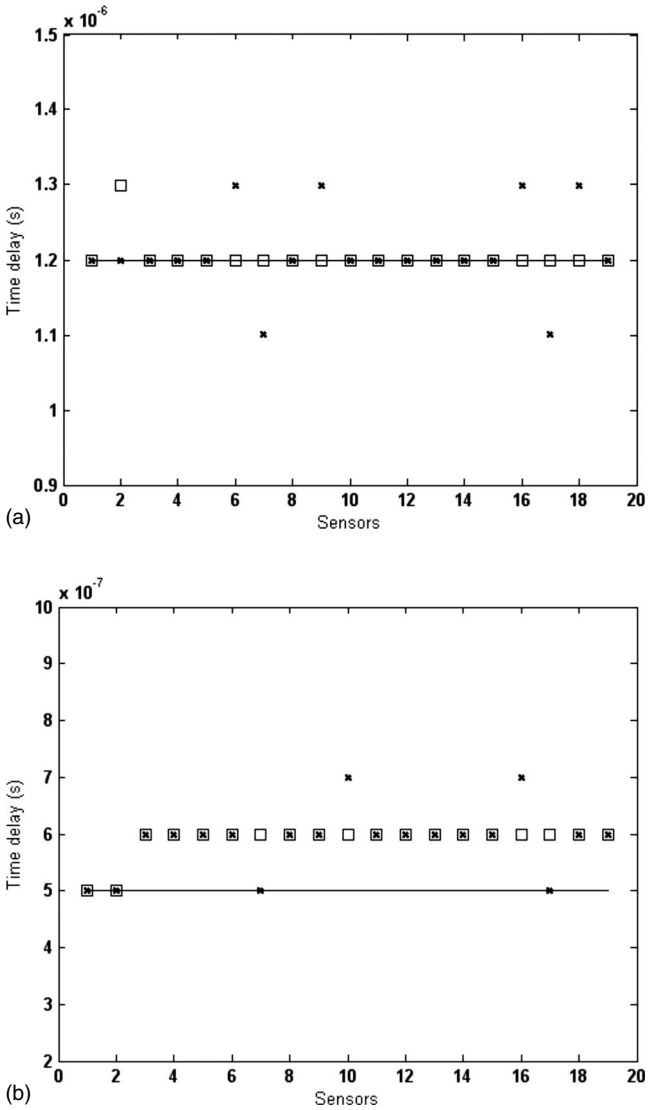


FIG. 9. TD estimation by theoretical, MBF, and MF methods. Estimated TD versus the sensors position indexes along the array in Datal (a) and Data2 (b), respectively (∗: MF, □: MBF, –: Theoretical methods).

ACKNOWLEDGMENTS

Authors would like to thank the Associate Editor and the anonymous reviewers for their helpful valuable comments and suggestions that helped to improve the paper. We would also like to thank Dr. J. P. Sessarego from CNRS-LMA, Marseille (France) and Dr. Z. Saidi from Ecole Navale, Brest (France) for making data available. Further, they express their gratitude to Dr. L. Guillon from Ecole Navale for helpful discussions about buried objects and multipath environments.

[1]S. Haykin, *Adaptive Filter Theory* (Prentice–Hall, Englewood Cliffs, NJ) (1996).

[2]J. Cexus and A. Boudraa, "Link between cross-Wigner distribution and cross-Teager energy," IEE Electronics Lett. **40**, 778–780 (2004).

[3]A. Boudraa, J. Cexus, K. Abed-Meraim, and Z. Saidi, "Interaction measure of AM-FM signals by cross-ψ_b-operator," in *Proc. IEEE ISSPA*, 775–778 (2005).

[4]A. Boudraa, J. Cexus, and H. Zaidi, "Functional segmentation of dynamic nuclear images by cross-ψ_b-operator," Comput. Methods Programs Biomed. **84**, 146–152 (2006).

[5]J. Kaiser, "Some useful properties of Teager's energy operators," in *Proc. ICASSP*, 149–152 (1993).

[6]P. Maragos and A. Potamianos, "Higher order differential energy operators," IEEE Signal Process. Lett. **2**, 152–154 (1995).

[7]However, we do not claim that MBF is an optimum detector of signal in white noise as the Matched Filter (MF). The main difference between the MBF and the MF is that the MF is linear measure while MBF is a nonlinear one.

[8]R. McDonough and A. Whalen, *Detection of Signals in Noise* (Academic, New york, 1995).

[9]A. Savitzky and M. Golay, "Smoothing and differentiation of data by simplified least squares procedures," Anal. Chem. **36**, 1627–1639 (1964).

A localization algorithm based on head-related transfer functions[a)]

Justin A. MacDonald[b)]
U.S. Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, Maryland 21005

(Received 27 September 2007; revised 22 March 2008; accepted 24 March 2008)

Two sound localization algorithms based on the head-related transfer function were developed. Each of them uses the interaural time delay, interaural level difference, and monaural spectral cues to estimate the location of a sound source. Given that most localization algorithms will be required to function in background noise, the localization performance of one of the algorithms was tested at signal-to-noise ratios (SNRs) from 40 to −40 dB. Stimuli included ten real-world, broadband sounds located at 5° intervals in azimuth and at 0° elevation. Both two- and four-microphone versions of the algorithm were implemented to localize sounds to 5° precision. The two-microphone version of the algorithm exhibited less than 2° mean localization error at SNRs of 20 dB and greater, and the four-microphone version committed approximately 1° mean error at SNRs of 10 dB or greater. Potential enhancements and applications of the algorithm are discussed.
 [DOI: 10.1121/1.2909566]

PACS number(s): 43.60.Jn, 43.66.Pn, 43.66.Qp [EJS] Pages: 4290–4296

I. INTRODUCTION

An ongoing project at the U.S. Army Research Laboratory has been to develop a biologically inspired algorithm to localize sounds in noisy environments in near real time. The motivations for this project are twofold: first, the algorithm could be implemented in autonomous robots or other unmanned vehicles to allow for accurate navigation and environment monitoring. The human listener provides an excellent example of an autonomous system that provides accurate location estimates in a wide variety of suboptimal environments. Second, a biologically inspired sound localization algorithm could be integrated into a computational auditory scene analysis (CASA) framework to segregate concurrent sounds based on the spatial locations of the sound sources. Location-based CASA approaches rely on localization algorithms to estimate sound source positions.

Machine-based sound localization systems take the input from two or more microphones to estimate the azimuth and elevation of a sound source. The localization algorithm must somehow extract location cues from the inputs to the sensors and determine the sound source location most likely to have produced the observed cues. Development of a new localization algorithm requires identification of the cues to be extracted as well as the decision process to be used to produce a location estimate. Which cues and decision processes are chosen depends on the goal of the algorithm designer. The human listener achieves accurate localization performance by using three types of location cues: the interaural time difference (ITD), the interaural level difference (ILD), and monaural spectral cues resulting from the irregular shape of the head and torso of the listener. Many sound localization algorithms utilize only the time delay cue to estimate the location of the sound source presumably because it is the easiest cue to extract from an incoming signal. For example, Calmes *et al.* (2007) constructed a neurally inspired model to detect ITDs to localize pure tones and wideband stimuli. The model performed well for wideband stimuli when the domain of potential locations was restricted to the front hemisphere. Viera and Almeida (2003) constructed a two-sensor system that localized sound sources between +60° and −60° azimuth to a precision of 9°. The restriction of possible locations to a single hemisphere is common to all localization algorithms that exclusively rely on the time delay between two microphones to estimate the location of the sound (e.g., Lotz *et al.*, 1989; Halupka *et al.*, 2005). The performance of these algorithms will suffer if sounds located in the rear hemisphere are included because any two-sensor system that exclusively relies on time delay cues will suffer from frequent front/back confusions. A time delay measured between the sensors will identify a subset of potential locations that lie on the surface of a cone extending outward from one of the sensors (Blauert, 1989). The subset of potential locations can be further reduced in two ways: by restricting the range of potential locations [the approach taken by Calmes *et al.* (2007), Viera and Almeida (2003), and others] or by using additional cues to estimate the source location. Chung *et al.* (2000) sought to resolve the ambiguity of the time delay cue by including monaural spectral cues in the localization algorithm. The model exhibited approximately 10° of localization error when broadband stimuli were presented from either hemisphere. Zakarauskas and Cynader (1993) developed an algorithm that compared the frequency spectrum of the incoming stimulus to the head-related transfer function (HRTF) (see Wightman and Kistler, 1989) in the frequency

[a)]Portions of this work were presented in "A sound localization algorithm for use in unmanned vehicles," Papers from the American Association for Artificial Intelligence Symposium on Aurally Informed Performance, Technical Report No. FS-01-01, Arlington, VA, October 2006.

[b)]Present address: New Mexico State University, P.O. Box 30001/MSC 3452, Las Cruces, NM 88003. Electronic mail: jmacd@nmsu.edu

domain, which is equivalent to using the ILD and monaural spectral cues to localize the sound. The mean error ranged from 0.29° to 25.4° depending on the stimulus being localized. Several other models have relied on ILD and monaural spectral cues to localize stimuli as well [Neti *et al.*, 1992; Middlebrooks, 1992; Chau and Duda, 1996]. Lim and Duda (1995) constructed a localization algorithm that utilized ITD, ILD, and monaural spectral cues to accurately estimate the location of an impulse source in an anechoic environment. The algorithm functioned by estimating the ITD and ILD from an incoming signal and comparing the estimates to a set of ITDs and ILDs from known source locations. The algorithm was able to perform quite accurately, exhibiting 0.8° of azimuth error in an anechoic environment.

The real-world performance of these algorithms is difficult to determine, however, given that nearly all of them were tested in a quiet environment. Most localization systems will be required to operate in a noisy environment, and the type of stimuli to be localized could considerably vary from the ideal. To this end, this paper details the design and subsequent testing of a biologically inspired sound localization algorithm that uses ITD, ILD, and monaural spectral cues to estimate the locations of real-world sounds. The human listener also takes advantage of other location cues during real-world sound localization tasks, including cues based on head movement, knowledge of the environment, or previous exposure to the stimulus being localized. Ideally, these location cues would have been included in the algorithm as well. Given the difficulty in extracting these cues, however, the localization algorithm was constructed to take advantage of only those cues available to a stationary, naive listener. The performance of the algorithm was measured across a range of signal-to-noise ratios (SNRs) to estimate the performance under suboptimal conditions.

The algorithm was designed to function using two or more microphones mounted in nearly arbitrary locations. The number of microphones included in the array depends on the application of the algorithm. If the algorithm is to be part of a system in which biological plausibility is required (in CASA applications or human localization modeling, for example), only two microphones are appropriate. If one is not subject to this restriction, however, then the number of microphones included in the array will be determined by the accuracy required and the computational resources that are available. Increasing the number of microphones from two to four, for example, is likely to improve the performance by reducing the number of front/back confusions exhibited by the localization system. This will be accomplished at the cost of increased computational requirements of the algorithm. Given that one of the eventual goals of this effort is to develop a location-based approach to CASA, the microphones were mounted to the Knowles electronics mannequin for acoustic research (KEMAR). The KEMAR is a human model that mimics the effects of the head and torso on an incoming sound wave. Both two- and four-microphone implementations of the algorithm were tested to determine the increased performance gained when another pair of microphones is added to the array.

II. LOCALIZATION ALGORITHMS

Consider an array of m microphones mounted at arbitrary locations whose center is at point P. Imagine a sound that originates from azimuth θ and elevation ϕ relative to P. The task of any localization algorithm is to process each of the m microphone inputs $\{I_1, \ldots, I_m\}$ to generate azimuth and elevation estimates $\hat{\theta}$ and $\hat{\phi}$, respectively. Ideally, the algorithm should utilize all available location cues to maximize accuracy. Differences in times of arrival between the microphones will vary with the location of the sound source and can therefore be utilized to generate location estimates. Additional location cues are available if the frequency content of the microphone inputs varies with the location of the sound source. This can be achieved by inserting an object centered at P into the listening environment so that the filtering properties of the object will vary with the orientation of the sound source.

For illustrative purposes, consider the situation in which $m=2$ microphones are mounted at the opening to each ear canal of a KEMAR. Let the center of the head of the KEMAR be located at P. Consider a sound that originates at azimuth θ and elevation ϕ relative to P. The sound is altered by the head and torso of the KEMAR before it arrives at the microphones. If I_j is a digital recording of the input to the jth microphone, then

$$I_j = O * F_j^{(\theta,\phi)}, \tag{1}$$

where O is the sound that would arrive at point P if the KEMAR were absent, $*$ is the convolution operator, and $F_j^{(\theta,\phi)}$ is the head-related impulse response (HRIR) for microphone j when a sound originates from (θ,ϕ). The HRIR is a representation of the HRTF in the time domain rather than the frequency domain and can therefore include both the time- and frequency-based filtering effects of the head and torso.

Consider the result when I_1 is convolved with $[F_1^{(\theta,\phi)}]^{-1}$, which is the inverse of the HRIR associated with (θ,ϕ) at microphone 1. In this case,

$$I_1 * [F_1^{(\theta,\phi)}]^{-1} = (O * F_1^{(\theta,\phi)}) * [F_1^{(\theta,\phi)}]^{-1} = O \tag{2}$$

due to the associativity of the convolution operator. In other words, if the effects of the head and torso of the KEMAR are removed from the recordings, the stimulus that would have arrived at P if the KEMAR was absent is the result. Similarly,

$$I_2 * [F_2^{(\theta,\phi)}]^{-1} = (O * F_2^{(\theta,\phi)}) * [F_2^{(\theta,\phi)}]^{-1} = O. \tag{3}$$

In both cases, if the inverse of the HRIR associated with the actual location of the sound source is chosen, then the original unaltered stimulus is the result. However, if the inverse of the HRIR associated with some other location (θ',ϕ') is convolved with the microphone inputs, then

$$I_1 * [F_1^{(\theta',\phi')}]^{-1} = (O * F_1^{(\theta,\phi)}) * [F_1^{(\theta',\phi')}]^{-1} \tag{4}$$

and

$$I_2 * [F_2^{(\theta',\phi')}]^{-1} = (O * F_2^{(\theta,\phi)}) * [F_2^{(\theta',\phi')}]^{-1}. \quad (5)$$

In this case, the convolution does not lead to the same result for I_1 and I_2. This suggests a method for determining the location of the sound source (θ, ϕ) from the microphone inputs I_1 and I_2: choose $(\hat{\theta}, \hat{\phi})$ to maximize the similarity between $I_1 * [F_1^{(\hat{\theta},\hat{\phi})}]^{-1}$ and $I_2 * [F_2^{(\hat{\theta},\hat{\phi})}]^{-1}$. Of course, a wide variety of similarity metrics are available; a moderate amount of testing suggested that the Pearson correlation maximized the accuracy and reliability of the "inverse" localization algorithm. Formally, the inverse algorithm chooses $(\hat{\theta}, \hat{\phi})$ according to the following equation:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_1 * [F_1^{(\hat{\theta},\hat{\phi})}]^{-1}, I_2 * [F_2^{(\hat{\theta},\hat{\phi})}]^{-1}). \quad (6)$$

Another localization algorithm that does not require inverse filters was also developed. Continuing the example of two microphones mounted at the openings of the ear canals of the KEMAR, each input is convolved with the HRIR associated with the opposite microphone. If the HRIR associated with the correct location (θ, ϕ) is used, then

$$I_1 * F_2^{(\theta,\phi)} = (O * F_1^{(\theta,\phi)}) * F_2^{(\theta,\phi)} = O * F_1^{(\theta,\phi)} * F_2^{(\theta,\phi)} \quad (7)$$

and

$$I_2 * F_1^{(\theta,\phi)} = (O * F_2^{(\theta,\phi)}) * F_1^{(\theta,\phi)} = O * F_2^{(\theta,\phi)} * F_1^{(\theta,\phi)} = O * F_1^{(\theta,\phi)} * F_2^{(\theta,\phi)}. \quad (8)$$

This follows from the commutativity and associativity of the convolution operator. As with the inverse algorithm, if the correct location is chosen, then the operation will lead to the same result for both microphone inputs. If the HRIR associated with some other location (θ', ϕ') is chosen, however, then the results will differ:

$$I_1 * F_2^{(\theta',\phi')} = (O * F_1^{(\theta,\phi)}) * F_2^{(\theta',\phi')} = O * F_1^{(\theta,\phi)} * F_2^{(\theta',\phi')} \quad (9)$$

and

$$I_2 * F_1^{(\theta',\phi')} = (O * F_2^{(\theta,\phi)}) * F_1^{(\theta',\phi')} = O * F_2^{(\theta,\phi)} * F_1^{(\theta',\phi')} = O * F_1^{(\theta',\phi')} * F_2^{(\theta,\phi)}. \quad (10)$$

As before, the "cross-channel" algorithm uses the Pearson correlation coefficient as the similarity metric, choosing $(\hat{\theta}, \hat{\phi})$ as follows:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_1 * F_2^{(\hat{\theta},\hat{\phi})}, I_2 * F_1^{(\hat{\theta},\hat{\phi})}). \quad (11)$$

The generalization of these algorithms to more than two microphones is relatively straightforward. A microphone array with $2N$ microphones is arbitrarily partitioned into N microphone pairs. Let the first and second microphones in the kth pair be denoted by k_1 and k_2, respectively. The associated microphone inputs will be denoted by I_{k_1} and I_{k_2} by using this notation. For the inverse algorithm, the first microphone input in each pair is convolved with the associated inverse impulse response, as in Eq. (2). The results of the N convolutions are then concatenated:

$$I_{1_1} * [F_{1_1}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_1} * [F_{N_1}^{(\hat{\theta},\hat{\phi})}]^{-1}, \quad (12)$$

where & is the concatenation operator. Similarly, the second microphone input in each pair is convolved with the associated inverse impulse response, and the results of the N convolutions are concatenated:

$$I_{1_2} * [F_{1_2}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_2} * [F_{N_2}^{(\hat{\theta},\hat{\phi})}]^{-1}. \quad (13)$$

The concatenated results are then correlated to determine $(\hat{\theta}, \hat{\phi})$:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_{1_1} * [F_{1_1}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_1} * [F_{N_1}^{(\hat{\theta},\hat{\phi})}]^{-1},$$

$$I_{1_2} * [F_{1_2}^{(\hat{\theta},\hat{\phi})}]^{-1} \,\&\, \cdots \,\&\, I_{N_2} * [F_{N_2}^{(\hat{\theta},\hat{\phi})}]^{-1}). \quad (14)$$

Note that Eq. (14) simplifies to Eq. (6) when only one microphone pair is used. The multichannel implementation of the cross-channel algorithm is structured in a similar fashion. In this case, $(\hat{\theta}, \hat{\phi})$ is chosen as follows:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_{1_1} * F_{1_2}^{(\hat{\theta},\hat{\phi})} \,\&\, \cdots \,\&\, I_{N_1} * F_{N_2}^{(\hat{\theta},\hat{\phi})},$$

$$I_{1_2} * F_{1_1}^{(\hat{\theta},\hat{\phi})} \,\&\, \cdots \,\&\, I_{N_2} * F_{N_1}^{(\hat{\theta},\hat{\phi})}). \quad (15)$$

Equation (15) reduces to Eq. (11) when only one pair of microphones is used.

Of the two algorithms presented here, it is likely that the cross-channel algorithm will be preferred for most applications. The inverse algorithm is likely to require greater computational resources than the cross-channel algorithm. Appropriate inverse filters that account for the magnitude and phase portions of the HRIR are typically of greater complexity than the original filter. For example, by using the method detailed by Greenfield and Hawksford (1991), an inverse filter that accounts for both magnitude and phase response will be approximately three times longer than the original HRIR. Considering that these inverted HRIRs are convolved with the microphone inputs, the computational requirements of the inverse algorithm will be substantial. In addition, methods to compute inverse filters produce filters that are only an approximate inverse of the original (Rife and Vanderkooy, 1989). Because the accuracy of the inverse filter increases with its length, one must consider the trade-off between the accuracy of the inverse filter and computational requirements of the algorithm. Fortunately, the cross-channel algorithm does not suffer from these drawbacks: inverse filters are not required. For this reason, the cross-channel algorithm was chosen for further testing.

Our initial test of the cross-channel algorithm examined the performance of a two-microphone implementation (MacDonald, 2005). Real-world, broadband sounds were recorded at 5° intervals around the head of the KEMAR. Noise was added to each recording to obtain SNRs from 40 to −40 dB, and the cross-channel algorithm estimated the location of the sound source from the noisy recordings. The algorithm performed well beyond expectations: the localization error in quiet was measured at 2.9° using only two

microphones, and above-chance performance was observed at greater than or equal to −10 dB SNRs. Front/back confusions occurred in approximately 5% of the trials at the higher SNRs.

These promising initial results prompted a larger-scale test using both two- and four-microphone versions of the algorithm. Accordingly, two additional microphones were mounted on the front and rear of the head of the KEMAR. The additional microphones should allow for a reduced number of front/back confusions and an increased localization accuracy at the expense of an increased computation time.

III. SIMULATION METHOD

A. Stimuli

Ten naturally occurring sounds were chosen as the test signals: the sounds of breaking glass, a speech stimulus, the insertion of an M-16 magazine, a camera shutter release sound, machine gun fire, a cough, a dog bark, a door being slammed, a water dripping noise, and the sound of a heavy object being dropped into a body of water. Sounds ranged from 400 to 600 ms in duration and were stored in a 16 bit Microsoft WAV format with a sampling rate of 44.1 kHz.

B. Stimulus recording apparatus

Stimuli were presented using the Army Research Laboratory Human Research and Engineering Directorate's RoboArm 360 system. This system consists of a speaker attached to a computer-controlled robotic arm. The stimuli were output through a Tucker–Davis Technologies (TDT) System II DD1 digital to analog converter, which was amplified using a TDT System 3 SA1 amplifier, and presented from a GF0876 loudspeaker (CUI, Inc.) at the end of the robotic arm. Stimuli were presented at approximately 75 dB (A) measured 1 meter from the loudspeaker. The arm positioned the loudspeaker at 5° intervals around the KEMAR (a total of 72 positions). The loudspeaker was located 1 m from the center of the head of the KEMAR and at 0° elevation for all stimulus presentations.

Two EM-125 miniature electret microphones (Primo Microphones, Inc.) were used to record the stimulus presentations. Recordings were made in two sessions. In the first, the pair of microphones was mounted in foam inserts at the entrance of the ear canals of the KEMAR. In the second, the microphones were placed at the front and rear of the head of the KEMAR. The front microphone was attached to the center of the forehead just above the bridge of the nose, and the rear microphone was attached at the same elevation at the rearmost part of the head. Inputs to the microphones were amplified by a TDT System 3 MA3 microphone amplifier before being sent to a TDT System II DD1 analog to digital converter. The digital output of the DD1 was sent to a computer for storage in a 44.1 kHz, 16 bit Microsoft WAV format. By combining across recording sessions, a total of 720 four-channel recordings were made, one for each position/sound combination.

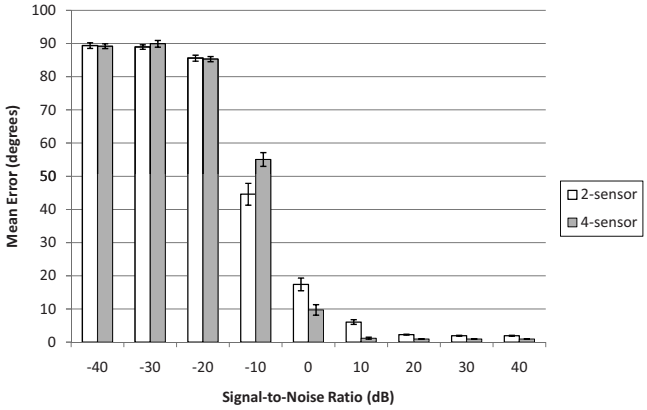


FIG. 1. Mean localization error in each SNR condition. The bars indicate the standard error associated with each mean. The location estimates were not corrected for front/back confusions. Chance performance in this localization task corresponds to a 90° mean error.

C. HRIR measurement

The HRIR of the KEMAR was measured using the same presentation and recording apparatus detailed above. The maximum-length sequence (see Rife and Vanderkooy, 1989) stimuli were presented at 5° intervals around the head of the KEMAR and the signals recorded at the microphones determined the HRIR of the KEMAR at each location. As with the stimulus recordings, the front/back and left/right impulse responses were separately estimated. Each HRIR was stored as a 256-tap finite impulse response digital filter.

D. Procedure

Simulations were conducted using a script written in MATLAB (The Mathworks, Natick, MA) to estimate the performance of both the two- and four-sensor versions of the cross-channel algorithm. In the two-sensor simulation, the algorithm utilized the HRIRs associated with the left and right microphones to process the recordings made at those locations, and estimates were produced using Eq. (11). The four-sensor simulation used Eq. (15) to apply the four-channel HRIRs to the four-channel recordings. Locations were estimated with 5° precision in both simulations. A random sample of Gaussian noise was added to each channel of each recording to obtain SNRs ranging from 40 to −40 dB in 10 dB increments. The SNR for each trial was calculated based on the signal channel with the greatest root-mean-squared amplitude. The algorithm was required to localize each recording ten times; a different sample of Gaussian noise was added on each attempt. This resulted in a total of 64 800 localization attempts for each of the simulations (9 SNRs × 720 recordings × 10 trials each).

IV. RESULTS

All location estimates in the simulation were left uncorrected: estimates were not reflected across the interaural axis when a front/back confusion occurred. The absolute error (the absolute value of the angular distance between the estimated and actual sound locations, in degrees) was used as the error measure. The mean error observed at each SNR (collapsed across the ten sound stimuli) is shown in Fig. 1.

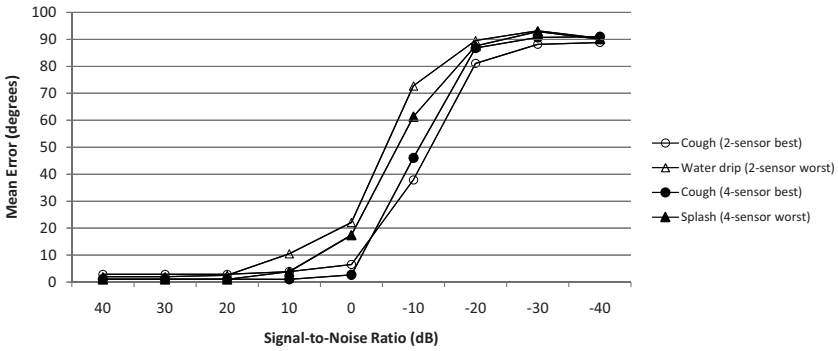


FIG. 2. Errors associated with the best- and worst-localized sounds in each SNR condition. The performance of both algorithms tended to increase with the bandwidth of the stimulus.

The two-microphone implementation exhibited approximately 2° localization error when the SNR was greater than 10 dB and performed well above chance levels to −20 dB. The four-microphone implementation exhibited an even greater accuracy, maintaining a mean error of approximately 1° in SNRs of 10 dB and greater and performing above chance to −20 dB. The performance of the algorithm varied somewhat across stimuli; the error bars in the figure indicate the standard error of the mean calculated across the stimulus set. The effect of the stimulus on the performance of the algorithm is illustrated in greater detail in Fig. 2: the errors associated with the best- and worst-localized sounds are shown for the two- and four-sensor versions of the algorithm. In general, performance increased with the bandwidth of the stimulus.

The mean localization error observed in the 10 dB SNR condition at each location is shown is Fig. 3. The 10 dB condition was chosen so that a sufficient number of errors could be included in the figure. The two-sensor implementation of the algorithm exhibited systematic errors at higher noise levels when sounds were located just behind the interaural axis. The large majority of these errors were back-to-front confusions. There is a slight asymmetry in the two-sensor error pattern; this is likely due to the acoustic properties of the room in which the sound recordings were made. The four-sensor implementation was much less susceptible to these errors, illustrating the benefits of including an additional two sensors in the array.

The proportion of front/back confusions in each SNR condition is shown in Fig. 4. A front/back confusion occurred when the estimated and actual locations of the sound source were on opposite sides of the interaural axis. Two-microphone systems that exclusively rely on time-of-arrival differences will exhibit a 50% confusion rate. The inclusion of the ILD and monaural cues in the cross-channel algorithm led to a significant reduction in the number of confusions: fewer than 5% of the trials resulted in confusions in the 40, 30, and 20 dB SNR conditions, and performance was well above chance to −20 dB. As expected, the addition of two microphones in the four-sensor implementation led to an increased performance: confusions were reduced to a trivial level (0.28%) at 10 dB and were entirely eliminated in the 20, 30, and 40 dB conditions.

V. DISCUSSION

These simulations demonstrate the extremely high accuracy that can be achieved with the cross-channel algorithm. The two-microphone implementation exhibited a mean localization error of less than 2° despite the addition of a moderate amount of Gaussian noise. The accuracy of the algorithm is especially impressive considering that sounds were allowed to originate from the rear hemisphere, thereby allow-

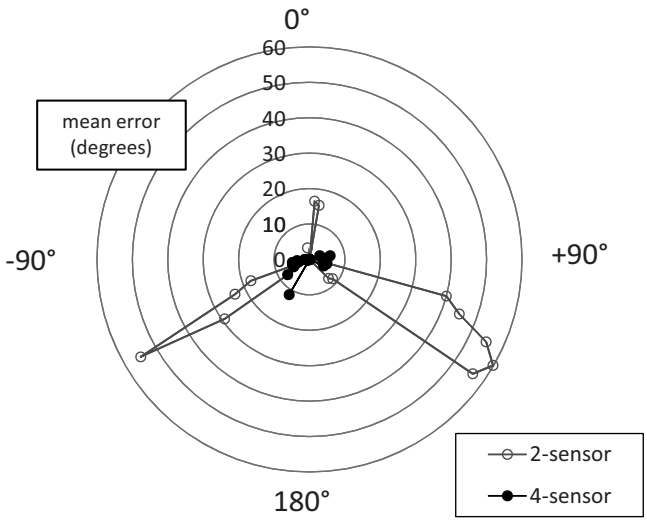


FIG. 3. Mean error for each location in the 10 dB SNR condition. The two-sensor version of the algorithm exhibited frequent back-to-front confusions at moderate and high noise levels for sounds located just behind the interaural axis. These errors were not observed in the four-sensor version of the algorithm.

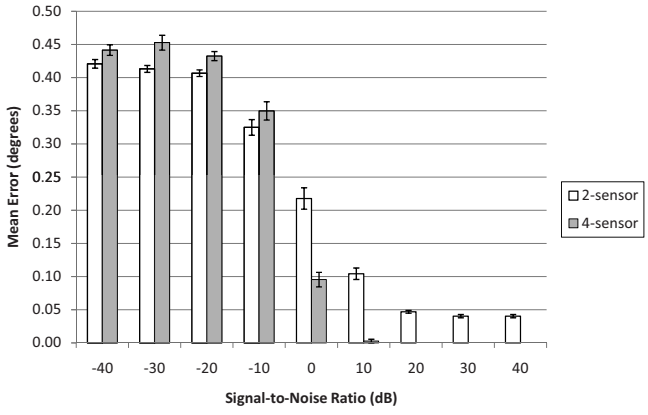


FIG. 4. Proportion of front/back confusions in each SNR condition. A confusion occurred when the estimated and actual source locations were on opposite sides of the interaural axis. Chance performance corresponds to a 50% confusion rate. No confusions (out of 21 600 trials) were observed for the four-sensor implementation of the algorithm at SNRs of 40, 30, and 20 dB.

ing for the possibility of front/back confusions. The inclusion of frequency-based location cues allowed for a severe reduction in the number of front/back confusions. As expected, the four-microphone implementation of the algorithm exhibited an even better performance, committing almost no reversals in all SNRs greater than 0 dB.

It is worth noting that the performance of the algorithm depends on the transfer function of the structure to which the microphones are mounted. Asymmetrical structures with maximally separated microphone mounting points should possess transfer functions that exhibit considerable variance across sound source locations, thereby increasing the performance of the algorithm. The KEMAR is likely to be a suboptimal choice in this regard: it is relatively symmetric with respect to both the medial and interaural axes and there is only a short distance between the mounting points for the microphones. Despite this handicap, the KEMAR-based implementation compares favorably to the large majority of other localization algorithms, exhibiting a mean localization error of 1.9 degrees in the 40 dB SNR condition when localizing ten different real-world sounds. In comparison, Berdugo *et al.* (1999) reported errors of approximately five degrees in quiet when using an array of seven microphones to localize a 20 s speech signal. Viera and Almeida (2003) reported a mean localization error of approximately nine degrees when source locations were restricted to the front hemisphere. Schauer and Gross (2001) observed a mean localization error of approximately ten degrees under the same restriction. Zakarauskas and Cynader (1993) measured localization errors between 0.29 and 25.4 degrees depending on the stimulus being localized. Neti *et al.* (1992) reported a mean localization error of 6.3° when source locations were restricted to the range between −30° and +30°. The strongest performance was reported by Lim and Duda, who observed a 0.8° mean error in azimuth when localizing an ideal broadband stimulus (an impulse) in anechoic conditions. The performance of the algorithm in suboptimal (noisy) conditions was not reported.

An analysis of the results of the two-microphone implementation of the algorithm can provide insight into the performance of the human listener. By assuming that the KEMAR is an accurate model of the human head and torso, the results of the simulation indicate that a highly accurate localization performance is possible using information that is available at the entrance to the ear canal. Accurate localization is quite possible in noisy environments without previous exposure to the stimulus. The inferior performance of the human listener in these conditions must arise from the following: either the information available at the ear canal is not available in the central nervous system where the location estimate is made, or the decision process used to produce the location estimate is suboptimal, or (most likely) both. The former possibility can be partially tested by filtering the recording through a model of the auditory periphery and by using the cross-channel algorithm to localize sounds based on the output of the model.

It is clear that several questions remain to be answered about the performance of the algorithm. As with all localization algorithms, performance is likely to decrease in reverberant environments. In addition, the performance of the algorithm is unknown when the elevation of the sound source is allowed to vary. It seems likely that the accuracy of elevation judgments would improve with the four-microphone version of the algorithm, but that remains to be investigated. An examination of the accuracy of the cross-channel algorithm across elevations is currently underway. In addition, localization accuracy in a multisound environment must be investigated especially if the localization algorithm is to be integrated into a CASA algorithm.

Both the inverse and cross-channel algorithms could be altered in a variety of ways to determine if the accuracy of the algorithm can be improved. For example, the Pearson correlation is only one of the many possible similarity metrics that could be used in the inverse and cross-channel algorithms. Several other metrics were considered during the initial testing of the algorithms, including using the sum of the squared deviations rather than the Pearson correlation. Of the metrics considered, however, the Pearson correlation led to the best localization performance in an initial test and was therefore chosen for use in the subsequent full-scale evaluation. In addition, the computational requirements of the algorithm could be reduced using shortcuts to eliminate potential source locations. In a quiet environment, for example, all locations in the right hemisphere could be eliminated as potential source locations if the system determined that the sound arrived at the left microphone before the right. Many possible compromises between the two- and four-microphone algorithm implementations are worth investigating as well. For example, the algorithm could use the left and right channels to generate location estimates that are modified based on the relative intensity of the input to the front and back microphones. In addition, the locations of the microphones were somewhat arbitrarily chosen; it is quite possible that other locations will lead to better performance. Finally, it is likely that other mounting structures could be found that introduce greater variation in the HRTF across sound source locations, thereby increasing the performance of the algorithm. Refinements such as these will be explored in future work.

ACKNOWLEDGMENTS

The author wishes to thank Phuong Tran for her help in making the four-channel recordings used in the testing of the algorithm.

Berdugo, B., Doron, M. A., Rosenhouse, J., and Azhari, H. (**1999**). "On direction finding of an emitting source from time delays," J. Acoust. Soc. Am. **105**, 3355–3363.

Blauert, J. (**1989**). *Spatial Hearing* (MIT Press, Cambridge, MA).

Calmes, L., Lakemeyer, G., and Wagner, H. (**2007**). "Azimuthal sound localization using coincidence of timing across frequency on a robotic platform," J. Acoust. Soc. Am. **121**, 2034–2048.

Chau, W., and Duda, R. O. (**1996**). "Combined monaural and binaural localization of sound sources," IEEE Proceedings of the 29th Asilomar Conference on Signals, Systems, and Computers, pp. 1281–1285.

Chung, W., Carlile, S., and Leong, P. (**2000**). "A performance adequate computational model for auditory localization," J. Acoust. Soc. Am. **107**, 432–445.

Greenfield, R., and Hawksford, M. O. (**1991**). "Efficient filter design for loudspeaker equalization," J. Audio Eng. Soc. **39**, 739–751.

Halupka, D., Mathai, N. J., Aarabi, P., and Sheikholeslami, A. (**2005**). "Ro-

bust sound localization in 0.18 μm CMOS," IEEE Trans. Signal Process. **53**, 2243–2250.
Lim, C., and Duda, R. O. (**1995**). "Estimating the azimuth and elevation of a sound source from the output of a cochlear model," IEEE Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers, pp. 399–403.
Lotz, K., Bölöni, L., Roska, T., and Hámori, J. (**1999**). "Hyperacuity in time: A CNN model of a time-coding pathway of sound localization," IEEE Trans. Circuits Syst., I: Fundam. Theory Appl. **46**, 994–1002.
MacDonald, J. A. (**2005**). "An algorithm for the accurate localization of sounds," Proceedings of the NATO HFM-123 Symposium on New Directions for Improving Audio Effectiveness, Paper no. 28. Available: http://www.rta.nato.int/pubs/rdp.asp?RDP=RTO-MP-HFM-123.
Middlebrooks, J. C. (**1992**). "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**, 2607–2624.
Neti, C., Young, E. D., and Schenider, M. H. (**1992**). "Neural network models of sound localization based on directional filtering by the pinna," J. Acoust. Soc. Am. **92**, 3140–3156.
Rife, D. D., and Vanderkooy, J. (**1989**). "Transfer-function measurement with maximum-length sequences," J. Audio Eng. Soc. **37**, 419–444.
Schauer, C., and Gross, H. M. (**2001**). "Model and application of a binaural 360° sound localization system," IEEE Proceedings of the International Joint Conference on Neural Networks, Vol. **2**, pp. 1132–1137.
Viera, J., and Almeida, L. (**2003**). "A sound localizer robust to reverberation," Proceedings of the 115th Convention of the Audio Engineers Society, Paper No. 5973.
Wightman, F. L., and Kistler, D. J. (**1989**). "Headphone simulation of free-field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867.
Zakarauskas, P., and Cynader, M. S. (**1993**). "A computational theory of spectral cue localization," J. Acoust. Soc. Am. **94**, 1323–1331.

The acoustical cues to sound location in the rat: Measurements of directional transfer functions

Kanthaiah Koka
Department of Physiology and Biophysics, University of Colorado Health Sciences Center, Stop 8307, P.O. Box 6511, Aurora, Colorado 80045

Heather L. Read
Department of Psychology, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, Connecticut 06269-1020

Daniel J. Tollin[a)]
Department of Physiology and Biophysics, University of Colorado Health Sciences Center, Stop 8307, P.O. Box 6511, Aurora, Colorado 80045

(Received 15 February 2008; revised 3 April 2008; accepted 4 April 2008)

The acoustical cues for sound location are generated by spatial- and frequency-dependent filtering of propagating sound waves by the head and external ears. Although rats have been a common model system for anatomy, physiology, and psychophysics of localization, there have been few studies of the acoustical cues available to rats. Here, directional transfer functions (DTFs), the directional components of the head-related transfer functions, were measured in six adult rats. The cues to location were computed from the DTFs. In the frontal hemisphere, spectral notches were present for frequencies from ~16 to 30 kHz; in general, the frequency corresponding to the notch increased with increases in source elevation and in azimuth toward the ipsilateral ear. The maximum high-frequency envelope-based interaural time differences (ITDs) were 130 μs, whereas low-frequency (<3.5 kHz) fine-structure ITDs were 160 μs; both types of ITDs were larger than predicted from spherical head models. Interaural level differences (ILDs) strongly depended on location and frequency. Maximum ILDs were <10 dB for frequencies <8 kHz and were as large as 20–40 dB for frequencies >20 kHz. Removal of the pinna eliminated the spectral notches, reduced the acoustic gain and ILDs, altered the acoustical axis, and reduced the ITDs.
 [DOI: 10.1121/1.2916587]

PACS number(s): 43.64.Ha, 43.66.Pn, 43.66.Qp [JCM] Pages: 4297–4309

I. INTRODUCTION

Whether predator or prey, sound localization is critical to the survival of most animals. The acoustical cues for sound source location are generated by the spatial- and frequency-dependent filtering of the propagating sound waves by the head and external ears (Figs. 1 and 2). Therefore, the linear dimensions of the head and pinnae are critical factors in determining the magnitude and frequency ranges of the resultant cues to location. There are three main cues: interaural differences in time (ITD) and level (ILD) and monaural spectral shape cues (Fig. 2).

Behavioral, anatomical, and physiological studies of the mechanisms of sound localization have suggested that there can be considerable differences (Irving and Harrison, 1967; Kelly, 1980; Heffner, 1997) and similarities (Brand *et al.*, 2002; McAlpine and Grothe, 2003) among different species making cross species generalizations difficult and perhaps even misleading. These differences might be reconciled by taking into consideration the sound localization cues that are actually available for each particular species. In other words, proper interpretation of the data from anatomical, physiological, and behavioral studies of sound localization requires detailed knowledge of the properties of acoustical information available to the species under study. Unfortunately, for some common species, such as the rat, the cues to location have not been measured. Yet in other species, there has been considerable study of the cues, including human (Wightman and Kistler, 1989; Middlebrooks *et al.*, 1989; Middlebrooks and Green, 1990), cat (Wiener *et al.*, 1966; Moore and Irvine, 1979; Roth *et al.*, 1980; Phillips *et al.*, 1982; Irvine, 1987; Musicant *et al.*, 1990; Rice *et al.*, 1992; Xu and Middlebrooks, 2000), monkey (Spezio *et al.*, 2000), ferrets (Carlile, 1990; Schnupp *et al.*, 2003), tammar wallaby (Coles and Guppy, 1986), various species of bat (Jen and Chen, 1988; Obrist *et al.*, 1993; Fuzessery, 1996; Firzlaff and Schuller, 2003; Aytekin *et al.*, 2004), guinea pig (Carlile and Pettigrew, 1987; Sterbing *et al.*, 2003), gerbil (Maki and Furukawa, 2005), mouse (Chen *et al.*, 1995), and barn owl (Moiseff, 1989; Keller *et al.*, 1998).

In this paper, we extend these studies to include the albino rat (*Rattus norvegicus*). Rats have been a common model system for the behavior (Beecher and Harrison, 1971; Burlile *et al.*, 1985; Heffner and Heffner, 1985; Harrison, 1988; Heffner *et al.*, 1994), anatomy (Beyerl, 1978; Kelly, 1980; Kelly and Kavanagh, 1986; Kelly and Glazier, 1978), physiology (Flammino and Clopton, 1975; Inbody and Feng,

[a)]Electronic mail: daniel.tollin@uchsc.edu.

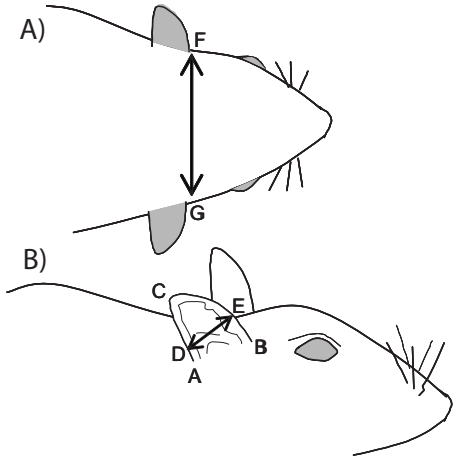


FIG. 1. The dimensions of the animal subjects. Across the six rats, the average head diameter (FG) was 29.6 ± 1.3 mm, the average pinna width (DE) was 11.4 ± 0.5 mm, and the average pinna lengths (AC) and (BC) were 16.8 ± 0.8 and 17.7 ± 1.7 mm, respectively.

1981; Kelly and Sally, 1988; Finlayson and Caspary, 1991; Kelly and Phillips, 1991; Li and Kelly, 1992; Irvine *et al.*, 1995; Kelly *et al.*, 1998; Irvine *et al.*, 2001), and development (Silverman and Clopton, 1977; Clopton and Silverman, 1977; Kelly and Judge, 1985; Kandler and Gillespie, 2005) of sound localization mechanisms. However, aside from some spatially and spectrally sparse measurements of the ILD cues by Harrison and Downey (1970), little is actually known about the localization cues available to rats. To fill this void, here we measured the directional transfer functions (DTFs) (Middlebrooks and Green, 1990), the directional components of the head-related transfer functions (HRTFs), for adult rats. From the DTFs, the primary acoustical cues to location were computed and examined, including the spectral "notches" (Rice *et al.*, 1992), ITDs, and the spatial and frequency dependence of the ILD cues. The acoustic gains of the DTFs were examined at individual frequencies to determine the "acoustical axis," the spatial location of the maximum gain at a particular frequency. Finally, the role of the pinna in generating the cues was examined by repeating the acoustical measurements after surgical removal of the pinna. Our measurements confirm that the three typical acoustical cues that are available to most mammals are also available in rats. Moreover, we demonstrate that the pinna substantially contributes to the formation of all three of the primary cues to location.

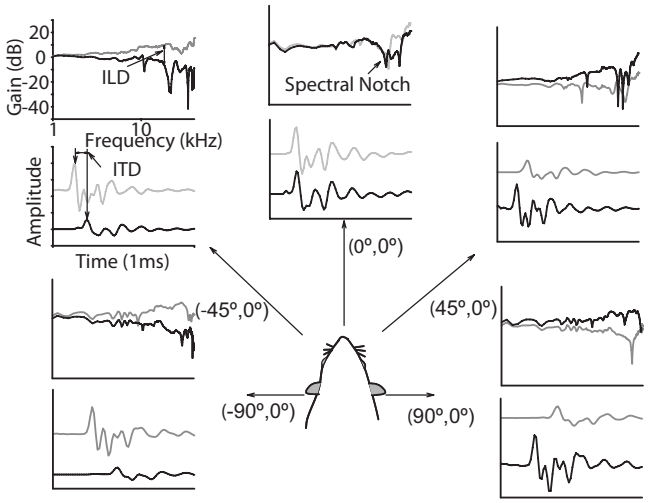


FIG. 2. Examples of the three primary acoustical cues to sound location, spectral notches, interaural time differences (ITDs), and interaural level differences (ILDs), for five sources in azimuth along the horizontal plane. At each location, the directional transfer functions (DTFs, top panels) and impulse responses (lower panels) are shown for the left (light gray lines) and right (dark lines) ears. Spectral notches are apparent in the DTFs, ITDs are given by the delay between the left- and right-ear impulse responses, and ILDs are given by the difference in the left- and right-ear DTFs.

II. METHODOLOGY

A. Animal preparation

Six adult female Sprague Dawley rats were used (mean weight of 287 ± 12 g). All surgical and experimental procedures complied with the guidelines of the University of Colorado Health Science Center Animal Care and Use Committees and the National Institutes of Health. Rats were euthanized with sodium pentobarbital (100 mg/kg, i.p.) prior to acoustic measurements. We adapted the technique of Obrist *et al.* (1993) and Wotton *et al.* (1995) in their study of the acoustical cues in bats in which the euthanized animals were frozen prior to the acoustic measurements. Here, the euthanized rats were frozen before performing the experiments. Care was taken to ensure that the head position relative to the body and the pinna position relative to the head were maintained in a natural posture during freezing. Freezing the tissue facilitated the consistency of the insertion of the probe tube microphones to deeper portions of the ear canal. The weight, head diameter, and pinna height and width [Figs. 1(a) and 1(b)] of each animal were measured after freezing. Measurements of four animals before and after freezing showed that these dimensions were not significantly changed by freezing (paired $t_{25}=-1.26$, $p=0.22$). The acoustic measurement procedure lasted ~1 hour during which the animals remained frozen. Similar techniques have been used to measure the acoustical cues in frozen (Obrist *et al.*, 1993; Wotton *et al.*, 1995), formaline-fixed (Fuzessery, 1996; Aytekin *et al.*, 2004; Firzlaff and Schuller, 2003; Koay *et al.*, 1998), alcohol-fixed (Obrist *et al.*, 1993), and cadaver (Harrison and Downey, 1970; Moore and Irvine, 1979; Middlebrooks and Pettigrew, 1981; Coles and Guppy, 1986; Martin and Webster, 1989; Chen *et al.*, 1995; Maki and Furukawa, 2005) animals.

After the linear measurements of the head and pinna, a small hole was made in the wall of posterior aspect of pinna with a 14-gauge needle. A 50-mm-long flexible probe tube (Bruel and Kjaer, part no. AF-0555, 1.65 mm outer diameter) was inserted into the hole so that the tip of the tube was just within the ear canal and fixed in place with super glue. The animal was then placed in the center of the sound-attenuating room (see below), with its interaural axis aligned in the arc of loudspeakers by using three lasers, two at the poles and one at (0°,0°). To examine the role of the pinna, after taking the first set of acoustic measurements, the pinnae of three animals were removed and the measurements were repeated.

Removal of the pinna did not alter the position of the probe tube microphone in the ear canal and animals remained centered in the loudspeaker arc.

B. Experimental setup

All experiments were performed in an $\sim 3 \times 3 \times 3$ m^3 (interior dimensions) double-walled, sound-attenuating room (IAC, Bronx, NY), where the walls and equipment were lined with 4-in.-thick reticulated wedged acoustic foam (Sonex Classic). Stimuli were presented from 25 loudspeakers (Morel MDT-20) attached to a custom-built horizontally oriented [i.e., "single-pole" coordinate system (Middlebrooks and Pettigrew, 1981; Leong and Carlile, 1998)] semicircular boom. The 25 loudspeakers were spaced in azimuth along the arc at 7.5° spacing, from −90° (left) to +90° (right). The imaginary axis of rotation of the arc was aligned with the interaural axis of the animal (i.e., through the ears). The radius of the arc was 1 m. The 25 loudspeakers were selected from a larger set (~100) on the basis of best-matching frequency responses. A stepper motor (Advanced Micro Systems AMH34-1303-3) and motor controller/power supply (Advanced Micro Systems CMAX-810) under computer control could position the arc in elevation with a precision of <1°. The semicircular arc was moved in steps of 7.5° along the elevation by using the stepper motor controlled via personal computer by custom written software in MATLAB (Mathworks, Natick, MA). Stimuli were presented from a total of 625 different locations, covering azimuth and elevation. The elevation spanned −45° to +225°.

The measurement stimuli consisted of 11th order maximum length sequences (MLSs) (Rife and Vanderkooy, 1989) repeated without interruption 128 times from each loudspeaker. The MLS sequence was presented at full 24-bit resolution at a rate of 97 656.25 Hz (Tucker-Davis Technologies, TDT RP2.1, Alachua, FL). A single sequence of the 11th order MLS is comprised of 2047 samples ($2^{11}-1$) and is 20.96 ms in duration. Loudspeakers for stimulus presentation were selected via two daisy-chained TDT multiplexers (TDT PM2R) and the stimulus was amplified (TDT SA1) before being presented to the loudspeaker. The resulting acoustic waveforms in the ear canals of the left and right ears were simultaneously recorded through two probe tube microphones (Bruel and Kjaer, Type-4182), amplified (TDT MA2), and collected by using two analog to digital converter channels at 97 656.25 Hz (TDT RP2.1). Stimulus presentation, acquisition, analysis, and movement of the speaker arc were controlled by custom software in MATLAB. In all experiments, a calibration measurement was also made for each of the 25 loudspeakers in the absence of the animal by placing the tips of the probe tubes so that they corresponded to the location where the center of the head of the rats would be located. The calibration measurements capture the spectral characteristics of the loudspeakers and microphones for later processing.

C. Data processing and data analysis

The acoustic impulse response for each ear and each location was calculated by circular cross correlation of the original 11th order MLS stimulus and the in-ear recording from the probe tube microphone (Rife and Vanderkooy, 1989). The impulse responses were then truncated to 512 points (5.12 ms duration) by a 512-point Hanning window where the center of the window was set to approximately coincide with the point of maximum amplitude in the impulse response. This windowing procedure removes the small-amplitude reflections that may be contained in the impulse response. Next, the HRTFs were derived by dividing the frequency response of the in-ear recording by that of appropriate loudspeaker calibration measurement. This procedure removes the loudspeaker and microphone frequency response from each in-ear measurement. The resulting function is referred to as the HRTF, as it represents the acoustical gain and delay introduced by the head and the pinna. However, the resulting HRTF can be highly dependent on the exact placement of the tip of the probe tube microphone in the ear canal relative to the tympanic membrane (Middlebrooks *et al.*, 1989; Chan and Geisler, 1990). To reduce the confounding effects of the probe tube placement in the ear canal, for each ear the DTFs were then calculated from the HRTFs by dividing the HRTF made at each spatial location by the geometrical mean of all the measured HRTFs across all measurement locations for that ear. The spectral features resulting from the exact placement of the probe tube microphone in the ear canal are expected to be similar for all measurement locations (i.e., they are not dependent on spatial location), so this "common" spectral feature is removed from the HRTFs, resulting in the DTFs (Middlebrooks *et al.*, 1989). In essence, the DTFs are the sound source direction-dependent components of HRTFs.

The amplitude spectra of the DTFs were calculated after the spectra were passed through a bank of 300 bandpass filters, the center frequencies of which were spaced at intervals of 0.0143 octave spanning from 2 to 40 kHz. The 3 dB bandwidth of filters was held constant across all frequencies at 0.12 octaves, and the upper and lower slopes of the filters fell off at ~105 dB/octave. These filters have properties similar to the bank of bandpass filters that have been used elsewhere to filter DTFs (Middlebrooks, 1999; Xu and Middlebrooks, 2000; Schnupp *et al.*, 2003). The filter bandwidths used here were comparable to neural frequency tuning curves found in various auditory nuclei in the rat (Hernandez *et al.*, 2005).

Two binaural cues to sound location were studied here. The ILD spectrum was derived by computing the differences (in decibels) in the DTFs, frequency by frequency between right and left ears at all elevations, and for all azimuth angles. Positive ILD indicates that the decibel level at the left ear was higher than the decibel level at the right ear. The ILDs for particular frequencies and locations were extracted from the ILD spectra. The ITDs in this paper were measured in two ways. First, ITDs based on the envelopes of the left- and right-ear impulse responses were computed. We choose to focus on envelope ITDs because the rat audiograms reveal that they have poor hearing for frequencies <2 kHz (Heffner *et al.*, 1994) where ongoing ITDs in the fine structure would be useful. And a recent study has suggested that rats might not be able to use ongoing ITDs in the fine structure for

sound localization of low-frequency pure tones (Wesolek *et al.*, 2007). For each sound location, the envelopes of the left and right ear impulse responses were extracted by computing the magnitudes of their respective Hilbert transforms. These envelopes were then cross correlated and the delay corresponding to the maximum point was taken as the ITD for that particular location (Middlebrooks and Green, 1990). This process was repeated for each location. We also computed low-frequency ongoing ITDs in the fine structure by first low-pass filtering the impulse responses at 3.5 kHz, cross correlating the left and right impulse responses for each spatial location, and then defining the ITD at that location as the delay corresponding to the maximum in the correlation function.

For spatial plotting purposes, the data were displayed as Aitov projections. In this projection the nose of the animal is considered to be projecting out of the page at 0° azimuth and 0° elevation, as if the animals were looking at the reader. The spatial plots were plotted for frontal data for elevations from −45° to +90° and azimuth from +90° to −90°.

III. RESULTS

The results are based on DTF measurements from six adult female rats (Sprague Dawley). The mean weight was 287 ± 12 g ($n=6$). Measurements of several linear dimensions of the head and pinna were made (Fig. 1). Across the six animals, the average head diameter FG was 29.6 ± 1.3 mm, the average pinna width DE was 11.4 ± 0.5 mm, and the average pinna lengths AC and BC were 16.8 ± 0.8 and 17.7 ± 1.7 mm, respectively (both the left and right pinna were measured, yielding a total of 12 measurements of AC, BC, and DE for the six animals). The six rats were remarkably similar in size and weight and the acoustical properties were also very similar. For this reason, we display the results from only three of the animals. Average cue values are based on all six animals, unless otherwise noted. Figure 2 shows the left and right ear impulse responses and their corresponding DTFs from one animal at five azimuthal locations at 0° elevation. Figure 2 illustrates the three main acoustical cues to location that are of particular interest in this paper; the ITDs, ILDs, and spectral notches.

A. Monaural aspects of the DTFs

1. Frequency range and spatial-location dependence of spectral notches

We observed a systematic change in the frequency of the first (lowest frequency) spectral notch with changes in source location (e.g., Rice *et al.*, 1992). Figure 3 shows DTFs for the left ear of one animal (R016) for elevations ranging from −45° to 90° in 7.5° steps for 15° azimuth (panel A) and for azimuths ranging from −90° (ipsilateral ear) to 90° in 7.5° steps at 0° elevation (panel B). Similar results are shown for a different animal (R004, panels C and D). Spectral notches were observed for frequencies above 15 kHz in all animals. At a given sound source azimuth, the frequencies of the first spectral notch generally increased with increasing source elevation. The first notch frequency was easily detectable and

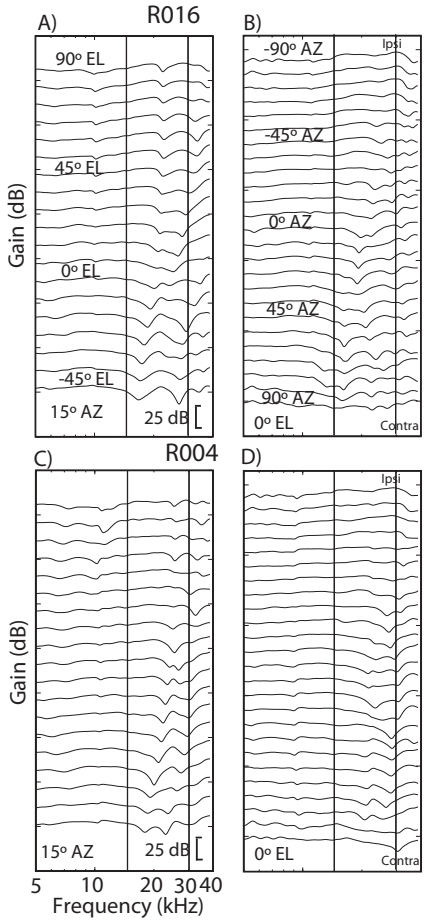


FIG. 3. The DTF gains for the left ear of two animals (R016 and R004) for sources at 15° azimuth and varying in elevation from −45° to +90° (panels A and C) and at 0° elevation varying from −90° (ipsilateral) to +90° (contralateral) azimuth (panels B and D). The two vertical bars in each figure bound the approximate frequency range of the first spectral notches, 16–30 kHz.

systematically moved with elevation for sources in the frontal hemisphere (e.g., Fig. 3). However, the notch was difficult to detect or erratically moved with source locations above and behind the animals.

The first notch frequency also tended to increase with changing source azimuth toward the ipsilateral ear. This is evident for R016 [Fig. 3(b)] for changing source azimuth from the contralateral hemisphere toward the ipsilateral ear (−90°). As was the case for source elevation, the first notch frequency movements with source azimuth occurred for sources in the frontal hemisphere, but were difficult to detect or did not move in an orderly manner for sources behind the animal. Across the six rats, the notches tended to increase from ~16 to 30 kHz as the source increased in elevation from −45° to 90° and from 17 to 23 kHz as the source azimuth went from ~+30° to −90° (i.e., toward the ipsilateral ear).

Figure 4 shows the iso-first-notch frequency contours for the left ear (−90° is ipsilateral) for two animals (R0016 and R004) for sound sources in the frontal hemisphere. An automated procedure based on that described by Maki and Furukawa (2005) was used to find the frequency corresponding to the spectral notch for each ear and each spatial location.

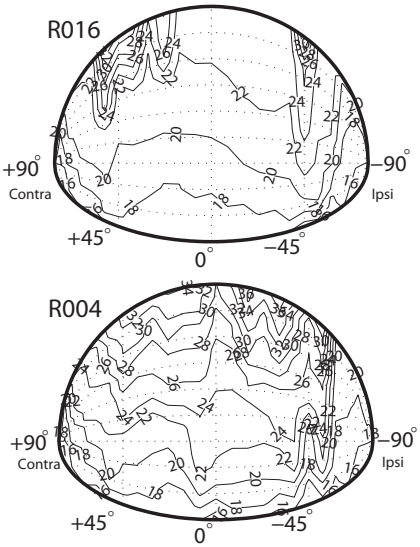


FIG. 4. A plot of the isofrequency contours of the first notch frequencies for sources in the frontal hemisphere for the left ears of two animals (R016 and R004). First, notch frequencies increase with source elevation and azimuth toward the ipsilateral ear (−90° in this figure).

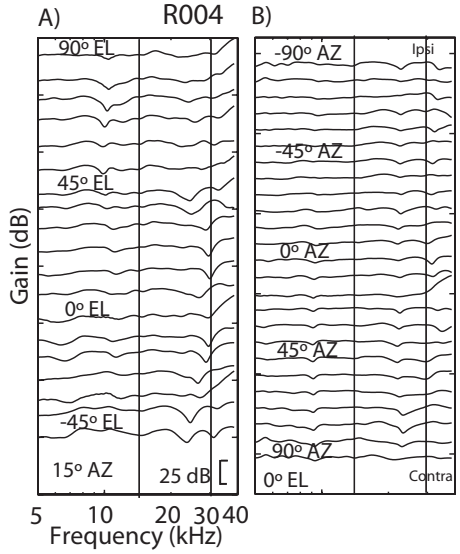


FIG. 5. The DTFs for the left ear of one animal (R004) after the pinna were removed for sources at 15° azimuth and varying in elevation from −45° to +90° (panel A) and at 0° elevation varying from −90° (ipsilateral) to +90° (contralateral) azimuth (panel B). The spectral notches apparent with the pinna (Figs. 3(c) and 3(d)) were no longer present after pinna removal. The vertical bars bound the approximate frequency range of the first spectral notches, 16–30 kHz (see Fig. 3).

As detailed above, the first notch frequencies were either often difficult to detect or erratically moved with source location for extreme lateral azimuths and for elevations above and behind the animal. Yet in the middle part of the frontal hemisphere, the notch frequency contours were largely mirror symmetrical about the midline between right and left ears. The notches with lower frequencies were distributed on the contralateral azimuths and lower elevations, and the notch frequencies generally increased as the sources were ipsilaterally moved in azimuth and toward higher elevations. Although data are only shown in Figs. 3 and 4 for two animals, similar results were observed in the other animals.

2. The role of the pinna in generating the spectral notches

Many studies have speculated that the spectral features, such as the notch, are generated by the pinna. To test this hypothesis, the pinna on both sides were surgically removed and the acoustic measurements were repeated in three animals (R001, R002, and R004). Removal of the pinna did not alter the position of the probe tube in the ear canal and care was taken to ensure that the animal remained centered in the loudspeaker arc (see Methods). Figure 5 shows DTFs for left ear of one animal [R004, Figs. 3(c) and 3(d)] after pinna removal; the DTFs are plotted in the same way as Fig. 3. The spectral notches apparent when the pinna were intact [Figs. 3(c) and 3(d)] were no longer apparent when the pinna were removed [Figs. 5(a) and 5(b)]. Removal of the pinna also eliminated the spectral notch cue in the other two animals (e.g., see the "head only" data in Fig. 6). This experiment supports the hypothesis that the spectral notches occurring from ~16 to 30 kHz in rats are exclusively produced by the pinna.

3. Spatial distribution of DTF amplitude gain

The acoustical gain of the DTF varied with source direction and frequency. Figure 6 (left column, "head+pinna") shows the distribution of DTF gains for seven frequencies for sources in the frontal hemisphere for the right ear (+90° is ipsilateral) of one intact animal (R002). DTF gain was plotted only for the right ear as it was almost mirror symmetric with that of the left ear in most of the animals (this will be apparent when the ILD cue to location is discussed). The amplitude gain contained nonsystematic peaks and dips for sources behind the animal, so these sources were not considered further. Moreover, even in the frontal hemisphere, the DTF gains were also complicated for the frequencies corresponding to the first spectral notch (~16–30 kHz); the spectral notches in Fig. 6 are particularly apparent for this animal from 20 to 30 kHz. The maximum DTF gains observed for the group of six animals tended to increase as a function of frequency; the gains were ~4–5 dB at 5 kHz, 6 dB at 10 kHz, 12 dB at 15 kHz, and approached 15–20 dB for frequencies higher than 15 kHz. For frequencies higher than ~20 kHz, there were often two or more distinct regions of high gain. For frequencies <5 kHz, the gains were <4 dB.

Since DTFs were computed, any acoustical gains that were nondirectional were removed from the data. For example, there is a large nondirectional gain associated with the resonance frequency of the ear canal. The gain of the canal resonance was estimated from the common components of the HRTFs (see Methods). The common component of the HRTFs averaged across the six rats revealed a peak in the gain of 24.1 ± 3.3 dB occurring at 17.2 kHz.

4. Acoustical axis

The direction of maximum acoustical gain at a given frequency in the DTF gain spectra is known as the acoustical

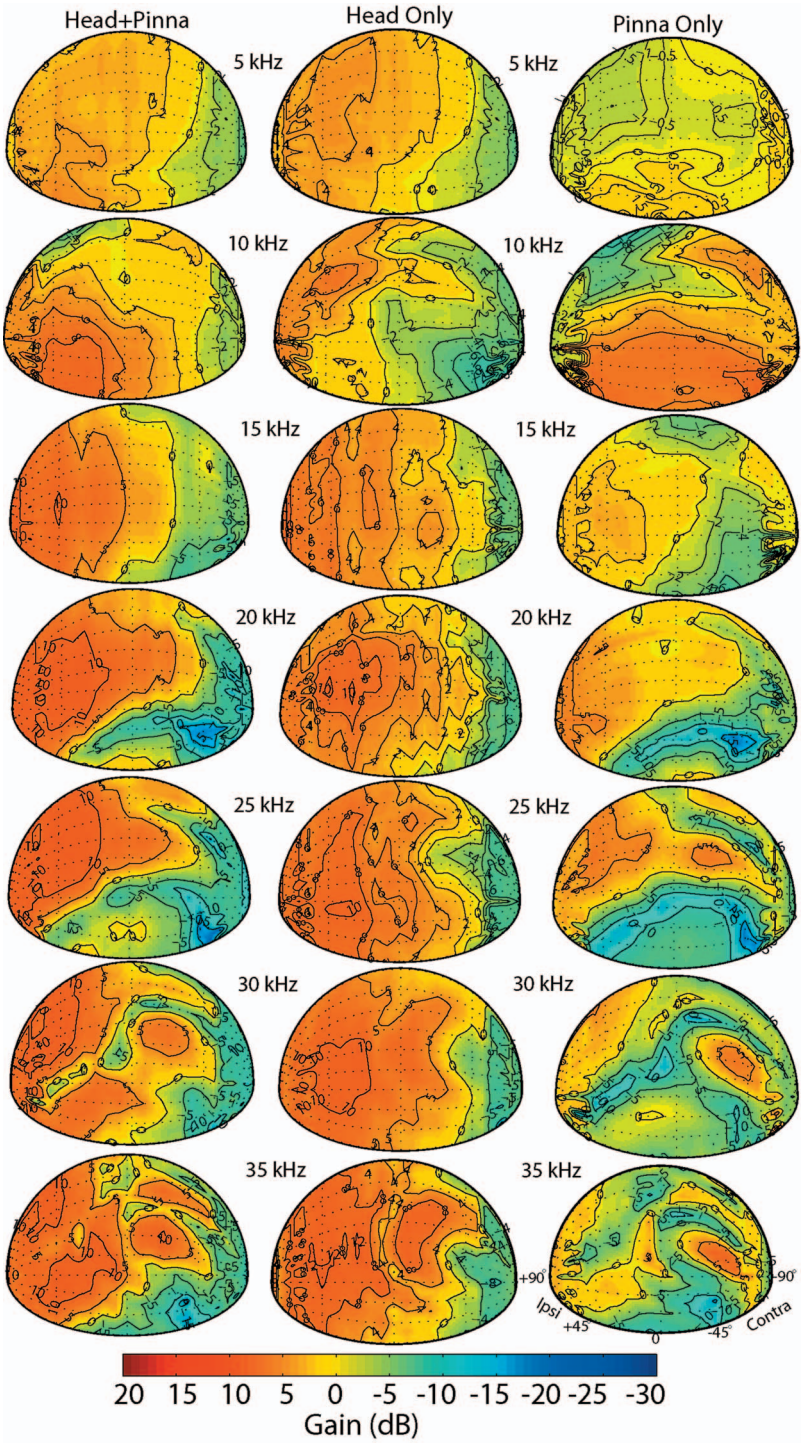


FIG. 6. Spatial distribution of DTF gains for seven frequencies for the right ear (+90° is ipsilateral) of one animal (R002) for three different conditions; intact animal with head and pinna ("head+pinna"), head only after the pinna were removed ("head only"), and the contribution of the pinna ("pinna only") which was computed as the difference of the "head+pinna" and "head only" measurements. The color bar indicates the gain in decibels.

axis (Middlebrooks and Pettigrew, 1981; Phillips *et al.*, 1982). Figure 6 (left column, "head+pinna") shows the acoustical gain for one animal (R002) from which the acoustical axis could be found. The acoustical axis for each animal did not systematically vary as a function of frequency, but rather often made sudden and spatially distant jumps from one location to another. In all animals, the spatial directivity and the acoustical axis were not well defined for frequencies below ~10 kHz. For the group of six animals, the acoustical axis for each frequency was observed after the spatial DTF gains were averaged across the animals. On average, for 10 kHz the acoustical axis occurred at (45°, −30°), for 15 kHz (22.5°, 15°), for 20 kHz (45°, 30°), for 25 kHz (37.5°, 37.5°), for 30 kHz (45°, 60°), and for 35 kHz the acoustical axis was (7.5°, 22.5°) (for these data, positive azimuths correspond to the ipsilateral hemisphere). The results based on the across-animal average DTF gains are in agreement with the individual results shown in Fig. 6 (left column).

5. The contribution of the pinna to acoustical gain and the acoustical axis

Figure 6 shows the DTF gain of the right ear (+90° is ipsilateral) for one animal (R002) at seven frequencies for

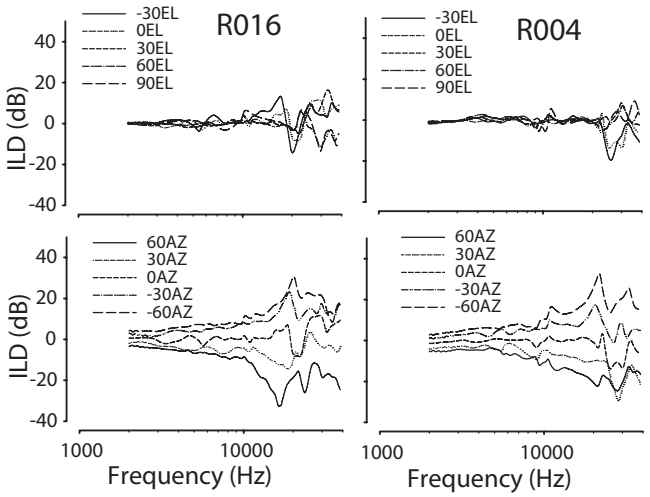


FIG. 7. The ILD spectrum for two animals (R016 and R004). ILD spectrum is the frequency-by-frequency difference between left- and right-ear DTFs at a given location. Positive ILD indicates higher gain at the left ear than the right. ILDs do not change as with source elevation along the midsagittal plane (top row) but do substantially change with source azimuth along the horizontal plane (bottom row). ILDs for some frequencies and some locations may be as large at 40 dB.

three conditions; the original intact measurements (head +pinna), measurements after pinna removal (head only), and the difference between the intact and pinna-removed measurements ("pinna only"). The pinna only data show that the pinna significantly contributed to the DTF gain, but only at high frequencies above 20 kHz. The pinna by themselves generate up to 5–12 dB of gain for frequencies from 20 to 35 kHz. Below 20 kHz, the gain produced by the pinna is generally small, <2 dB. The pinna also produced substantial attenuation, some of which was shown in Fig. 5 where pinna removal eliminated the spectral notch cues. The spectral notches (i.e., regions of low gain surrounded by relatively higher gain) are particularly evident for frequencies of 20, 25, and 30 kHz in the head+pinna and pinna only acoustical gains (Fig. 6, left and right columns, respectively). The notches are not apparent in the head only condition (Fig. 6, middle column).

The pinna also contributed to the location of the acoustic axis. Examination of the head only condition shows that the gain produced by the head is largely symmetrical about the midsagittal plane and tends to monotonically increase as source location moves in azimuth from contralateral to ipsilateral. The head only acoustical axis tends to occur in the area from ~45° to 90° at the ipsilateral ear and around 0° elevation. The acoustic axis with the pinna present does not obey this symmetry, suggesting that the directivity of the pinna at high frequencies determines the acoustic axis.

B. ILDs

1. Variation of ILD with frequency

The difference between left and right ear DTF gains (e.g., Fig. 6) results in the ILD spectra. ILD cues varied with frequency and source location. Positive and negative ILDs indicate higher DTF gain for left and right ears, respectively. Figure 7 shows for two animals (left column, R016; right column, R004) the ILDs at five different elevations when the azimuth was fixed at 0° (top row) and for five azimuths when the elevation was fixed at 0° (bottom row). ILDs varied with changes in source azimuth and frequency, but not with changes in source elevation. The distinctive positive and negative peaks in the ILDs between 15 and 30 kHz for sources varying in elevation are due to slight asymmetries in the spectral notch frequencies at the left and right ears (Figs. 3, 4, and 6). These positive and negative peaks systematically moved to higher frequencies with changing azimuth angle and constant elevation consistent with the movements of the first notch frequencies (Figs. 3, 4, and 6) Also, as a function of azimuth, ILDs are small, on the order of a few decibels, for low frequencies, and become systematically larger for frequencies up to ~20 kHz. For frequencies from ~20 to 30 kHz, the spectral notches that are primarily present in the contralateral ear (farthest from the source) help to create very large ILDs approaching 40 dB for some frequencies. Note also that for some source azimuths and frequencies, ILD was negative due to the first spectral notch at the ipsilateral ear.

2. Variation of ILD with azimuth and frequency

Figure 8(a) illustrates the way that the ILD cues for different frequencies vary with azimuth along the horizontal plane (i.e., at a constant elevation of 0°). The data are the mean (± standard error of the mean) ILDs computed across the six animals as a function of azimuth at six different frequencies. In general, ILDs symmetrically varied about the midline. At 5 kHz, the ILD steadily increased from −7 to +7 dB as the azimuth changed from −90° to +90°. For 10, 15, and 20 kHz, the ILDs varied from ~±10, ~±17, and ~±26 dB, respectively, for azimuths between ~±70°, but then tended to slightly decrease for larger azimuths. For 25 and 30 kHz, the ILDs varied from ~±30 dB as the azimuth changed from −90° to +90°. Finally, the ILDs at 35 kHz were ±15 dB. At 35 kHz, the ILD cues were quite variable from animal to animal, leading to smaller average ILDs at each azimuth. The maximum mean ILDs across the six animals for sources along the horizontal plane in the frontal hemisphere were 7 dB at 5 kHz, 10 dB at 10 kHz, 17 dB at 15 kHz, 26 dB at 20 kHz, 28 dB at 25 kHz, 27 dB at 30 kHz, and 15 dB at 35 kHz.

For most frequencies examined, the ILD cue was largely symmetric in azimuth about the midline (at 0° elevation) and also varied approximately linearly for sources between ±30°. We noticed that as the frequency increased, the rate of change of the ILD cue with changes in source azimuth, the ILD slope, also increased. Figure 8(b) plots the slope of the ILD cue (dB/deg) for sources between ±30°. The slope steadily increased from ~0.03 dB/deg for frequencies <1.5 kHz up to 0.42 dB/deg by 18.5 kHz. For higher frequencies, particularly between 18 and 30 kHz where the spectral notches occur, the ILD slope considerably varied from animal to animal. This can also be appreciated in the across-animal variability in ILD in Fig. 8(a).

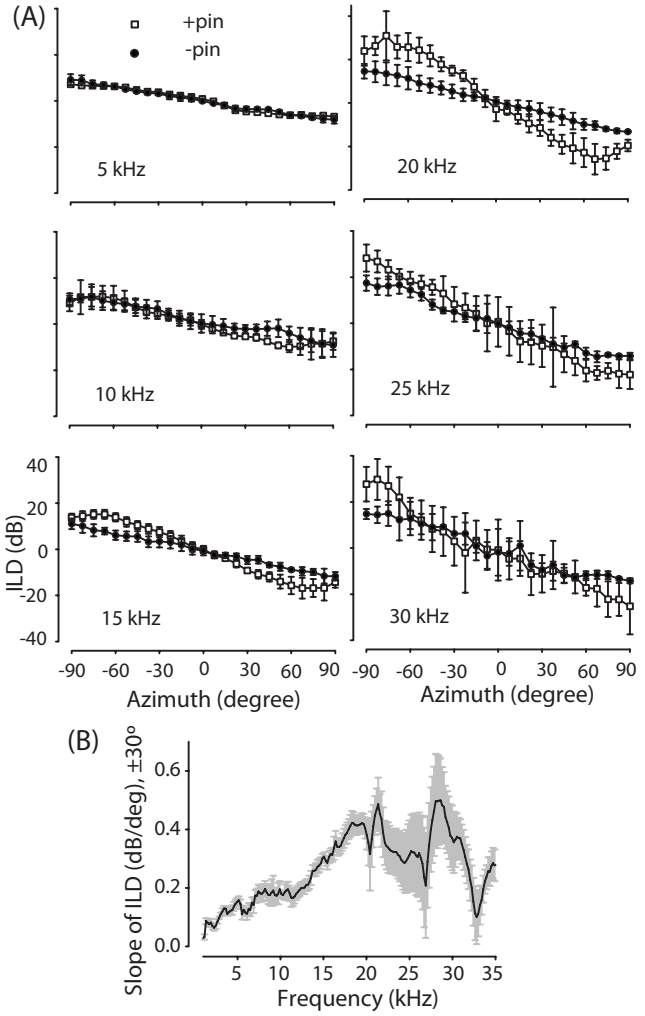


FIG. 8. (A) The mean ILD cue (□, +pin) computed across the six intact animals varies as a function of azimuth along the horizontal plane and frequency. Error bars indicated ±1 standard error of the mean (SEM). The mean ILDs are shown for six frequencies from 5 to 30 kHz. The mean ILD cue for three animals after removal of the pinna is also shown (●, −pin). Removal of the pinna reduced the ILDs for frequencies >10 kHz. (B) Rate of change of the ILD cue (i.e., ILD slope) with changes in source azimuth along the horizontal plane between ±30°. The ILD slope was obtained from the mean ILD cues computed across the six animals (panel *A*). Error bars plot ±1 SEM.

3. Spatial distribution of ILD at various frequencies

The data presented above indicate how ILD varies as a function of azimuth and frequency along the cardinal azimuthal dimension. Similar to the DTF gains, however, the ILD cue is actually a complex function of azimuth, elevation, and frequency. The spatial distributions of ILDs at various frequencies were studied by similarly plotting them to the monaural DTF gains (e.g., Fig. 6). Figure 9 shows the ILDs calculated at seven frequencies for two animals (R016 and R004). As was seen in Fig. 8(a) for the across-animal average ILDs, the ILDs were, in general, symmetrical between the right and left hemispheres for frequencies up to ~15–20 kHz, but were more complex for higher frequencies. This is due, in part, to the role of the pinna, which generate the spectral notches (Figs. 3, 4, and 6) and contribute to the DTF gain and location of the acoustic axis (Fig. 6).

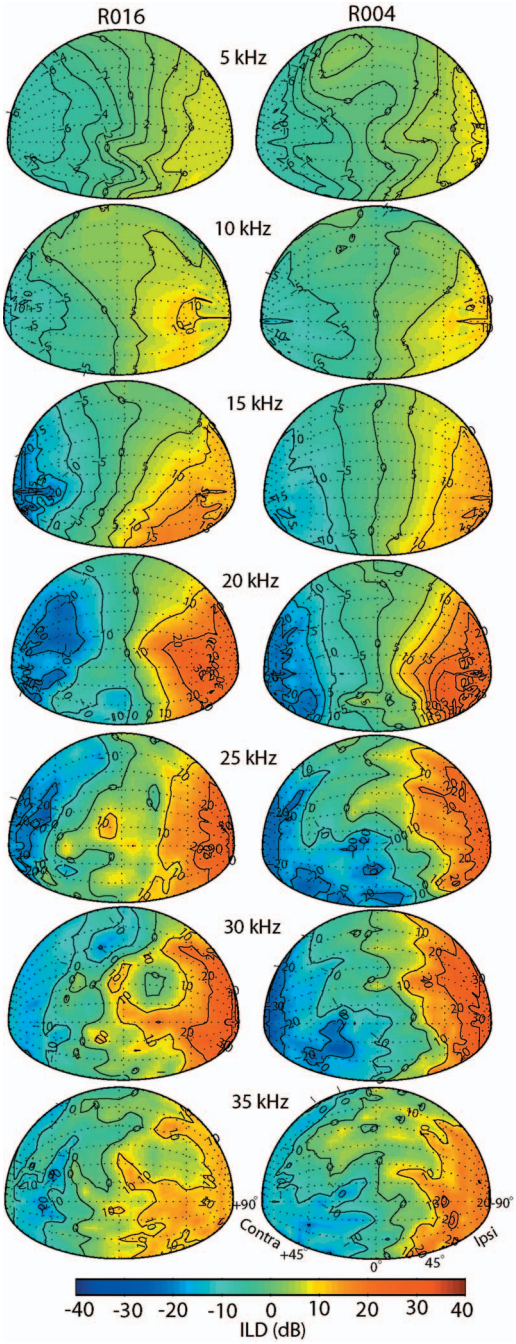


FIG. 9. Spatial distribution of ILDs at seven different frequencies for two animals (R016 and R004) for locations in the frontal hemisphere. Positive ILDs indicate higher gain at the left ear (−90°) than the right ear (+90°). Color bar indicates ILD magnitude in dB.

In particular, for the frequency ranges of the prominent first notches (~15–30 kHz), the spatial distribution of ILDs can be quite complex.

4. Spatial distribution of directions with maximum and minimum ILDs

Like the monaural acoustical axis, the spatial location of the maximum ILD was dependent on frequency. Figure 9 shows the directions of maxima and minima in the ILD cue for different frequencies for animals R016 and R004. We defined the minimum of the ILD cue at each frequency as the

binaural acoustical axis (using the maximum ILD yielded comparable results, as expected given the symmetry of the ILD computation). The binaural acoustic axes for R016 for 10 kHz were ~(45°,0°), for 15 kHz ~(52.5°,−7.5°), and ~(52.5°,30°) for 20 and 25 kHz. Due to the complexity of the ILD cue for higher frequencies, the determination of the binaural acoustic axis was not always clear. As was done for Fig. 6, the spatial distributions of the ILD spectra were averaged across the six animals. On average, the binaural axes were (90°, 0°) for 5 kHz, (75°, 0°) for 10 kHz, (75°,−22.5°) for 15 kHz, (75°, 0°) for 20 kHz, (75°, 30°) for 25 kHz, (90°, 0°) for 30 kHz, and (45°,−15°) for 35 kHz.

5. The contributions of the pinna to the ILD cues

The importance of pinna in generating acoustical gain and the resulting ILD cues was examined by computing the spatial distributions of ILD before and after pinna removal. Figure 8(a) shows plots of the average ILDs as a function frequency and azimuth along the horizontal plane recorded in three animals (R001, R002, and R004) in two conditions: original measurements in the intact animals (+pin) and measurements after pinna removal (−pin). The contribution of the pinna themselves to ILD (−pin) can be seen where the two functions differ. This generally occurred for frequencies of 10 kHz and above; for frequencies <10 kHz, the two functions were nearly identical. For the higher frequencies, particularly those where the spectral notches were observed (16–30 kHz, Figs. 3, 4, and 6) and where the pinna themselves introduce considerable acoustical gain (e.g., Fig. 6, "pinna only"), the pinna substantially contributed to the ILD. From the difference of the functions in Fig. 8(a), the maximum contributions of the pinna to ILD were 2.2, 7.4, 7.8, 17.5, 10.9, 15.4, and 8.4 dB for frequencies of 5, 10, 15, 20, 25, 30, and 35 kHz, respectively.

C. ITDs

1. Spatial distribution of ITDs

The ITD cue was dependent mostly on sound source azimuth. Positive and negative ITD values indicate that sound was leading at left and right ears respectively. The detailed frequency dependence of ongoing ITDs in the fine structure (or equivalent interaural phase differences) was not examined in detail in this paper. Rather, we focus here on the interaural envelope delays conveyed by the impulse responses of the DTFs (e.g., Middlebrooks and Green, 1990). Figure 10(a) shows spatial distribution of envelope ITDs in the frontal hemisphere for two animals (R016 and R004). For both animals, the ITD was symmetric about the midline. Although not shown, comparable results were observed for ITDs in the rear hemisphere. The same trends were observed for the other four animals. Across the six animals in this study, the maximum ITD averaged 127±14 μs. The maximum ITDs based on five of six animals (one animal had empirical ITDs that were anomalously larger than expected based on its head diameter) were positively correlated (r =0.79) with the linear head diameter. Maximum ITDs in all animals tended to occur within ±10° of the most lateral azimuth (90°).

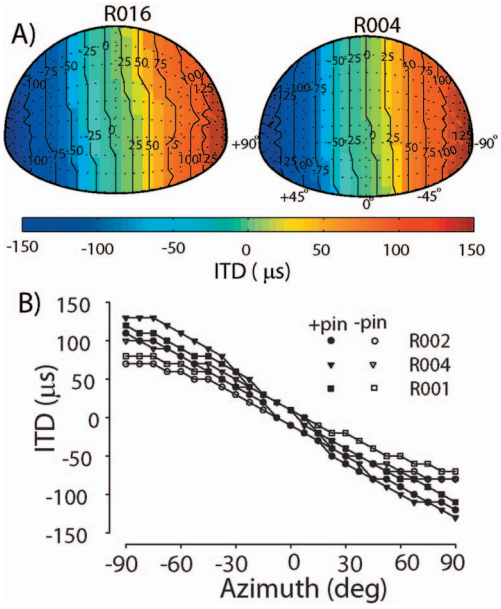


FIG. 10. (A) Spatial distribution of envelope ITDs for two animals (R016 and R004) for locations in the frontal hemisphere. Positive ITDs indicate that the signal leads to the left ear. Color bar and countour lines indicate the ITD in microseconds. (B) The envelope ITDs in three animals (R001, R002, and R004) as a function of azimuth along the horizontal plane in two conditions: intact (filled symbols, +pin) and after removal of the pinna (open symbols, −pin). Removal of the pinna reduced the ITDs at all azimuthal locations.

Examination of the spatial dependence of ITD in Fig. 10(a) shows that envelope-based ITDs essentially follow the contours of azimuth and show little to no dependence on elevation. Results such as this would be expected from a spherical head model of ITDs. As such, similar results on maximum envelope ITD were obtained by first collapsing the ITD measurements from all elevations onto a single dimension of azimuth and then fitting the Woodworth (1938) spherical head model to these data. The Woodworth model is given by

$$\mathrm{ITD} = r/c[\theta + \sin(\theta)], \tag{1}$$

where r is the radius of the head (in meters), c is the speed of sound (340 m/s), and θ is the lateral angle of the source relative to the median plane. In the model, the radius r is the only free parameter. For all animals, the fitted model produced correlation coefficients >0.95. Finding the maximum ITD by using this procedure allows the ITD data from all spatial locations to contribute to the estimate.

By using this same procedure, low-frequency ongoing ITDs in the fine structure were also calculated as described in the Methods. In short, the left and right ear impulse responses at each location were low-pass filtered (3.5 kHz), cross correlated, and the ITD was taken as the delay corresponding to the maximum in the correlation function. ITDs measured in this way were larger than the envelope-based high-frequency ITDs. As before, the Woodworth model fits all produced correlation coefficients >0.95. Across the six

rats, the maximum low-frequency ITDs estimated by using the fitting procedure described above was 158 ± 8 μs. This value is 1.25 times larger than the high-frequency envelope ITDs.

2. Pinna contribution to high-frequency envelope-based ITDs

It is often believed that the pinna do not play a large part in determining the ITD cue to sound location, particularly for low-frequency ongoing ITDs (e.g., Roth *et al.*, 1980). For the high-frequency envelope-based ITD cues measured here, however, we hypothesized that the pinna does play a role. For high-frequency sounds, whose wavelengths are on the order of, or smaller, than the linear dimensions of the pinna, the pinna may present a significant obstacle for the sound. Thus, we expected the pinna to increase the magnitude of the envelope-based ITDs. To test this hypothesis, for three animals (R001, R002, and R004), the pinna were completely removed and ITDs were remeasured.

Figure 10(b) shows the envelope ITDs measured with (filled symbols) and without the pinna (open symbols) for the three animals for sources varying in azimuth at 0° elevation. The maximum ITDs for these three animals were determined in both conditions by using the fitting procedure described in the previous section. The mean maximum ITDs were significantly reduced from 120 to 77 μs after removing the pinna. The position of the probe tube microphone was not altered in any way by removing the pinna due to its deep placement in the ear canal. Thus, removal of the pinna in rats reduced the maximal high-frequency envelope-based ITDs by 36%. The ITDs in the two conditions were also computed at each azimuth for the three animals, and the average reduction after pinna removal (excluding 0° azimuth) was 27 ± 0.13%. The pinna in the rat appear to alter the ITDs more for lateral angles near the poles than for source locations in front of the animal. The removal of the pinna also reduced the low-frequency ITDs by 32% to an average of 108 ± 3 μs. Thus, for both low- and high-frequency envelope-based ITD cues, the pinna effectively increases the functional diameter of the head, thereby increasing the magnitude of the ITD.

IV. DISCUSSION

A. Comparison of present study with other animal studies

1. Monaural aspects of the acoustical cues: spectral notches, gain, and acoustic axis

The spectral notch cues were found in the rat primarily for source locations in the frontal hemisphere. First, notch frequencies increased from ~16 to 30 kHz as the source elevation increased from −45° to +90° and from ~17 to 23 kHz as the source azimuth moved from −30° to +90° toward the ipsilateral ear. In general, the first notch frequencies increased as location varied from low elevations and contralateral azimuths to high elevations and ipsilateral azimuths. Similar patterns of first notch frequency changes with source azimuth and elevation changes have been observed in other species, albeit at different rates and over different frequency ranges (Rhesus monkey, 5–15 kHz: Spezio *et al.*, 2000; human, 6–12: Middlebrooks, 1999; bat 30–50 kHz: Wotton *et al.*, 1995; Fuzessery, 1996; Firzlaff and Schuller, 2003; Aytekin *et al.*, 2004). In the cat, where spectral cues have been studied in great detail, first, notch frequencies have been found to vary from ~8 to 18 kHz (Musicant *et al.*, 1990; Rice *et al.*, 1992; Young *et al.*, 1996; Xu and Middlebrooks, 2000; Tollin, 2004).

In the gerbil, which has slightly smaller pinna and head size than the rat, spectral notch cues have also been observed to move in a similar fashion with azimuth and elevation as the rat, but over a frequency range of ~25–45 kHz (Maki and Furukawa, 2005). It is likely that the linear dimensions of the pinna determine the frequency range for spectral notches. Maki and Furukawa reported that the pinna height [e.g., (A)–(C), Fig. 1] for their adult gerbils was ~12 mm, which is 67% of the rat pinna dimensions measured here. If we assume a linear scaling of first notch frequency with pinna size (e.g., Middlebrooks, 1999), then scaling the observed range of gerbil spectral notch cues (25–45 kHz) by 67% yields a predicted spectral notch cue range of 16–30 kHz for the rats, which is what we found here. As expected, species with larger pinna (human, cat, and monkey cited above) produce notches at lower frequencies.

We showed here that the pinna of the rat were essential for generating the spectral notch cues to sound location. Similar observations have been made in other species where the spectral notch cues have also been observed and then eliminated after pinna removal (bat: Wotton *et al.*, 1995; Aytekin *et al.*, 2004; cat: Musicant *et al.*, 1990; ferret: Parsons *et al.*, 1999). Further evidence that the spectral notches are created by the pinna comes from examination of the changes in first notch frequencies with changes in the positions of the pinna on the head (Young *et al.*, 1996; Xu and Middlebrooks, 2000). In these latter studies, systematic rotations of the pinna were accompanied by predictable shifts in the spatial locations of the first notch frequencies. Finally, examination of the audiograms of rats reveals a prominent increase in thresholds of up to 16 dB, relative to neighboring frequencies, for frequencies between 16 and 32 kHz (Kelly and Masterton, 1977; Heffner *et al.*, 1994). This frequency range overlaps that for which we found spectral notches for sources in the frontal hemisphere.

Studies in a variety of species have shown that the acoustic axis changes for different frequencies, but often in complicated and nonsystematic ways. In the rat, the acoustical axis generally changed from lower to higher elevations as the source frequency increased from 10 to 30 kHz. For frequencies >35 kHz, the acoustic axis decreased somewhat in elevation. In some animals, for frequencies >~20 kHz, there were sometimes two or more distinct spatial locations associated with high gain. This observation is consistent with the splitting of the acoustic axis for frequencies above 8 kHz in humans (Middlebrooks *et al.*, 1989) and above 55 kHz in bats (Firzlaff and Schuller, 2003). Comparable observations on the complexities of the acoustic axis have been reported in other species (cat: Musicant *et al.*, 1990; Phillips *et al.*, 1982; Middlebrooks and Knudsen, 1987; Martin and Webster, 1989; Calford and Pettigrew, 1984; Tammar wallaby: Coles and Guppy, 1986; bat: Obrist *et al.*, 1993; Firzlaff and

Schuller, 2003; Fuzessery, 1996; Aytekin *et al.*, 2004; gerbil: Maki and Furukawa, 2005; Rhesus monkey: Spezio *et al.*, 2000; owl: Keller *et al.*, 1998; ferret: Carlile, 1990).

There have been some models of localization based on a cue similar to the acoustical axis (e.g., Musicant and Butler, 1984). However, in general, the acoustical axis in the rat, like that in the cat (Phillips *et al.*, 1982) and gerbil (Maki and Furukawa, 2005), changes in such complex ways with frequency, and often with multiple peaks, that we believe it to be a poor cue for sound source localization. Rather, as suggested by Middlebrooks and Pettigrew (1981), Coles and Guppy (1986), and Young *et al.* (1996), the frequency dependence of the acoustic axis along with the ability of some animals with mobile pinna to independently manipulate the left and right pinna positions might allow the pinna to operate like an acoustical antenna, increasing the signal-to-noise ratio for sounds of interest, at the expense of other sounds. The increased acoustical gain due to the pinna might then be more useful for source detection than for localization. The pinna of the rat are indeed independently mobile (Li and Frost, 1996).

Finally, we studied the role of the pinna in the monaural DTF gain and the acoustical axis by making acoustic measurements before and after removing the pinna. Similar experiments have been done in other species (cat: Wiener *et al.*, 1966; Phillips *et al.*, 1982; Musciant *et al.*, 1990; guinea pig: Palmer and King, 1985; Carlile and Pettigrew, 1987; Tammar wallaby: Coles and Guppy, 1986; bat: Orbist *et al.*, 1993; Fuzessery, 1996; Aytekin *et al.*, 2004; ferret: Carlile and King, 1994). In agreement with Carlile and Pettigrew (1987) and Carlile and King (1994), after pinna removal the acoustical axis, for virtually all frequencies, was oriented along the ipsilateral interaural axis, supporting the hypothesis that the pinna play a critical role in establishing the acoustic axis for high frequencies. Moreover, the overall gains of the DTFs were reduced for most high frequencies, confirming that the pinna is responsible for part of the gain at high frequencies. In the rat, the pinna themselves generated ~5–12 dB of gain for frequencies from 20 to 35 kHz. Below 15 kHz, the gain due to the pinna was small. Pinna-only gains of 3–15 dB, with the higher gains occurring at higher frequencies, have been reported in other species (Coles and Guppy, 1986; Carlile and King, 1994; Schnupp *et al.*, 1998).

Finally, since DTFs were computed in this paper, any acoustical gains that were largely nondirectional, such as the gain associated with the resonance frequency of the ear canal, were removed from the data. Examination of the common component of the HRTFs averaged across the six rats revealed a peak in the gain of 24.1 ± 3.3 dB occurring at 17.2 kHz, which is comparable to the recent estimates by Gratton *et al.* (2008) of 19.1 kHz and ~27 dB.

2. Binaural aspects of the acoustical cues

ILDs were highly dependent on both frequency and spatial location. ILDs were small for low frequencies <5 kHz, but rapidly increased for frequencies up to 20 kHz. Up to ~20 kHz, the ILDs systematically varied with azimuth and would likely provide a very stable cue for localization. Figure 8(b) shows that the rate of change of the ILD cue with changes in azimuth (i.e., the ILD slope) systematically increased with frequency up to ~20 kHz. For frequencies from ~20 to 35 kHz, the ILD cues were complicated due to the presence of the spectral notches. The deep notches resulted in large ILDs that for some frequencies and locations could approach 35 dB or more. For frequencies >35 kHz, the spatial distribution of ILDs was quite variable.

In the cat, the ILD slope has been shown to monotonically increase from ~0.17 dB/deg from 2 kHz to ~0.6 dB/deg by 8 kHz (Irvine, 1987; Martin and Webster, 1989). These range and magnitude of ILD slope values are comparable to that found here in the rat. In the cat, above 8 kHz, the ILD slopes were also large, but did not systematically vary from frequency to frequency. Thus, as was the case here in the rat, there appears to be a range of frequencies (<8 kHz) in the cat where the ILD cue might also provide a stable cue for localization. Rice *et al.* (1992) refered to this region as the ΔL region. Interestingly, in the cat, the spectral notches begin to occur first for frequencies at ~8 kHz. In the rat, the notches first begin at about 16 kHz. In both species, it appears that ILDs systematically vary with azimuth and also increase their slope for frequencies up to where the spectral notches first occur.

We showed here that the pinna play an important role in the formation and spatial distribution of the ILD cue in rats. Removal of both pinna changed the acoustical gain and the acoustic axis for frequencies >10 kHz, so it should not be surprising that the pinna also contribute to the ILD cue. From the difference of the functions in Figure 8(a), the maximum contributions of the pinna to ILD were 2.2, 7.4, 7.8, 17.5, 10.9, 15.4, and 8.4 dB for frequencies of 5, 10, 15, 20, 25, 30, and 35 kHz, respectively. At locations nearer to the acoustic axis, for particular frequencies, the contribution of the pinna to ILD could be even larger. Contributions of the pinna to ILD by up to 9 dB have been reported in the ferret (Carlile and King, 1994; Schnupp *et al.*, 1998; Parsons *et al.*, 1999). In guinea pig (Carlile and Pettigrew, 1987) pinna removal reduced ILDs for frequencies >5 kHz; similar to that found here in the rat, the largest ILD reduction occurred for those locations and frequencies associated with the acoustic axis.

Finally, the monaural gains and ILDs measured in rats by Harrison and Downey (1970) for a few frequencies and locations along the horizontal plane were remarkably similar to the data we recorded here by using DTFs.

Both low-frequency ongoing fine-structure ITDs and high-frequency envelope-based ITDs were measured. These ITDs systematically and symmetrically varied about the midsagittal plane [Fig. 10(a)]. Maximum envelope-based ITDs averaged 127 μs ($n=6$ rats). This value is 13% larger than the predicted maximum ITD of 111 μs based on the Woodworth (1938) spherical head model by using the across-rat average head diameter of 29.6 mm ($n=6$). By using the same average head diameter, a maximum value of only 87 μs is predicted based on Kuhn's (1977) model for high-frequency ITDs [$\mathrm{ITD}(\theta)=2(r/c)\sin(\theta)$]. Our empirical measurements of low-frequency ongoing ITDs, which averaged 158 μs ($n=6$ rats), were 1.25 times larger than the empirical high frequency ITDs. A maximum value of 131 μs is predicted

based on Kuhn's (1977) model for low-frequency ITDs [$\mathrm{ITD}(\theta)=3(r/c)\sin(\theta)$]. In all cases, the empirical maximum ITD values were substantially larger than the appropriate spherical head model predictions. None of the models takes into account the possible role of the pinna. Maki and Furukawa (2005) found maximum high-frequency ITDs in gerbils that were ~30% larger than those predicted based on the Woodworth (1938) model and the measured head diameters.

The finding that the low-frequency ongoing ITDs were 1.25 times larger than high-frequency envelope-based ITDs is in agreement with the theoretical spherical head model of ITDs by Kuhn (1977). The measurement in humans and subsequent modeling by Kuhn suggests low-frequency ITDs should be ~1.5 times larger than high-frequency ITDs due to the dispersion that occurs when sound waves encounter an object, such as the head. Similar differences between low- and high-frequency ITDs and the discrepancies between empirical ITDs and those predicted based on spherical head models have been observed (cat: Roth *et al.*, 1980; monkey: Spezio *et al.*, 2000; gerbil: Maki and Furukawa, 2005; guinea pig: Sterbing *et al.*, 2003).

For sources between ±30° along the horizontal plane, the rates of change (i.e., ITD slope) of the ITD cue with azimuth were 1.72 and 2.2 μs/deg for the high-frequency envelope ITD and low-frequency ITD, respectively. The ITD cues change nearly linearly for azimuths within ~±30° of the midsagital plane.

We examined the role of the pinna in the formation of the ITD cue in three animals. The maximum high-frequency ITDs in those animals were reduced by 36%, from 120 to 77 μs, after removing the pinna. On average, across azimuths (excluding 0°), the ITDs were reduced by 27%. The maximum ITDs of 77 μs without the pinna (i.e., head-only ITDs) are comparable to the predictions of the maximum ITDs based on Kuhn's high-frequency ITD model of 87 μs. These results together reveal that the pinna substantially contributed to the magnitude of high-frequency envelope ITDs, and do so more at larger lateral angles. Few studies have examined the role of the pinna in ITD cue generation. Roth *et al.* (1980) in the cat found that folding back the pinna in the cat had little effect on low-frequency ongoing ITDs, but did quite substantially reduce the high-frequency ITDs. We were surprised here to find that low-frequency ITDs in the rat were reduced 32% by pinna removal from an average of 158 to 108 μs. Our result might be due to different sound stimuli and/or to the different methods of computing ITD relative to Roth *et al.* (1980). Alternatively, Roth *et al.* (1980) may have underestimated the role of the pinna because they did not completely remove the pinna, but rather only folded it back on itself. This procedure likely still leaves a considerable portion of the pinna and distal parts of the auditory meatus in the path of the sound. In future studies of the contribution of the pinna to ITDs, the pinna should be completely removed to avoid this potential confound.

ACKNOWLEDGMENTS

Special thanks to Staci Stanford and Dr. Karl Pfenninger for providing the animals for these experiments. Thanks to Dr. John Middlebrooks, Dr. Henry Heffner, Dr. Tom Yin, and Dr. Philip Joris and also Heath Jones and Eric Lupo for comments on the manuscript. This work was supported by National Institutes of Deafness and Other Communicative Disorders Grant No. DC-006865 to DJT and the National Institutes of Child and Health Development Grant No. HD-2080 to HLR.

Aytekin, M., Grassi, E., Sahota, M., and Moss, C. F. (**2004**). "The bat head-related transfer function reveals binaural cues for sound localization in azimuth and elevation," J. Acoust. Soc. Am. **116**, 3594–3605.

Beecher, M. D., and Harrison, J. M. (**1971**)."Rapid acquisition of an auditory localization discrimination by rats," J. Exp. Anal Behav. **16**, 193–199.

Beyerl, B. D. (**1978**). "Afferent projections to the central nucleus of the inferior colliculus in the rat," Brain Res. **145**, 209–223.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D., and Grothe, B. (**2002**). "Precise inhibition is essential for microsecond interaural time difference coding," Nature (London) **417**, 543–547.

Burlile, C. J., Feldman, M. L., Craig, C., and Harrison, J. M. (**1985**). "Control of responding by the location of sound: Role of binaural cues," J. Exp. Anal Behav. **43**, 315–319.

Calford, M. B., and Pettigrew, J. D. (**1984**). "Frequency dependence of directional amplification at the cat's pinna," Hear. Res. **14**, 13–19.

Carlile, S. (**1990**). "The auditory periphery of the ferret. I: Directional response properties and the pattern of interaural level differences," J. Acoust. Soc. Am. **88**, 2180–2195.

Carlile, S., and Pettigrew, A. G. (**1987**). "Directional properties of the auditory periphery in the guinea pig," Hear. Res. **31**, 111–122.

Carlile, S., and King, A. J. (**1994**). "Monaural and binaural spectrum level cues in the ferret: acoustics and the neural representation of auditory space," J. Neurophysiol. **71**, 785–801.

Chan, J. C. K., and Geisler, C. D. (**1990**). "Estimation of tympanic membrane acoustic pressure and of ear canal length from remote points in the canal," J. Acoust. Soc. Am. **87**, 1237–1247.

Chen, Q.-C., Cain, D., and Jen, P. H.-S. (**1995**). "Sound pressure transformation at the pinna of *Mus Domesticus*," J. Exp. Biol. **198**, 2007–2023.

Clopton, B. M., and Silverman, B. M. (**1977**). "Plasticity of binaural interaction. II. Critical period and changes in midline response," J. Neurophysiol. **40**, 1275–1280.

Coles, R. B., and Guppy, A. (**1986**). "Biophysical aspects of directional hearing in the tammar wallaby, *Macropus eugenii*," J. Exp. Biol. **121**, 371–394.

Finlayson, P. G., and Caspary, D. M. (**1991**). "Low-frequency neurons in the lateral superior olive exhibit phase-sensitive binaural inhibition," J. Neurophysiol. **65**, 598–605.

Firzlaff, U., and Schuller. G., (**2003**). "Spectral directionality of the external ear of the lesser spear-nosed bat, Phyllostomus discolor," Hear. Res. **181**, 27–39.

Flammino, F., and Clopton, B. M. (**1975**). "Neural responses in the inferior colliculus of albino rat to binaural stimuli," J. Acoust. Soc. Am. **57**, 692–695.

Fuzessery, Z. M. (**1996**), "Monaural and binaural spectral cues created by the external ears of the pallid bat," Hear. Res. **95**, 1–17.

Gratton, M. A., Bateman, K., Cannuscio, J. F., and Saunders, J. C. (**2008**). "Outer- and middle-ear contributions to presbycusis in the brown norway rat," Audiol. Neuro-Otol. **13**, 37–52.

Harrison, J. M., and Downey, P. (**1970**). "Intensity changes at the ear as a function of the azimuth of a tone source: a comparative study," J. Acoust. Soc. Am. **47**, 1509–1518.

Harrison, J. M. (**1988**). "Control of responding by sounds of different quality: An evolutionary analysis," J. Exp. Anal Behav. **50**, 521–539.

Heffner, H. E., and Heffner, R. S. (**1985**). "Sound localization in wild Norway rats (*Rattus norvegicus*)," Hear. Res. **19**, 151–155.

Heffner, H. E., Heffner, R. S., Contos, C., and Ott, T. (**1994**). "Audiogram of the hooded Norway rat," Hear. Res. **73**, 244–247.

Heffner, R. S. (**1997**). "Comparative study of sound localization and its anatomical correlates in mammals," Acta Oto-Laryngol., Suppl. **532**, 46–53.

Hernández, O., Espinosa, N., Pérez-González, D., and Malmierca, M. S. (**2005**). "The inferior colliculus of the rat: A quantitative analysis of monaural frequency response areas," Neuroscience (Oxford) **132**, 203–217.

Inbody, S. B., and Feng, A. S. (**1981**). "Binaural response characteristics of

single neurons in the medial superior olivary nucleus of the albino rat," Brain Res. **210**, 361–366.
Irvine, D. R. (**1987**). "Interaural intensity differences in the cat: Changes in sound pressure level at the two ears associated with azimuthal displacements in the frontal horizontal plane," Hear. Res. **26**, 267–286.
Irvine, D. R., Park, V. N., and Mattingley, J. B. (**1995**). "Responses of neurons in the inferior colliculus of the rat to interaural time and intensity differences in transient stimuli: Implications for the latency hypothesis," Hear. Res. **85**, 127–141.
Irvine, D. R., Park, V. N., and McCormick, L. (**2001**). "Mechanisms underlying the sensitivity of neurons in the lateral superior olive to interaural intensity differences," J. Neurophysiol. **86**, 2647–2666.
Irving, R., and Harrison, J. M. (**1967**). "The superior olivary complex and audition: a comparative study," J. Comp. Neurol. **130**, 77–86.
Jen, P. H., and Chen, D. M. (**1988**). "Directionality of sound pressure transformation at the pinna of echolocating bats," Hear. Res. **34**, 101–117.
Kandler, K., and Gillespie, D. C. (**2005**). "Developmental refinement of inhibitory sound-localization circuits," Trends Neurosci. **28**, 290–296.
Keller, C. H., Hartung, K., and Takahashi, T. T. (**1998**). "Head-related transfer functions of the barn owl: Measurement and neural responses," Hear. Res. **118**, 13–34.
Kelly, J. B. (**1980**). "Effects of auditory cortical lesions on sound localization by the rat," J. Neurophysiol. **44**, 1161–1174.
Kelly, J. B., and Glazier, S. J. (**1978**). "Auditory cortex lesions and discrimination of spatial location by the rat," Brain Res. **145**, 315–321.
Kelly, J. B., and Masterton, B. (**1977**). "Auditory sensitivity of the albino rat," J. Comp. Physiol. Psychol. **91**, 930–936.
Kelly, J. B., and Kavanagh, G. L. (**1986**). "Effects of auditory cortical lesions on pure-tone sound localization by the albino rat," Behav. Neurosci. **100**, 569–575.
Kelly, J. B., and Judge, P. W. (**1985**). "Effects of medial geniculate lesions on sound localization by the rat," J. Neurophysiol. **53**, 361–372.
Kelly, J. B., and Sally, S. L. (**1988**). "Organization of auditory cortex in the albino rat: binaural response properties," J. Neurophysiol. **59**, 1756–1769.
Kelly, J. B., and Phillips, D. P. (**1991**). "Coding of interaural time differences of transients in auditory-cortex of rattus-norvegicus—implications for the evlolution of mammalian sound localization," Hear. Res. **55**, 39–44.
Kelly, J. B., Buckthought, A. D., and Kidd, S. A. (**1998**). "Monaural and binaural response properties of single neurons in the rat's dorsal nucleus of the lateral lemniscus," Hear. Res. **122**, 25–40.
Koay, G., Kearns, D., Heffner, H. E., and Heffner, R. S. (**1998**). "Passive sound-localization ability of the big brown bat (*Eptesicus fuscus*)," Hear. Res. **119**, 37–48.
Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157–167.
Leong, P., and Carlile, S. (**1998**). "Methods for spherical data analysis and visualization," J. Neurosci. Methods **80**, 191–200.
Li, L., and Frost, B. J. (**1996**). "Azimuthal sensitivity of rat pinna reflex: EMG recordings from cervicoauricular muscles," Hear. Res. **100**, 192–200.
Li, L., and Kelly, J. B. (**1992**). "Binaural responses in rat inferior colliculus following kainic acid lesions of the superior olive: interaural intensity difference functions," Hear. Res. **61**, 73–85.
Maki, K., and Furukawa, S. (**2005**). "Acoustical cues for sound localization by the Mongolian gerbil, *Meriones unguiculatus*," J. Acoust. Soc. Am. **118**, 872–886.
Martin, R. L., and Webster, W. R. (**1989**). "Interaural sound pressure level differences associated with sound-source locations in the frontal hemifield of the domestic cat," Hear. Res. **38**, 289–302.
McAlpine, D., and Grothe, B. (**2003**). "Sound localization and delay lines—do mammals fit the model?," Trends Neurosci. **26**, 347–350.
Middlebrooks, J. C., and Pettigrew, J. D. (**1981**). "Functional classes of neurons in primary auditory cortex of the cat distinguished by sensitivity to sound location," J. Neurosci. **1**, 107–120.
Middelbrooks, J. C., and Knudsen, E. I. (**1987**). "Changes in external ear position modify the spatial tuning of auditory units in the cat's superior colliculus," J. Neurophysiol. **57**, 672–687.
Middlebrooks, J. C., Makous, J. C., and Green, D. M. (**1989**). "Directional sensitivity of sound-pressure levels in the human ear canal," J. Acoust. Soc. Am. **86**, 89–108.
Middlebrooks, J. C., and Green, D. M. (**1990**). "Directional dependence of interaural envelope delays," J. Acoust. Soc. Am. **87**, 2149–2162.
Middlebrooks, J. C. (**1999**). "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Am. **106**, 1480–1492.
Moiseff, A. (**1989**). "Binaural disparity cues available to the barn owl for sound localization," J. Acoust. Soc. Am. **59**, 1222–1226.
Moore, D. R., and Irvine, D. R. F. (**1979**). "A developmental study of the sound pressure transformation by the head of the cat," Acta Oto-Laryngol. **87**, 434–440.
Musicant, A. D., and Butler, R. A. (**1984**). "The psychophysical basis of monaural localization," Hear. Res. **14**, 185–190.
Musicant, A. D., Chan, J. C., and Hind, J. E. (**1990**). "Direction-dependent spectral properties of cat external ear: New data and cross-species comparisons," J. Acoust. Soc. Am. **87**, 757–781.
Obrist, M. K., Fenton, M. B., Eger, J. L., and Schlegel, P. A. (**1993**). "What ears do for bats: A comparative study of pinna sound pressure transformation in chiroptera," J. Exp. Biol. **180**, 119–152.
Parsons, C. H., Lanyon, R. G., Schnupp, J. W. H., and King, A. J. (**1999**). "Effects of altering spectral cues in infancy on horizontal and vertical sound localization by adult ferrets," J. Neurophysiol. **82**, 2294–2309.
Palmer, A. R., and King, A. J. (**1985**). "A monaural space map in the guinea-pig superior colliculus," Hear. Res. **17**, 267–280.
Phillips, D. P., Calford, M. B., Pettigrew, J. D., Aitkin, L. M., and Semple, M. N. (**1982**). "Directionality of sound pressure transformation at the cat's pinna," Hear. Res. **8**, 13–28.
Rice, J. J., May, B. J., Spirou, G. A., and Young, E. D. (**1992**). "Pinna-based spectral cues for sound localization in cat," Hear. Res. **58**, 132–152.
Rife, D. D., and Vanderkooy, J. (**1989**). "Transfer-function measurement with maximum-length sequences," J. Audio Eng. Soc. **37**, 419–444.
Roth, G. L., Kochhar, R. K., and Hind, J. E. (**1980**). "Interaural time differences: Implications regarding the neurophysiology of sound localization," J. Acoust. Soc. Am. **68**, 1643–1651.
Schnupp, J. W. H., King, A. J., and Carlile, S. (**1998**). "Altered spectral localization cues disrupt the development of the auditory space map in the superior colliculus of the ferret," J. Neurophysiol. **79**, 1053–1069.
Schnupp, J. W. H., Booth, J., and King., A. J. (**2003**). "Modeling individual differences in ferret external ear transfer functions," J. Acoust. Soc. Am. **113**, 2021–2030.
Silverman, M. S., and Clopton, B. M. (**1977**). "Plasticity of binaural interaction. I. Effect of early auditory deprivation," J. Neurophysiol. **40**, 1266–1274.
Spezio, M. L., Keller, C. H., Marrocco, R. T., and Takahashi, T. T. (**2000**). "Head-related transfer functions of the rhesus monkey," Hear. Res. **144**, 73–88.
Sterbing, S. J., Hartung, K., and Hoffmann, K.-P. (**2003**). "Spatial tuning to virtual sounds in the inferior colliculus of the guinea pig," J. Neurophysiol. **90**, 2648–2659.
Tollin, D. J. (**2004**). "The development of the acoustical cues to sound location in cats," Assoc. Res. Otolaryngol. Abstr. **27**, 56.
Wesolek, C. M., Koay, G., and Heffner, H. E. (**2007**). "Laboratory rats do not use binaural time cues to localize sound," J. Acoust. Soc. Am. **121**, 3093 (Abstract).
Wiener, F. M., Pfeiffer, R. R., and Backus, A. S. N. (**1966**). "On the sound pressure transformation by the head and auditory meatus of the cat," Acta Oto-Laryngol. **61**, 255–269.
Wightman, F. L., and Kistler, D. J. (**1989**). "Headphone simulation of free-field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867.
Woodworth, R. S. (**1938**). *Experimental Psychology* (Henry Holt and Co., New York).
Wotton, J. M., Haresign, T., and Simmons, J. A. (**1995**). "Spatially dependent acoustic cues generated by the external ear of the big brown bat, Eptesicus fuscus," J. Acoust. Soc. Am. **98**, 1423–1445.
Xu, L., and Middlebrooks, J. C. (**2000**). "Individual differences in external-ear transfer functions of cats," J. Acoust. Soc. Am. **107**, 1451–1459.
Young, E. D., Rice, J. J., and Tong, S. C. (**1996**). "Effects of pinna position on head-related transfer functions in the cat," J. Acoust. Soc. Am. **99**, 3064–3076.

Distortion product otoacoustic emission fine structure is responsible for variability of distortion product otoacoustic emission contralateral suppression

Xiao-Ming Sun[a)]
Department of Communication Sciences and Disorders, Wichita State University, 1845 Fairmount St., Wichita, Kansas 67260-0075

(Received 2 November 2007; revised 17 March 2008; accepted 28 March 2008)

Alteration of the distortion product otoacoustic emission (DPOAE) level by a contralateral sound, which is known as DPOAE contralateral suppression, has been attributed to the feedback mechanism of the medial olivocochlear efferents. However, the limited dynamic range and large intra- and intersubject variabilities in the outcome of the measurement restrict its application in assessing the efferent function. In this study, the $2f_1-f_2$ DPgram was measured with a high frequency resolution in six human ears, which exhibits a fine structure with the quasiperiodic appearance of peaks and dips. In the presence of contralateral noise, the DPOAE level increased, decreased, or remained unchanged depending on the frequency. At the peaks, DPOAEs were mostly suppressed with a larger change, while those at the dips had greater variance, with increased occurrence of enhancement or no change. The difference between the peak and dip frequencies in the DPOAE-level change was significant. A switch from suppression to enhancement of the DPOAE level or vice versa with a small change in frequency was noted. These results imply that the DPOAE fine structure is a main source of the variability in DPOAE contralateral suppression measurement. The study suggests that the DPOAE contralateral suppression test would be improved if it is conducted for frequencies at major peaks of the DPOAE fine structure. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912434]

PACS number(s): 43.64.Jb, 43.64.Kc, 43.64.Ri [BLM] Pages: 4310–4320

I. INTRODUCTION

The distortion product otoacoustic emission (DPOAE) is a low-level acoustic energy emanating from the cochlea in response to a two-tone (f_1 and f_2, where $f_1<f_2$) stimulation and is recordable in the ear canal. The generation of DPOAEs has been linked to the mechanics of the cochlea and the activity of the outer hair cells (OHCs) in the amplification of basilar-membrane motion (e.g., Kim, 1980; Kemp, 1986; Brownell, 1990). DPOAE measurement provides a tool for monitoring the function of the olivocochlear (OC) efferent system. The OC bundle originates from the superior olivary complex of the brainstem and projects to the cochlea (e.g., Warr and Guinan, Jr., 1979). The nerve fibers of the neurons in the medial part of the OC system directly innervate the OHCs of the cochlea. Previous studies with a direct electrical excitation of the medial olivocochlear (MOC) nerve fibers in animals revealed the alteration in the DPOAE level (e.g., Siegel and Kim, 1982) and basilar-membrane displacement as well as the velocity response (e.g., Murugasu and Russell, 1996; Dolan *et al.*, 1997), indicating the consequence of the MOC activation. More studies demonstrated that the presence of acoustic stimulation in one ear resulted in a level change of DPOAEs and transient-evoked otoacoustic emissions (TEOAEs) (recorded with a transient sound stimulus) in the opposite ear in animals (e.g., Puel and Rebillard, 1990; Sun and Kim, 1999b) and humans (e.g., Collet *et al.*, 1990; Moulin *et al.*, 1992; Berlin *et al.*, 1993). This phenomenon has been conventionally known as DPOAE or TEOAE contralateral suppression since a level reduction was observed in the majority of cases. Investigations using a surgical section of MOC nerve fibers (e.g., Puel and Rebillard, 1990; Liberman *et al.*, 1996) and the application of antagonists of the OC efferents (e.g., Kujawa *et al.*, 1993) have suggested that DPOAE contralateral suppression is the consequence of a feedback or control mechanism of the auditory system mediated by MOC efferent neurons. Hence, the acoustically induced DPOAE suppression has been attributed to the MOC reflex.

While DPOAE contralateral suppression measurement shows the potential to be a noninvasive tool for assessing the functional status of the MOC system, its clinical application seems to be restricted due to some weaknesses. Previous studies, which were noted above, have shown that the magnitude of DPOAE contralateral suppression in humans is small, typically up to a few decibels, with relatively large intra- and intersubject variabilities. This may be partially due to the various contralateral sound-induced changes in the DPOAE level—decrease, no change, or even increase—that occurred under some stimulus conditions (e.g., Siegel and Kim, 1982; Moulin *et al.*, 1992; Williams and Brown, 1997; Sun and Kim, 1999b). Because of difficulty in interpreting the various potential outcomes, doubts have arisen regarding the clinical value of the measurement (e.g., Lisowska *et al.*, 2002).

[a)]Electronic mail: xiao-ming.sun@wichita.edu.

 0001-4966/2008/123(6)/4310/11/$23.00

The inconsistent results in DPOAE contralateral suppression measurement may be a result of the commonly observed variability in the DPOAE level with stimulus parameters. DPOAE measured at the $2f_1-f_2$ frequency as a function of stimulus frequency, which is known as a DPgram, is the most commonly used clinical protocol. The DPOAE amplitude-versus-frequency function tested with a high frequency resolution exhibits a quasiperiodic variation in level, which is characterized by consistent peaks and dips and referred to as the DPOAE fine structure. The fine structure in a DPgram has inspired extensive interest (e.g., Gaskill and Brown, 1990; He and Schmiedt, 1993; Talmadge *et al.*, 1999; Reuter and Hammershoi, 2006). It provides a clue in elucidating the generation mechanism of DPOAEs, as will be discussed later. It also motivates researchers to reexamine the validation of some conventionally utilized procedures in DPOAE measurement. Williams and Brown (1997) measured DPOAEs as a function of the f_2/f_1 ratio by using the protocol of sweeping f_1 with f_2 fixed to study the effect of contralateral broadband noise (CBBN). Most of the DPOAE amplitude-versus-f_2/f_1 functions also exhibited a fine structure. Interestingly, a trend was noted in one subject: the DPOAE level mostly decreased at peaks and increased at dips or remained unchanged. This implies that the sound-induced MOC efferent effect differentially alters DPOAEs depending on the primary frequency. A systematic investigation in this area is essential.

A recently reported study (Zhang *et al.*, 2007) extended the investigation to the DPOAE amplitude-versus-frequency function, which was measured with a frequency resolution of 8–17 points per octave, and demonstrated that contralateral suppression of DPOAEs at peak frequencies in the DPgram was significantly larger than that at other frequencies. The objective of the present study is to provide a more lucid and complete picture of the effect of the DPOAE fine structure on DPOAE contralateral suppression measurement in humans. The DPgram with a high frequency resolution is measured across a wide frequency range. The change in the DPOAE level by CBBN at a frequency corresponding to the peak in the DPOAE fine structure is compared to that at the dip frequency. The results are expected to contribute to a better understanding of DPOAE contralateral suppression and an improved application for assaying the MOC efferent function.

II. METHODS

A. Subjects

Six female adults, with ages at 19 to 25 years, participated in this study. Subject inclusion criteria included no major pathologies in either ear in their medical history, normal middle-ear function as assessed by otoscopy and tympanometry (a peak pressure of ± 25 daPa in tympanogram), and audiometric pure-tone thresholds of 20 dB hearing level (HL) or better at octave frequencies from 0.5 to 8 kHz. The study was approved by the university's Institutional Review Board. All participants signed an informed consent.

B. Apparatus and procedure

The contralateral acoustic reflex threshold (ART) to broadband noise of all the subjects was tested in both ears. A Madsen Zodiac 901 middle-ear analyzer (GN Otometrics, Denmark) was used for this procedure as well as for tympanometry mentioned above. ART, which is defined as the lowest noise level that produces a measurable change in acoustic admittance, was automatically determined by the system. The subjects' ARTs ranged from 75 to 90 dB sound pressure level (SPL) (mean: 83 dB SPL), which was employed as a reference in determining the level of CBBN presented as the contralateral elicitor sound in the nontest ear in the following experiments.

An ILO96 OAE system (Otodynamics Ltd.) was used to measure the $2f_1-f_2$ DPOAE as a function of stimulus frequency, i.e., DPgram. The stimulus tones, f_1 and f_2, were presented through the original earphones of the system with a fixed f_2/f_1 at 1.22. The levels of stimuli were $L_1=65$ and $L_2=50$ dB SPLs. Before each of the measurement stated below, a probe checkfit procedure was performed for the purpose of in-the-ear calibration. The recorded signals for each test frequency underwent an averaging process for a predetermined period of time (see the tests described below) and a spectrum analysis with a fast Fourier transform (FFT) (frequency resolution: 12.2 Hz). The amplitude and phase of DPOAEs as well as background noise level were then extracted. The background noise in the test ear was estimated from the ten FFT frequency components nearest the $2f_1-f_2$ frequency—five above and five below. The noise level represents the mean plus one standard deviation from the signal average. As a preliminary test, a DPgram for f_2 at octave frequencies from 1 to 8 kHz was measured in both ears of each subject. The ear of each subject showing higher DPOAE levels was used as the test ear. A total of four left and two right ears were selected.

As a major component of the experimental test, DPgrams were measured for the f_2 frequency in the range of 684–6201 Hz in the test ear. A protocol "microstructure" of the ILO system provided a frequency resolution of 12.2 Hz for $f_2<3$ kHz and 24.4 Hz for $f_2>3$ kHz. That is, 17 data points were collected within 196 Hz and 392 Hz frequency segments, respectively. The high-resolution DPgram for the whole frequency range was composed of 19 consecutive frequency segments. To improve the signal-to-noise ratio as well as shorten the test time, the number of signal averages was selected at 48 and 32 for the recordings, with $f_2<3$ kHz and $f_2>3$ kHz, respectively, corresponding to a recording time of approximately 3–4 s for one data point. For each 17-point frequency segment, DPOAEs were measured in the absence and presence of CBBN. The probe was not removed between all the measurements.

The DPgram with a frequency resolution of 17 points per octave was also measured in the test ears of all the subjects by using the protocol of "fine structure" provided by the system. Data were collected with and without CBBN in the f_2 frequencies from 500 to 7000 Hz. The number of signal averages was 48. DPOAEs as a function of time were measured in some subjects for some frequencies at which the fine

structure was distinct. The time resolution was approximately 0.7 s, corresponding to eight signal averages for each data point. In the 30 s DPOAE time-course measurement, the CBBN was turned on at the 11th second and off at the 20th second. For all of the tests described above, CBBN was presented at a level that was 20 dB below the contralateral ART of the test ear, which were for all subjects in the range of 55–70 dB SPL (mean: 63 dB SPL). A Madsen Aurical PC-based audiometer and an E-A-R 3A insert earphone were used for presenting the CBBN. In-the-ear calibration was performed before each test session.

Spontaneous otoacoustic emissions (SOAEs) were also measured. Five subjects had measurable SOAEs (criterion: signal-to-noise ratio >6 dB). Three of them had a few SOAEs that were above 0 dB SPL. However, only one subject (044MB-r) had strong SOAEs, which were up to 15 dB SPL, for every octave frequency. The interaction between DPOAEs and SOAEs are beyond the scope of this study. These data are not displayed here; however, some are briefly discussed in the following sections. The experiment for each subject lasted for approximately 180 min. All testing was conducted within a double-walled, sound-treated booth (Industrial Acoustics Co.).

C. Data analysis

In offline data analysis, DPOAEs were accepted as valid only when the level was 3 dB above background noise. In most cases, all data points in the recorded DPgrams, except for those at frequencies below about 700–800 Hz, met the criterion. The effect of CBBN was expressed as the change in the DPOAE level and phase in the presence of CBBN. Both were computed by subtracting the base line values from those with CBBN.

For the high-resolution DPgram, two types of data points were classified according to their positions in the fine structure: peaks and dips. A peak was defined as a data point with a DPOAE level at least 1 dB higher than those of two adjacent points that could be determined as dips on both sides. A dip was defined as a data point with a DPOAE level at least 1 dB lower than those of two adjacent points that could be determined as peaks on both sides. The rationale for applying these definitions is that, with the frequency resolution used here, the DPOAE-level variation may occur due to noise and/or measurement errors. As a result, the immediate adjacent data points of a peak in the DPgram may not be dips, and those of a dip may not be peaks. The first and last data points were excluded in the identification of peaks and dips because the reference data point was only available on one side.

The change in the DPOAE level by CBBN for the peak frequency (f_{peak}) is referred to as $\Delta L_{\text{DP,peak}}$ and that for the dip frequency (f_{dip}) is referred to as $\Delta L_{\text{DP,dip}}$. For each subject, the $\Delta L_{\text{DP,peak}}$ (or $\Delta L_{\text{DP,dip}}$) for three octave frequencies, which were 1, 2, and 4 kHz, were respectively calculated by averaging the DPOAE-level changes for three peaks (or dips) near these frequencies. Across the subjects, the mean $\Delta L_{\text{DP,peak}}$ and mean $\Delta L_{\text{DP,dip}}$ were compared for the three frequencies. For evaluating the significance of the difference, a two-way repeated analysis of variance (ANOVA) in a general linear model was performed with a fine structure (peak and dip) and frequency as the within-subject effects. A p value of <0.05 was considered significant.

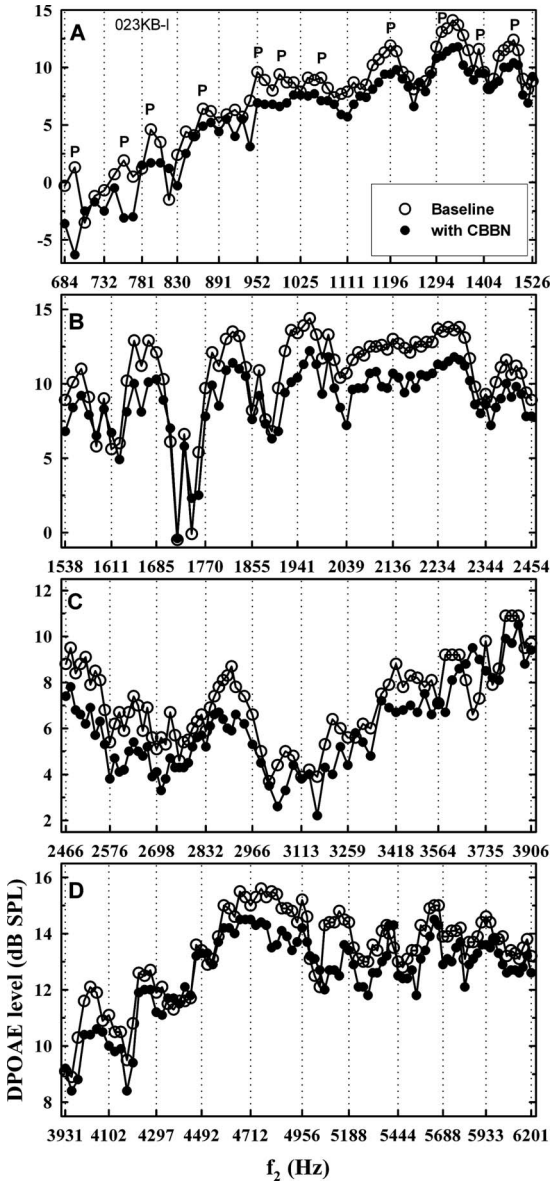


FIG. 1. Example of high-resolution $2f_1-f_2$ DPgrams measured without and with CBBN from one subject. The DPgram was broken into four portions and displayed in frequency sequence in four panels. Frequency resolution: 12.2 Hz for $f_2<3000$ Hz and 24.4 Hz for $f_2>3000$ Hz. Stimulus parameters: $f_2/f_1=1.22$, $L_1=65$ dB, and $L_2=50$ dB SPL. The peak is marked with the letter P only for a part of the base line DPgram as an example.

III. RESULTS

An example of the high-resolution DPgrams in the absence and presence of CBBN obtained from one subject, who showed a larger effect of CBBN, is displayed in Fig. 1. The base line DPOAE (open circles) shows a fine structure, quasiperiodic variation in the DPOAE level with frequency. It is obvious that the peaks of the fine structure are relatively broad, and the dips are sharp notches. Since a detailed account of the DPOAE fine structure was not the objective of the present study, only certain basic characteristics of the fine

TABLE I. Basic characteristics of the DPOAE fine structure for each subject. The right two columns list mean $\Delta L_{DP,peak}$ and $\Delta L_{DP,dip}$ by CBBN with the standard deviation (SD) in parentheses.

Subject	Prevalence of peak/dip (0. per octave)	Peak height (dB)		Dip height (dB)		$\Delta L_{DP,peak}$ (dB)	$\Delta L_{DP,dip}$ (dB)
		Mean (SD)	Range	Mean (SD)	Range		
010LS-l	16.9	3.0 (1.6)	1.3–8.3	3.0 (1.6)	1.2–7.1	−0.2 (0.5)	0.2 (0.7)
021SS-r	14.4	3.7 (1.3)	1.1–7.0	3.8 (1.9)	1.1–8.6	−0.9 (0.6)	0.1 (1.1)
023KB-l	12.6	3.4 (1.9)	1.2–9.5	3.4 (2.1)	1.3–10.7	−2.1 (1.2)	−0.4 (1.3)
032JS-l	13.2	3.4 (1.5)	1.4–6.4	3.1 (1.8)	1.4–9.1	−0.8 (0.6)	0.2 (0.8)
037VS-l	13.5	3.1 (1.5)	1.0–7.9	3.0 (1.5)	1.0–7.7	−0.9 (0.8)	0.0 (0.8)
044MB-r	10.5	6.7 (5.4)	1.5–21.3	7.0 (6.0)	1.2–28.7	−0.2 (0.8)	0.5 (1.5)

structure for this and all other subjects are listed in Table I. Valid DPOAEs were within the frequency range of around 2.7–3.1 octaves in these subjects. The prevalence of a peak/dip ranged from 10.5 to 16.9 per octave. The height of the peak was defined by the mean difference between a peak and two adjacent dips in the DPOAE level and that of the dip by the mean difference between a dip and two adjacent peaks. The mean height of the peak cross frequency was similar to that of the dip for all subjects, which could be up to 8–10 dB with a mean of approximately 3 dB except for 044MB-r.

In the presence of CBBN, DPOAEs were reduced (suppressed) in level for most frequencies. A rough inspection of the base line DPgram and that with CBBN in Fig. 1 reveals a phenomenon that DPOAE suppression is greater for the frequencies at or around peaks. The mean $\Delta L_{DP,peak}$ and $\Delta L_{DP,dip}$ across frequency of each subject are listed in Table I. It appears that the reduction in the DPOAE level by CBBN at the peak is larger than that at the dip for all of the subjects. Actually, for almost all subjects, the mean $\Delta L_{DP,dip}$ is a positive value, indicating an increase (enhancement). The change in the DPOAE level by CBBN as a function of frequency for a portion of the DPgram in Fig. 1 is illustrated in the upper panel of Fig. 2. Please note that the CBBN-induced DPOAE-level change also shows a fine structure, quasiperiodic variation with frequency. Three types of outcomes can be seen: suppression (negative values), no change (approximately zero), and enhancement (positive values). For the same frequency range, the change in the DPOAE phase by CBBN with frequency is displayed in the lower panel of Fig. 2 (to be described later).

The basic features of the DPOAE fine structure and the effect of CBBN for almost all subjects were similar, as shown in Table I. Subject 044MB-r, however, had a lower prevalence of peak/dip and a greater peak/dip height, which were up to 28 dB in some frequencies, but a smaller $\Delta L_{DP,peak}$ (negative value) and larger $\Delta L_{DP,dip}$ (positive value). This subject showed a strong and large number of SOAEs in the test ear (see Sec. II) and higher overall DPOAE levels (not shown). Subject 010LS-l also had minimal CBBN-induced DPOAE-level change, although the prevalence and height of peak/dip of this subject was similar to those of the other subjects.

The distribution of the CBBN-induced DPOAE-level change over frequency for f_{peak} and f_{dip} for the case in Fig. 1 is shown in the upper panel of Fig. 3. While the DPOAE level for all peaks was reduced (negative values) in this subject, the reduction in the DPOAE level was smaller for dips, and even no change or enhancement (positive values) occurred. It appears that the change is associated with frequency for the peak ($r=0.61$) but not for the dip ($r=0.00$). While the correlation between the DPOAE-level change and height of peak as well as dip were not strong in this case (lower panel of Fig. 3), a trend could be observed—that DPOAE suppression increased with the increase in peak height, whereas DPOAE suppression diminished and turned into enhancement with increase in dip height. All of the subjects exhibited this trend (not shown) except for 044MB-r, who had a larger variance in the distribution in the DPOAE change over frequency.

To examine the difference between f_{peak} and f_{dip} in the effect of CBBN on the DPOAE level across the subjects, the mean $\Delta L_{DP,peak}$ and $\Delta L_{DP,dip}$ for 1, 2, and 4 kHz were computed (Table II). A two-way repeated ANOVA showed a significant main effect of the fine structure (peak versus dip) ($p<0.01$) but no significant effect of frequency ($p>0.05$). No significant interaction between the effects was found ($p>0.05$). In considering the assumption that both DPOAE

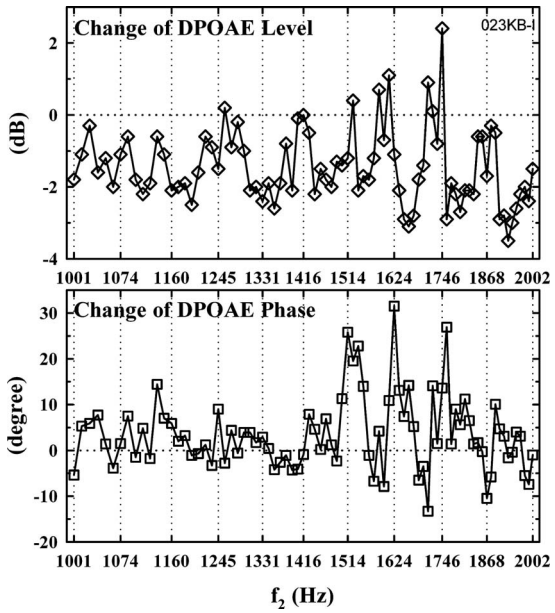


FIG. 2. Example of the effect of CBBN on DPOAEs for a portion of the high-resolution DPgram displayed in Fig. 1. Upper panel: Change in the DPOAE level for a frequency segment from 1000 to 2000 Hz. Lower panel: Change in the DPOAE phase for the same frequency segment.

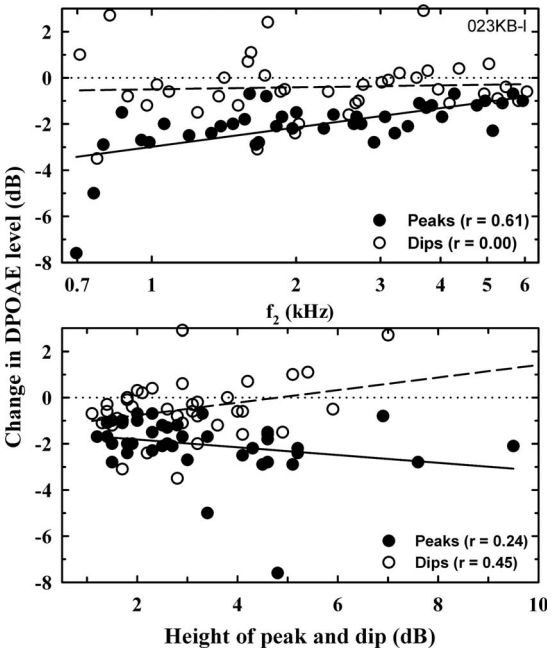


FIG. 3. Data analysis of the high-resolution DPgram shown in Fig. 1. Upper panel: Change in the DPOAE level by CBBN as a function of frequency for peak (closed circles) and dip (open circles) in the DPgram fine structure. Lower panel: Change in the DPOAE level as a function of the height of the peak (closed circles) and dip (open circles). Correlation coefficients (r values) are displayed.

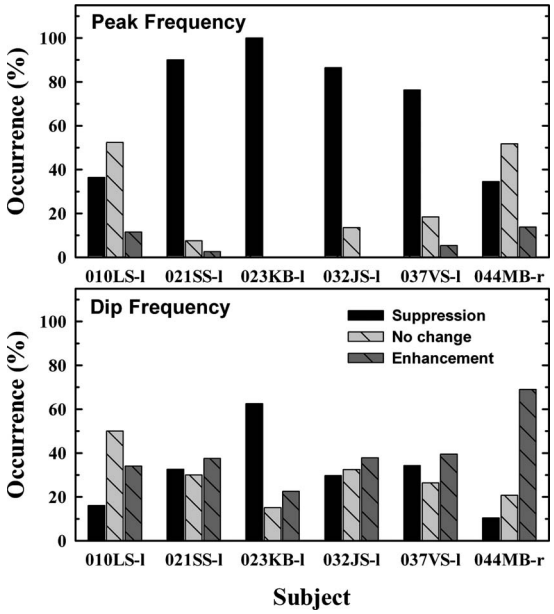


FIG. 4. Occurrence of DPOAE changes by CBBN in percentage for all of the subjects. Three possible outcomes of the change are suppression, no change, and enhancement. Upper panel: data for peak frequency in the DPOAE fine structure. Lower panel: data for dip.

suppression and enhancement are related to the MOC reflex, to be discussed later, the absolute value of the DPOAE change was employed to reexamine the data (right columns in Table II). Compared to the data in the left columns, the mean DPOAE change in absolute value becomes larger for some cases, particularly for $\Delta L_{\mathrm{DP,dip}}$. This is due to the higher occurrence of enhancement. A two-way repeated ANOVA showed a significant main effect of both the fine structure ($p<0.05$) and frequency ($p<0.05$) but no significant interaction between the two effects ($p>0.05$).

For showing the variability in the DPOAE-level change by CBBN, the occurrence of three possible outcomes at f_{peak} and f_{dip} were determined for each subject (Fig. 4). For f_{peak} (upper panel), suppression of DPOAEs was obviously dominant in four subjects. For f_{dip} (lower panel), the percentages of both no change and enhancement became considerably higher in almost all cases. Subject 010LS-l had a higher prevalence of no-change outcome for both f_{peak} and f_{dip} (Fig. 4), which resulted in much smaller mean changes, as displayed in Table I. Subject 044MB-r showed a high prevalence of no-change outcome for f_{peak} (Fig. 4), hence, smaller mean DPOAE changes (see Table I). However, the occurrence of enhancement for f_{dip} dramatically increased in this subject.

TABLE II. Effect of CBBN on the DPOAE level for f_{peak} and f_{dip}. Mean $\Delta L_{\mathrm{DP,peak}}$ and $\Delta L_{\mathrm{DP,dip}}$ across the subjects as well as absolute values are listed with the standard deviation (SD) in parentheses.

f_2 (Hz)	$\Delta L_{\mathrm{DP,peak}}$ (dB)	$\Delta L_{\mathrm{DP,dip}}$ (dB)	Absolute value	
			$\Delta L_{\mathrm{DP,peak}}$	$\Delta L_{\mathrm{DP,dip}}$
1000	−1.3 (1.0)	0.1 (0.7)	1.4 (0.8)	0.5 (0.4)
2000	−0.9 (0.8)	−0.2 (0.9)	1.0 (0.7)	0.6 (0.6)
4000	−0.6 (0.4)	0 (0.2)	0.6 (0.4)	0.1 (0.1)

Figure 5 displays several examples of the DPOAE level as a function of time measured in one subject for the frequencies at which the fine structure in the DPgram was distinct. In the 30 s DPOAE time-course measurement, CBBN was turned on during the middle 10 s. It is obvious that the DPOAE level decreases after CBBN is on for the frequencies showing relatively high base line levels (i.e., higher positions in the figure, corresponding to peaks in the DPgram, which are indicated by the dotted line) and increases for the frequencies with lower base line levels (i.e., dips). Certain variation may be evident in the time-course measures. For instance, the maximum DPOAE change after onset of CBBN seems to not sustain for most of the frequencies. Basic features of the DPOAE time course could not be derived because this measurement was not specifically designed with a finer time resolution in this study. For all tested frequencies in this measurement, the first 10 s of data (base line level) of the DPOAE time course and the middle 10 s of data (CBBN effect) were averaged, respectively, and plotted in Fig. 6, where a DPgram (upper panel) and change in DPOAE level (middle panel) are shown. A general trend that DPOAEs are suppressed for f_{peak} and enhanced for f_{dip} is notable.

More interesting in Fig. 6 (lower panel) is the change in the DPOAE phase. It can be seen that the DPOAE-phase change varies in the direction opposite to what the DPOAE level does for most frequencies (compare the middle and lower panels of Fig. 6). That is, in the presence of CBBN, the DPOAE phase increases (phase lead) as its level is suppressed and decreases (phase lag) as its level is enhanced. In the lower panel of Fig. 2, it is also apparent that the CBBN-induced DPOAE-phase change fluctuates with frequency and seems to be associated with the level change. Unfortunately,

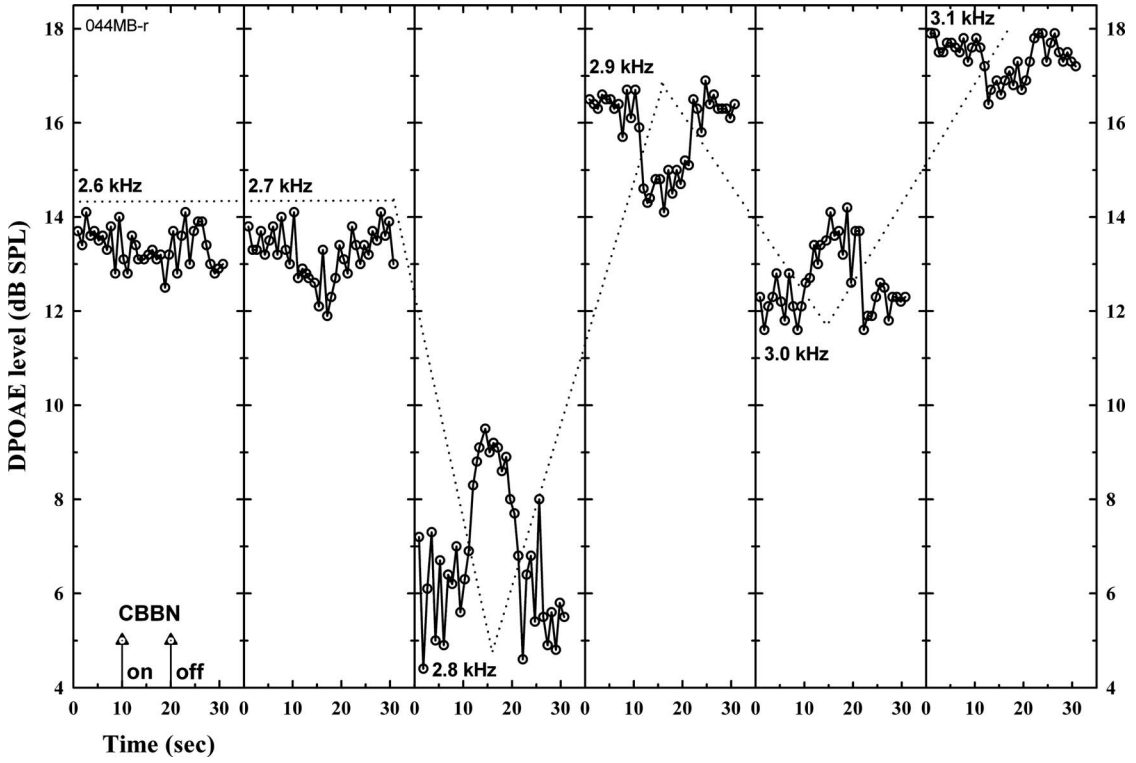


FIG. 5. Examples of the DPOAE time course (30 s) measured for a range of frequencies. The time resolution is approximately 0.7 s. Stimulus parameters: $f_2/f_1=1.22$, $L_1=65$ dB, and $L_2=50$ dB SPL. CBBN was turned on at the 11th second and off at the 20th second. The dotted line approximately indicates the base line DPOAE levels for the test frequencies.

a strong correlation between the level change and phase change was not observed for the pooled data across frequency in each subject as well as that across subjects (not shown). Since the data for only f_{peak} and f_{dip} across subjects was separately analyzed for frequency segments, some frequency-related trends can be seen in Fig. 7. For frequencies below 1 kHz, data points for f_{peak} mainly distribute in the upper left quadrant, indicating that phase lead (positive values for the ordinate) occurred as its level was suppressed (negative values for the abscissa), and data points for f_{dip} mainly distribute in the lower right quadrant, indicating that phase lag (negative values) occurred as the level was enhanced (positive values). Apparently, this tendency disappeared for higher frequencies.

Examples of the repeated measures of the DPOAE fine structure from individuals are illustrated in Fig. 8. In the subject with data in the left column, the first measure [Fig. 8, panel (A1)] reveals a large and sharp dip in the DPgram. Please note the presence of SOAEs at nearby frequencies. The change in the DPOAE level by CBBN [panel (B1)] follows the same trend as observed in other cases—suppression for f_{peak} and enhancement for f_{dip}. In the repeated measure [panel (A2)], SOAEs are not present. The DPOAE levels are significantly higher than those in the first measure, and all peaks and dips move to higher frequencies [panel (A2) versus (A1)]. However, the pattern of the effect of CBBN on the DPOAE level over frequency looks similar [panel (B2)] to that in the first measure [panel (B1)]. In another subject (right column), the two DPgrams measured on different days had minimal variation [panels (C1) and (C2)]. While the amount of CBBN-induced DPOAE-level change is not ex-

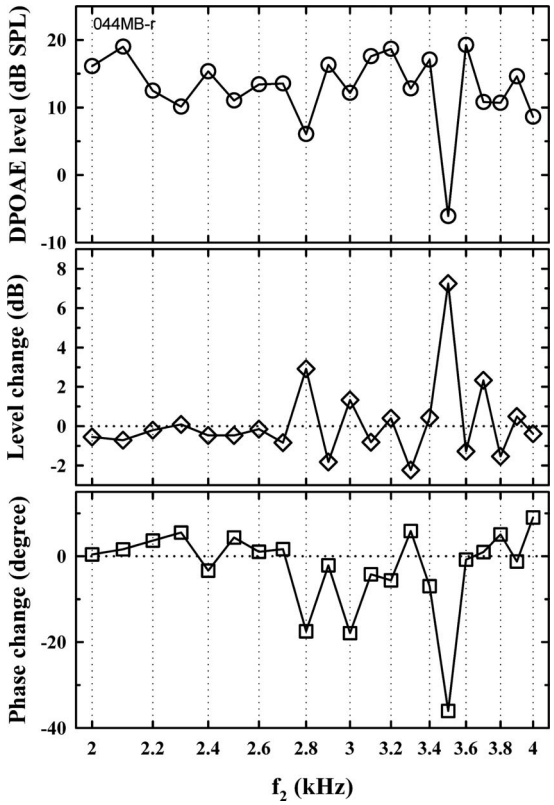


FIG. 6. Upper panel: DPOAE level derived from the DPOAE time-course measurement (see Fig. 5 and text) for a range of frequencies in one subject. Middle panel: Change in the DPOAE level by CBBN obtained from the DPOAE time-course measurement. Lower panel: Corresponding change in the DPOAE phase for the same frequencies.

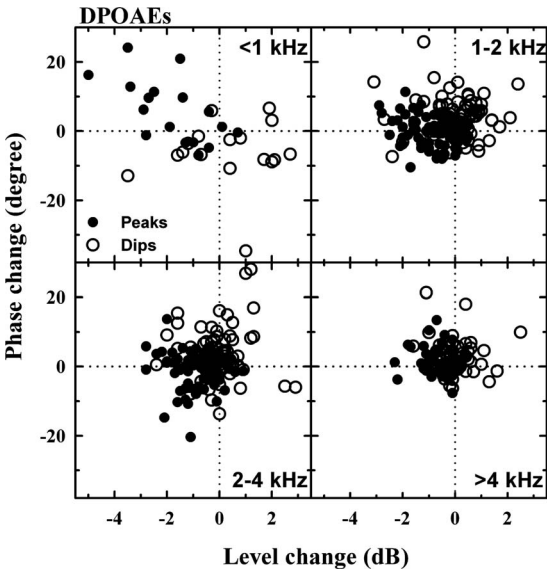


FIG. 7. Correlation analysis between change in the DPOAE level by CBBN and that of the phase across the subjects. Data for four frequency ranges are respectively displayed in four panels. Data for the peak and dip are differentially symbolized.

actly the same for the two measures [panels (D1) and (D2)], the trend of suppression in f_{peak} and enhancement in f_{dip} is still evident. At or near major dips in the DPgrams [panels (B1), (B2), (D1), and (D2)] please note that a small shift in frequency yields a large sign shift in the DPOAE change, which is from suppression to enhancement, or vice versa.

An example of the DPgram measured with the resolution of 17 points per octave is shown in Fig. 9. Although the DPgram with a reduced resolution here does not exhibit the same fine structure quality as that with a high resolution displayed in Fig. 1, the major peaks and dips are still evident (open circles). It is apparent that the DPOAE levels in the presence of CBBN were reduced (closed diamonds, below the dotted line) for f_{peak} of the DPgram. For f_{dip}, the DPOAE levels had less suppression, no change, or even enhancement (above the dotted line). The same measure for all subjects exhibited a similar trend (not shown).

IV. DISCUSSION

The major finding of the present study is that the CBBN-induced change in the DPOAE level varies with frequency depending on the DPOAE fine structure. The data also imply that the effect of CBBN on the DPOAE level is larger and less variant at the peak than that at the dip in the DPOAE fine structure.

A. DPOAE fine structure and contralateral suppression test

The $2f_1-f_2$ DPOAE measured in the ear canal is believed to be a consequence of the vector sum of two or more DPOAE components. Two major components have been widely accepted and named "wave fixed" and "place fixed" components (Kemp, 1986) since they are linked to two distinct DPOAE generation sources in the cochlea, one of which is the overlap region of the traveling waves of two

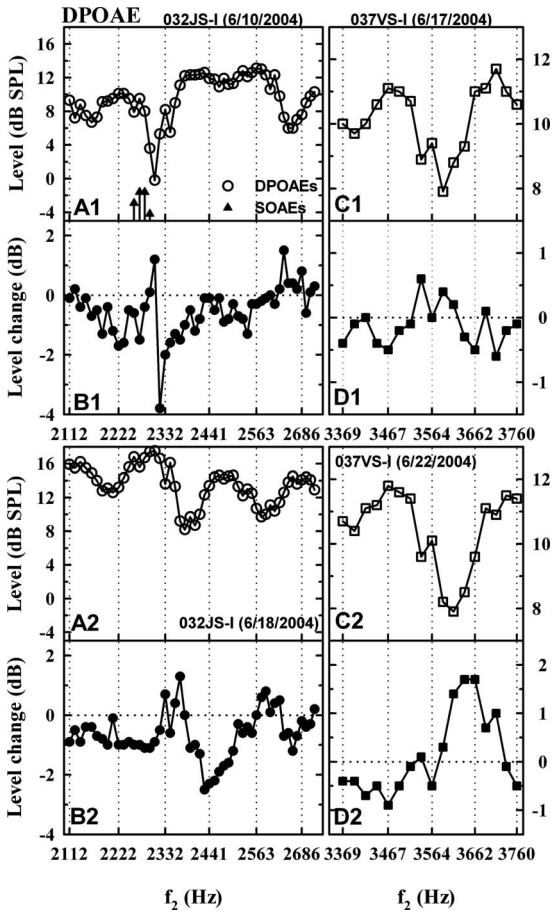


FIG. 8. Examples of DPOAE fine structure and DPOAE-level change by CBBN for a frequency segment as well as repeated recordings from two subjects (right and left columns). Stimulus parameters: $f_2/f_1=1.22$, $L_1=65$ dB, and $L_2=50$ dB SPL. First and third rows: Base line DPgrams. Second and fourth rows: Change in the DPOAE level by CBBN.

stimulus tones near the f_2 place and the other of which is the characteristic frequency place of the generated distortion product (f_{dp} place) (e.g., Kim, 1980; Kummer *et al.*, 1995; Martin *et al.*, 1999; Talmadge *et al.*, 1999). These two components are also called "nonlinear distortion" and "linear reflection" components (Shera and Guinan, Jr., 1999) since they are linked to two generation mechanisms for all types of

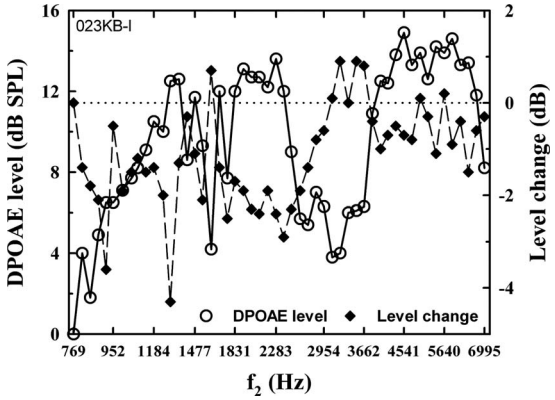


FIG. 9. Example of $2f_1-f_2$ DPgram (frequency resolution: 17/octave) and corresponding change in the DPOAE level by CBBN from one subject whose high-resolution DPgram is displayed in Fig. 1. Stimulus parameters: $f_2/f_1=1.22$, $L_1=65$ dB, and $L_2=50$ dB SPL.

otoacoustic emissions—intermodulation distortion due to cochlear nonlinearity and coherent reflection due to impedance irregularity along the basilar membrane (e.g., Kemp, 1986; Zweig and Shera, 1995).

The source of the $2f_1-f_2$ DPOAE fine structure, as measured with a fixed f_2/f_1, has been attributed to the interaction between the reflection and distortion components, as suggested by the following observations: (1) the phase of the reflection component depended on frequency, whereas the phase of the distortion component almost did not (e.g., Brown *et al.*, 1996; Knight and Kemp, 2001); (2) the fine structure could be largely flattened or removed by tonal suppression of the reflection components (e.g., Heitmann *et al.*, 1998; Kalluri and Shera, 2001; Konrad-Martin *et al.*, 2001); (3) no fine structure appeared as DPOAE was measured with a fixed DPOAE frequency paradigm (Mauermann *et al.*, 1999a); (4) the fine structure diminished or disappeared when the DPOAE frequency was located in the frequency range with hearing loss (Mauermann *et al.*, 1999b). According to the two-source or two-mechanism model, it is convincing that, as the DPOAE is measured by varying the stimulus frequency, the phase of the reflection component rapidly varies with frequency. Thus, the phase relation of the two components cyclically alters between in and out of phase. As the two components arrive at the ear canal, the difference in relative phases between them causes either constructive or destructive interaction [see Fahey and Allen (1997) for a detailed discussion]. The vector sum of these fluctuating values would result in a gradual, periodic alteration (increase and decrease) in the DPOAE level, exhibiting peaks and dips in a DPgram. Of course, the DPOAE fine structure cannot be truly periodic at the ear canal due to such factors as the interference of other possible components from the cochlea and acoustic modification from the middle ear and ear canal.

The present study demonstrates in detail the effect of the MOC reflex activated by contralateral sound on the DPOAE fine structure in a high-resolution DPgram and over a wide frequency range (Fig. 1). The CBBN-induced change in the DPOAE level varies with frequency (Fig. 2). The DPOAE level may be suppressed, enhanced, or remain unchanged depending on the DPOAE fine structure (Figs. 1, 6, and 9). In previous studies on DPOAE contralateral suppression measurement, the primary frequencies were arbitrarily determined, which is most often at octave or half-octave intervals for clinical tests. It is conceivable that all three possible outcomes can be seen across subjects. This most likely is a major cause of the small mean DPOAE suppression values and large intersubject variance observed in those studies. This implies, as well, that conclusions regarding the MOC reflex in some of those studies need to be revalidated.

B. DPOAE suppression and enhancement in contralateral stimulation

Suppression in the DPOAE level by contralateral sounds has been extensively investigated, as discussed in Sec. I. The enhancement phenomenon in the DPOAE level has been noted at certain frequencies in some cases in this line of study (e.g., Moulin *et al.*, 1992; Williams and Brown, 1997; Sun and Kim, 1999b) as well as in studies on DPOAE ipsilateral adaptation (e.g., Liberman *et al.*, 1996; Sun and Kim, 1999a). DPOAE enhancement by CBBN was also observed to last for minutes (Sun, 2003), which is a feature similar to that shown in DPOAE suppression. Furthermore, studies on direct electrical stimulation of the MOC system demonstrated the enhancement in the DPOAE level at some frequencies (e.g., Siegel and Kim, 1982; Sun and Dolan, 2000). Both DPOAE suppression and enhancement were linked to activities of the MOC efferents because the electrically induced changes could be significantly diminished by injecting strychnine, which is a known antagonist of the OC efferents (e.g., Sun and Dolan, 2000), and the acoustically induced changes could be largely eliminated by cutting the OC fibers (e.g., Kujawa and Liberman, 2001). These studies have proposed that the mechanism underlying the differential changes in the DPOAE level in contralateral stimulation is the differential effects of the MOC reflex on the two components of DPOAEs.

If the reflection component from the f_{dp} place is the major contributor of the DPOAE fine structure, as discussed above, it could be plausibly speculated that alteration of the fine structure under contralateral sound can be ascribed to a large extent to the MOC reflex effect on the reflection component of DPOAEs. The results of Williams and Brown (1997) and the present studies provide indirect evidence for testing this hypothesis. When the MOC reflex is activated, the cochlear amplifier gain at the f_{dp} place, which originated from the OHCs, would be inhibited. As a result, the reflection component would be primarily suppressed in amplitude and may also be shifted in phase. A vector sum of the two components at the ear canal is expected to modify the outcomes of the constructive and destructive interactions between them in different ways. Thus, the total DPOAE amplitude is altered depending on frequency, as shown in the present results (see Figs. 1 and 9). That is, DPOAEs would be decreased to the greatest amount at peaks in a DPgram and increased the most at dips. For some frequencies around peaks and dips, DPOAEs would be decreased or increased to various extents and, unsurprisingly, would show no change for some other frequencies (see Figs. 2 and 8). The phenomena of DPOAE suppression at f_{peak} and enhancement at f_{dip} in the presence of CBBN were clearly exhibited in DPOAE time courses (see Figs. 5 and 6). If this view is true, the results of this study present additional evidence of fully understanding the inhibitory role of the MOC system in modulating the activities of the cochlea. Both DPOAE-level suppression and enhancement by CBBN originate from the inhibitory effect of the MOC reflex on the OHC function. Accordingly, DPOAE contralateral suppression is still a valid term.

C. Peak frequency versus dip frequency in DPgram

The present study compared the difference between the peak and dip frequencies of a DPgram in DPOAE contralateral suppression. At f_{peak}, the DPOAE level is reduced by CBBN on average across frequency in all subjects (see Table I), and the occurrence of suppression is extremely high for

most subjects (see Figs. 3 and 4). In contrast, at f_{dip}, the DPOAE level shows a minimal change on average across frequency, with a positive value for most of the subjects; the occurrence of enhancement and no change dramatically increases, which gives rise to a larger variance in the outcome across the subjects. The difference between f_{peak} and f_{dip} in terms of CBBN-induced DPOAE-level change is statistically significant across subjects (see Table II). These results are in agreement with what was observed by Zhang *et al.* (2007) using DPgram measures with much lower frequency resolutions. Since both suppression and enhancement are supposed to be a result of the MOC reflex, using the absolute value in data analysis may be "fair" for f_{dip}, where there are more data with the opposite sign. In the absolute value, the change for f_{peak} was still significantly higher than that for f_{dip} across subjects. The consequence of the differential change in the DPOAE level at f_{peak} and f_{dip} is a reduced peak-to-dip height in the fine structure. This is comparable to the data from one subject in the study of Williams and Brown (1997) using the DPOAE amplitude-versus-f_2/f_1 function. The present results reveal that the effect of the MOC reflex induced by contralateral sound on the DPOAE level is larger at the peak frequency in the DPOAE fine structure than that at the dip.

A lower variability of contralateral sound-induced DPOAE-level change at f_{peak} than that at f_{dip} is also expected according to the profile of the DPOAE fine structure. Both the peak and dip in the DPOAE fine structure should not be as sharp as shown in some studies (e.g., Gaskill and Brown, 1990; He and Schmiedt, 1993). Instead, the fine structure is defined by broad peaks and narrow or sharp dips if the frequency resolution is sufficiently high, as shown in the present study (Fig. 1) and other studies (e.g., Heitmann *et al.*, 1998; Reuter and Hammershoi, 2006), and as predicted with models (e.g., Fahey and Allen, 1997; Talmadge *et al.*, 1999). It is analogous to a complex sound wave created by mixing two sine waves and gradually varying their relative phase. Consequently, in the repeated measures of the DPgram with and without contralateral sound, the likelihood of measurement error for a peak would be much lower than that for a dip. Furthermore, higher levels of DPOAEs at the peak than those at the dip may also contribute to a lower variability. It is indeed difficult to obtain a good signal-to-noise ratio at large dips in a DPgram (e.g., Reuter and Hammershoi, 2006). Reliability of DPOAEs measured at the dip in the fine structure would be lessened. However, more studies are needed for testing the variability of DPOAE contralateral suppression at f_{peak} and f_{dip}.

D. Individual differences in measures

In the present study, not all subjects displayed comparable results in all measures. One subject (010LS-l) showed much smaller mean CBBN-induced changes in the DPOAE level (see Table I) and a high prevalence of a no-change outcome (see Fig. 4) for both f_{peak} and f_{dip}, although the subject had the DPOAE fine structure similar to that of most others in terms of morphology, and prevalence and height of peaks and dips. This case might be an example of individual differences in the strength of the MOC reflex. However, individual differences in DPOAE contralateral suppression outcomes, which are indicated in some previous studies, may be due to the existence of SOAEs. One subject (044MB-r) in this study had very strong SOAEs for every octave frequency and displayed higher overall DPOAE levels and large peaks and dips but had a lower prevalence of the peak and dip (see Table I). This subject showed a smaller CBBN-induced change in the DPOAE level, a higher prevalence of the no-change outcome for f_{peak}, and a dramatically increased occurrence of enhancement for f_{dip}. In another subject (032JS-l), the appearance of SOAEs, although not very strong, also altered the DPOAE fine structure as well as outcomes of the contralateral suppression measurement (see Fig. 8). Whereas this study was not designed to examine the interaction between DPOAEs and SOAEs, these cases alert us to the possible interference of SOAEs with the DPOAE fine structure measurement and DPOAE contralateral suppression test. More cases should be investigated to characterize this effect.

E. Phase change of DPOAEs by contralateral sound

As discussed above, the change in the DPOAE level under contralateral sound is a consequence of the inhibitory effect of the MOC reflex, presumably to a larger extent on the reflection component. It is expected that the DPOAE-level change is accompanied with a change in its phase because the DPOAE phase was observed to vary with the level change as the stimulus level was varied (Fahey and Allen, 1997). A previous study verified the phase increase (phase lead) of TEOAEs under CBBN (Giraud *et al.*, 1996). Kim *et al.* (2001) observed that a decreasing ipsilateral adaptation of DPOAE was accompanied with phase lead in some cases. In another study (Sun, 2008), the DPOAE phase has been noted to increase under CBBN for 1 kHz, whereas the phase change turned negative (phase lag) with increasing frequency. However, outcomes from some studies showed that contralateral sound caused no effect on the DPOAE phase (e.g., Williams and Brown, 1997). The present study has demonstrated that the change in the DPOAE phase under CBBN varies with frequency in all the subjects and seems to be associated with the change in its level (e.g., Figs. 2 and 6). Unfortunately, a significant correlation between them was not seen. Only the data for frequencies below 1 kHz show a trend that phase change is opposite to level change in sign (Fig. 7), which is in accordance with the results from some studies just mentioned. Most likely, the truth is obscured in signal averaging due to the larger variability of the DPOAE phase compared to its level since a conventional recording approach is employed. A systematic investigation with the DPOAE time-course recording may help uncover the truth.

F. Bipolar change of the DPOAE level by contralateral sound

The phenomenon that contralateral sound-induced change in the DPOAE level switches sign, from negative (suppression) to positive (enhancement), or vice versa, was observed in DPOAE time-course recordings for some frequencies, as one of the primary tones was varied in level

(i.e., DPOAE versus L_2/L_1 ratio function) in cats (Liberman *et al.*, 1996) and in mice (Sun and Kim, 1999a), and in DPOAE level versus L_1 and L_2 function (e.g., Moulin *et al.*, 1992; Sun and Kim, 1999b). This sign-flipping occurred even as the primary level was varied by a few decibels, which is known as a bipolar change. This phenomenon has been systematically investigated by Kujawa and Liberman (2001) with ipsilateral adaptation and contralateral suppression measures in guinea pigs. Results showed that, at moderate levels and high frequencies, abrupt transitions of sign from positive to negative occurred, which were always associated with a large notch in the DPOAE versus L_2/L_1 ratio function.

In two recently reported studies (Muller *et al.*, 2005; Wagner *et al.*, 2007), a similar paradigm was employed for the test frequency corresponding to a large dip in DPgram because it was believed that the dip in the fine structure and the notch in the DPOAE versus L_2/L_1 function mentioned above were based on the same mechanism as discussed earlier regarding two DPOAE components. In both studies, the test frequency and the control frequency were determined in the DPgrams (resolution: 47 Hz) that were measured with L_2=20, 30, and 40 dB SPLs. The bipolar change in the DPOAE level by CBBN has been confirmed in humans (Muller *et al.*, 2005, Fig. 3). However, recordings with 99 L_2/L_1 combinations are apparently excessive because no large notch would occur for L_2 at higher levels so that the bipolar change would not appear. This is predictable based on the fact that the DPOAE fine structure is highly dependent on stimulus parameters, e.g., level and f_2/f_1 ratio (e.g., Gaskill and Brown, 1990; He and Schmiedt, 1993; Heitmann *et al.*, 1998). The results reported by Wagner *et al.* (2007) appeared exciting for improving the DPOAE contralateral suppression test. However, their conclusion that the MOC reflex effect on DPOAEs is largest at frequencies with distinct fine structure dips seems to be inappropriate. In their investigation, the majority of the primary level combinations for measuring the DPOAE versus L_2/L_1 function (to determine the MOC effect) and those for measuring DPgrams (to determine the test frequency) were dissimilar. This means that the dips emerging in the DPgrams may not always show up as notches in the DPOAE versus L_2/L_1 functions, as suggested in the studies just mentioned. Similarly, a measure at 4 kHz, which is the control frequency used, may represent the outcome for a flat point, a peak, or a dip in the fine structure since the primary level varies. This certainly increases the variability of the outcome, as suggested by the present study. While the protocol they used may work in a laboratory study, measurement with 121 L_2/L_1 combinations plus three high-resolution DPgram recordings is time consuming for clinical applications.

In the present study, the bipolar change in the DPOAE level by CBBN was observed in the DPOAE versus frequency function (see Figs. 2, 6, and 8). Essentially, it is a "group picture" of DPOAE suppression and enhancement by CBBN for the adjacent peak and dip in the DPOAE fine structure within a small frequency range. The bipolar change has also been noted in other forms of DPOAE functions (Sun, 2007). It is hypothesized that the bipolar change could be measured in all forms of the DPOAE functions—versus frequency, versus f_2/f_1 ratio, versus L_1 and L_2, and versus L_2/L_1 ratio—since all of these functions exhibit nonmonotonicity ("dip" or "notch"), as observed in many previous studies, and all have presumably been attributed to the same mechanism, i.e., acoustic interference between two DPOAE components. Because the bipolar change in the DPOAE level enlarges the dynamic range of the DPOAE contralateral suppression test, as shown in the aforementioned studies, a systematic investigation on this issue is necessary.

G. Implication for applications

The present results suggest that application of the DPOAE contralateral suppression test is improved if it is conducted only at or near the peak frequency in the DPOAE fine structure. The first argument is the superiority of the peak frequency over the dip frequency in displaying the change in the DPOAE level, as confirmed in the present study and discussed above. One may point out that the mean CBBN-induced change in the DPOAE level in some individuals and across subjects seems so small that the measurement may not be practical. Please note that the mean data in individuals were computed for tens of frequencies in the range of more than three octaves. Testing these many frequencies in a clinical application is certainly unnecessary. The present results also showed the trend that the change in the DPOAE level by CBBN increased with the height of the peak (see Fig. 3). This indicates that the magnitude in the DPOAE change would be boosted upward if only the frequencies for several major peaks are selected in a test. In addition, lowering the stimulus level increases the magnitude in the DPOAE-level change, which has been verified by many aforementioned studies. However, we may have to trade off the single-to-noise ratio of DPOAE measures for it, which is an issue that needs more investigations.

To improve the DPOAE contralateral suppression measurement for assaying the MOC reflex effect, an outline procedure is proposed here: measure the DPOAE-level change under CBBN in several interesting frequency segments with a high resolution and then determine the change at the largest peak or compute the mean change for all peaks in each frequency segment. An alternative would be to take the measurement in a broader frequency range with a reduced resolution, for instance, 17 points per octave, as used in this study (Fig. 9), and then determine the DPOAE change at the largest peak for each of the interesting frequency ranges. It would be helpful to count both suppression and enhancement as an index of the MOC reflex effect and apply their absolute values or the range between a maximum suppression and a maximum enhancement (bipolar change), as suggested in previous studies (e.g., Kujawa and Liberman, 2001; Muller *et al.*, 2005; Wagner *et al.*, 2007). The present study did not intend to establish norms for utilization of this measurement. Further efforts are needed to validate the recommended procedure in a larger sample of subjects.

ACKNOWLEDGMENTS

These data were collected at the University of South Alabama. The author is grateful to the subjects for their patient participation in the experiment. A preliminary account of parts of this work was presented at the Association for Research in Otolaryngology Mid-Winter Meeting, New Orleans, LA, February 2005.

Berlin, C. I., Hood, L. J., Wen, H., Szabo, P., Cecola, R. P., Rigby, P., and Jackson, D. F. (**1993**). "Contralateral suppression of non-linear click-evoked otoacoustic emissions," Hear. Res. **71**, 1–11.

Brown, A. M., Harris, F. P., and Beveridge, H. A. (**1996**). "Two sources of acoustic distortion products from the human cochlea," J. Acoust. Soc. Am. **100**, 3260–3267.

Brownell, W. E. (**1990**). "Outer hair cell electromotility and otoacoustic emissions," Ear Hear. **11**, 82–92.

Collet, L., Kemp, D. T., Veuillet, E., Duclaux, R., Moulin, A., and Morgon, A. (**1990**). "Effect of contralateral auditory stimuli on active cochlear micro-mechanical properties in human subjects," Hear. Res. **43**, 251–261.

Dolan, D. F., Guo, M. H., and Nuttall, A. L. (**1997**). "Frequency-dependent enhancement of basilar membrane velocity during olivocochlear bundle stimulation," J. Acoust. Soc. Am. **102**, 3587–3596.

Fahey, P. F., and Allen, J. B. (**1997**). "Measurement of distortion product phase in the ear canal of the cat," J. Acoust. Soc. Am. **102**, 2880–2891.

Gaskill, S. A., and Brown, A. M. (**1990**). "The behavior of the acoustic distortion product, $2f1-f2$, from the human ear and its relation to auditory sensitivity," J. Acoust. Soc. Am. **88**, 821–839.

Giraud, A. L., Perrin, E., Chery-Croze, S., Chays, A., and Collet, L. (**1996**). "Contralateral acoustic stimulation induces a phase advance in evoked otoacoustic emissions in humans," Hear. Res. **94**, 54–62.

He, N. J., and Schmiedt, R. A. (**1993**). "Fine structure of the $2f1-f2$ acoustic distortion product: Changes with primary level," J. Acoust. Soc. Am. **94**, 2659–2669.

Heitmann, J., Waldmann, B., Schnitzler, H.-U., Plinkert, P. K., and Zenner, H.-P. (**1998**). "Suppression of distortion product otoacoustic emissions (DPOAE) near $2f1-f2$ removes DP-gram fine structure-evidence for a secondary generator," J. Acoust. Soc. Am. **103**, 1527–1531.

Kalluri, R., and Shera, C. A. (**2001**). "Distortion-product source unmixing: A test of the two-mechanism model for DPOAE generation," J. Acoust. Soc. Am. **109**, 622–637.

Kemp, D. T. (**1986**). "Otoacoustic emissions, travelling waves and cochlear mechanisms," Hear. Res. **22**, 95–104.

Kim, D. O. (**1980**). "Cochlear mechanics: Implications of electrophysiological and acoustical observations," Hear. Res. **2**, 297–317.

Kim, D. O., Dorn, P. A., Neely, S. T., and Gorga, M. P. (**2001**). "Adaptation of distortion product otoacoustic emission in humans," J. Assoc. Res. Otolaryngol. **2**, 31–40.

Knight, R. D., and Kemp, D. T. (**2001**). "Wave and place fixed DPOAE maps of the human ear," J. Acoust. Soc. Am. **109**, 1513–1525.

Konrad-Martin, D., Neely, S. T., Keefe, D. H., Dorn, P. A., and Gorga, M. P. (**2001**). "Sources of distortion product otoacoustic emissions revealed by suppression experiments and inverse fast Fourier transforms in normal ears," J. Acoust. Soc. Am. **109**, 2862–2879.

Kujawa, S. G., Glattke, T. J., Fallon, M., and Bobbin, R. P. (**1993**). "Contralateral sound suppresses distortion product otoacoustic emissions through cholinergic mechanisms," Hear. Res. **68**, 97–106.

Kujawa, S. G., and Liberman, M. C. (**2001**). "Effects of olivocochlear feedback on distortion product otoacoustic emissions in guinea pig," J. Assoc. Res. Otolaryngol. **2**, 268–278.

Kummer, P., Janssen, T., and Arnold, W. (**1995**). "Suppression tuning characteristics of the $2f1-f2$ distortion-product otoacoustic emission in humans," J. Acoust. Soc. Am. **98**, 197–210.

Liberman, M. C., Puria, S., and Guinan, Jr., J. J. (**1996**). "The ipsilaterally evoked olivocochlear reflex causes rapid adaptation of the $2f1-f2$ distortion product otoacoustic emission," J. Acoust. Soc. Am. **99**, 3572–3584.

Lisowska, G., Smurzynski, J., Morawski, K., Namyslowski, G., and Probst, R. (**2002**). "Influence of contralateral stimulation by two-tone complexes, narrow-band and broad-band noise signals on the $2f1-f2$ distortion product otoacoustic emission levels in humans," Acta Oto-Laryngol. **122**, 613–619.

Martin, G. K., Stagner, B. B., Jassir, D., Telischi, F. F., and Lonsbury-Martin, B. L. (**1999**). "Suppression and enhancement of distortion-product otoacoustic emissions by interference tones above f(2). I. Basic findings in rabbits," Hear. Res. **136**, 105–123.

Mauermann, M., Uppenkamp, S., van Hengel, P. W., and Kollmeier, B. (**1999a**). "Evidence for the distortion product frequency place as a source of distortion product otoacoustic emission (DPOAE) fine structure in humans. I. Fine structure and higher-order DPOAE as a function of the frequency ratio f2/f1," J. Acoust. Soc. Am. **106**, 3473–3483.

Mauermann, M., Uppenkamp, S., van Hengel, P. W., and Kollmeier, B. (**1999b**). "Evidence for the distortion product frequency place as a source of distortion product otoacoustic emission (DPOAE) fine structure in humans. II. Fine structure for different shapes of cochlear hearing loss," J. Acoust. Soc. Am. **106**, 3484–3491.

Moulin, A., Collet, L., and Morgon, A. (**1992**). "Influence of spontaneous otoacoustic emissions (SOAE) on acoustic distortion product input/output functions: Does the medial efferent system act differently in the vicinity of an SOAE?," Acta Oto-Laryngol. **112**, 210–214.

Muller, J., Janssen, T., Heppelmann, G., and Wagner, W. (**2005**). "Evidence for a bipolar change in distortion product otoacoustic emissions during contralateral acoustic stimulation in humans," J. Acoust. Soc. Am. **118**, 3747–3756.

Murugasu, E., and Russell, I. J. (**1996**). "The effect of efferent stimulation on basilar membrane displacement in the basal turn of the guinea pig cochlea," J. Neurosci. **16**, 325–332.

Puel, J. L., and Rebillard, G. (**1990**). "Effect of contralateral sound stimulation on the distortion product $2F1-F2$: Evidence that the medial efferent system is involved," J. Acoust. Soc. Am. **87**, 1630–1635.

Reuter, K., and Hammershoi, D. (**2006**). "Distortion product otoacoustic emission fine structure analysis of 50 normal-hearing humans," J. Acoust. Soc. Am. **120**, 270–279.

Shera, C. A., and Guinan, Jr., J. J. (**1999**). "Evoked otoacoustic emissions arise by two fundamentally different mechanisms: A taxonomy for mammalian OAEs," J. Acoust. Soc. Am. **105**, 782–798.

Siegel, J. H., and Kim, D. O. (**1982**). "Efferent neural control of cochlear mechanics? Olivocochlear bundle stimulation affects cochlear biomechanical nonlinearity," Hear. Res. **6**, 171–182.

Sun, X.-M. (**2003**). "Distinctive contributions of olivocochlear efferent and middle-ear muscle reflexes to the alteration of distortion product otoacoustic emissions by contralateral noise," Assoc. Res. Otolaryngol. Abstr. **26**, p. 101.

Sun, X.-M. (**2007**). "A study on improving the DPOAE contralateral suppression test in human ears: Effect of middle-ear muscle reflex and measurement of phase change," Assoc. Res. Otolaryngol. Abstr. **30**, p. 277.

Sun, X.-M. (**2008**). "Contralateral suppression of distortion product otoacoustic emissions and the middle-ear muscle reflex in human ears," Hear. Res. **237**, 66–75.

Sun, X.-M., and Dolan, D. F. (**2000**). "Characteristics of efferent mediated enhancement of distortion product otoacoustic emissions," J. Acoust. Soc. Am. **107**, p. 2916.

Sun, X.-M., and Kim, D. O. (**1999a**). "Adaptation of $2f1-2f2$ distortion product otoacoustic emission in young-adult and old CBA and C57 mice," J. Acoust. Soc. Am. **105**, 3399–3409.

Sun, X.-M., and Kim, D. O. (**1999b**). "Effects of contralateral stimulation on distortion product otoacoustic emissions in CBA and C57 mice of various ages," Assoc. Res. Otolaryngol. Abstr. **22**, p. 97.

Talmadge, C. L., Long, G. R., Tubis, A., and Dhar, S. (**1999**). "Experimental confirmation of the two-source interference model for the fine structure of distortion product otoacoustic emissions," J. Acoust. Soc. Am. **105**, 275–292.

Wagner, W., Heppelmann, G., Muller, J., Janssen, T., and Zenner, H. P. (**2007**). "Olivocochlear reflex effect on human distortion product otoacoustic emissions is largest at frequencies with distinct fine structure dips," Hear. Res. **223**, 83–92.

Warr, W. B., and Guinan, Jr., J. J. (**1979**). "Efferent innervation of the organ of corti: Two separate systems," Brain Res. **173**, 152–155.

Williams, D. M., and Brown, A. M. (**1997**). "The effect of contralateral broad-band noise on acoustic distortion products from the human ear," Hear. Res. **104**, 127–146.

Zhang, F., Boettcher, F. A., and Sun, X.-M. (**2007**). "Contralateral suppression of distortion product otoacoustic emissions: Effect of the primary frequency in Dpgrams," Int. J. Audiol. **46**, 187–195.

Zweig, G., and Shera, C. A. (**1995**). "The origin of periodicity in the spectrum of evoked otoacoustic emissions," J. Acoust. Soc. Am. **98**, 2018–2047.

Estimates of compression at low and high frequencies using masking additivity in normal and impaired ears

Christopher J. Plack[a)]
Department of Psychology, Lancaster University, Lancaster LA1 4YF, United Kingdom

Andrew J. Oxenham and Andrea M. Simonson
Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455

Catherine G. O'Hanlon
Institute of Neuroscience and School of Psychology, Newcastle University, Henry Wellcome Building, Newcastle Upon Tyne, Tyne NE2 4HH, United Kingdom

Vit Drga
School of Psychology, University of St. Andrews, St. Mary's College, South Street, St. Andrews, Fife KY16 9JP, United Kingdom

Dhany Arifianto
Department of Psychology, Lancaster University, Lancaster LA1 4YF, United Kingdom

(Received 28 January 2008; revised 20 March 2008; accepted 21 March 2008)

Auditory compression was estimated at 250 and 4000 Hz by using the additivity of forward masking technique, which measures the effects on signal threshold of combining two temporally nonoverlapping forward maskers. The increase in threshold in the combined-masker condition compared to the individual-masker conditions can be used to estimate compression. The signal was a 250 or 4000 Hz tone burst and the maskers (M1 and M2) were bands of noise. Signal thresholds were measured in the presence of M1 and M2 alone and combined for a range of masker levels. The results were used to derive response functions at each frequency. The procedure was conducted with normal-hearing and hearing-impaired listeners. The results suggest that the response function in normal ears is similar at 250 and 4000 Hz with a mid level compression exponent of about 0.2. However, compression extends over a smaller range of levels at 250 Hz. The results confirm previous estimates of compression using temporal masking curves (TMCs) without assuming a linear off-frequency reference as in the TMC procedure. The impaired ears generally showed less compression. Importantly, some impaired ears showed a linear response at 250 Hz, providing a further indication that low-frequency compression originates in the cochlea.
 [DOI: 10.1121/1.2908297]

PACS number(s): 43.64.Kc, 43.66.Dc, 43.66.Sr [BCM] Pages: 4321–4330

I. INTRODUCTION

It is now well established that the response of the basilar membrane (BM) in the base of the cochlea [the region tuned to high characteristic frequencies (CFs)] is highly compressive. This has been confirmed by direct measurements of BM displacement or velocity in nonhuman mammals (Murugasu and Russell, 1995; Rhode, 1971; Rhode and Recio, 2000; Robles *et al.*, 1986; Ruggero *et al.*, 1997; Russell and Nilsen, 1997) and by indirect behavioral measurements in humans (Nelson *et al.*, 2001; Oxenham and Plack, 1997; Plack and O'Hanlon, 2003). The estimated compression exponent at mid to high levels is typically 0.2 or less, meaning that a 1 dB change in input sound level leads to a 0.2 dB (or less) change in BM response. However, there is some uncertainty regarding the response in the apex of the cochlea (low CFs). Direct measurements in nonhuman mammals suggest that compression may be much reduced in this region with a compression exponent of 0.5 or greater for CFs in the region of 400–800 Hz (Rhode and Cooper, 1996; Zinn *et al.*, 2000).

Most of the recent behavioral techniques have compared the effects of forward maskers at the signal frequency with those well below the signal frequency. Because the response to tones well below the CF of a given place in the base of the cochlea is linear (Ruggero *et al.*, 1997; Russell and Nilsen, 1997), it is assumed that the response to a forward masker well below the frequency of the signal can be used as a linear reference. This is the basis of the growth of masking technique, in which masker level at threshold is measured as a function of signal level or vice versa (Hicks and Bacon, 1999; Moore *et al.*, 1999; Oxenham and Plack, 1997; Rosengard *et al.*, 2005), and the temporal masking curve (TMC) technique, in which the masker level needed to just mask a low-level signal is measured as a function of the temporal gap between the masker and the signal (Lopez-Poveda *et al.*, 2003; Nelson *et al.*, 2001; Plack and Drga, 2003; Rosengard *et al.*, 2005). In both cases, the BM response to the on-frequency masker can be derived by a comparison of the on-

a)Author to whom correspondence should be addressed. Electronic mail: c.plack@lancaster.ac.uk

and off-frequency masking functions. The comparison leads to an estimate of on-frequency compression by assuming that the BM response to the off-frequency masker is linear and that all other effects (such as neural adaptation and other mechanisms involved in forward masking) are the same for both the on- and off-frequency maskers.

The first assumption, involving linear BM processing of the off-frequency masker, is problematic at low signal frequencies, which are represented near the apex of the cochlea. This is because compression appears to be less frequency selective at the apical end of the cochlea, with compression apparently being applied to a wider range of frequencies above and below the CF of a particular BM location (Plack and Drga, 2003; Rhode and Cooper, 1996). It follows that the off-frequency masker cannot be used as a linear reference at low CFs because the masker may also be compressed to some extent. To circumvent this problem, researchers have used the off-frequency TMC for a high signal frequency (e.g., 4000 Hz) as a linear reference for estimating compression from the TMC for a low on-frequency masker and signal (Lopez-Poveda *et al.*, 2003; Lopez-Poveda *et al.*, 2005; Nelson and Schroder, 2004; Plack and Drga, 2003; Williams and Bacon, 2005). When this is done, estimates of on-frequency compression are similar across a wide range of CFs. Thus, there appears to be a discrepancy between physiological measures in animals, suggesting more linear processing in the apical regions (Rhode and Cooper, 1996; Zinn *et al.*, 2000), and behavioral estimates in humans, suggesting that compression in the apex is similar to that in the base.

Possible reasons for the discrepancy include between-species differences, errors in the physiological measurements, and errors in the behavioral estimates or their underlying assumptions. In the case of the physiology, accessing the apex of the cochlea is technically very challenging and may have led to some damage before measurements were taken. In the case of the psychoacoustic estimates, a number of the assumptions underlying the estimates of compression are subject to challenge, although it is noteworthy that the compression estimates for frequencies in the base of the cochlea are relatively close to those found from direct physiological measurements.

One assumption necessary for using the TMC technique to derive estimates of compression in the apical portion of the cochlea is that the postcochlear decay of forward masking is independent of CF, so that the only process producing differences in the slopes of the TMCs across frequency is the frequency-specific cochlear response to the maskers. Stainsby and Moore (2006) have cast doubt on this assumption recently in a study on listeners with sensorineural hearing loss. Sensorineural hearing loss is often a consequence of dysfunction of the outer hair cells (OHCs) in the organ of Corti. The OHCs are involved in an "active" mechanism that effectively applies gain to stimulation at frequencies close to the CF of each place on the BM (see Yates, 1995). The gain is greatest at low levels and diminishes as the level is increased, resulting in a compressive response function (Murugasu and Russell, 1995; Robles *et al.*, 1986; Ruggero *et al.*, 1997; Yates *et al.*, 1990). OHC dysfunction leads to an increase in absolute threshold and a more linear (less compressive) response function (Ruggero and Rich, 1991; Ruggero *et al.*, 1997). Linearization has been observed in hearing-impaired human listeners by using behavioral measures (Lopez-Poveda *et al.*, 2005; Moore *et al.*, 1999; Nelson *et al.*, 2001; Oxenham and Plack, 1997; Plack *et al.*, 2004). Stainsby and Moore found that the slopes of TMCs for their hearing-impaired listeners were greater at 500 and 1000 Hz than they were at higher frequencies. However, the degree of hearing loss (40–75 dB) suggested that these listeners had lost most, if not all, of their OHC function. Stainsby and Moore argued that the steep TMCs could not be explained by greater cochlear compression at low CFs compared to high CFs. Instead, they suggested that the results could be explained if the rate of decay of forward masking were greater at low CFs. If this were true, it would imply that previous studies using TMCs have overestimated the degree of compression at low CFs since the steeper TMCs at low CFs could be caused in part by the greater rate of decay of forward masking, rather than just by cochlear compression.

A behavioral technique for measuring compression that does not depend on a comparison of the effects of on- and off-frequency maskers is the additivity of nonsimultaneous masking technique, which can involve one forward and one backward masker (Oxenham and Moore, 1995) or two forward maskers (Plack *et al.*, 2007; Plack and O'Hanlon, 2003; Plack *et al.*, 2006). In the additivity of forward masking (AFM) technique, signal threshold is measured in the presence of two temporally nonoverlapping forward maskers and compared to the threshold for each masker presented individually. Compression, as applied to the signal, will influence the amount by which the signal level at threshold increases when the effects of the two maskers are combined: The greater the compression, the more the physical signal level has to increase to produce the same change in internal excitation. There is evidence that the effects of the two maskers add linearly (Plack *et al.*, 2007; Plack *et al.*, 2006); hence, the increase in threshold in the combined case can be used to estimate auditory compression. Plack and O'Hanlon (2003) used the AFM technique to estimate compression at 250, 500, and 4000 Hz at two overall levels, although their results were slightly equivocal. The mean data showed midlevel compression exponents of 0.29 at 250 Hz and 0.34 at 500 Hz, both greater than the exponent of 0.17 at 4000 Hz. However, there was considerable variability between the listeners, such that the effect of signal frequency on compression was not significant. The first aim of the present study was to use the AFM technique to estimate the response function at 250 and 4000 Hz over a wider range of levels to provide a more rigorous test of the hypothesis that compression at low CFs is similar to compression at high CFs.

The second aim of the study was to determine if the compression observed at low CFs originates in the cochlea. For high signal frequencies, the growth of forward masking with masker level is greater for off-frequency maskers than for on-frequency maskers, implying that the on- and off-frequency maskers are compressed differently in the neural frequency channel tuned to the signal. As described above, the BM response to a forward masker well below the fre-

quency of the signal is usually assumed to be linear. Since each auditory nerve fiber responds to the activity at a single CF in the cochlea, this would seem to imply that the on-frequency compression occurs before neural transduction. It is difficult to see how the two maskers (or the off-frequency masker and the signal) could be differentially compressed by subsequent processing in the same neural frequency channel. However, there is no such differential masking growth at low CFs, so it is possible that the site of the compression observed psychophysically is postcochlear. In fact, postcochlear compression provides an alternative explanation for the results of Stainsby and Moore (2006). If a component of the compression at low CFs is postcochlear, then it should not be affected by cochlear hearing loss. Hence, the compression would still be reflected in steep TMCs. In the present study, the hypothesis was tested by using the AFM technique to estimate compression in listeners with normal hearing and listeners with sensorineural hearing loss. If listeners with low-frequency hearing loss show a linearization of the response at low CFs, then this makes a cochlear origin for the compression more likely.

II. METHOD

A. Stimuli and equipment

The sinusoidal signal had a frequency (f_s) of either 250 or 4000 Hz. The maskers were bands of noise, low pass filtered at 1 kHz (3 dB cutoff, 90 dB/octave) for the 250 Hz conditions and bandpass filtered between 2800 and 5600 Hz (3 dB cutoffs, 90 dB/octave) for the 4000 Hz conditions. For the 250 Hz conditions, the signal had a total duration of 10 ms, which consisted of 5 ms raised-cosine onset and offset ramps (no steady state). Masker 1 (M1) had a total duration of 200 ms, including 5 ms onset and offset ramps and 190 ms steady state. Masker 2 (M2) had a total duration of 10 ms, which consisted of 5 ms onset and offset ramps (no steady state). For the 4000 Hz conditions, the signal had a total duration of 4 ms, which consisted of 2 ms onset and offset ramps (no steady state). M1 had a total duration of 200 ms, including 2 ms onset and offset ramps and 196 ms steady state. M2 had a total duration of 6 ms, including 2 ms onset and offset ramps and 2 ms steady state. At both frequencies, the offset of M1 coincided with the onset of M2, and the silent interval between the end of M2 and the start of the signal (0 V points) was 4 ms. When one or the other masker was not present, it was replaced by silence of the same duration, so that the temporal relationships between the remaining stimuli remained the same. The temporal and spectral parameters were chosen based on pilot data, so that at each frequency, the two maskers (M1 and M2) would be roughly equally effective when presented at the same spectrum level. Since the masker bandwidth was much greater than that of the signal at both frequencies, it is unlikely that the signal spectral splatter provided a useful cue.

The data were collected in two different laboratories (UK and US). In both locations, the experiment was run by using custom-made software on a personal computer workstation located outside a double-walled sound-attenuating booth. For the normal ears (UK), stimuli were generated digitally and were output by using an RME Digi96/8 PAD 24 bit sound card set at a clocking rate of 48 kHz. The sound card included an antialiasing filter. The headphone output of the sound card was fed via a patch panel in the sound booth wall, without filtering or amplification, to Sennheiser HD 580 circumaural headphones. All stimuli were presented to the right ear. Listeners viewed a computer monitor through a window in the sound booth. Lights on the monitor display flashed on and off concurrently with each stimulus presentation and provided feedback at the end of each trial. Responses were recorded via a computer keyboard.

For the hearing-impaired ears (US), the stimuli were generated digitally at a clocking rate of 32 kHz and were played out via a LynxStudio LynxOne sound card at 24 bit resolution. The stimuli were passed through a programable attenuator (TDT PA4) and a headphone buffer (TDT HB6) before being fed to Sennheiser HD 580 circumaural headphones. The stimuli were presented monaurally to either the right or left ear in a double-walled sound-attenuating booth. Lights on a flat-panel monitor located inside the booth flashed on and off concurrently with each stimulus presentation and provided feedback at the end of each trial. Responses were made via the computer keyboard or mouse.

B. Procedure

The procedure was based on that described by Plack and O'Hanlon (2003). A three-interval, three-alternative, forced-choice adaptive tracking procedure was used with a 300 ms interstimulus interval. In the masking conditions, all three intervals contained either one or both maskers. One of the intervals (chosen at random) contained the signal. Threshold was determined by using a two-up one-down (masker thresholds) or a two-down one-up (signal thresholds) adaptive procedure that tracked the 70.7% correct point on the psychometric function (Levitt, 1971). In the UK setup, the step size was 4 dB up to the fourth turn point, which was reduced to 2 dB for 12 subsequent turn points. The mean level at the last 12 turn points was taken as the threshold estimate for each block of trials. At least four estimates were made for each condition and the results averaged. In the US setup, the step size was 8 dB up to the first turnpoint, which was reduced to 4 dB for the following two turn points and reduced to 2 dB for six subsequent turnpoints. The mean level at the last six turnpoints was taken as the threshold estimate for each block of trials. At least three estimates were made for each condition and the results averaged.

First, the absolute threshold for the signal in the absence of maskers was determined. The main experiment was then conducted in two phases. In phase 1, the signal was presented at a range of sensation levels, chosen separately for each listener and each frequency (limited by the need to avoid clipping when the masker level approached the maximum output of the apparatus and to avoid discomfort for listeners if levels became uncomfortably loud). At each sensation level, the signal was presented with either M1 or M2, and the masker level was varied adaptively to determine the level required to mask the signal. In this way, phase 1 generated pairs of roughly equally effective maskers for each

TABLE I. Age and audiometric thresholds (dB HL) for each of the six hearing-impaired listeners. Test ear is indicated by *.

			Frequency (Hz)					
Listener	Age	Ear	250	500	1000	2000	4000	8000
I1	57	L	35	30	30	45	50	50
		R*	30	30	25	45	45	55
I2	67	L*	35	30	30	40	50	50
		R	35	30	30	30	45	45
I3	33	L*	30	35	35	35	55	70
		R	20	25	25	25	50	60
I4	40	L*	60	60	60	60	40	5
		R	60	55	60	50	55	30
I5	77	L	40	25	20	35	55	65
		R*	40	20	25	30	55	60
I6	52	L	50	55	75	80	70	65
		R*	55	60	70	65	60	55

signal level. For some hearing-impaired listeners (I2 and I4 at 250 Hz; I1, I2, and I6 at 4000 Hz), the M2 thresholds could not be determined at the highest signal sensation level due to discomfort. In these cases, M2 was set to 60 dB spectrum level for the highest sensation level in phase 2, except for listener I6 at 4000 Hz, in which case M2 was set to 68 dB spectrum level for the highest sensation level in phase 2.

In phase 2, for each pair of equally effective maskers, the signal threshold was measured in the presence of M1 alone, M2 alone, and M1 and M2 combined. For these conditions, the masker levels were fixed and the signal level was varied adaptively to determine threshold. Thresholds were measured at the two frequencies in separate sessions. In each phase, the conditions were presented in a random order.

C. Listeners

For the UK study, three normal-hearing listeners (ages 25–34) were tested at both 250 and 4000 Hz, and an additional listener (age 26) was added to make four listeners at 4000 Hz. For the US study, six listeners with mild-moderate sensorineural hearing loss of unknown etiology were tested at both 250 and 4000 Hz. Audiometric thresholds and ages for the individual hearing-impaired subjects are provided in Table I. All listeners were given several hours of training on the tasks before data collection.

III. RESULTS AND ANALYSIS

A. Absolute thresholds

The thresholds for the signal in quiet are shown in Table II. For the normal-hearing listeners, thresholds are higher at 250 Hz than those at 4000 Hz, despite the longer-duration signal used at 250 Hz. The hearing-impaired listeners show a range of threshold elevations, relative to the normal-hearing listeners, from just 12 dB above the highest normal threshold (I5 at 250 Hz) to 58 dB above the highest normal threshold (I6 at 4000 Hz).

B. Results of phase 1

The results of phase 1 of the experiment are shown in Figs. 1 and 2. The masker spectrum level at threshold for each of the two maskers is plotted as a function of the signal sensation level. The results for the normal-hearing listeners (Fig. 1) show that in most cases, the two maskers were roughly equally effective when they had the same spectrum level, although this was not the case for listeners N1 and N2 at 4000 Hz. In these cases, M2 needed to be higher in level than M1 to mask the signal; hence, M2 was relatively less effective. The slopes of the masking functions are sometimes greater than unity at low levels, as might be expected for cases in which the masker level is higher than the signal level, and hence the masker is subject to more compression (Plack and Oxenham, 1998). For some cases at 4000 Hz, and at higher masker levels (above about 30–40 dB spectrum level, equivalent to 63–73 dB SPL overall), the slope is less than 1. This could indicate that the masker is entering the more linear high-level region of the response function that is observed in some listeners (Nelson *et al.*, 2001; Oxenham and Plack, 1997). In this case, the masker may be compressed less than the signal and, hence, the masker level at threshold grows more slowly than the signal level (Plack and Oxenham, 1998).

Some of the hearing-impaired listeners (Fig. 2) also show similar levels for the two maskers, again indicating that they were roughly equally effective at the same spectrum level. For the impaired listeners, however, the slopes of the masking functions were often close to unity at all levels,

TABLE II. Absolute thresholds for the signals (dB SPL) used in the experiment for each normal-hearing listener (N1–N4) and each hearing-impaired listener (I1–I6) at the two test frequencies.

Frequency	Listener									
(Hz)	N1	N2	N3	N4	I1	I2	I3	I4	I5	I6
250	22	35	28		48	52	54	73	47	66
4000	13	10	13	10	52	53	69	46	58	71

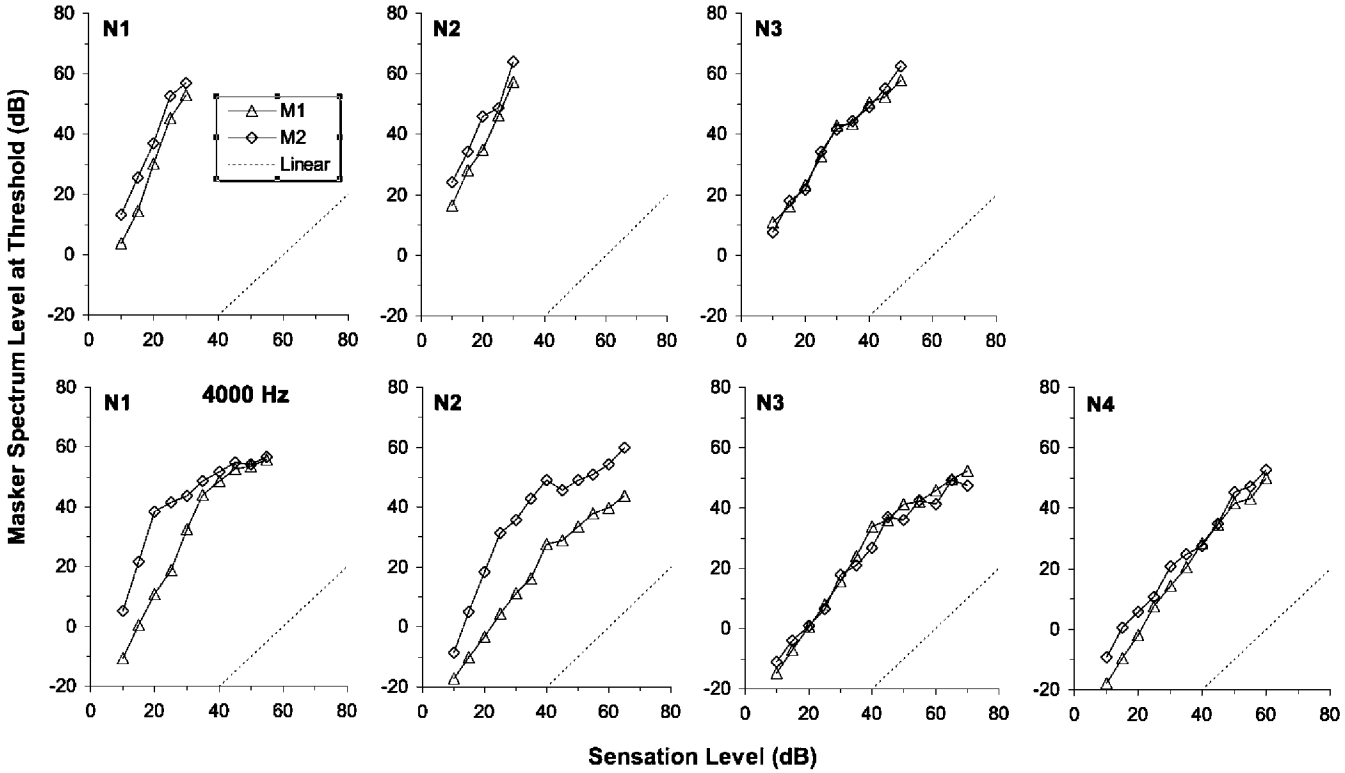


FIG. 1. The results of phase 1 of the experiment for the normal-hearing listeners. Masker spectrum level at threshold is plotted against the signal sensation level for maskers M1 (open triangles) and M2 (open diamonds). The dotted lines show linear masking growth (unity slope). The upper panels show the results for the 250 Hz conditions and the lower panels show the results for the 4000 Hz conditions.

consistent with the interpretation that compression was similar for both maskers and signal and also consistent with a more linear response function overall.

C. Results of phase 2

The results of phase 2 of the experiment are shown in Figs. 3 and 4. The signal level at threshold is plotted as a function of the sensation level of the signal used to derive the masker levels in phase 1. The figures show signal thresholds in the presence of M1 alone, M2 alone, and M1 and M2 combined. The dashed lines show predictions of the model for the combined thresholds, which will be described in the next section.

The results for the normal-hearing listeners are shown in

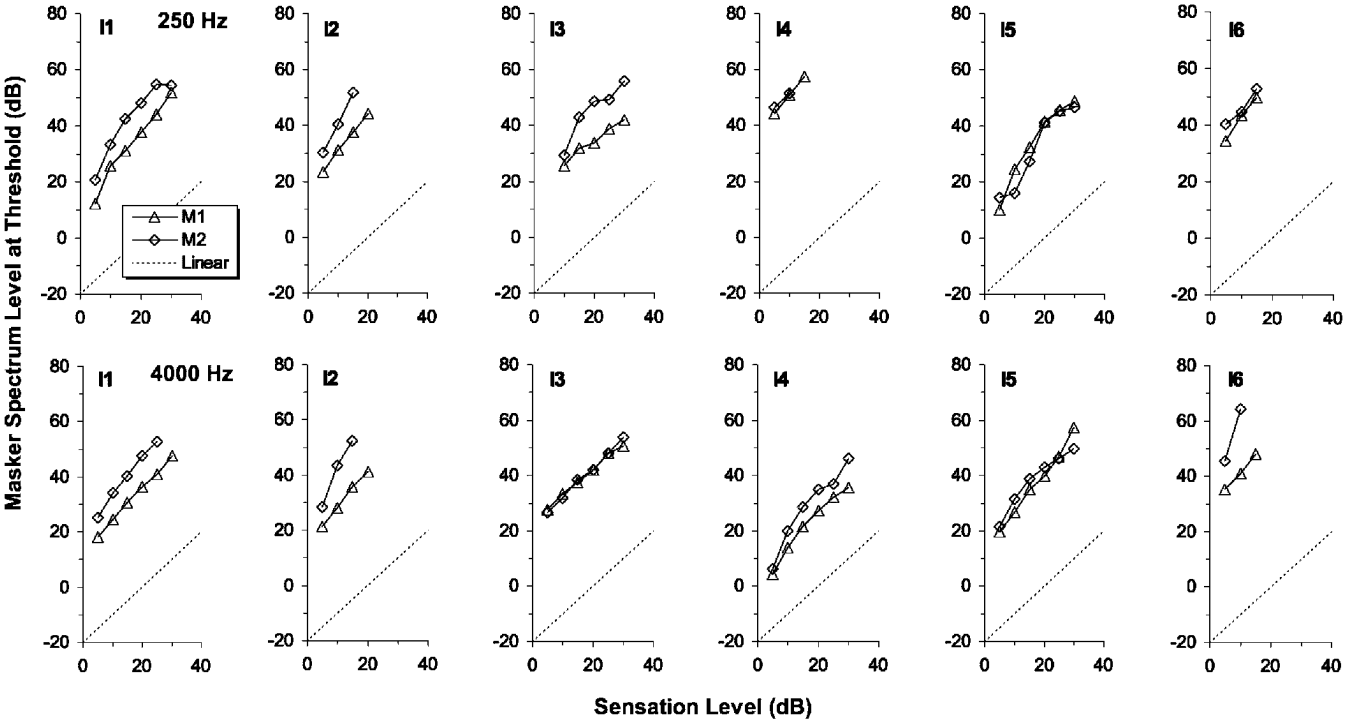


FIG. 2. As Fig. 1, except showing the results for the hearing-impaired listeners.

Fig. 3. For the single masker conditions (open symbols), thresholds are generally similar in the presence of M1 and M2, as might be expected since the levels of the maskers were chosen in phase 1 so that they were equally effective. At low levels, thresholds for the combined-masker conditions (filled circles) are only slightly above those for the single-masker conditions. Linear intensity summation predicts a 3 dB increase when two equally effective maskers are combined, and the low-level results are not far removed from this prediction. At higher levels, however, the thresholds for the combined-masker conditions are well above those for the single-masker conditions. This is indicative of a compressive system (Oxenham and Moore, 1995; Penner and Shiffrin, 1980; Plack and O'Hanlon, 2003). The amount of "excess" masking is broadly similar at the two frequencies, suggesting similar amounts of compression.

Some of the hearing-impaired listeners (Fig. 4) also show considerable excess masking, notably I1, I2, and I3 at 250 Hz and I1, I2, and I4 at 4000 Hz. However, in most cases, the effect on threshold of combining the two maskers was less for the impaired ears than for the normal ears, indicative of a less compressive, or more linear, system.

D. Response functions

Response functions were derived from the results of phase 2 by using the procedure described by Plack *et al.* (2006). The response function was modeled by a third-order polynomial in dB/dB coordinates with three parameters. In units of intensity, this becomes

$$f(x) = 10^{(a(10\log_{10}(x))^3 + b(10\log_{10}(x))^2 + c(10\log_{10}(x))/10}, \quad (1)$$

where x is the input intensity and a, b, and c are the coefficients of the polynomial. (The constant or intercept in the equation is not constrained by the data and does not affect the predictions of the model.) A separate polynomial was derived for each listener.

After preprocessing by the response function, the responses to the stimuli (maskers and signal) were assumed to add linearly. Detection of the signal was based on the signal-to-masker ratio after preprocessing and summation, and this ratio was assumed to be constant at threshold for all conditions. This means that a measure of the masking effect can be taken as the signal intensity at masked threshold after preprocessing,

$$E = f(S), \quad (2)$$

where E is the masking effect and S is the signal intensity at threshold. Assuming that the effects of two maskers sum linearly,

$$E_{M1+M2} = E_{M1} + E_{M2}, \quad (3)$$

where E_{M1}, E_{M2}, and E_{M1+M2} are the masking effects produced by M1, M2, and M1 and M2 combined. Substituting from Eq. (2) and solving for S gives

$$S_{M1+M2} = f^{-1}(f(S_{M1}) + f(S_{M2})), \quad (4)$$

where S_{M1} and S_{M2} are the signal intensities at threshold in the presence of M1 and M2, respectively, and S_{M1+M2} is the signal intensity at threshold in the presence of M1 and M2

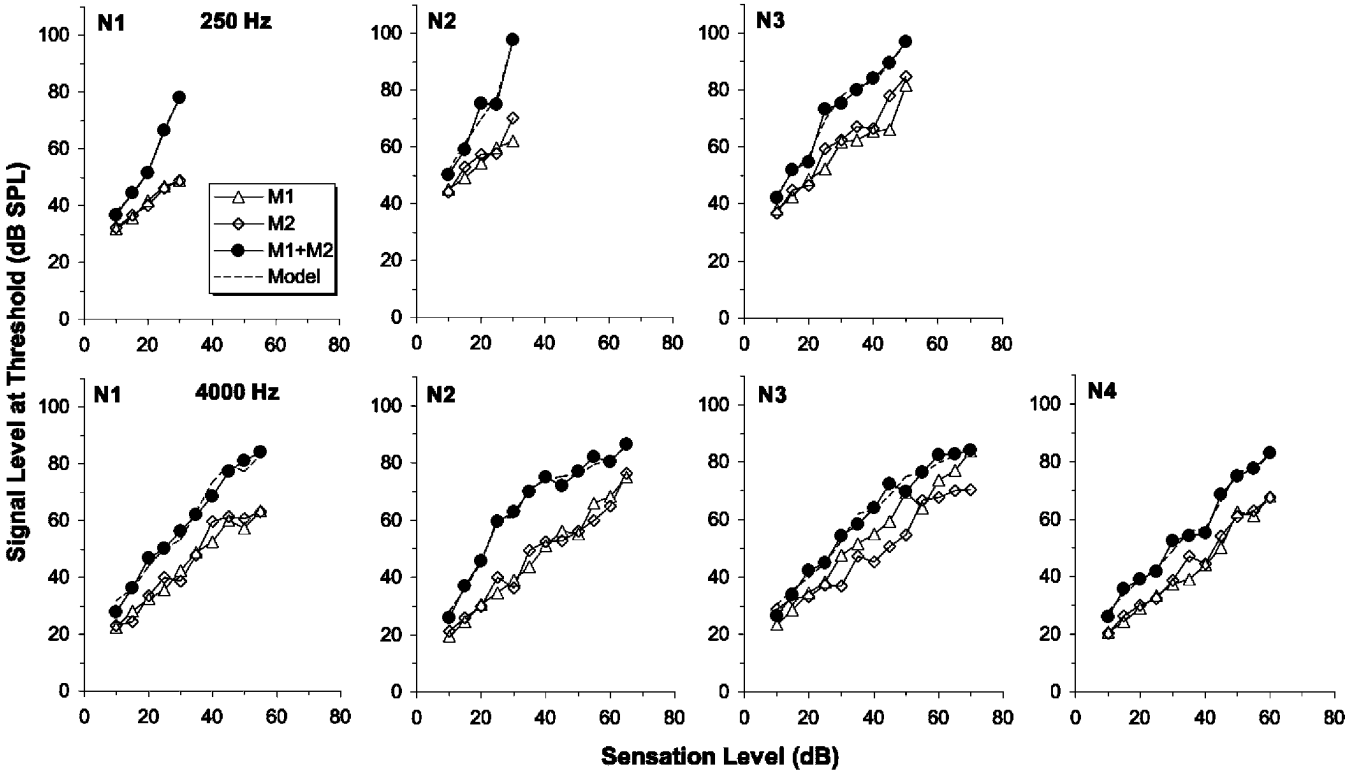


FIG. 3. The results of phase 2 of the experiment for the normal-hearing listeners. The signal level at threshold is plotted against the signal sensation level used to generate the masker levels in phase 1. The signal threshold is shown in the presence of M1 (open triangles), M2 (open diamonds), and M1 and M2 combined (filled circles). The dashed lines show predictions for the combined masker conditions generated by the model described in the text. The upper panels show the results for the 250 Hz conditions and the lower panels show the results for the 4000 Hz conditions.

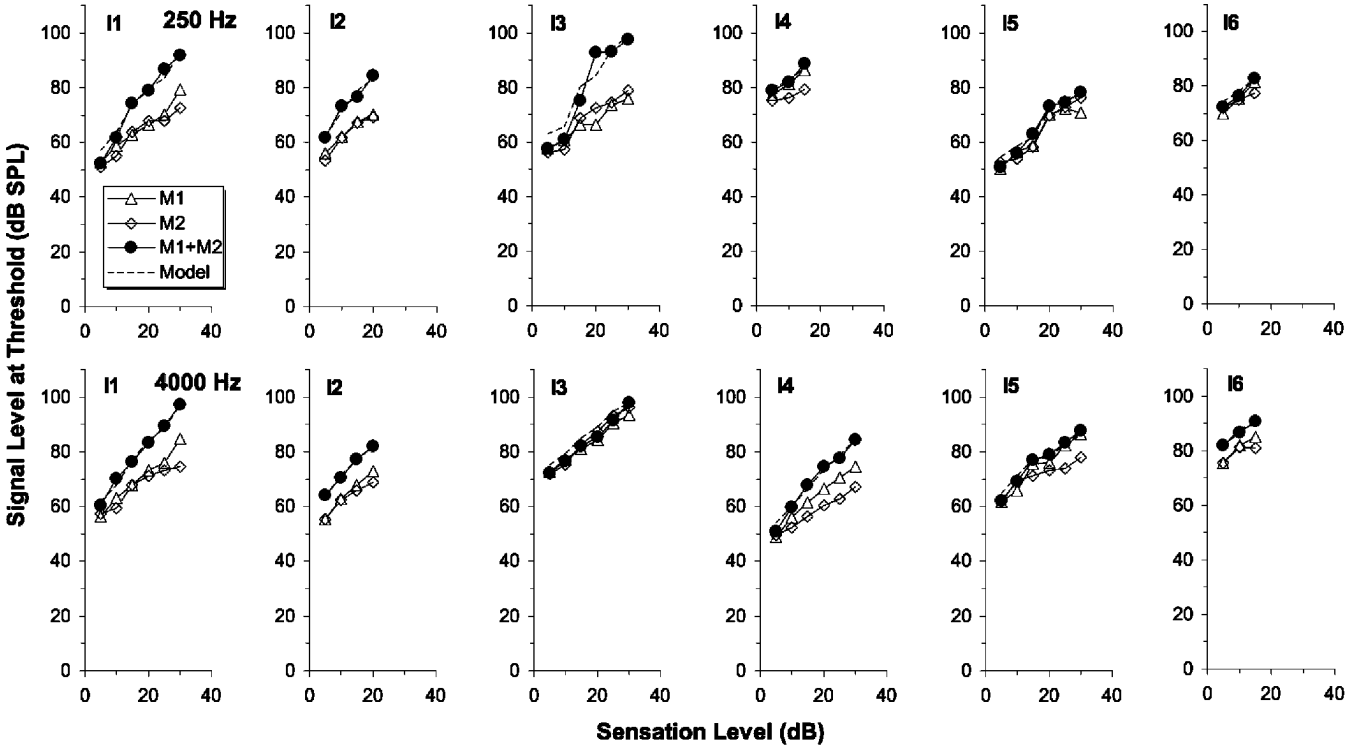


FIG. 4. As Fig. 3, except showing the results for the hearing-impaired listeners.

combined. Using this equation and Eq. (1) as the function f, the thresholds from each masker alone in phase 2 (S_{M1} and S_{M2}) were used as the input to the model, and the thresholds in the presence of both maskers (S_{M1+M2}) were predicted. For each listener and each frequency independently, the coefficients of Eq. (1) (a, b, and c) were selected to minimize the sum of the squared deviations of the model predictions from the thresholds in the combined-masker conditions, under the constraint that the differential of the function f was not permitted to be less than 0 or greater than 1 over the range of signal thresholds measured in each case. The predictions of the best-fitting models are illustrated by the dashed lines in Figs. 3 and 4. The model generally provides an accurate account of the data, suggesting that a third-order polynomial provides a good approximation to the shape of the response function.

The derived response functions are shown in Fig. 5. Calibration along the y axis (i.e., the vertical position of the functions) is arbitrary and is not constrained by the data. The response functions are calibrated to give a 100 dB output for a 90 dB SPL input. For the normal-hearing listeners, the functions are quite shallow (i.e., compressive) at both frequencies, although the functions are steeper at low levels and in some cases at high levels than at mid levels. It is interesting to note that the listeners who show a steepening in the response function at high levels (N3 at 250 Hz and N2 and N3 at 4000 Hz) also show a reduction in the slope of the masking function in phase 1 at high levels, consistent with

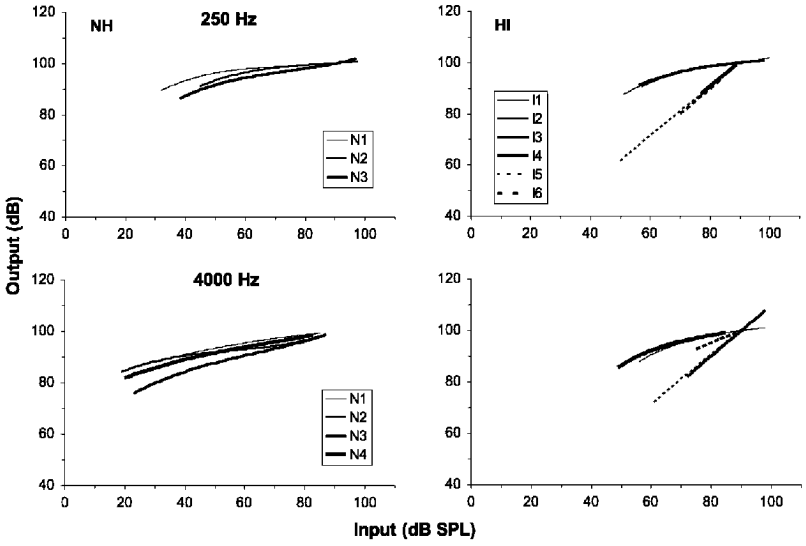


FIG. 5. Response functions derived from the results of phase 2 using the model described in the text. The left panels show the functions for the normal-hearing listeners (NH) and the right panels show the results for the hearing-impaired listeners (HI). The upper panels show the functions for the 250 Hz conditions and the lower panels show the functions for the 4000 Hz conditions.

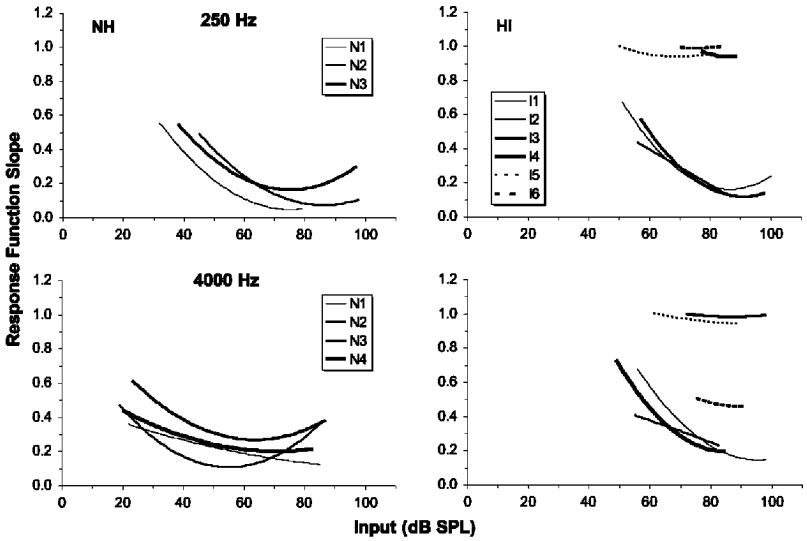


FIG. 6. The slopes of the response functions presented in Fig. 5. These plots were produced by taking the first derivative of each polynomial.

the explanation for the reduction in terms of the reduced compression of the masker compared to the signal (see Sec. III A). The exception is listener N1 at 4000 Hz, who shows a reduction in the slope of the masking function in Fig. 1 but a response function slope that decreases monotonically with level.

Figure 6 shows the slope of each response function for each listener (i.e., the derivative of each function shown in Fig. 5), which is the compression exponent value for any given input level. Between input levels of 50 and 80 dB SPL, the slopes averaged across levels and then across listeners are similar at the two frequencies for the normal-hearing listeners: 0.17 at 250 Hz and 0.21 at 4000 Hz (across-listener standard errors of 0.021 and 0.014, respectively). The *minimum* slope averaged across listeners is actually substantially less at 250 Hz (0.09) than that at 4000 Hz (0.18), although a *t* test showed that this difference is not significant. Hence, there is no evidence for a reduction in midlevel compression (which would correspond to an *increase* in the minimum slope value) at 250 Hz compared to 4000 Hz. However, the region of high compression extends to lower input levels at 4000 Hz, so the range of levels that are strongly compressed is greater at 4000 Hz than that at 250 Hz. This is consistent with some previous TMC studies that also reported a smaller range of compressed levels at low frequencies (Nelson and Schroder, 2004; Williams and Bacon, 2005).

The response functions for the hearing-impaired listeners (Fig. 5) are much more variable. Some listeners (I1, I2, and I3 at 250 Hz and I1 and I4 at 4000 Hz) show regions with compression comparable to that for the normal-hearing listeners, although the range of levels that are compressed is typically smaller. The other listeners show a more linear response with some listeners showing an almost complete loss of compression (I4, I5, and I6 at 250 Hz and I3 and I5 at 4000 Hz). The listener-frequency combinations with the highest absolute thresholds tend to show the most linear response functions (see Table II), although listener I5 at 250 Hz has a relatively low threshold but an almost linear response function. Combining the results from the normal-hearing and hearing-impaired listeners revealed a significant positive correlation between the signal absolute threshold and minimum response function slope at both 250 Hz [$r(7)=0.70$, $p=0.037$, two tailed] and 4000 Hz [$r(8)=0.64$, $p=0.048$, two tailed]. At both frequencies, high absolute thresholds are associated with more linear response functions.

To provide a summary of the response function results, the coefficients of the third-order polynomials were averaged across each listener group at each frequency. The resulting polynomials are shown in Fig. 7 together with plots of the slopes of the response functions in each case. The mean functions for the impaired listeners are clearly steeper (more linear) than those for the normal-hearing listeners at both frequencies.

IV. DISCUSSION

A. Compression at low CFs

For the normal-hearing listeners, the response functions are similar at 250 and 4000 Hz. Estimates of average midlevel compression are similar at the two frequencies, although the range of levels that are compressed is smaller at 250 Hz. This implies that the maximum gain of the active mechanism is less at 250 Hz. The estimated exponent of about 0.2 is similar to previous estimates of compression at low and high CFs using TMCs (Lopez-Poveda *et al.*, 2003; Nelson and Schroder, 2004; Plack and Drga, 2003; Williams and Bacon, 2005). As described in the Introduction, Stainsby and Moore (2006) found that TMC slopes were steeper at low frequencies than at high frequencies for listeners whose hearing loss was consistent with a complete loss of OHC function. They suggested that the postcochlear decay of forward masking may be more rapid at low than at high CFs, so that the use of an off-frequency TMC reference from a high signal frequency would produce an overestimate of low-CF compression. However, the present compression estimates, which do not depend on a linear off-frequency reference, are consistent with the previous TMC estimates. This result supports the assumption that the postcochlear decay of forward masking is similar at low and high CFs and suggests that the use of an off-frequency reference from a higher signal frequency does not lead to an overestimate of compression.

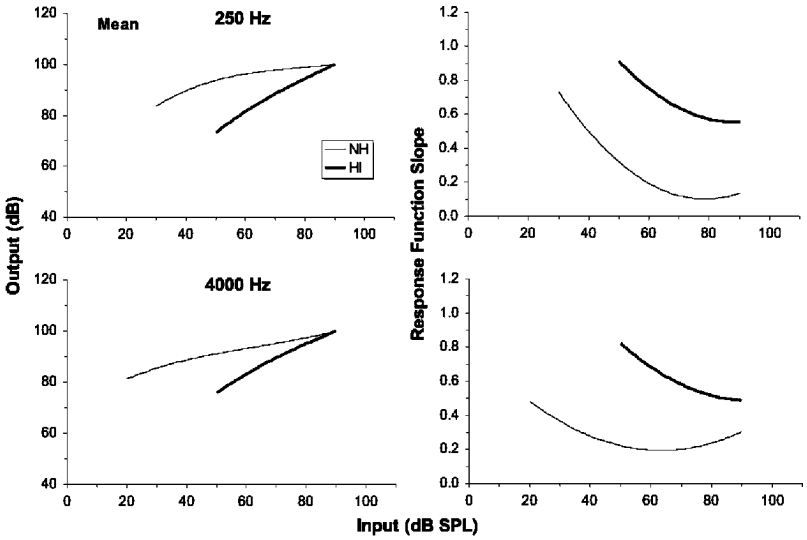


FIG. 7. The left panels show response functions generated by taking the mean of the polynomial coefficients for each group of listeners. The right panels show the slopes of these functions.

Results from a different masking procedure also suggest strong compression at low CFs. Oxenham and Dau (2001, 2004) measured the amount of masking produced by harmonic tone complex maskers as a function of the phase relation between the harmonics. When the phase relation is such that the envelope of the response at the signal place in the cochlea has a high peak factor, compression is assumed to reduce the effectiveness of the masker. This is because compression reduces the overall level of a stimulus with a high peak factor compared to a stimulus with a flat temporal envelope with the same rms level. Oxenham and Dau demonstrated substantial phase effects at 250 Hz, suggesting that compression is strong in this CF region.

The present results are also consistent with a recent study of cochlear nonlinearity using distortion-product otoacoustic emissions (DPOAEs). In the DPOAE technique, two pure tones are presented to the ear, with frequencies f_1 and f_2 ($f_2 > f_1$) and levels L_1 and L_2. The level of the $2f_1$-f_2 distortion product generated in the cochlea is measured as a function of L_2 with L_1 set to maximize the distortion product. This produces an estimate of the BM response function. Gorga *et al.* (2007) used this technique to estimate response functions at 500 and 4000 Hz. They found similar high-level slopes (approximately 0.25) at the two frequencies but that the compression region extended to lower input levels at 4000 Hz. These indirect physiological findings are similar to the present indirect psychophysical results, suggesting a common (cochlear) origin.

B. Compression in impaired ears

The results from the hearing-impaired listeners are variable, but it is clear that for most listeners, hearing loss is associated with a partial or complete linearization of the response function. This result is consistent with the findings of Oxenham and Moore (1995) using a forward and a backward masker. In their study, a linear response function was derived from the data of all three hearing-impaired listeners, all of whom had a more severe hearing loss than those tested here. In the present study, linearization was observed at 250 Hz, suggesting that compression at low CFs is susceptible to hearing loss, presumably of a cochlear origin. It seems reasonable to assume that the underlying cause of hearing loss is the same at the two frequencies and that, for the mild-to-moderate impairment of the listeners tested here, the cause is primarily the dysfunction of the OHCs (Plack *et al.*, 2004). Hence, the present results provide evidence that normal compression at low CFs is a consequence *primarily* of OHC activity, rather than compression in the inner hair cells (IHCs) (Cheatham and Dallos, 2001; Lopez-Poveda *et al.*, 2005; Patuzzi and Sellick, 1983) or postcochlear compression. Supporting this conclusion, Oxenham and Dau (2004) found reduced effects of harmonic phase at 250 Hz for hearing-impaired listeners in their study of masking by harmonic complexes. These results also suggest that cochlear dysfunction at low CFs is associated with a reduction in cochlear compression. Finally, as mentioned above, since DPOAEs are generated by cochlear processes, the results of Gorga *et al.* (2007) seem to confirm the presence of compression that is cochlear in origin.

For several of the listeners tested in the present study, the hearing loss can be categorized as mild at one or both frequencies (I1, I2, I3, and I5 at 250 Hz and I1, I2, I4, and I5 at 4000 Hz, see Table I). Plack *et al.* (2004) showed that for listeners with a mild cochlear loss, the response function shows a reduction in gain at *low levels only*, such that the compression at high levels is unaffected. The response function appears to be shifted to the right. These characteristics can be observed in some of the listeners tested here. Compared to the normal response functions, the response functions for I1, I2, and I3 at 250 Hz and I1 and I4 at 4000 Hz show a linearization at low-medium levels but comparable compression at high levels. These listeners had relatively low absolute thresholds at the specified frequencies compared to the others.

As described in the Introduction, an alternative explanation for the results of Stainsby and Moore (2006) is that a contribution to low-CF compression arises from a process that is not affected by OHC dysfunction, perhaps because it arises from some aspect of IHC function or a postcochlear neural mechanism. The effect of frequency on the TMC

slope for the three impaired listeners of Stainsby and Moore was not large. The average TMC slope ratio between 250 and 4000 Hz was 1.6. This could imply a low-CF compression component with an exponent of 0.6 that is not sensitive to OHC dysfunction. However, it is also conceivable that despite the high thresholds and low DPOAE levels, the ears tested by Stainsby and Moore had residual OHC activity at low CFs that could account for the difference in slopes.

V. CONCLUSIONS

Response functions, estimated using the AFM technique, were similar at 250 and 4000 Hz for the normal-hearing listeners with a midlevel compression exponent of about 0.2. However, compression extended over a smaller range of levels at 250 Hz, implying that the maximum gain of the active mechanism is reduced at low CFs.

Response functions for the hearing-impaired listeners were generally more linear at both frequencies, although some mildly impaired listeners showed residual high-level compression similar to that for the normal-hearing listeners.

The findings suggest that maximum compression is similar at low and high CFs in humans and are consistent with the idea that the compression at both low and high CFs is primarily cochlear in origin.

ACKNOWLEDGMENTS

The authors thank the Associate Editor and two anonymous reviewers for helpful comments on an earlier version of the manuscript. The research was supported by BBSRC (UK) Grant No. BB/D012953/1, by EPSRC (UK) Grant No. GR/N07219, and by NIH Grant No. R01 DC 03909.

Cheatham, M. A., and Dallos, P. (**2001**). "Inner hair cell response patterns: Implications for low-frequency hearing," J. Acoust. Soc. Am. **110**, 2034–2044.

Gorga, M. P., Neely, S. T., Dierking, D. M., Kopun, J., Jolkowski, K., Groenenboom, K., Tan, H., and Stiegemann, B. (**2007**). "Low-frequency and high-frequency cochlear nonlinearity in humans," J. Acoust. Soc. Am. **122**, 1671–1680.

Hicks, M. L., and Bacon, S. P. (**1999**). "Psychophysical measures of auditory nonlinearities as a function of frequency in individuals with normal hearing," J. Acoust. Soc. Am. **105**, 326–338.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Lopez-Poveda, E. A., Plack, C. J., and Meddis, R. (**2003**). "Cochlear nonlinearity between 500 and 8000 Hz in listeners with normal hearing," J. Acoust. Soc. Am. **113**, 951–960.

Lopez-Poveda, E. A., Plack, C. J., Meddis, R., and Blanco, J. L. (**2005**). "Cochlear compression between 500 and 8000 Hz in listeners with moderate sensorineural hearing loss," Hear. Res. **205**, 172–183.

Moore, B. C. J., Vickers, D. A., Plack, C. J., and Oxenham, A. J. (**1999**). "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism," J. Acoust. Soc. Am. **106**, 2761–2778.

Murugasu, E., and Russell, I. J. (**1995**). "Salicylate ototoxicity: The effects on basilar membrane displacement, cochlear microphonics, and neural responses in the basal turn of the guinea pig cochlea," Aud. Neurosci. **1**, 139–150.

Nelson, D. A., and Schroder, A. C. (**2004**). "Peripheral compression as a function of stimulus level and frequency region in normal-hearing listeners," J. Acoust. Soc. Am. **115**, 2221–2233.

Nelson, D. A., Schroder, A. C., and Wojtczak, M. (**2001**). "A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **110**, 2045–2064.

Oxenham, A. J., and Dau, T. (**2001**). "Towards a measure of auditory filter phase response," J. Acoust. Soc. Am. **110**, 3169–3178.

Oxenham, A. J., and Dau, T. (**2004**). "Masker phase effects in normal-hearing and hearing-impaired listeners: Evidence for peripheral compression at low signal frequencies," J. Acoust. Soc. Am. **116**, 2248–2257.

Oxenham, A. J., and Moore, B. C. J. (**1995**). "Additivity of masking in normally hearing and hearing-impaired subjects," J. Acoust. Soc. Am. **98**, 1921–1934.

Oxenham, A. J., and Plack, C. J. (**1997**). "A behavioral measure of basilar-membrane nonlinearity in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **101**, 3666–3675.

Patuzzi, R., and Sellick, P. M. (**1983**). "A comparison between basilar membrane and inner hair cell receptor potential input-output funcitons in the guinea pig cochlea," J. Acoust. Soc. Am. **74**, 1734–1741.

Penner, M. J., and Shiffrin, R. M. (**1980**). "Nonlinearities in the coding of intensity within the context of a temporal summation model," J. Acoust. Soc. Am. **67**, 617–627.

Plack, C. J., Carcagno, S., and Oxenham, A. J. (**2007**). "A further test of the linearity of temporal summation in forward masking," J. Acoust. Soc. Am. **122**, 1880–1883.

Plack, C. J., and Drga, V. (**2003**). "Psychophysical evidence for auditory compression at low characteristic frequencies," J. Acoust. Soc. Am. **113**, 1574–1586.

Plack, C. J., Drga, V., and Lopez-Poveda, E. A. (**2004**). "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss," J. Acoust. Soc. Am. **115**, 1684–1695.

Plack, C. J., and O'Hanlon, C. G. (**2003**). "Forward masking additivity and auditory compression at low and high frequencies," J. Assoc. Res. Otolaryngol. **4**, 405–415.

Plack, C. J., and Oxenham, A. J. (**1998**). "Basilar-membrane nonlinearity and the growth of forward masking," J. Acoust. Soc. Am. **103**, 1598–1608.

Plack, C. J., Oxenham, A. J., and Drga, V. (**2006**). "Masking by inaudible sounds and the linearity of temporal summation," J. Neurosci. **26**, 8767–8773.

Rhode, W. S. (**1971**). "Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique," J. Acoust. Soc. Am. **49**, 1218–1231.

Rhode, W. S., and Cooper, N. P. (**1996**). "Nonlinear mechanics in the apical turn of the chinchilla cochlea in vivo," Aud. Neurosci. **3**, 101–121.

Rhode, W. S., and Recio, A. (**2000**). "Study of mechanical motions in the basal region of the chinchilla cochlea," J. Acoust. Soc. Am. **107**, 3317–3332.

Robles, L., Ruggero, M. A., and Rich, N. C. (**1986**). "Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input-output functions, tuning curves, and phase responses," J. Acoust. Soc. Am. **80**, 1364–1374.

Rosengard, P. S., Oxenham, A. J., and Braida, L. D. (**2005**). "Comparing different estimates of cochlear compression in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **117**, 3028–3041.

Ruggero, M. A., and Rich, N. C. (**1991**). "Furosemide alters organ of Corti mechanics: Evidence for feedback of outer hair cells upon the basilar membrane," J. Neurosci. **11**, 1057–1067.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (**1997**). "Basilar-membrane responses to tones at the base of the chinchilla cochlea," J. Acoust. Soc. Am. **101**, 2151–2163.

Russell, I. J., and Nilsen, K. E. (**1997**). "The location of the cochlear amplifier: Spatial representation of a single tone on the guinea pig basilar membrane," Proc. Natl. Acad. Sci. U.S.A. **94**, 2660–2664.

Stainsby, T. H., and Moore, B. C. J. (**2006**). "Temporal masking curves for hearing-impaired listeners," Hear. Res. **218**, 98–111.

Williams, E. J., and Bacon, S. P. (**2005**). "Compression estimates using behavioral and otoacoustic emission measures," Hear. Res. **201**, 44–54.

Yates, G. K. (**1995**). "Cochlear structure and function," in *Hearing*, edited by B. C. J. Moore (Academic, San Diego), pp. 41–73.

Yates, G. K., Winter, I. M., and Robertson, D. (**1990**). "Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range," Hear. Res. **45**, 203–220.

Zinn, C., Maier, H., Zenner, H.-P., and Gummer, A. W. (**2000**). "Evidence for active, nonlinear, negative feedback in the vibration response of the apical region of the in-vivo guinea-pig cochlea," Hear. Res. **142**, 159–183.

Doppler-shift compensation behavior by Wagner's mustached bat, *Pteronotus personatus*

Michael Smotherman[a)]
Department of Biology, Texas A&M University, College Station, Texas 77843-3258

Antonio Guillén-Servent[b)]
Instituto de Ecología, A. C., Km. 2.5 Ctra. Antigua a Coatepec 351, Xalapa 91070, Veracruz, Mexico

(Received 28 September 2007; revised 28 March 2008; accepted 28 March 2008)

Doppler-shift compensation behavior (DSC) is a highly specialized vocal response displayed by bats that emit pulses with a prominent constant frequency (CF) component and adjust the frequency of their CF component to compensate for flight-speed induced Doppler shifts in the frequency of the returning echoes. DSC has only been observed in one member of the Neotropical Mormoopidae, a family of bats that use pulses with prominent CF components, leading researchers to suspect that DSC is a uniquely derived trait in the single species *Pteronotus parnellii*. Yet recent phylogenetic data indicate that the lineage of *P. parnellii* originates from the most basal node in the evolutionary history of the genus *Pteronotus*. DSC behavior was investigated in another member of this family, *Pteronotus personatus*, because molecular data indicated that this species stems from the second most basal node in *Pteronotus*. DSC was tested for by swinging the bats on a pendulum. *P. personatus* performed DSC as well as *P. parnellii* under identical conditions. Two other closely related mormoopids, *Pteronotus davyi* and *Mormoops megalophylla*, were also tested and neither shifted the peak frequency of their pulses. These results shed light on the evolutionary history of DSC among the mormoopids. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912436]

PACS number(s): 43.64.Tk, 43.66.Gf, 43.66.Fe, 43.66.Hg [JAS] Pages: 4331–4339

I. INTRODUCTION

Doppler-shift compensation (DSC) is a highly specialized vocal behavior exhibited by selected groups of bats that rely upon a very narrowly tuned auditory system to discriminate fine acoustic details of their prey and to navigate through dense foliage (Schnitzler, 1967; Simmons, 1974; Simmons *et al.* 1979). The echolocation pulses of Doppler-shift compensating bats are distinguished by their prominent constant-frequency (CF) components. Because their auditory systems are precisely tuned to a narrow bandwidth around the CF frequency of their outgoing pulse, flight-induced Doppler shifts in the frequency of the returning echoes are canceled out by a systematic adjustments of subsequent pulse frequencies, which serves to maintain the bandwidth of the returning echo within the range of frequencies to which their ears are most sensitive. DSC has been reported for several species of the old world families Rhinolophidae and Hipposideridae (Schuller, 1980; Habersetzer *et al.*, 1984; Neuweiler *et al.*, 1987; Hiryu *et al.*, 2005), but only one mormoopid species, *Pteronotus parnellii* (Schnitzler, 1970; Schnitzler and Henson, 1980; Gaioni *et al.*, 1990; Keating *et al.*, 1994). *P. parnellii* is one of the eight members of the family Mormoopidae, two of which are in the genus *Mormoops*, and six belong to the genus *Pteronotus* (Fig. 1). Although all of the *Pteronotus* species incorporate a CF component into their pulses, *P. parnellii* is the only member of this genus to use particularly long (≈25 ms) CF pulses and it is the only *Pteronotus* species believed to possess the narrowly tuned auditory system typical of other Doppler-shift compensating bats (Suga, 1989; Kossl *et al.*, 1999); thus its DSC performance has been considered a derived trait, probably unique among the mormoopids. Long CF pulses are not necessarily a prerequisite for DSC, however, as several species of hipposiderids emit short CF pulses similar in structure to the pulses uttered by the smaller *Pteronotus* species and perform DSC (Schuller, 1980; Habersetzer *et al.*, 1984; Hiryu *et al.*, 2005).

Recent phylogenetic evidence has indicated that *P. parnellii* stems from the most basal node in the *Pteronotus* lineage (Fig. 1) (Van Den Bussche and Weyandt, 2003), in which case it would be somewhat surprising if none of the more recently originated members of the genus maintained the ability to perform DSC despite continuing to use prominent CF components in their pulses. In light of this evidence, one member of the *Pteronotus* genus, *Pteronotus personatus*, becomes particularly interesting with regard to the evolution of DSC among the mormoopids because (1) this species stems from the second most basal node (so, after *P. parnellii*) in the *Pteronotus* lineage (Van Den Bussche and Weyandt, 2003) and (2) recent behavioral observations (Guillén-Servent, 2005) suggest that it forages in a manner similar to another Doppler-shift compensating neotropical bat, *Noctilio albiventris* (Roverud and Grinnell, 1985a; Kalko *et al.*, 1998)}. The echolocation pulses of *P. personatus* are typically 5–8 ms long and include not one but two prominent narrow bandwidth components separated by a short ~15 kHz downward FM sweep. Based on the above observations, we hypothesized that *P. personatus* possessed the

a)Author to whom correspondence should be addressed.
b)Electronic mail: antonio.guillen@inecol.edu.mx

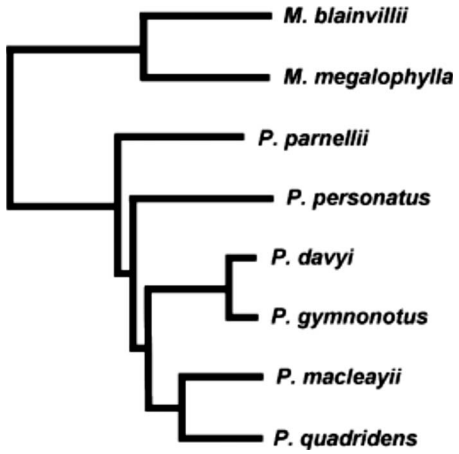


FIG. 1. Phylogenetic relationships among mormoopid bats according to the maximum likelihood analyses of the concatenated nucleotide sequences of the mitochondrial ribosomal and cytochrome *b* and the nuclear RAG-2 genes. Branch lengths are proportional to the amount of evolution in substitutions per site. All nodes had 1.00 Bayesian posterior probabilities (modified from Fig. 3 of Van Den Bussche and Weyandt, 2003).

ability to perform DSC behavior, and if so we would be able to make significant inferences about the evolution of this specialized behavior among mormoopids. We tested for DSC in *P. personatus* by swinging the bats on a pendulum similar to the previous studies of DSC in *P. parnellii* (Gaioni *et al.*, 1990). We also tested in the same setting *Pteronotus davyi*, a member of the most recent radiation within *Pteronotus*, originating from an ancestor sister to *P. personatus*, and *Mormoops megalophylla*, a species in the genus sister to *Pteronotus*, in order to get information on the evolutionary history of DSC in the Mormoopidae. Here, we present the results of our analysis of the compensation performance and other vocal characteristics of these bats, and, in particular, we provide a comparative description of DSC performance by *P. personatus* in side-by-side trials with *P. parnellii* under the same conditions. Our conclusions provide insight into the origin and evolution of DSC among the Mormoopidae.

II. METHODS

The bats used in this study were captured by a harp trap as they emerged from a cave in the state of Veracruz, Mexico (Emiliano Zapata municipality, 19° 21′N 96° 42′W) just after sunset on the evening of June 8, 2005. Ten bats of each of four species of the family Mormoopidae (Fig. 1), *Pteronotus parnellii, Pteronotus davyi, Pteronotus personatus*, and *Mormoops megalophylla*, were captured and placed in individual temporary holding cloth bags. Table I provides the representative morphometric data (body mass and forearm length) for the four species tested. All of the bats captured at this study site were females. All procedures were in accordance with the National Institutes of Health guidelines for the care and use of research animals, and were preapproved by both institutional animal care and use committees. Upon completion of the experiments, animals were released unharmed at the site where they were captured before midnight of the same evening. To test for DSC behavior, a pendulum was constructed of heavy-duty PVC irrigation pipe and placed at a distance of roughly 10 m from the entrance to the cave. The pendulum was positioned to swing toward a large concrete wall intended to serve as a target. The arm of the pendulum was 3.05 m long and swung through a maximum arc of approximately 100°, reaching a maximum velocity of 6 m/s. At this velocity, the maximum Doppler shifts in the echo frequencies ranged from 2.0 to 2.95 kHz, depending on the species-specific frequencies and bandwidths of the emitted pulses. Bats were secured facing forward in a soft foam body mold that was then secured within a hard plastic box attached to the end of the pendulum arm. Each bat was swung on the pendulum ten times, and each swing consisted of two to three forward and reverse cycles, although only the vocalization data from the first cycle of all ten swings were included in this analysis. All four of the species tested here continued to spontaneously vocalize at high rates while being restrained in the pendulum; a few individuals of the species *Pteronotus davyi* stopped calling after their first swing on the pendulum and these bats were released and replaced by other more vocal representatives of their species.

An externally polarized condenser microphone (Avisoft Bioacoustic, Berlin Germany, model CM16) facing the bat was attached to a rod extending approximately 10 cm in front of and 5 cm above the head of the bat. The frequency response of the microphone spans 10–200 kHz and is flat ±3 dB over the range of 30–140 kHz (manufacturer's specifications). The microphone recorded both the bat vocalizations and resulting echoes reflected off the ground during the swing, which made it possible to use pulse-echo time disparities in the recording to infer the position of the pendulum arm at each vocalization. Recorded vocalizations were digitized at 250 kHz sampling rate using the Avisoft UltrasoundGate hardware (Avisoft Bioacoustics, model 116-200) attached to a personal computer running the accompanying AVISOFT-RECORDER software v. 2.9. Data were analyzed offline using the software AVISOFT-SASLAB PRO V. 4.3. For each

TABLE I. Body size and basic echolocation call parameters among four mormoopids.

	M. megalophylla	*P. parnellii*	*P. personatus*	*P. davyi*
Weight (g)	14.9 ± 1.2	14.6 ± 1.7	7.4 ± 1.2	7.4 ± 1.1
Forearm length (mm)	55.8 ± 1.4	57.0 ± 1.3	42.7 ± 1.2	44.0 ± 2.0
CF2[a] frequency (kHz)	53.9 ± 0.88	64.9 ±	85.1 ± 1.3	73.6 ± 2.0
Duration (ms)	5.1 ± 1.1	19.7 ± 6.2	4.8 ± 0.9	4.9 ± 0.6
Bandwidth (kHz)	6.5 ± 1.9	10.7 ± 1.8	15.1 ± 1.5	16.2 ± 1.8

Since the echolocation calls of *M. megalophylla* do not exhibit a true CF, values presented as CF2 frequency measurements refer to the frequency of the loudest portion of their quasi-CF echolocation calls (F_{peak}) for this species, and are provided as a point of comparison across the genus. Calls were recorded from stationary bats.

bat, spectral parameters of the echolocation pulses, including the average value of the CF component of the dominant second harmonic (CF2), the bandwidth of the terminal FM (tFM) component, and for *P. personatus* and *P. davyi*, the average value of the terminal CF (tCF2) component, were taken from a spectrogram created with a 1024-point fast Fourier transform (FFT). Since the echolocation pulses of *Mormoops megalophylla* may exhibit a quasi-CF component (in search phase) but never a true CF component, we defined the frequency of the loudest portion of its vocalization as identified in the magnitude power spectrum as the peak frequency of the second harmonic (F_{peak}), and the maximum (highest starting) and minimum (lowest ending) frequencies (F_{max} and F_{min}) were defined as the frequencies of the upper and lower bandwidths of the power spectrum surrounding the F_{peak} of the vocalization at −40 dB relative to the loudness of the peak frequency. For the purpose of comparisons across species, we speculate that the quasi-CF component represented by F_{peak} in M. megalophylla's search phase calls is ancestral to the CF2 component that characterizes the echolocation calls of the entire *Pteronotus* genus, and therefore this measurement is the most appropriate value to compare to the CF2 of the three *Pteronotus* species. For the tFM sweep bandwidth, we measured the bandwidth of the power spectrum extending below the CF2 at −40 dB relative to the loudness of the CF2. For temporal analyses, pulse durations and intervals were measured from a 256-point FFT. For automated measurements of pulse duration and intervals, the thresholds for pulse onset and offset were defined as the time points at which the rising and falling amplitudes of the pulse passed a value −20 dB relative to the peak pulse amplitude. Initial pulse values prior to being swung on the pendulum were calculated from 1 min of pulses (typically about 100 pulses) recorded from the restrained bat while the pendulum hung straight down in a stationary position. For each of the species described here, echolocation pulses consisted of multiple prominent harmonics; however, the second harmonic was the dominant harmonic for all four bats, and therefore the measurements presented for comparison were taken from the second harmonic. Pulse amplitudes on the pendulum varied from 85 to 115 dB SPL. Statistical analyses were performed using the commercial statistical software package SIGMASTAT V. 3.1 (Systat Software, Inc.). For statistical comparisons, either a paired t-test or a nonparametric one-way repeated measure ANOVA was used to assess the significance of changes in pulse parameters between data sets. Data are presented as means ± SD unless indicated otherwise.

To quantify each species' DSC behavior, ten bats of each of the four species were swung on the pendulum ten times. For each of the ten swings, the CF2 (or for *M. megalophylla* the peak frequency) of the last pulse occurring prior to when the pendulum was released was compared to the lowest recorded CF2 values taken from pulses emitted while the pendulum was traveling at its fastest velocity, i.e., near the halfway point of the forward swing, as the bat passed closest to the ground. From these measurements, the maximum observed change in CF2 frequency on each swing was determined for each animal.

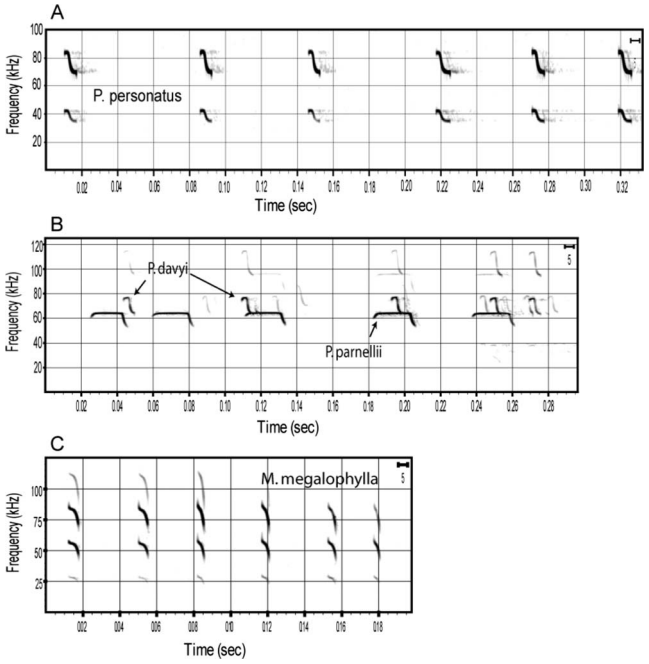


FIG. 2. Spectrograms of representative echolocation pulse sequences for the four species of bats in flight near the site of capture, including examples of (A) *Pteronotus personatus*, (B) *Pteronotus davyi* flying alongside *Pteronotus parnellii*, and (C) *Mormoops megalophylla* flying past the corner of a large brick wall.

Representative recordings of echolocation pulses emitted while the bats were in free flight were obtained by positioning a microphone 1 m above ground, facing upwards at a 45° angle, approximately 10 m away from the cave opening. Although each of the four species tended to follow a slightly different trajectory upon emerging from the cave, they all stayed with 2–3 m off the ground as they passed over our recording equipment.

III. RESULTS

A. A comparison of the echolocation pulse structures of all four species

Figure 2 illustrates the orientation sounds emitted by the four species of bats included in this study in free flight as they exited the cave. Since previous reports have provided reliable descriptions of the echolocation pulse structures of flying *Pteronotus davyi, Pteronotus personatus, and Pteronotus parnellii* (Griffin and Novick, 1955; Novick and Vaisnys, 1964; O'Farrell and Miller, 1997; Ibanez *et al.*, 1999; Macias and Mora, 2003), we restrict our description here to the vocal behavior of these animals on the pendulum. *P. parnellii* was included in the study because its DSC behavior on a pendulum is already well documented and thus could serve as a point of reference between these and prior experiments (Gaioni *et al.*, 1990; Keating *et al.*, 1994). We have included additional details regarding the vocal characteristics of *Mormoops megalophylla* because the vocal characteristics of this species are sparsely represented in the literature. Table I provides a comparison of basic call features for all four species of bats.

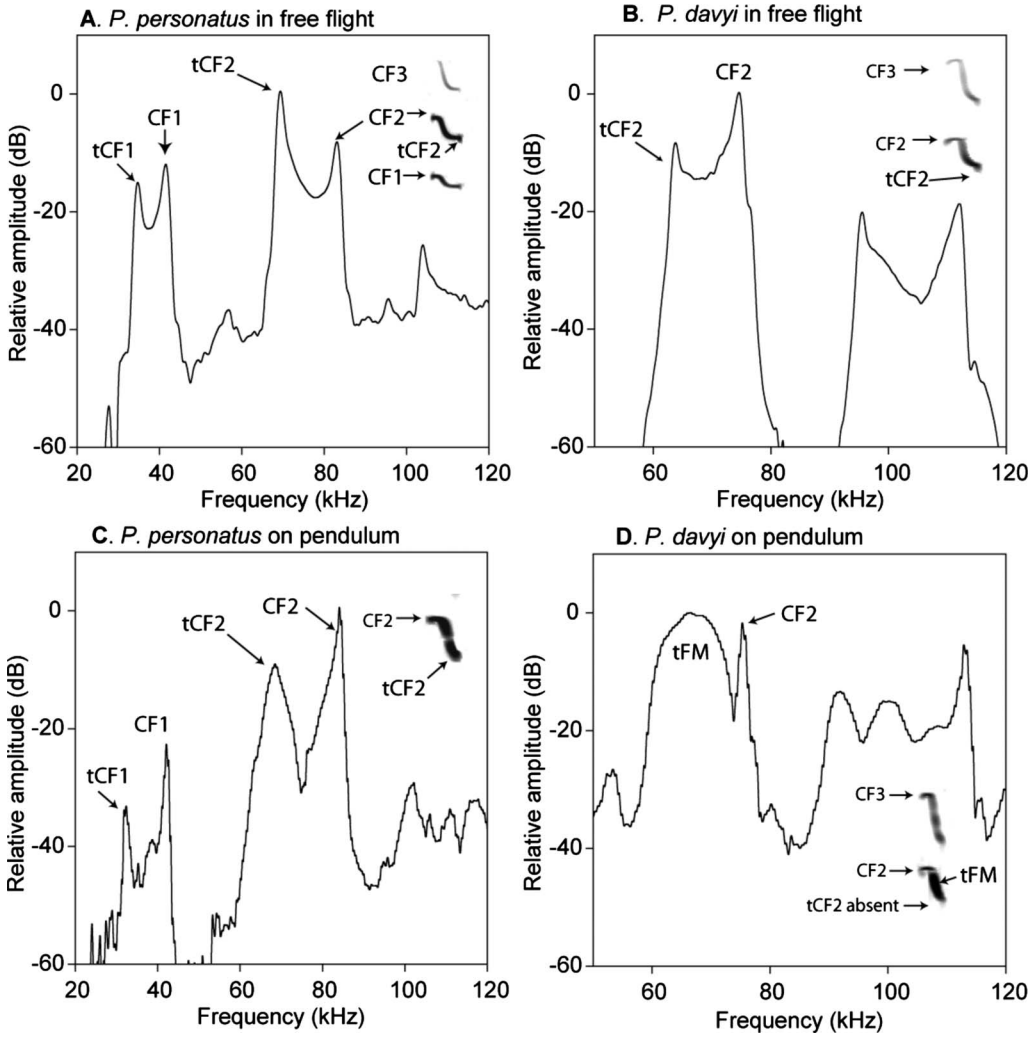


FIG. 3. Representative power spectra of individual calls emitted by (A) *P. personatus* and (B) *P. davyi* (B) in flight versus [(C) and (D)] when restrained. Both bats emit short (5 ms) calls characterized by brief initial and terminal CF components. In both species, the second harmonic is the dominant harmonic component of the call, but often both lower and higher harmonic components are also detectable in recordings. Insets show calls represented in power spectrums. Note that the CF is substantially reduced in both species when held stationary, but a prominent peak representing the CF was still distinguishable in the power spectrum for (C) *P. personatus* but not in (D) *P. davyi*, where the FM component of the call was emphasized relative to the rest of the call.

1. Mormoops megalophylla

In flight [Fig. 2(c)], *M. megalophylla* was observed emitting quasi-CF-FM pulses of approximately 6.2 ± 1.2 ms ($n=250$ pulses). When approaching obstacles such as nearby trees and buildings [as seen in Fig. 2(c)], *M. megalophylla* was observed to increase pulse bandwidth by increasing the starting frequency (F_{max}) and decreasing the ending frequency (F_{min}) while also shortening the pulse duration. These broadband calls began with steeply sloping downward FM sweeps which became shallowly frequency modulated in the center of the pulse for a brief period of 1–2 ms before resuming a second rapid drop in frequency, thus producing an S-shaped pulse with a peak intensity centered in the middle of the pulse. Restrained *M. megalophylla* slightly emitted shorter pulses (5.0 ± 1.1 ms, 6.5 ± 1.9 kHz bandwidth; $n=1000$ pulses from ten bats) than those observed in open flight. The F_{peak} of the narrow bandwidth pulses emitted in flight (54.5 ± 0.85 kHz, $n=250$ pulses) closely corresponded with the central F_{peak} recorded from restrained bats on the pendulum (53.9 ± 0.88 kHz, $n=1000$ pulses), which indicated that a change in pulse bandwidth was not normally accompanied by a change in the peak frequency of the pulse in this species.

2. Pteronotus parnellii, Pteronotus personatus, and Pteronotus davyi

The echolocation pulses of each of the Pteronotus species were essentially identical to previous descriptions in the literature (Griffin and Novick, 1955; Novick and Vaisnys, 1964; O'Farrell and Miller, 1997; Ibanez *et al.*, 1999; Macias and Mora, 2003). Echolocation pulses of restrained *P. parnellii* averaged 19.7 ± 6.2 ms long ($n=1000$) and had a mean CF2 value of 64.9 ± 0.9 kHz. The pulses of *P. personatus* displayed an average duration of 4.8 ± 0.9 ms ($n=1000$). In flight [Figs. 2(a) and 3(a)], the pulses of *P. personatus* exhibited two separate CF components; an initial CF (CF2) and a terminal CF (tCF2), separated by a brief downward FM sweep. For stationary *P. personatus*, the initial CF2 prior to being swung on the pendulum averaged 85.1 ± 1.3 kHz and the tCF2 averaged 70.0 kHz ± 1.2 kHz. On the pendulum,

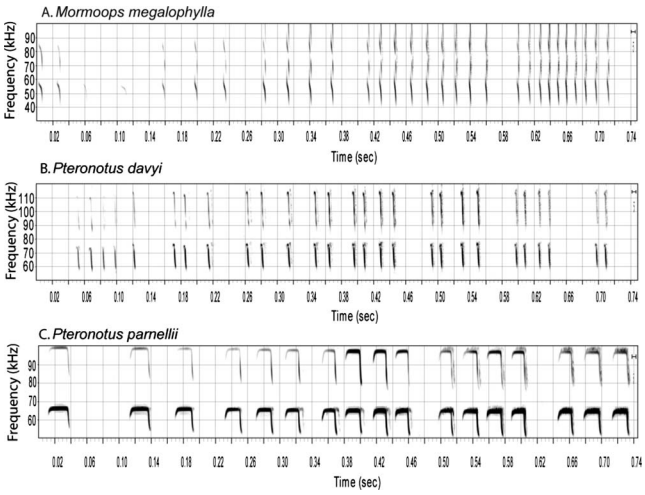


FIG. 4. Spectrograms of sequences of pulses emitted during the initial 0.75 s of the forward swing of the pendulum by (A) *Mormoops megalophylla*, (B) *Pteronotus parnellii*, and (C) *Pteronotus davyi*. To serve as a reference, a dotted line is placed in each graph just above the CF2, or for Mormoops the peak frequency, of pulses emitted before the swing began. The second and third harmonics are visible in all three panels.

the initial CF2 was maintained while the tCF2 became much shorter [Fig. 3(c), inset], although evidence of the tCF2 remained prominent in the magnitude power spectra [Fig. 3(c)].

In free flight, the echolocation pulses of *P. davyi* were also observed to include a second, CF2 component at the end of the pulse that fell roughly 10 kHz below the initial CF2 [Fig. 2(b)]; however, the tCF2 component was completely absent from the spectrograms of pulses recorded from restrained *P. davyi* on the pendulum [Fig. 3(d), inset] and a tCF2 was not consistently visible in the power spectra [Fig. 3(d)]. Prior to being swung on the pendulum, the echolocation pulses of restrained *P. davyi* averaged 4.9 ± 0.6 ms ($n = 1000$) and the initial CF2 averaged 73.6 ± 2.0 kHz.

B. Doppler-shift compensation behavior on the pendulum

Figures 4 and 5 present the representative examples of call sequences emitted by each of the four species of bats while swinging forward on the pendulum. Entire forward swings lasted approximately 1.5 s, but for the sake of clarity we show here only brief sections of the swing corresponding to the period of maximum acceleration (the first 0.75 s). The four species studied differed in their vocal responses on the pendulum, and, in particular, in the average change in the CF2 or F_{peak} frequency while swinging on the pendulum (Fig. 6). As mentioned, although the echolocation pulses of *Mormoops megalophylla* did not include a true CF component, a prominent F_{peak} is clearly visible in the power spectrum (Fig. 7), which we used here for purposes of comparing *Mormoops megalophylla* to the other mormoopids. All four bats responded to forward pendulum swings with rapid bursts of calls (Fig. 4). Of the four bats tested, *P. personatus*

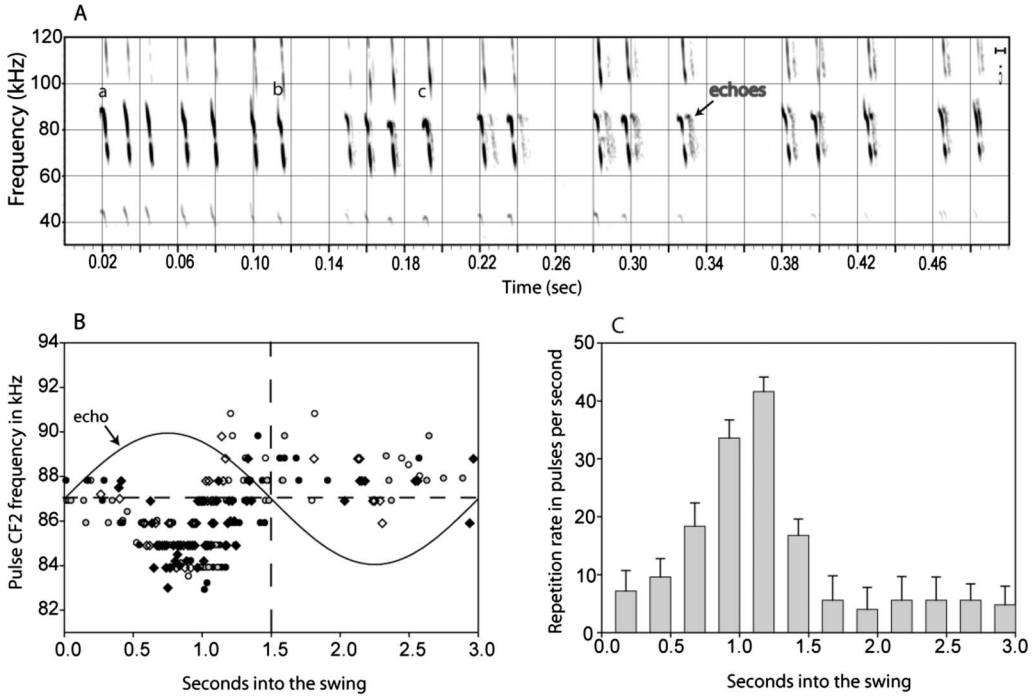


FIG. 5. Spectrogram of a sequence of pulses emitted during the initial forward swing of the pendulum by (A) *Pteronotus personatus*. In panel B, pulse CF2 frequency data from five complete swings from five different bats, each represented by a different symbol, are shown to illustrate the typical pattern of changes in pulse frequency. Estimated echo frequencies during pendulum swings shown in panel B (solid line) are based on the empirically determined speed of the pendulum and the average CF2 frequency of the bats before the swing began. Time 0–1.5 s represents the time course of the forward swing and the time period from 1.5 to 3.0 s represents the backwards swing. Observe that since the bats rapidly raised their pulse CF2 frequencies between 1.0 and 1.5 s while still moving forward, they were likely focusing their attention on the receding ground beneath them rather than the intended forward target. Since most of the pulse CF2 frequencies recorded during the backward swing were higher than the initial CF2, it may be concluded that the bats partly compensated for negative changes in echo frequency. Panel C illustrates the average pulse emission rate of the same five bats (five swings per bat) during the pendulum swings divided into 250 ms time bins.

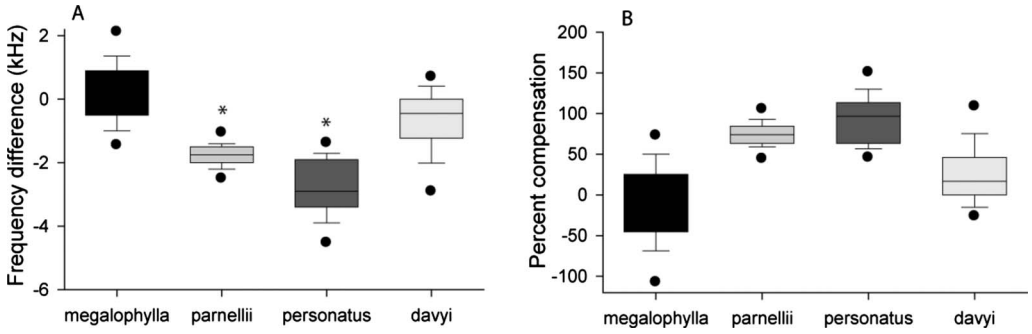


FIG. 6. (A) Box plot of the average difference in the peak or CF2 frequency (or F_{peak} for *M. megalophylla*) of pulses uttered at the midway point of the forward pendulum swing relative to their average frequency at the start of the swing. Only *P. parnellii* and *P. personatus* displayed a statistically significant shift (*) in mean pulse frequency when swung on the pendulum. The solid line denotes the mean, the box denotes the first standard deviation, the whiskers denote the 10th/90th percentiles, and the black dots denote the 5th/95th percentiles. $N=100$ (ten forward swings from each of ten bats). (B) Here, the average frequency differences are expressed as a percentage of the estimated maximum Doppler shift based on the intitial CF2 frequency for each bat. *P. personatus* showed the highest relative compensation, but *P. parnelli* exhibited on average a much less variable compensation performance than *P. personatus*.

displayed the greatest shifts in its CF2 on the pendulum, on average lowering their frequencies by 2.75 ± 0.99 kHz, which means that they compensated for approximately 94% of the maximum Doppler shifts appearing in the echo. Figure 5(b) shows the pattern of changes in *P. personatus*' CF2 value throughout the entire swing for five swings representing five different bats. In Fig. 5(b), it can been seen that the bats rapidly lowered their CF2 values early in the swing, but also started to rapidly raise their CF2 values in the second half of the forward part of the swing, which indicates that the bats were turning their attention to echoes deflected off the ground (which was receding) rather than echoes deflected off the intended forward target. From Fig. 5(b), it can also be observed that *P. personatus* raised and on average held the CF2 values of their calls above their starting CF2 values for the duration of the backwards swing, suggesting that these bats at least partly compensated for negative shifts in echo frequency. Figure 5(c) plots the average call rate over the time course of the complete swing for the same five bats; *P. personatus* rapidly increased call rate during the forward swing, but call rate held steady at or below the initial rate during the return swing.

As expected, *P. parnellii* also performed DSC well, on average lowering the frequency of their CF2 by 1.79 kHz, which represented a compensation level of approximately 88%. *P. davyi* exhibited a small change in CF2 on the pendulum, on average lowering their frequencies by 709 ± 969 Hz (change not significant, $P > 0.05$). *M. megalophylla* made significant changes in pulse structure and bandwidth (described below) but we did not find a significant change in the F_{peak} of their pulses when swinging on the pendulum; the average F_{peak} changed only slightly (220 ± 974 Hz) and insignificantly ($P > 0.05$; Fig. 6).

C. Changes in pulse bandwidth

During the forward swing *M. megalophylla* increased pulse rate, shortened pulse durations, and increased the bandwidth of its pulses by increasing F_{max} and decreasing F_{min} (Fig. 7). *M. megalophylla* increased F_{max} by adding a very steep downward FM sweep to the beginning of its pulse, and decreased F_{min} by exaggerating the FM portion that is present in all of its pulses [Fig. 7(a)]. F_{max} increased from an average of 56.3 ± 1.0 kHz at the beginning of the swing to 60.8 ± 3.3 kHz at the midway point, and F_{min} decreased from 49.8 ± 1.7 to 44.4 ± 1.5 kHz. This amounted to an approximate threefold increase in bandwidth (Fig. 7, panels C and D), which was also accompanied by a relative increase in the loudness of the higher third and fourth harmonics of the pulses relative to the dominant second harmonic (Fig. 7, panel B).

In the case of *P. personatus*, both the initial CF2 portion of the pulse and CF2 were shifted down in frequency during the forward swing of the pendulum [Figs. 8(b) and 8(c)]. Notably, the tCF2 was not lowered as much as the initial CF2 component, but instead went down by an average of 1.85 kHz [75% compensation, Figs. 8(c) and 8(d)], meaning that unlike *M. megalophylla, P. personatus* ultimately reduced the bandwidth of its pulses while swinging on the pendulum. By comparison, *P. parnellii* significantly increased the bandwidth of the tFM component of their pulses while swinging forward on a pendulum; tFM increased from 10.7 ± 1.8 kHz before the swing to 14.6 ± 2.0 kHz (pairwise comparison, Signed-Rank test, $P < 0.01$, $n=100$) at the midway point of the forward swing, and returned to an average of 11.4 ± 1.8 kHz at the midway point of the backwards swing.

Although less pronouncedly, our results indicated that *P. davyi* also manipulated the bandwidth of the FM components on the pendulum. Prior to swinging, *P. davyi*'s tFM bandwidth averaged 16.2 ± 1.8 kHz (median of 15.8 kHz), and during the swing it slightly increased to 16.7 ± 1.9 kHz (median of 17.4 kHz), while on the return swing was reduced to 15.1 ± 1.9 kHz (median of 14.8 kHz). The mean data for before swinging and during the forward swing were not significantly different from one another ($P > 0.05$), but both data sets were significantly broader than the mean tFM bandwidth of pulses emitted on the return swing ($P < 0.01$).

IV. DISCUSSION

The DSC behavior occurs in bats that have narrowly tuned yet extremely sensitive auditory systems. In general, the use of CF-type echolocation pulses is one aspect of the

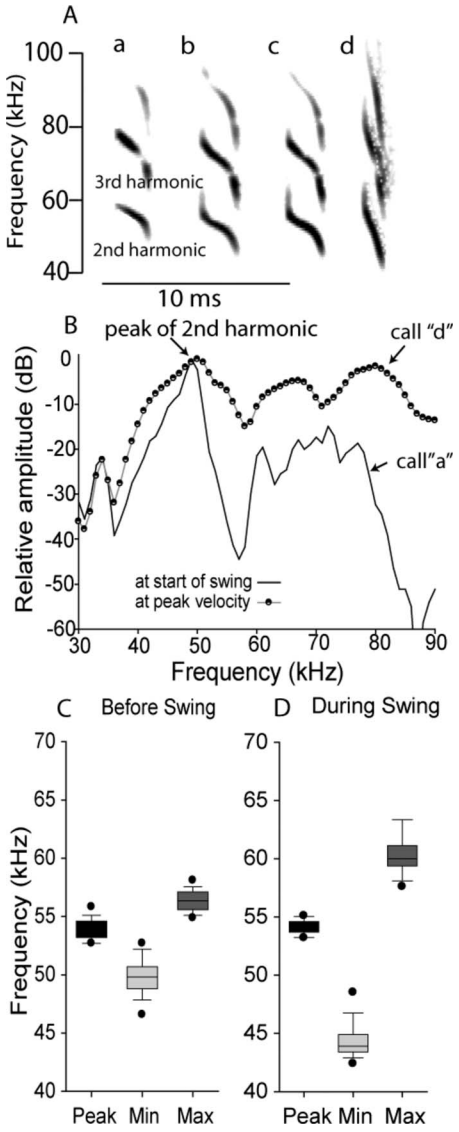


FIG. 7. Changes in the acoustic structure of echolocation pulses emitted by *M. megalophylla* while swinging forward on the pendulum. Panel A shows a representative collection of pulses taken at different succeeding time points during a forward swing; call a was taken before the swing began, calls b and c were emitted early in the swing, and call d is representative of calls emitted while the swing was moving at its highest velocity. The power spectra for two of these pulses (calls a and d) are shown in (B) to demonstrate the increase in bandwidth and the increase in the loudness of the higher frequency harmonic components of the pulse. (C) and (D) are the box plots of the mean peak, minimum (ending), and maximum (starting) frequencies of the pulses ($n=100$).

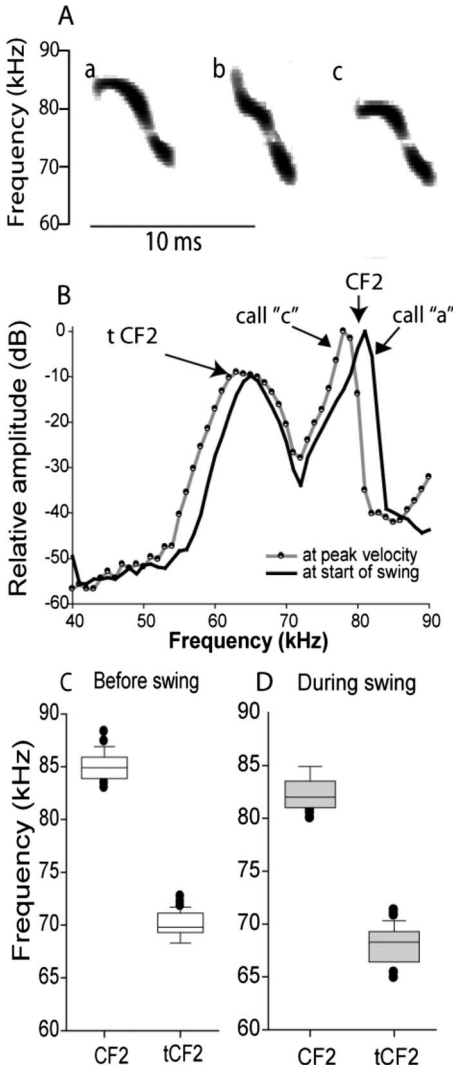


FIG. 8. Changes in acoustic structure of echolocation pulses emitted by *P. personatus* while performing DSC on a pendulum. The pulses in A were selected from the sequence shown in Fig. 5 (A). (B) Power spectra of calls a and c shown in panel A demonstrate the bat lowering its CF2 frequency as well as shifting the entire pulse bandwidth downward. (C) and (D) are the box plots of the frequencies of the CF2 and CF component of the pulses for all ten *P. personatus* tested ($n=100$) before and during the pendulum swing.

specialized sensorial strategies employed by bats that fly and hunt within dense vegetation (Schnitzler and Kalko, 2001; Neuweiler, 2003), but CF pulses and DSC behavior have also been associated with fishing behavior in some bats (Roverud and Grinnell, 1985a, 1985b). While relatively little is known about the foraging behaviors of any Mormoopidae, an examination of the auditory systems of several mormoopids (but not including *P. personatus*) concluded that *P. parnellii* might be the only mormoopid possessing an auditory system so finely tuned that it would benefit from DSC behavior (Kossl *et al.*, 1999). Recent well resolved phylogenetic data indicating that *P. parnellii* and *P. personatus* stem, respectively, from the two most basal nodes at the base of the *Pteronotus* lineage provided an opportunity to address questions of when DSC may have appeared during the evolution of the moormopids. Our observation that *P. personatus* performs DSC is consistent with a conclusion that DSC behavior may have been an ancestral characteristic in the *Pteronotus* lineage.

While the two members of the genus *Mormoops, M. blainvillii* (Kossl *et al.*, 1999) and *M. megalophylla*, do, in fact, use narrowband pulses in open flight (own observations), these bats appear to control the sound of their voices in a manner more similar to FM bats in that they transition to using short broadband pulses while approaching targets. Alternatively, these bats appear distinct from other FM bats in the way they produce almost symmetrical increases and decreases in the beginning and ending frequencies of their pulses, respectively, and in that the peak energy of the pulse is maintained in the center of the pulse's bandwidth, whereas other FM bats only increase the initial frequency while keep-

ing the ending frequency almost invariable (Kalko and Schnitzler, 1998). This dependence on a central peak frequency in *Mormoops* undoubtedly played a role in the evolution of the CF pulse structure in the *Pteronotus* lineage. This bat seems to forage in more open space than other mormoopids (Guillén-Servent, 2005). The evolution of enhanced sensitivity to the narrowband component of the signal could have triggered the adaptation to forage near vegetation aided by narrow frequency analysis echolocation in the sister lineage, *Pteronotus*, in a process similar to what has been suggested for other bats (Guillén-Servent and Ibáñez, 2007), and not necessarily through an ontogenetic accident as suggested by other authors (Kossl *et al.*, 1999). However, given the evidence that *M. megalophylla* uses broadband echolocation pulses, it is not surprising that this bat did not exhibit a vocal response similar to DSC. That *Mormoops* does not perform DSC suggests that the behavior evolved after the separation between the *Mormoops* and *Pteronotus* lineages.

While swinging forward on the pendulum, *Pteronotus davyi* maintained a prominent CF throughout the swing and yet made only minor changes to its CF2 frequency and tFM bandwidth. From our results, we conclude that *P. davyi* does not exhibit DSC behavior on the pendulum, although it remains possible that *P. davyi* may yet perform the behavior under more natural conditions. If, however, we accept the conclusion that *P. davyi* does not or cannot perform DSC, then it is reasonable to conclude that the DSC behavior may have been lost at some point along the evolutionary lineage leading to *P. davyi*. The most parsimonious hypothesis would be that the foraging strategy adopted by *P. davyi* relies upon a more broadly tuned auditory system and, like most FM bats, can tolerate modest Doppler effects. *P. davyi* similarly behaves to other FM bats in the shortening of the FM and the ending narrowband tail of the pulses when approaching targets or flying near background clutter, but differs from them in the multiharmonic nature of the pulses and in keeping an almost fixed bandwidth and maintaining a short CF element at the beginning of the pulses during all the phases of the echolocation behavior (Ibarra-Alvarado and Guillén-Servent, 2005). Questions remain on the functional meaning of the short CF element and the physiological base for the tolerance to Doppler shifts in it. We would hypothesize that this bat's auditory system exhibits tuning characteristics similar to those reported for *P. quadridens* and *P. macleyii*, which use echolocation pulses with similar design. Likewise, we would predict that the auditory system of *M. megalophylla* is similar to that of *M. blainvillii*. No prominent narrow peaks of highly enhanced sensitivity to narrow frequency bands, such as those present in *P. parnellii*, appear in the audiograms of any of these species (Kossl *et al.*, 1999). The putative loss of DSC in the lineage leading to *P. davyi* adds further puzzle to the different success of narrow frequency echolocation in the Old World (Hipposiderids and Rhinolophids have radiated in some 150 species using this echolocation system) versus the New World tropics [only one species in the mormoopid family (Neuweiler, 2003)]. Secondary loss of the capacity in the most recently evolved lineages of the Mormoopidae points to a possible limit to evolutionary diversification of bats using this sonar system imposed by differences between the two biogeographical realms in the ecological space available for realizing the foraging strategy associated with narrow frequency analysis echolocation.

The observation that *P. personatus* not only performs DSC but that it was the only mormoopid tested that decreased rather than increased the overall pulse bandwidth while swinging forward on the pendulum suggests that this bat may be under some pressure to maintain both the initial CF and tCF portions of its pulses within narrow ranges of acoustic sensitivity. In some situations, increased call bandwidth may be a by-product of increased call intensity, and since we observed increases in call intensity during the forward swing for all four species, this may explain the observed increases in bandwidth in those species that did so. Both horseshoe bats (Tian and Schnitzler, 1997) and the mustached bat *P. parnellii* increase the bandwidth of the tFM component of their pulses by lowering the F_{min} while approaching the targets, which raises several questions about why *P. personatus* would not do so. This may imply that their auditory system is finely tuned to the bandwidth of the second, lower CF, or it may reflect a more benign mechanical constraint associated with producing this particular pulse structure. DSC behavior could be part of the adaptations that this bat uses for the particular foraging behavior over water (Guillén-Servent, 2005), when it uses echolocation pulses with a prominent narrowband tail often lacking the initial CF and most of the FM sweep (Guillén-Servent, unpublished data). Further studies on the behavior, ecology, and auditory physiology of *P. personatus* may hold the answers to these questions.

Finally, we address the unique sequence of changes in pulse structure that *P. personatus* appeared to use as it performed DSC (Fig. 8, panel A). The long CF pulses used by horseshoe bats (for example, *Rhinolophus ferrumequinum*), the lesser bulldog bat (*Noctilio albiventris*), and the mustached bat *P. parnellii* (Fig. 2, panel B) include initial upward FM sweeps, and as these bats lower their pulse frequency during DSC the initial upward sweep is maintained throughout the DSC behavior. Although not as prominent, the echolocation pulses of *P. personatus* sometimes included short upward FM sweeps (see call "a" of Fig. 8, panel A), but during their initial response to the Doppler-shifted echoes on the pendulum, the acoustic structure of the pulses changed with the appearance of an initial downward sweep that swept down to a plateau representing a slightly lower initial CF2 than the previous pulse (compare calls "a" and "b" of Fig. 8, panel A). As the pendulum reached its peak velocity, the initial downward sweeps disappeared as the bat reestablished its normal pulse structure at a lower CF2 frequency (call "c" of Fig. 8, panel A). The appearance of this transitional pulse shape (call "b" of Fig. 8, panel A) during DSC performance is interesting not just because it reflects a pattern of pulse adjustments not seen in other Doppler-shift compensating bats but also because of how closely it resembles some of the transitional pulse shapes used by *M. megalophylla* on the pendulum. This sequence of transitions in pulse structure exhibited by *P. personatus* on the pendulum could be interpreted as a reflection of the evolutionary history of CF pulse types in the mormoopids. Thus, at least from a behavioral

standpoint, *P. personatus* would seem to be endowed with some vocal characteristics that probably were present in the ancient common ancestor with *M. megalophylla* and others that can be traced to the more recent common ancestor with *P. parnellii* and the other Pteronotus.

ACKNOWLEDGMENTS

We are most grateful to Mari Carmen García-Escalona and Carlos Enrique Ibarra-Alvarado for their help during fieldwork in Veracruz, Mexico, and we thank Dr. Robert Manson and Dr. Renée González Montagut for their kind hospitality. Funding was provided by NIH NIDCD Grant No. DC007962 to Michael Smotherman, and Grant No. 39709 from the Mexican National Council for Science and Technology (CONACyT), and Institutional Grant No. 902-07-1054 from Instituto de Ecología to A.G.-S. Permission for capturing and handling bats was granted to A.G.-S. by the Mexican Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT).

Gaioni, S. J., Riquimaroux, H., and Suga, N. (**1990**). "Biosonar behavior of mustached bats swung on a pendulum prior to cortical ablation," J. Neurophysiol. **64**, 1801–1817.

Griffin, D. R., and Novick, A. (**1955**). "Acoustic orientation of Neotropical bats," J. Exp. Zool. **130**, 251–300.

Guillén-Servent, A. (**2005**). "Diversity of echolocation and foraging behavior of mormoopid bats in an evolutionary context," Bat Res. News **46**, 176–177.

Guillén-Servent, A., and Ibáñez, C. (**2007**). "Unusual echolocation behavior in a small molossid bat, *Molossops temminckii*, that forages near background clutter," Behav. Ecol. Sociobiol. **61**, 1599–1613.

Habersetzer, J., Schuller, G., and Neuweiler, G. (**1984**). "Foraging behavior and Doppler-shift compensation in echolocating hipposiderid bats, *Hipposideros bicolor* and *Hipposideros speoris*," J. Comp. Physiol. [A] **155**, 559–567.

Hiryu, S., Katsura, K., Lin, L. K., Riquimaroux, H., and Watanabe, Y. (**2005**). "Doppler-shift compensation in the Taiwanese leaf-nosed bat (Hipposideros terasensis) recorded with a telemetry microphone system during flight," J. Acoust. Soc. Am. **118**, 3927–3933.

Ibanez, C., Guillén, A., Juste, B. J., and Perez-Jorda, J. L. (**1999**). "Echolocation calls of Pteronotus davyi (Chiroptera: Moormopidae) from Panama," J. Mammal. **80**, 924–928.

Ibarra-Alvarado, C. E., and Guillén-Servent, A. (**2005**). "The echolocation behavior of Davy's naked-backed bat, *Pteronotus davyi* (Chiroptera: Mormoopidae)," Bat Res. News **46**, 183–184.

Kalko, E., and Schnitzler, H. U. (**1998**). "How echolocating bats approach and aquire food," *Bat Biology and Conservation*, edited by T. H. Kunz and P. A. Racey (Smithsonian Institution, Washington), pp. 197–204.

Kalko, E. K. V., Schnitzler, H. U., Kaipf, I., and Grinnell, A. D. (**1998**). "Echolocation and foraging behavior of the lesser bulldog bat, *Noctilio albiventris*: Preadaptations for piscivory?," Behav. Ecol. Sociobiol. **42**, 305–319.

Keating, A. W., Henson, O. W., Jr., Henson, M. M., Lancaster, W., and Xie, D. H. (**1994**). "Doppler-shift compensation by the mustached bat: Quantitative data," J. Exp. Biol. **188**, 115–129.

Kossl, M., Mayer, F., Frank, G., Faulstich, M., and Russell, I. J. (**1999**). "Evolutionary adaptations of cochlear function in Jamaican mormoopid bats," J. Comp. Physiol. [A] **185**, 217–228.

Macias, S., and Mora, E. (**2003**). "Variation of echolocation calls of Pteronotus quadridens (Chiroptera: Mormoopidae) in Cuba," J. Mammal. **84**, 1428–1436.

Neuweiler, G. (**2003**). "Evolutionary aspects of bat echolocation," J. Comp. Physiol., A **189**, 245–256.

Neuweiler, G., Metzner, W., Heilmann, U., Rubsamen, R., Eckrich, M., and Costa, H. H. (**1987**). "Foraging behaviour and echolocation in the rufous horseshoe bats, *Rhinolophus rouxi*, of Sri Lanka," Behav. Ecol. Sociobiol. **20**, 53–67.

Novick, A. R., and Vaisnys, J. R. (**1964**) "Echolocation of flying insects by the bat, Chilonycteris parnellii," Biol. Bull. **127**, 478–488.

O'Farrell, M. J., and Miller, B. W. (**1997**). "A new examination of echolocation calls of some neotropical bats (Emballonuridae and Mormoopidae)," J. Mammal. **78**, 954–963.

Roverud, R. C., and Grinnell, A. C. (**1985a**) "Frequency tracking and Doppler shift compensation in response to an artificial CF/FM echolocation sound in the CF/FM bat, *Noctilio albiventris*," J. Comp. Physiol., A **156**, 471–475.

Roverud, R. C., and Grinnell, A. D. (**1985b**). "Discrimination performance and echolocation signal integration requirements for target detection and distance determination in the CF/FM bat, Noctilio albiventris," J. Comp. Physiol., A **156**, 447–456.

Schnitzler, H. U. (**1970**). "Echoortung bei der Fledermaus, *Chilonycteris rubiginosa*," Zeitschrift fur Vergleichende Physiologie **68**, 25–38.

Schnitzler, H. U., and Kalko, E. K. V. (**2001**). "Echolocation by insect-eating bats." BioScience **51**, 557–569.

Schnitzler, H. U. (**1967**). "Compensation of Doppler effects in horseshoe bats," Naturwiss. **54**, 523.

Schnitzler, H. U., and Henson, O. W. J. (**1980**). "Performance of animal sonar systems I. Microchiroptera," *Animal Sonar Systems*, edited by R. G. Busnel and J. F. Fish (Plenum, New York), pp. 109–181.

Schuller, G. (**1980**). "Hearing characteristics and Doppler-shift compensation in South Indian CF-FM bats," J. Comp. Physiol. [A] **139**, 349–356.

Simmons, J. A. (**1974**). "Response of the Doppler echolocation system in the bat, Rhinolophus ferrumequinum," J. Acoust. Soc. Am. **56**, 672–682.

Simmons, J. A., Fenton, M. B., and O'Farrell, M. J. (**1979**). "Echolocation and pursuit of prey by bats," Science **203**, 16–21.

Suga, N. (**1989**). "Principles of auditory information-processing derived from neuroethology," J. Exp. Biol. **146**, 277–286.

Tian, B., and Schnitzler, H. U. (**1997**). "Echolocation signals of the greater horseshoe bat (*Rhinolophus ferrumequinum*) in transfer flight and during landing," J. Acoust. Soc. Am. **101**, 2347–2364.

Van Den Bussche, R. A., and Weyandt, S. E. (**2003**). "Mitochondrial and nuclear DNA sequence data provide resolution to sister-group relationships within Pteronotus (Chiroptera: Mormoopidae)," ACTA Chiropterol. **5**, 1–13.

Effects of frequency disparities on trading of an ambiguous tone between two competing auditory objects

Adrian K. C. Lee and Barbara G. Shinn-Cunningham[a)]
Hearing Research Center, Boston University, Boston, Massachusetts 02215, and Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139

(Received 24 July 2007; revised 17 March 2008; accepted 19 March 2008)

Listeners are relatively good at estimating the true content of each physical source in a sound mixture in most everyday situations. However, if there is a spectrotemporal element that logically could belong to more than one object, the correct way to group that element can be ambiguous. Many psychoacoustic experiments have implicitly assumed that when a sound mixture contains ambiguous sound elements, the ambiguous elements "trade" between competing sources, such that the elements contribute more to one object in conditions when they contribute less to others. However, few studies have directly tested whether such trading occurs. While some studies found trading, trading failed in some recent studies in which spatial cues were manipulated to alter the perceptual organization. The current study extended this work by exploring whether trading occurs for similar sound mixtures when frequency content, rather than spatial cues, was manipulated to alter grouping. Unlike when spatial cues were manipulated, results are roughly consistent with trading. Together, results suggest that the degree to which trading is obeyed depends on how stimuli are manipulated to affect object formation.
 [DOI: 10.1121/1.2908282]

PACS number(s): 43.66.Ba, 43.66.Lj, 43.66.Mk [RYL] Pages: 4340–4351

I. INTRODUCTION

Sound arriving at our ears is a sum of acoustical energy from all the auditory sources in the environment. In order to make sense of what we hear, we must group related elements from a source of interest and perceptually separate these elements from the elements originating from other sources. Auditory scene analysis (Bregman, 1990; Darwin, 1997; Carlyon, 2004) depends on grouping together simultaneous sound energy as well as grouping energy across time (streaming or sequential grouping).

The perceptual organization of a sequence of tones depends on the frequency proximity of the component tones, the tone repetition rate, and the attentional state of the observer (Van Noorden, 1975). Specifically, when a sequence of tones alternates between two frequencies, the probability of perceiving two separate streams (corresponding to the two frequencies) increases as the frequency separation and/or the tone repetition rate increases. While there have been attempts to explain sequential streaming by considering only the peripheral processing of the auditory system (Hartmann and Johnson, 1991; Beauvois and Meddis, 1996; McCabe and Denham, 1997), such explanations cannot fully account for how even simple tone sequences are perceptually organized (Vliegen and Oxenham, 1999).

Simultaneous grouping associates sound elements that occur together in time, such as the harmonics of a vowel (Culling and Summerfield, 1995; Hukin and Darwin, 1995; Drennan *et al.*, 2003). Some of the cues that dominate how simultaneous sound elements are grouped together include common amplitude modulation (e.g., common onsets and offsets) as well as harmonic structure.

The relative potency of acoustical cues influencing sequential and simultaneous grouping has been measured by pitting different acoustic grouping cues against one another to determine which cue perceptually dominates. Frequency separation strongly influences sequential grouping, with the grouping strength decreasing as the frequency separation increases. Common onset/offset causes simultaneous elements to group together (Bregman and Pinker, 1978). Moreover, there is an interaction between sequential and simultaneous grouping cues (Dannenbring and Bregman, 1978; Steiger and Bregman, 1982; Darwin *et al.*, 1995). For instance, the contribution of one harmonic to a harmonic tone complex is reduced by the presence of a tone sequence surrounding the complex whose frequency matches that harmonic (Darwin and Sutherland, 1984; Darwin *et al.*, 1995; Darwin and Hukin, 1997, 1998).

Acoustically, if two sources (S_1 and S_2) are uncorrelated, the total energy in the sum of the sources at each frequency is expected to equal the sum of the energies in S_1 and S_2 at that frequency. Thus, if listeners form fixed, veridical estimates of uncorrelated sources in a mixture, the sum of the energies perceived in the two objects should equal the physical energy of the sound in a mixture, an idea we will call the "energy conservation" hypothesis. However, it is quite likely that listeners do not form perfect, veridical estimates of the sources in a mixture. Even if energy conservation fails, it seems reasonable to expect that if a sound mixture contains

a)Author to whom correspondence should be addressed. Present address: Department of Cognitive and Neural Systems, Room 311, Boston University, 677 Beacon St., Boston, MA 02215. Tel.: 617-353-5764. FAX: 617-353-7755. Electronic mail: shinn@cns.bu.edu

an ambiguous element that could logically belong to more than one object, the energy in that element should trade between objects. Specifically, when an ambiguous element perceptually contributes more to one object, it should contribute less to the competing object. We call this the "trading" hypothesis. One special form of trading would occur if the pressure amplitude, rather than the total energy, of the ambiguous element is conserved ("pressure conservation"). In such cases, the sum of the effective energies that an ambiguous element contributes to competing objects should be 3 dB less than the physical energy of the ambiguous element (e.g., see McAdams *et al.*, 1998). Pressure conservation would be veridical if the frequency components making up the ambiguous elements in a sound mixture are in phase (and therefore correlated) rather than independent.

There are many studies exploring how a sequential stream influences a simultaneously grouped harmonic complex. However, there are only a handful of studies exploring whether a simultaneous complex has reciprocal influences on a sequential stream and whether or not energy conservation or trading occurs for ambiguous sound mixtures (Darwin, 1995; McAdams *et al.*, 1998; Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). Some past studies have assumed that trading occurs without actually measuring the effective contribution of an ambiguous element to both objects in a sound mixture. For instance, a recent pair of studies tested how frequency proximity interacts with harmonicity and common onset/offset to influence the perceived content of a harmonic complex (Turgeon *et al.*, 2002; Turgeon *et al.*, 2005). In interpreting these results, it was explicitly assumed that when an ambiguous tone did not strongly contribute to the simultaneously grouped object, it strongly contributed to the ongoing stream even though the perceived spectral content of the ongoing stream was not tested.

The few studies that have explicitly tested whether the energy in an ambiguous element trades between competing objects give conflicting results (Darwin, 1995; McAdams *et al.*, 1998; Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008).

One study measured perception for a harmonic complex when one tone in the complex was turned on before the other harmonics (Darwin, 1995). The intensity of the portion of the ambiguous tone preceding the harmonic complex was manipulated to alter how much of the simultaneous portion was perceived in the complex. In general, the contribution of the simultaneous portion of the ambiguous harmonic to the complex was reduced when the ambiguous harmonic began before the other harmonics. Moreover, the amount by which the contribution of the ambiguous tone to the simultaneous complex was reduced increased as the intensity of the precursor portion of the ambiguous tone increased. While energy conservation failed, trading was observed: The sum of the ambiguous element's contribution to the two competing objects (the separate precursor tone and the harmonic complex) was roughly constant at about 3 dB less than the physical energy of the ambiguous tone present during the complex (Darwin, 1995). This result was roughly consistent with pressure rather than energy conservation.

Another study exploring perception of alternating low-intensity, narrow-band stimuli and high-intensity wider-band stimuli that overlapped in frequency found similar results (McAdams *et al.*, 1998). In this study, listeners generally perceived the low-level, narrow-band stimulus as continuous and the higher-intensity, broader-band stimulus as a pulsed stream. Thus, there was a band of energy during the high-intensity, wider-band stimuli that was ambiguous and perceptually contributed to both streams. Three alternative forms of trading were explicitly evaluated to determine whether trading occurred: Energy conservation, pressure conservation, and loudness conservation (where the total loudness of the ambiguous elements, in sones, was apportioned between the two competing streams). Results generally fell between energy and pressure conservations. Other studies of "homophonic induction" also suggest a form of trading (e.g., see Warren *et al.*, 1994; Kashino and Warren, 1996), although these studies did not quantify the contribution of the ambiguous sound energy to each of the competing objects.

While many past studies found results roughly consistent with pressure conservation, trading completely failed in two studies presenting a sound mixture consisting of a repeating tone stream and a harmonic complex whose fourth harmonic was one of the tone-stream components (Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). When spatial cues were manipulated to vary the perceptual organization of the scene, the sum of the energy contributions of the ambiguous element to the two objects dramatically varied across conditions. In fact, in one condition, the ambiguous element contributed almost nothing to either of the two competing objects.

The current study was designed to determine whether trading occurs for stimuli similar to those for which trading failed in previous studies. As in the earlier studies, the stimuli contained an ambiguous pure-tone element (the target) that could logically be heard as one tone in an isochronous stream of repeating tones and/or as part of a more slowly repeating harmonic complex. In the current experiment, frequency proximity rather than spatial cues were manipulated to affect perceptual organization. Specifically, the frequency of the repeating tones varied from trial to trial from below to above the frequency of the target, whose frequency equaled the fourth harmonic of the simultaneous harmonic complex.

A control experiment allowed us to compute the effective level of the target perceived in the two objects (tone stream and complex) to test how the ambiguous target was allocated across the competing objects. We performed another experiment to relate our results to past studies exploring how frequency proximity influences perception of a tone sequence. We find that for the current stimuli, energy conservation fails, but trading (roughly consistent with pressure conservation) is observed.

II. EXPERIMENT 1: COMPETING OBJECTS

A. Methods

1. Stimuli

Stimuli generally consisted of a repeating sequence of a pair of tones followed by a harmonic complex [Fig. 1(a)].

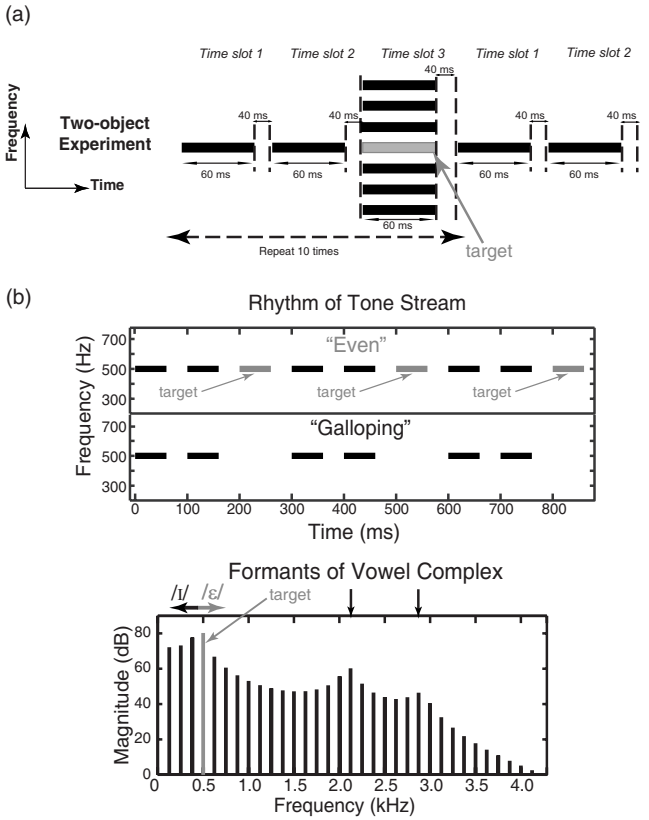


FIG. 1. (a) Two-object stimuli were created by repeating a three-item sequence consisting of a pair of pure tones followed by a harmonic complex. In the reference configuration, the tones in time slots 1 and 2 are at 500 Hz. Time slot 3 is made up of two components: a target tone at 500 Hz and a tone complex with fundamental frequency of 125 Hz (with the fourth harmonic at 500 Hz omitted). The tone complex is shaped by a synthetic vowel spectral envelope to make it sound like a short vowel (Darwin, 1995). Because the first formant of the vowel complex is near 500 Hz, the relative level of the target tone perceived in the vowel complex affects perception of the first formant frequency, which affects the perceived identity of the vowel. (b) Top panel: The perceived rhythm depends on whether or not the 500 Hz target tone is perceived in the sequential tone stream. If the target is grouped with the repeated tones, the resulting rhythmic percept is even; if the target is not grouped with the pair of tones, the resulting perceived rhythm is galloping. Bottom panel: The synthetic vowel spectral envelope is similar to that used by Hukin and Darwin (1995). The identity of the perceived vowel depends on whether or not the 500 Hz target is perceived in the complex. The vowel shifts to be more like /ɛ/ when the target is perceived as part of the complex and more like /ɪ/ when the target is not perceived in the complex. The arrows indicate the approximate locations of the first three formants of the perceived vowel.

The frequency of the pair of tones varied from trial to trial from two semitones below to two semitones above 500 Hz, taking on one of seven predetermined values (i.e., 0, ±0.5, ±1, and ±2 semitones relative to 500 Hz; also see Fig. 2, left panels).

The harmonic complex contained the first 39 harmonics of 125 Hz, excluding the fourth harmonic (500 Hz). The phase of each component was randomly chosen on each trial. The magnitudes of the harmonics were shaped to simulate the filtering of the vocal tract (Klatt, 1980). The first formant frequency (F1) was set to 490 Hz, close to the expected value for the American-English vowel /ɛ/ (Peterson and Barney, 1952). The second and third formants were fixed at 2100 and 2900 Hz, respectively. The half-power bandwidths of the

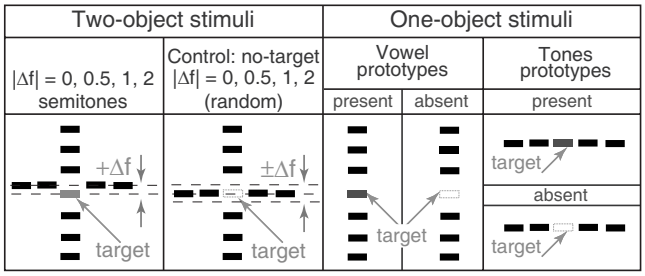


FIG. 2. Experimental conditions. Each block consists of seven two-object stimuli with the target present, a two-object control without the target present, and two one-object prototypes (see text for more details).

three formants were 90, 110, and 170 Hz [the parameters were chosen based on studies by Hukin and Darwin (1995)].

The target was a 500 Hz tone that was gated on and off with the harmonic complex. As a result of this structure, the target could logically be heard as the third tone in the repeating tone stream and/or as the fourth harmonic in the harmonic complex. The tones, the harmonic complex, and the target were all gated with a Blackman window of 60 ms duration.

The amplitude of the target and the tones was equal and matched the formant envelope of the vowel. There was a 40 ms silent gap between each tone and harmonic complex to create a regular rhythmic pattern with an event occurring every 100 ms. This basic pattern, a pair of repeating tones followed by the vowel complex/target, was repeated ten times per trial to produce a 3 s stimulus that was perceived as two objects: An ongoing stream of tones and a repeating vowel occurring at one-third that rate.

The rhythm of the tone sequence depends on the degree to which the target is perceived in the tone stream [Fig. 1(b), top panel]. The tone stream is heard as even when the target is heard in the stream and galloping when the target is not perceived in the stream. Similarly, the phonetic identity of the harmonic complex depends on whether or not the target is heard as part of the complex [Fig. 1(b), bottom panel; Hukin and Darwin, 1995]. The first formant of the complex (F1) is perceived as slightly higher when the target is perceived in the complex compared to when the target is not part of the complex. This slight shift of F1 causes the complex to be heard more like /ɛ/ when the target is part of the complex and more like /ɪ/ when it is not part of the complex.

Control stimuli consisted of one-object presentations with only the pair of tones or only the harmonic complex, either with the target ("target-present" one-object prototype) or without the target ("target-absent" one-object prototype; see Fig. 2, right panels). Finally, a two-object control was generated in which the repeating tones and complex were presented together, but there was no target ("no-target" control, see Fig. 2, second panel from the left).

2. Task

In order to assess the perceptual organization of the two-object mixture and how the frequency difference between the repeated tones and target affected the perceived structure of both the tone stream and vowel, the same physical stimuli were presented in two separate experimental blocks. In one

block, subjects judged the rhythm of the tone sequence ("galloping" or "even") by performing a one-interval, two-alternative-forced-choice task. In the other block, the same physical stimuli were presented in a different random order, and subjects judged the vowel identity ("/ɪ/" as in "bit" or "/ɛ/" as in "bet"). In order to control for the possibility that streaming changes over time, we asked subjects to attend to the object of interest throughout the 3-s-long stimulus but to make their judgments about the attended object based on what they perceived at the end of the stimulus presentation.

3. Environment

All stimuli were generated offline using the MATLAB software (Mathworks Inc.). Signals were processed with pseudoanechoic head-related transfer functions (HRTFs) [see Shinn-Cunningham (2005) for details] in order to make the stimuli similar to those used in companion studies that varied the source location (Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). In the current experiment, all components of the stimuli were processed by the same HRTFs, corresponding to a position straight ahead of and at a distance of 1 m from the listener.

Stimuli were generated at a sampling rate of 25 kHz and sent to Tucker-Davis Technologies hardware for digital to analog conversion and attenuation. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. All signals were presented at a listener-controlled, comfortable level that had a maximum value of 80 dB sound pressure level. The intensity of each stimulus was roved over a 14 dB range in order to discourage using the level as a cue to vowel identity or tone rhythm. Stimuli were presented over insertion headphones (Etymotic ER-1) to subjects seated in a sound-treated booth. Subjects responded via a graphical user interface.

B. Experimental procedure

1. Participants

Fourteen subjects (eight male, six female, aged 18–31) took part in the experiments. All participants had pure-tone thresholds in both ears within 20 dB of normal-hearing thresholds at octave frequencies between 250 and 8000 Hz and within 15 dB of normal-hearing thresholds at 500 Hz. All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board and the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology.

2. One-object prototype training

In each session of testing, each subject was familiarized with the one-object prototypes with and without the target (Fig. 2, right panels). During training, subjects were given feedback to reinforce the correct labeling of the one-object, target-present and target-absent prototypes. This feedback ensured that subjects learned to accurately label the rhythm of the sequence of tones and the phonetic identity of the harmonic complex for unambiguous, one-object stimuli. Subjects had to achieve at least 90% correct when identifying the two prototypes in the one-object pretest before proceeding to the main experiment.

3. Main experiment

Following training, listeners judged either the tone-stream rhythm or the vowel identity, depending on the experimental block. Both two-object stimuli and the appropriate one-object prototypes (see Fig. 2) were intermingled in each block. The one-object trials served as controls that allowed us to assess whether listeners maintained the ability to label the unambiguous stimuli throughout the run without feedback (see right side of Fig. 2). From trial to trial, the frequency of the repeating tones in the two-object stimuli randomly varied relative to the target (Fig. 2). Seven two-object conditions were tested in each block, with the frequency of the repeated tones ranging from two semitones below to two semitones above the target frequency ($\Delta f = 0$, ± 0.5, ± 1, ± 2 semitones). A control two-object condition was included in which the target was not presented. In this control, the frequency of the repeated tones was randomly selected from the seven possible frequencies used in the other two-object conditions (i.e., 0, ± 0.5, ± 1, ± 2 semitones from 500 Hz; Fig. 2 second panel from the left) to ensure that the subjects did not make rhythmic or vowel judgments based on the absolute frequency of the repeated tones.

In one block of the experiment, we presented eight two-object stimuli and the two one-object prototypes containing no vowel (target present and target absent). In this block, we asked the subjects to report the perceived rhythm of the tones. In a separate block of the experiment, we presented the same eight two-object stimuli intermingled with the two one-object vowel prototypes and asked the subjects to report the perceived vowel. Both blocks consisted of 30 repetitions of each stimulus in random order, for a total of 300 trials per block. We used the response to the prototype stimuli both for screening and in interpreting the results to the ambiguous two-object stimuli, as discussed below.

4. One-object control experiments

Two companion control experiments tested the subjective impressions of either the tone-stream rhythm or the vowel identity when there were no other objects present and the physical intensity of the target varied from trial to trial. In these control experiments, subjects were presented with one-object stimuli (tones in one experiment, harmonic complexes in the other) with a variable-level target. From trial to trial, the intensity of the target was attenuated by a randomly chosen amount ranging between 0 and 14 dB (in 2 dB steps) relative to the level of the target in the two-object experiments. In the one-object tone task, subjects reported whether the rhythm on a given trial was galloping or even. In the one-object harmonic complex task, subjects reported whether the complex was /ɪ/ or /ɛ/.

For both one-object control experiments, the percent responses (y) for each subject were related to the target attenuation (x) by fitting a sigmoidal function of the form

$$\hat{y} = \frac{1}{1 + e^{-a(x - x_0)}}, \tag{1}$$

where $\hat{y}$ is the estimated percent response, a is the best-fit slope parameter, and x_0 is the best-fit constant corresponding to the attenuation at which the function reaches 50% of its maximum value. The corresponding psychometric functions for each subject allowed us to map the percent response in the two-object experiment to an effective target attenuation based on the mapping between physical target attenuation and response percentages in the one-object control experiment. If 95% or more of a subject's responses for a given condition were target present (i.e., even or "/ɛ/" as in "bet") or target absent (i.e., galloping or "/ɪ/" as in "bit"), the effective attenuation was set to 0 or 16 dB, respectively.

5. Relative d′ calculation

Raw percent correct target-present responses (even for the tones, /ɛ/ for the vowel) were computed for each subject and condition. Because the raw percentage of responses does not give any insight into what differences were perceptually significant and which were perceptually small, we used decision theory to estimate the perceptual distance between the stimulus and the one-object target-absent prototypes (see Shinn-Cunningham *et al.*, 2007, Methods). This method is briefly summarized below.

In each block of the main experiment, one-object prototypes (with and without the target) were randomly intermingled with the ambiguous, two-object stimuli. We assumed that in judging the vowel identity or tone rhythm, listeners used an internal Gaussian-distributed decision variable whose mean depended on the stimulus and whose variance was independent of the stimulus. This internal decision variable was assumed to represent the perceptual continuum from target absent to target present. Listener responses on a given trial (either target absent or target present) were assumed to be the result of a comparison of a sample of this internal decision variable to a criterion that was constant throughout the block, enabling us to compute the relative perceptual separation of the means of the conditional distributions for the different stimulus conditions. In particular, conditioned on which stimulus was presented, the percent target-present responses were assumed to equal the portion of the conditional distribution of the decision variable falling to the appropriate side of an internal decision criterion. Differences in these conditional probabilities were used to compute the perceptual distances (d') between the distributions [Fig. 3(a)].

We use $d'_{\text{present:absent}}$ to denote the perceptual distance between the target-present and target-absent prototypes. By assuming the above decision model, $d'_{\text{present:absent}}$ is given as (Green and Swets, 1966; Macmillan and Creelman, 2005)

$$d'_{\text{present:absent}} = \Phi^{-1}[\Pr(\text{“target present”} \,|\, \text{target present})] - \Phi^{-1}[\Pr(\text{“target present”} \,|\, \text{target absent})], \tag{2}$$

where Φ^{-1} denotes the inverse of the cumulative Gaussian distribution and Pr("target present"|stimulus) represents the probability that the subject reports that the target is present in the specified stimulus. In order to avoid an incalculably large value of d' due to sampling issues, the number of responses in each possible category was incremented by 0.5 prior to computing the percentage of responses and the resulting values of d'. As a result of this adjustment, the maximum achievable d' value was 4.28. Values of $d'_{\text{present:absent}}$ were separately calculated for each subject.

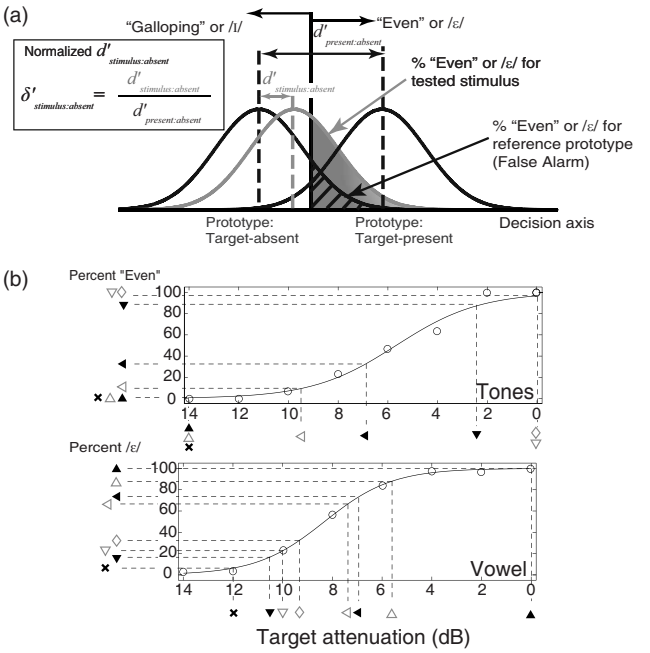


FIG. 3. (a) Schematics of the decision model assumed in computing $\delta'_{\text{stimulus:absent}}$. The decision axis (representing the decision variable for either the rhythmic or vowel identification space) is shown along the abscissa. The Gaussian distributions show the conditional probabilities of observing different values of the decision variable for the target-absent and target-present prototypes (left and right distributions, respectively) as well as for a particular two-object stimulus (middle distribution). (b) Computation of the effective target attenuation from the psychometric functions relating percent target-present responses to physical target attenuation for one-object stimuli for an example subject. The solid line shows the psychometric function fitted to the data points from the one-object control experiment, plotted as circles. The symbols on the ordinate and horizontal dashed lines represent the percentage of even (top panel) or /ɛ/ (bottom panel) responses for different stimuli. The vertical dashed lines and symbols along the abscissa show the effective target attenuation estimated from the control data.

The perceptual distance between any stimulus and the target-absent one-object controls was then individually calculated for each subject as

$$d'_{\text{stimulus:absent}} = \Phi^{-1}[\Pr(\text{“target present”} \,|\, \text{stimulus})] - \Phi^{-1}[\Pr(\text{“target present”} \,|\, \text{target absent})]. \tag{3}$$

In order to determine whether a particular stimulus was perceived as more similar to the target-present prototype or more like the target-absent prototype, for each subject, we computed a normalized sensitivity measure for each condition from the raw sensitivities as

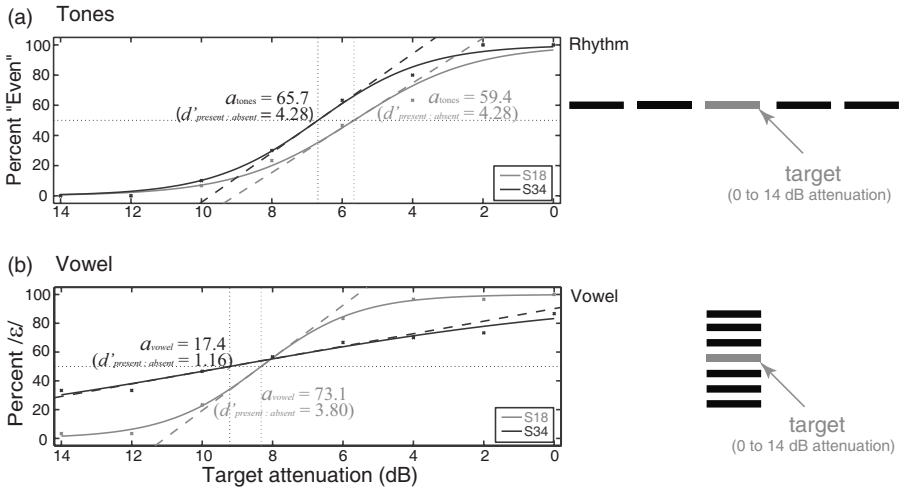


FIG. 4. Example psychometric functions for results of one-object experiments in which the target attenuation varied from 0 to 14 dB (in 2 dB steps), for two representative subjects (S18, a good subject, and S34, a subject who just passed our screening criteria). The dotted lines show the slope of each of the psychometric function at the 50% point. The raw percent responses (for tone-stream rhythm on top and vowel identity below) are shown for each subject as a function of target attenuation.

$$\delta'_{\text{stimulus:absent}} = \frac{d'_{\text{stimulus:absent}}}{d'_{\text{present:absent}}}. \quad (4)$$

A value of $\delta'_{\text{stimulus:absent}} < 0.5$ indicates that the stimulus was perceived as more like the target-absent than the target-present prototype. Conversely, a value of $\delta'_{\text{stimulus:absent}} > 0.5$ indicates that responses were more like those for the target-present than for the target-absent prototype [Fig. 3(a)].

6. Effective target level calculation

To quantify the effective level that the target contributed to each object, we analyzed the psychometric functions fit to the responses from the corresponding one-object control experiment (see Sec. II C 4), which relate the percentage of target-present responses to the physical intensity of the target actually present in the stimuli. By using the psychometric functions obtained for each individual subject, we mapped the percent response in the two-object experiment to the target intensity that would have produced that percentage of responses for the corresponding one-object stimuli [see Fig. 3(b)].

C. Results

1. Subject screening

Despite training, not all subjects could reliably label the one-object vowel prototype stimuli when they were presented in the main experiment, which provided no feedback and intermingled the prototype stimuli with ambiguous two-object stimuli. We adopted a screening protocol to exclude any subjects who could not accurately label the prototype stimuli during the main experiment. Specifically, we excluded data from subjects who failed to achieve $d'_{\text{present:absent}} > 1.0$.

We also excluded any subject for whom response percentages only weakly depended on the target attenuation in the one-object control experiments. Specifically, if the fitted slope parameter a in Eq. (1) was less than 10%/dB, the subject was excluded from further analysis. We also excluded any subject for whom the correlation coefficient (ρ) between the observed data (y) and the data fit ($\hat{y}$) was less than 0.9.

For all subjects, the slope relating the percentage of galloping responses to target attenuation was very steep and met our criterion. Thus, all subjects perceived consistent changes in the rhythm of the tone stream with attenuation of the target. Similarly, all $d'_{\text{present:absent}}$ scores were much greater than the criterion when listeners judged the tones' rhythm. Specifically, all subjects could maintain a consistent decision criterion for labeling the rhythm of the tones even without feedback when the prototypes were presented alongside ambiguous two-object stimuli. Thus, no subjects were excluded from the experiment based on poor performance in the tones task.

Six out of the 14 subjects (three male, three female) failed to meet our criteria for the vowel experiment and had their results excluded from further analysis. All data analyzed below are from the eight subjects who passed all criteria for both tone and vowel screenings.

Figure 4 shows example psychometric functions for the one-object control experiments for subjects S18 (a relatively good subject) and S34 (a subject who passed our screening criteria but was less consistent in labeling the vowels). The top panel in Fig. 4 shows results for the tone experiment. Both subjects responded "galloping" in conditions where the target tone was attenuated by about 12 dB or more relative to the repeating tones and "even" when the intensity of the target matched that of the repeating tones (i.e., 0 dB attenuation). Moreover, both subjects showed steep, monotonically increasing psychometric functions. The bottom panel in Fig. 4 shows the psychometric functions for the same two subjects for the vowel experiment. S18 shows a steeply increasing psychometric function relating percent correct responses to the target attenuation. In contrast, S34 has a very shallow function, demonstrating poor sensitivity to changes in the target attenuation as measured by vowel identification. Consistent with this, S34 also had a low $d'_{\text{stimulus:absent}}$ in the vowel task (1.16 compared to 3.80 for S18). Despite the relatively poor sensitivity of S34, this subject met the liberal inclusion criteria we imposed.

2. Rhythmic judgments

Figure 5 summarizes results of the main two-object experiment for the rhythm judgments (left column of Fig. 5;

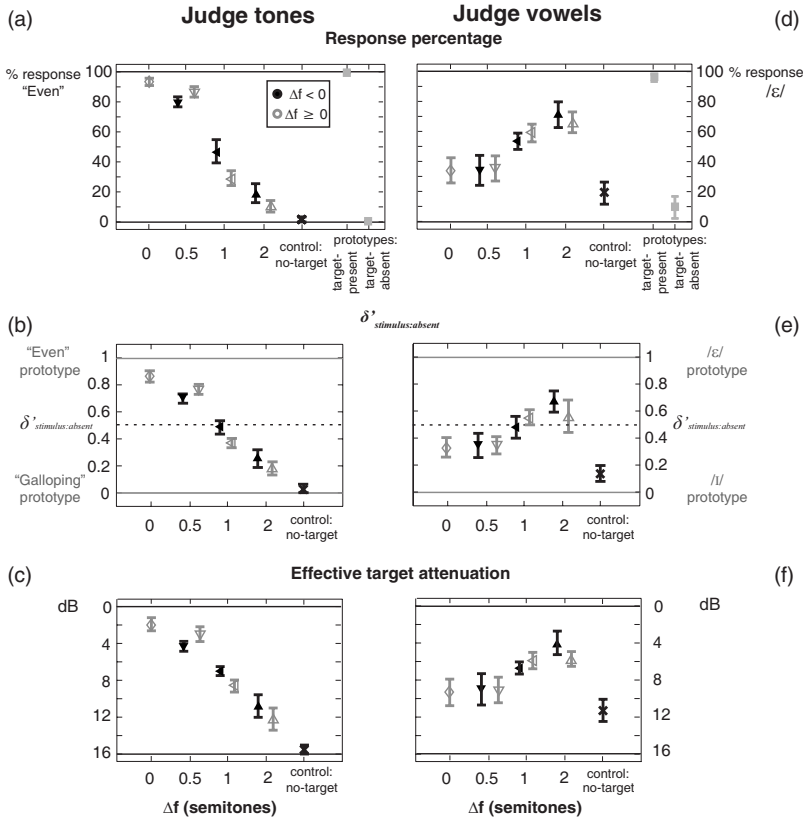


FIG. 5. Results of both rhythm judgments (left column) and vowel judgments (right column). [(a) and (d)] Raw response percentages. [(b) and (e)] $\delta'_{\text{stimulus:absent}}$ derived from raw results. [(c) and (f)] Effective target attenuation derived from the psychometric functions relating raw responses to effective target attenuations. Each marker represents the across-subject mean estimate and the error bar shows ± 1 standard error of the mean.

corresponding results for the vowel judgments are shown in the right column and are considered in the next section). Figure 5(a) shows the group mean percentages of even responses (error bars show the across-subject standard error of the mean). All but one subject correctly identified the even and galloping one-object prototypes with 100% accuracy [see squares at far right of Fig. 5(a)]. When the frequency of the repeating tones matched that of the target in the two-object condition (i.e., $\Delta f=0$ semitones), subjects generally reported an even percept (average target-present response rate was greater than 90%; diamond at left of plot). As the frequency difference between the repeating tones and target increased, the probability of responding as if the target was present in the tone stream decreased [see open and filled triangles in Fig. 5(a)]. As expected, there was a very low probability of reporting that the target was present in the two-object no-target control trials [i.e., the average percentage of target-present responses was 1.7%; see X in Fig. 5(a)]. The $d'_{\text{stimulus:absent}}$ values (not shown) range from a low of 0.136 (for the target-absent two-object stimuli) to a high of more than 3.5 (for the $|\Delta f|=0$ stimulus).

Figure 5(b) plots $\delta'_{\text{stimulus:absent}}$, which quantifies the perceptual distances between a given stimulus and the one-object prototypes (0, near the galloping prototype; 1, near the even prototype). As all subjects were nearly equal in their ability to properly label the two prototypes, the pattern of the mean δ' results looks very similar to the raw percent responses. A two-way repeated-measure analysis of variance (ANOVA) on the $\delta'_{\text{stimulus:absent}}$ scores was performed with factors of $|\Delta f|$ and $\text{sgn}(\Delta f)$. There was a significant main effect of $|\Delta f|$ [$F_{\text{GG}}(1.06, 7.42)=48.8$, $p_{\text{GG}}<0.000\,147$].[1] These results suggest that the bigger the frequency separation between the tones and the target, the more likely listeners are to report a galloping percept for the two-object stimuli. Although the ANOVA indicated there was a significant interaction between $|\Delta f|$ and $\text{sgn}(\Delta f)$ [$F(2,14)=8.85$, $p<0.003\,28$], paired-sample t tests (two-tailed with Dunn–Sidak *post hoc* adjustments for three planned comparisons) failed to support this result. Specifically, the paired t tests found no significant differences between positive and negative frequency differences for any of the frequency separations tested ($\Delta f=\pm 0.5$: $t_7=-2.61$, $p_{\text{DS}}=0.101$; $\Delta f=\pm 1$: $t_7=2.32$, $p_{\text{DS}}=0.152$; $\Delta f=\pm 2$ semitones: $t_7=1.69$, $p_{\text{DS}}=0.353$). Thus, there is little evidence that the sign of Δf influences the perceived rhythm of the tones.

3. Vowel judgments

Figure 5(d) shows the across-subject mean and the standard error of the raw response percentages for the vowel judgments. Unlike in the rhythmic judgment block of the experiment, there was a nonzero likelihood of subjects mislabeling the one-object prototypes [see squares to far right of Fig. 5(d)]. When the frequency difference between the tones and the target frequencies was zero, subjects were more likely to respond /ɪ/ [as if the target was not part of the vowel) than /ɛ/; (as if the target was part of the vowel; see diamond at far left of Fig. 5(d)]. As the magnitude of the frequency difference between the tones and the target increased, the probability of reporting an /ɛ/ increased (i.e., the target contributed more to the vowel; see open and filled triangles). As expected for the no-target control stimulus, subjects almost always responded /ɪ/, as if the target was not present [see X in Fig. 5(d)].

Because there were large individual differences in how consistently prototypes were labeled, transforming the data into d' scores increases the across-subject variability (not shown). Average d' values were lower overall than in the tone-rhythm experiment, consistent with the fact that listeners generally had more difficulty in identifying the vowel than labeling the tone rhythm (even for the one-object prototypes).

Transforming the results to δ' reduces the across-subject variability in d' by normalizing results by the differences in overall sensitivity [Fig. 5(e); comparison not shown]. In general, as the frequency difference between the repeated tones and the target increases, the likelihood that responses are like those to the /ɛ/ prototype increases. A two-way repeated-measure ANOVA was performed on the $\delta'_{\text{stimulus:absent}}$ results with factors of $|\Delta f|$ and sgn(Δf). The ANOVA found a significant main effect of $|\Delta f|$ [$F_{GG}(1.06, 7.39)=6.37$, $p_{GG}<0.0368$]. There were no main effect of sgn(Δf) [$F(1,7)=0.111$, $p=0.749$] and no significant interaction between $|\Delta f|$ and sgn(Δf) [$F(2,14)=1.55$, $p=0.246$]. Thus, as with the tone-rhythm task, we conclude that $|\Delta f|$ affects the contribution of the target to the attended object, but that there is no consistent effect of sgn(Δf).

4. Target trading

The percent responses found in the one-object control experiment provide mappings that allow us to evaluate whether there is a trading relationship between the level of target perceived in the tone stream and in the vowel. The individual psychometric functions that relate the target attenuation in a one-object stimulus to a percent response were used to find (for each subject and condition) the equivalent target attenuation for the perceived contribution of the target to the attended object [see Fig. 3(b)]. The across-subject means and standard errors of these mapped equivalent attenuations are plotted in Figs. 5(c) (for the tones) and 5(f) (for the vowel).

Conclusions drawn from analysis of raw percent responses and δ' (top two rows of Fig. 5) and of the effective attenuation of the target (bottom row of Fig. 5) are essentially the same. However, this final comparison enables a quantitative analysis of whether there is trading between the tones and the vowel.

Figure 6(a) plots the across-subject mean effective attenuation of the target in the tone stream against the mean attenuation of the target in the vowel. The plot shows all conditions that were common to the two experiments, including the two-object target-absent control. The solid curve in the figure shows the trading relationship that would be observed if energy is conserved, while the dashed line shows the trading relationship that corresponds to loss of 3 dB of target energy (consistent with pressure conservation).

Results for the target-absent control fall near the upper-right corner of the plot, as expected, indicating that the perceived qualities of the tone stream and vowel were consistent with a target that was strongly attenuated [see × in the top right of Fig. 6(a)]. For the ambiguous, two-object stimuli, trading occurs: The effective attenuation of the target in the tone stream is larger for stimuli that produce less attenuation

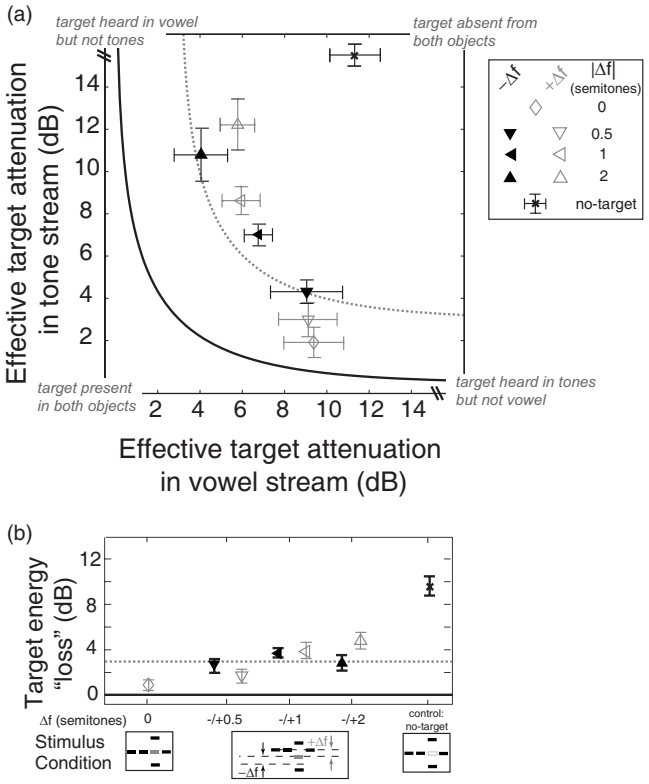


FIG. 6. (a) Scatter plot of the effective target attenuation in the tones vs the effective target attenuation in the vowel. Data would fall on the solid line if energy conservation holds. A trading relationship in which the total perceived target energy is 3 dB less than the physical target energy would fall on the dashed line (equivalent to conservation of pressure rather than energy; see Darwin, 1995). (b) The lost energy of the target for each condition, equal to the difference between the physical target energy and the sum of the perceived target energy in the tones and vowel. The solid line (0 dB lost energy) shows where results would fall if energy conservation held. The dashed line shows where results would fall if pressure, rather than energy, were conserved.

in the vowel [diamond and triangles fall on a monotonically decreasing curve in Fig. 6(a)]. However, the trading does not strictly follow energy conservation: For some conditions, the total of the sum of the effective energies of the target in the two streams is less than that actually present in the target (this can be seen in the fact that the data points fall above and right of the solid line in the figure).

To quantify the trading observed in Fig. 6(a), we computed the total effective energy of the target by summing, for each condition, its effective energy when subjects attended to the tones and its effective energy when subjects attended to the vowel. We then computed the "lost" target energy by subtracting the total effective target energy from the physical energy of the target. The across-subject means of these values are shown in Fig. 6(b).

In general, the total perceived target energy was less than the physical target energy in the stimuli (all symbols fall above 0 dB). The lost energy was near 3 dB for many of the stimuli (see dashed line at 3 dB, which is equivalent to pressure conservation; see also Darwin, 1995; McAdams *et al.*, 1998).

A two-way repeated-measure ANOVA on the total effective target energy lost found no effects of $|\Delta f|$ [$F(2,14)=4.29$, $p=0.0544$], sgn(Δf) [$F(1,7)=2.06$, $p=0.194$], or

their interaction [$|\Delta f|$ sgn(Δf), $F(2,14)=3.47$, $p=0.0599$]. One-sample t tests separately performed on the lost target energy values explored whether the lost energy was statistically significantly different from 0 dB (with Dunn–Sidak *post hoc* adjustments for seven planned comparisons). For conditions $\Delta f=-0.5$ ($t_7=-4.06$, $p_{DS}<0.0333$), $\Delta f=-1$ ($t_7=-9.30$, $p_{DS}<0.000\,242$), $\Delta f=+1$ ($t_7=-5.37$, $p_{DS}<0.007\,29$), and $\Delta f=+2$ semitones ($t_7=-6.13$, $p_{DS}<0.003\,33$), the lost energy was significantly greater than zero, supporting the conclusion that the total target energy allocated across the two competing objects is often less than the physical energy present in the target element of the sound mixture.

D. Discussion

In many past studies of this sort, adaptation in the periphery has been brought up as a possible explanation for the reduced contribution of the target to the harmonic complex (Darwin *et al.*, 1995). Peripheral adaptation could contribute to the lost target energy here, as well. However, such adaptation would be greatest when $|\Delta f|=0$ and the pair of tones excite the same neural population as the target. Instead, the amount of target energy that is lost is smallest, if anything, when $|\Delta f|=0$. Thus, adaptation does not fully account for the perceptual loss of target energy observed here. Moreover, adaptation was ruled out as the sole cause of lost target energy in similar studies when trading fails (also see the Discussion and Appendix in Shinn-Cunningham *et al.*, 2007).

While some target energy is not accounted for, we found trading: In general, the greater the contribution of the target to the tone stream, the smaller its contribution to the vowel. This finding is similar to past studies (Darwin, 1995; McAdams *et al.*, 1998) whose results were consistent with trading of an ambiguous element between two competing objects. However, this result contrasts with results using stimuli similar to the current stimuli with $|\Delta f|=0$ but in which the spatial cues of the tones and target were manipulated to change the relative strength of simultaneous and sequential groupings (Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). Thus, even though the stimuli and procedures employed in this study are very similar to those used when trading fails, results are more consistent with results of studies using very different methods. The difference between the current study and the past studies with similar stimuli suggests that the degree to which trading is observed depends on the way grouping cues are manipulated to alter perceptual organization.

None of these studies found energy conservation, where the perceived target energy in the competing objects sums to the physical energy of the target. In this respect, all results support the idea that the way in which the acoustic mixture is broken into objects is inconsistent with a veridical parsing of the acoustic mixture despite the intuitive appeal of this idea (see Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008).

III. EXPERIMENT 2: NO COMPETING OBJECTS

In the first experiment, subjects were less likely to hear the target as part of the tone stream as the frequency separation between the repeating tones and the target increased. In conditions where $|\Delta f|=2$ semitones, subjects reported a strong galloping percept. However, studies using one-object tone stimuli generally find that a two-semitone difference is not enough to cause a single object to break apart into two streams (Anstis and Saida, 1985; Vliegen and Oxenham, 1999; Carlyon *et al.*, 2001; Micheyl *et al.*, 2005). While it is likely that the presence of the vowel, which competes for ownership of the target, explains why relatively small $|\Delta f|$ lead to percepts of a galloping rhythm in our two-object stimuli, other differences between past experiments and our main experiment may also contribute. A follow-up experiment was conducted to directly assess whether the strong effect of a relatively small frequency separation on the perceived tone-stream rhythm was due to some procedural or stimulus differences between the current and past one-stream versus two-stream studies.

A. Methods

Stimuli were similar to the one-object tone stimuli used in the main experiment. In the main experiment, frequency separations of only up to two semitones were tested, as any bigger separation would make the repeating tones closer to neighboring harmonics than to the target (which could cause the repeating tones to capture those harmonics rather than the target). In this one-object experiment, there were no such constraints, and we tested separations between the repeating tones and the target of up to eight semitones (0, ± 0.5, ± 1, ± 2, ± 4, and ± 8 semitones relative to 500 Hz) to make results more comparable to previous one-stream versus two-stream experiments.

Subjects were instructed to judge the tone-stream rhythm (galloping versus even) after ten presentations of the pair-of-tones-target triplet. Eight subjects participated in this experiment, six of whom participated in the previous experiment and two who had previously participated in and passed the screening criteria in related experiments conducted in our laboratory.

B. Results

All subjects could nearly perfectly distinguish the galloping from the even prototypes. Figure 7(a) shows the raw percent response scores and Fig. 7(b) shows the $\delta'_{\text{stimulus:absent}}$ results, averaged across subjects. Both ways of considering the data show that the contribution of the target to the tones decreases as $|\Delta f|$ increases, as expected [data points fall along a monotonically decreasing curve in Figs. 7(a) and 7(b)].

A two-way repeated-measure ANOVA on the one-stream tone rhythm $\delta'_{\text{stimulus:absent}}$ found a significant main effect of $|\Delta f|$ [$F_{GG}(1.74,12.2)=67.0$, $p_{GG}<4.27\times10^{-7}$] but no significant effect of sgn(Δf) [$F(1,7)=0.078$, $p=0.789$] and no significant effect of the interaction between $|\Delta f|$ and sgn(Δf) [$F(4,28)=0.455$, $p=0.768$].

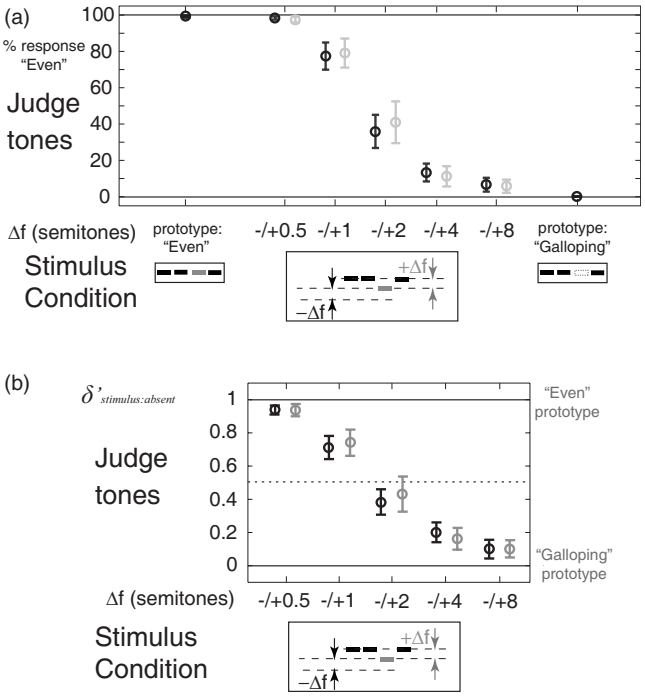


FIG. 7. (a) Raw percent responses for one-object stimuli as a function of Δf between the tones and the target. The target-present prototype is equivalent to the $\Delta f=0$ condition. Note that there is no equivalent vowel manipulation in this experiment. (b) Normalized $\delta'_{\text{stimulus:absent}}$ results, derived from the raw percent-category responses in (a).

Compared to the results for two-object conditions, the likelihood of hearing the repeating tones as galloping increased much more slowly with increasing $|\Delta f|$. Responses are perceptually only halfway between galloping and not galloping for separations of two semitones.

C. Discussion

The same $|\Delta f|$ was much less likely to lead to galloping responses in this one-object experiment than for two-object stimuli in our main experiment. In the main experiment, subjects judged the tone stream as galloping when the absolute frequency separation between the tones and the target was two semitones apart. In the absence of the competing harmonic complex, subjects only consistently judged the tone stream to be galloping when the absolute frequency separation between the tones and the target was four or more semitones apart. This suggests that the presence of the competing vowel made listeners more likely to report a galloping rhythm for a fixed $|\Delta f|$.

Past studies show that the potency of a particular frequency separation on streaming depends on the repetition rate, subject instructions, and even the musical training of the subjects (Vliegen and Oxenham, 1999). Along these lines, it is conceivable that the weaker effect of frequency separation observed in this experiment compared to that in the main experiment is wholly or partially due to the difference in the frequency separation ranges used (up to a maximum $|\Delta f|$ of two semitones in the main experiment but up to eight semitones in the one-object experiment). Regardless, the results of this control experiment show that our procedures produce results consistent with the literature when we use similar frequency separation ranges and one-object mixtures.

IV. GENERAL DISCUSSION

Many past studies of auditory object and stream formation investigated the potency of different acoustical grouping cues; however, the majority focused exclusively on either sequential grouping (Van Noorden, 1975; Anstis and Saida, 1985; Vliegen and Oxenham, 1999; Roberts *et al.*, 2002) or simultaneous grouping (Culling and Summerfield, 1995; Darwin and Hukin, 1997; Drennan *et al.*, 2003; Dyson and Alain, 2004). Most of these studies explored what acoustical parameters would lead sound elements to be heard as one object and what would lead the stimulus to break apart into two perceptual objects. However, in everyday complex settings, multiple acoustical objects often coexist. In such situations, it is more natural to ask how simultaneous objects interact and influence grouping of ambiguous elements that can logically belong to more than one object in the auditory scene rather than whether a mixture is heard as one or two objects.

In our main experiment, the presence of the tones at the same frequency as the target is enough to substantially remove the contribution of the target to the harmonic complex. Similarly, the presence of the harmonic complex reduces the contribution of the target to the tone stream. Results of experiment 2 suggest that the presence of the competing harmonic complex causes the target to "drop out" of the tone stream at smaller $|\Delta f|$ than when there is no competing object. These results are consistent with past work showing that the perceived content of an object depends on interactions with other objects in the mixture (e.g., influencing both perceived pitch as well as perceived vowel identity; Darwin *et al.*, 1995; Darwin and Hukin, 1997).

If the perceptual organization of the auditory scene is fixed and veridical, the sum of the energies at a given frequency in all of the perceived objects should, on average, equal the physical energy of that frequency in the signal reaching the ear, obeying the energy conservation hypothesis. A weaker hypothesis, of trading, is that when the perceptual contribution of the ambiguous elements to one object in a stimulus increases, the contribution to other objects decreases.

Current results are consistent with the target trading between the harmonic complex and tones. However, as in past studies, the contribution of the ambiguous target to the tones plus its contribution to the vowel is as much as 3 dB less than the energy that is physically present in the target. Nonlinearity in peripheral processing could partially explain trading in which energy in an ambiguous element is lost. For instance, if there is adaptation in the perceived level of the target due to the presence of the repeated tones, the total perceived energy in the two-object mixture could be less than the true physical energy of the target. However, peripheral adaptation cannot fully explain the current results. Peripheral adaptation of the target should be maximal when the tones and target have the same frequency. Instead, the target

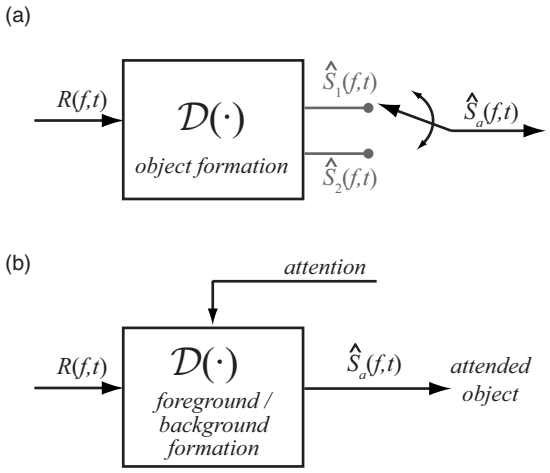


FIG. 8. Two possible models for how objects are formed. (a) A model in which the grouping of the scene depends only on the stimuli. (b) A model in which the grouping of the object in the foreground depends on top-down goals of the listener.

energy that is lost is smallest when the target frequency matches that of the repeated tones. Thus, as in past studies investigating the energy conservation hypothesis (Darwin, 1995; McAdams *et al.*, 1998; Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008), peripheral effects may contribute to, but cannot explain, results.

Although results of the current study and some past studies find trading that is roughly consistent with pressure conservation, not all past studies show trading. The current results, which are consistent with trading, use stimuli and methods that are closer to those used in studies in which trading fails rather than where trading occurs. The only difference between these studies is the grouping cues that were manipulated to affect the perceptual organization of the scene. Taken together, these results suggest that the fact that trading is roughly obeyed in some studies is coincidental, and is likely due to the specific stimulus manipulation employed in a study and not because trading is a rule governing auditory scene analysis. Instead, there are two possibilities that could explain current and past results, diagramed in Fig. 8 and considered below.

In Fig. 8(a), $\mathcal{D}(\cdot)$ represents an operator that, by utilizing all available information from the signal reaching the receiver, $R(f,t)$, yields estimates of the objects in the scene $\{\hat{S}_i(f,t)\}$, independent of the goals of the listener (such as the task s/he is performing or the object s/he is attending). In this scheme, attention simply selects one of the already-formed objects as the foreground object. If $\mathcal{D}(\cdot)$ operates in a purely bottom-up manner, then competition between different objects may alter the effective level of ambiguous sound energy through mutual inhibition. Such inhibition could cause the perceptual contribution of an ambiguous target to the objects in the scene to be less than the physical target energy through fixed interactions that depend only on the stimulus. To explain why trading sometimes fails, the strength of this mutual inhibition must depend on the balance between different competing grouping cues, such as spatial cues, frequency proximity, common modulation, and the like.

Alternatively, the goals of the listener may change how energy in the sound mixture is grouped, changing the operation of $\mathcal{D}(\cdot)$ [Fig. 8(b)]. If so, then which object or attribute we attend may alter how we group the mixture (Fig. 8; see also Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). If this is the case, then there is no reason to expect the perceived content of an attended foreground object to predict what a listener perceives when they switch attention or switch tasks.

The experimental goal of the current study is to estimate how $\mathcal{D}(\cdot)$ operates on the sound mixture. To probe this operation, we asked subjects to subjectively judge properties of an auditory object in a scene. However, by asking the subjects to make judgments about an object, we forced them to bring the object of interest into the auditory foreground. We cannot ask listeners to make judgments of objects in the perceptual background: Even if we did, they would undoubtedly focus attention on the background, which would change the old background into the new foreground. As a result, the current experiments cannot differentiate between the two possibilities diagramed in Fig. 8. Either possibility is consistent with the current results: Either a sound mixture is formed into objects in a way that depends only on the stimulus mixture but that does not obey energy conservation (or in some cases, trading) or the focus of attention and/or the goals of the listener affect how the sound mixture is grouped.

Attention modulates neural responses within the visual processing pathway (e.g., Reynolds *et al.*, 2000; Martinez-Trujillo and Treue, 2002; Reynolds and Desimone, 2003), including at the peripheral level (Carrasco *et al.*, 2004). In audition, spectrotemporal receptive fields in the primary auditory cortex change depending on the behavioral task (Fritz *et al.*, 2003; Fritz *et al.*, 2005). Attention has also been implicated in corticofugal modulation of cochlear function in awake mustached bats during vocalization (Suga *et al.*, 2002). These physiological results suggest that top-down processes alter how sound is represented even in relatively peripheral, sensory processing stages of the auditory pathway. While such results do not prove that the way we form auditory objects out of an ambiguous sound mixture depends on top-down factors, they are consistent with the idea that attention alters auditory scene analysis.

V. CONCLUSIONS

Competing objects alter the perceived content of an object in an auditory scene. In particular, there are reciprocal effects between simultaneously and sequentially grouped objects in our two-object mixtures, made up of a repeating tone sequence, a simultaneous vowel complex, and an ambiguous target tone that could logically belong to either object. When the frequency separation between the tone sequence and an ambiguous target tone is varied to alter the perceptual organization of the sound mixture, the contributions of the target to the tones and to the competing vowel obey a loose trading relationship. However, the trading is lossy rather than obeying energy conservation. When the repeated tones and target had slightly different frequencies, the total perceived target energy was roughly 3 dB less than the physical target energy.

While it is possible that some peripheral nonlinearity contributes to this loss of target energy, it cannot fully account for these findings.

These results, as well as past results, suggest either that (1) competing auditory objects mutually suppress ambiguous sound elements, leading to a reduction in the perceptual contribution of the ambiguous element to the objects in a sound mixture, or (2) how an auditory object is formed in a sound mixture depends on top-down goals of the listener. Further work is necessary to tease apart these two possibilities.

ACKNOWLEDGMENTS

This work was supported by a grant from the Office of Naval Research (N00014-04-1-0131) to B.G.S.C. Sigrid Nasser helped in the subject recruitment and the data collection process. Andrew J. Oxenham provided many helpful suggestions in the experimental design.

[1]Throughout, the subscript "GG" denotes that we used the Greenhouse–Geisser-corrected degrees of freedom when testing for significance to account for violations of the sphericity assumption under Mauchly's test. Where the subscript is left out, it signifies a condition for which the sphericity assumption was met.

Anstis, S., and Saida, S. (**1985**). "Adaptation to auditory streaming of frequency-modulated tones," J. Exp. Psychol. Hum. Percept. Perform. **11**, 257–271.

Beauvois, M. W., and Meddis, R. (**1996**). "Computer simulation of auditory stream segregation in alternating-tone sequences," J. Acoust. Soc. Am. **99**, 2270–2280.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge).

Bregman, A. S., and Pinker, S. (**1978**). "Auditory streaming and building of timbre," Can. J. Psychol. **32**, 19–31.

Carlyon, R. P. (**2004**). "How the brain separates sounds," Trends Cogn. Sci. **8**, 465–471.

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol. Hum. Percept. Perform. **27**, 115–127.

Carrasco, M., Ling, S., and Read, S. (**2004**). "Attention alters appearance," Nat. Neurosci. **7**, 308–313.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds-Absence of across—frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785–797.

Dannenbring, G. L., and Bregman, A. S. (**1978**). "Streaming vs. fusion of sinusoidal components of complex tones," Percept. Psychophys. **24**, 369–376.

Darwin, C. J. (**1995**). "Perceiving vowels in the presence of another sound: A quantitative test of the 'Old-plus-New heuristic'," in *Levels in Speech Communication: Relations and Interactions: A tribute to Max Wajskop*, edited by C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen (Elsevier, Amsterdam), pp. 1–12.

Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.

Darwin, C. J., and Hukin, R. W. (**1997**). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," J. Acoust. Soc. Am. **102**, 2316–2324.

Darwin, C. J., and Hukin, R. W. (**1998**). "Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony," J. Acoust. Soc. Am. **103**, 1080–1084.

Darwin, C. J., Hukin, R. W., and al-Khatib, B. Y. (**1995**). "Grouping in pitch perception: Evidence for sequential constraints," J. Acoust. Soc. Am. **98**, 880–885.

Darwin, C. J., and Sutherland, N. S. (**1984**). "Grouping frequency components of vowels—When is a harmonic not a harmonic," Q. J. Exp. Psychol. A **36**, 193–208.

Drennan, W. R., Gatehouse, S., and Lever, C. (**2003**). "Perceptual segregation of competing speech sounds: The role of spatial location," J. Acoust. Soc. Am. **114**, 2178–2189.

Dyson, B. J., and Alain, C. (**2004**). "Representation of concurrent acoustic objects in primary auditory cortex," J. Acoust. Soc. Am. **115**, 280–288.

Fritz, J., Shamma, S., and Elhilali, M. (**2005**). "One click, two clicks: The past shapes the future in auditory cortex," Neuron **47**, 325–327.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Neurosci. **6**, 1216–1223.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hartmann, W. M., and Johnson, D. (**1991**). "Stream segregation and peripheral channeling," Music Percept. **9**, 155–184.

Hukin, R. W., and Darwin, C. J. (**1995**). "Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel," J. Acoust. Soc. Am. **98**, 1380–1387.

Kashino, M., and Warren, R. M. (**1996**). "Binaural release from temporal induction," Percept. Psychophys. **58**, 899–905.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Lee, A. K. C., and Shinn-Cunningham, B. G. (**2008**). "Effects of reverberant spatial cues on attention-dependent object formation," J. Assoc. Res. Otolaryngol. **9**, 150–160.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide* (Erlbaum, Hillsdale, NJ).

Martinez-Trujillo, J. C., and Treue, S. (**2002**). "Attentional modulation strength in cortical area MT depends on stimulus contrast," Neuron **35**, 365–370.

McAdams, S., Botte, M. C., and Drake, C. (**1998**). "Auditory continuity and loudness computation," J. Acoust. Soc. Am. **103**, 1580–1591.

McCabe, S. L., and Denham, M. J. (**1997**). "A model of auditory streaming," J. Acoust. Soc. Am. **101**, 1611–1621.

Micheyl, C., Tian, B., Carlyon, R. P., and Rauschecker, J. P. (**2005**). "Perceptual organization of tone sequences in the auditory cortex of awake macaques," Neuron **48**, 139–148.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Reynolds, J. H., and Desimone, R. (**2003**). "Interacting roles of attention and visual salience in V4," Neuron **37**, 853–863.

Reynolds, J. H., Pasternak, T., and Desimone, R. (**2000**). "Attention increases sensitivity of V4 neurons," Neuron **26**, 703–714.

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. **112**, 2074–2085.

Shinn-Cunningham, B. G. (**2005**). "Influences of spatial cues on grouping and understanding sound," in Proceedings the Forum Acusticum 2005, Budapest, Hungary.

Shinn-Cunningham, B. G., Lee, A. K. C., and Oxenham, A. J. (**2007**). "A sound element gets lost in perceptual competition," Proc. Natl. Acad. Sci. U.S.A. **104**, 12223–12227.

Steiger, H., and Bregman, A. S. (**1982**). "Competition among auditory streaming, dichotic fusion, and diotic fusion," Percept. Psychophys. **32**, 153–162.

Suga, N., Xiao, Z. J., Ma, X. F., and Ji, W. Q. (**2002**). "Plasticity and corticofugal modulation for hearing in adult animals," Neuron **36**, 9–18.

Turgeon, M., Bregman, A. S., and Ahad, P. A. (**2002**). "Rhythmic masking release: Contribution of cues for perceptual organization to the cross-spectral fusion of concurrent narrow-band noises," J. Acoust. Soc. Am. **111**, 1819–1831.

Turgeon, M., Bregman, A. S., and Roberts, B. (**2005**). "Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping," J. Exp. Psychol. Hum. Percept. Perform. **31**, 939–953.

Van Noorden, L. P. A. S. (**1975**). *Temporal Coherence in the Perception of Tone Sequences*, Ph.D. Thesis (Institute for Perception Research, Eindhoven), pp. 1–127.

Vliegen, J., and Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am. **105**, 339–346.

Warren, R. M., Bashford, J. A., Healy, E. W., and Brubaker, B. S. (**1994**). "Auditory induction—Reciprocal changes in alternating sounds," Percept. Psychophys. **55**, 313–322.

The effect of a precursor on growth of forward masking[a)]

Vidya Krull and Elizabeth A. Strickland[b)]
Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana 47907-2038

(Received 10 September 2007; revised 25 March 2008; accepted 31 March 2008)

This study examined the effect of an on-frequency precursor on growth-of-masking (GOM) functions measured using an off-frequency masker. The signal was a 6-ms, 4-kHz tone. A GOM function was measured using a 40-ms, 2.8-kHz tone (the off-frequency masker). GOM functions were then measured with an on-frequency, fixed level precursor presented before the off-frequency masker. The precursor was 50 or 60 dB SPL, and 160 ms in duration. For the 60-dB SPL precursor, a 40-ms duration was also used. Two-line functions were fit to the GOM data to estimate the basilar membrane input-output function. The precursors reduced the gain of the input-output function, and this decrease was graded with precursor level. Both precursor durations had the same effect on gain. Changes in masking following a precursor were larger than would be predicted by additivity of masking. The observed decrease in gain may be consistent with activation of the medial olivocochlear reflex by the precursor.
 [DOI: 10.1121/1.2912440]

PACS number(s): 43.66.Dc, 43.66.Mk, 43.66.Ba [MW] Pages: 4352–4357

I. INTRODUCTION

Most sounds of interest are time varying in nature, so the response of the auditory system must be examined for dynamic as well as steady signals. This response changes for a short period after the presentation of a sound. This sound will be called a precursor. If the precursor is followed by a short-duration signal, it may increase the threshold for detecting the signal, an effect called forward masking. If the signal is presented with a simultaneous masker, the precursor may make the signal audible at a lower signal-to-masker ratio. This effect has been called overshoot (Zwicker, 1965a) or the temporal effect (Hicks and Bacon, 1992). That is, in simultaneous masking, a precursor may make a subsequent sound easier to hear, but in forward masking, a precursor may have the opposite effect. The point of this paper is to try to reconcile these two effects with one underlying mechanism.

The physiological bases of forward masking are not well understood, and indeed, forward masking probably depends on processes at several levels of the auditory system. One hypothesis is that forward masking is related to neural adaptation (Smith 1977, 1979). Neural firing decreases, or adapts, during the course of the precursor. If another sound, a signal, follows the precursor, the firing to the signal is decreased from what it would be if the signal were preceded by silence (Harris and Dallos, 1979). This decrease in firing might be expected to increase the threshold for the signal, i.e., the signal would be masked. An analysis of firing rates using signal detection theory predicts much less masking at the level of the auditory nerve than is seen in psychophysical tests (Relkin and Turner, 1988). Meddis and O'Mard (2005) were able to adjust an auditory model to predict psychophysical forward masking results from neural adaptation. Oxenham (2001) also showed that neural adaptation, modeled as a decrease in gain, was able to predict many forward masking results, if compression in the cochlea was also included in the model.

A second theory for forward masking is persistence of excitation. That is, the response to the precursor persists and overlaps with the response to the signal, making it harder to hear the signal. This is not seen at the level of the auditory nerve, but has been hypothesized to involve a "temporal window" at some higher level of the auditory system (Moore *et al.*, 1988). Oxenham (2001) found that if compression in the cochlea was included, the temporal window model was able to predict forward masking results.

A third hypothesis is that efferent feedback from the central auditory system is responsible for some forward masking. Shore (1998) observed changes in forward masking in the ventral cochlear nucleus after lesions to efferent pathways.

The present experiment was developed to examine the hypothesis that forward masking is partially due to efferent feedback, but at the level of the cochlea. There would be a known physiological basis in the medial olivocochlear reflex (MOCR). The MOCR is a decrease in the gain of the active process of the cochlea, caused by activation of the medial olivocochlear bundle of efferent fibers (Warr and Guinan, 1979; Warr, 1980; Liberman, 1989). Oxenham (2001) modeled neural adaptation as a decrease in gain, although at a postcochlear level, and the time course of recovery is in the same range as the offset of MOCR effects (Backus and Guinan, 2006). Temporal effects in simultaneous masking have been successfully modeled on the basis of a frequency-specific decrease in gain in the cochlea, which would be consistent with MOCR activation (Strickland 2001, 2004,

[a)]Portions of this research were presented at the 30th Midwinter Meeting of the Association for Research in Otolaryngology, Denver, CO, February 2007.
[b)]Electronic mail: estrick@purdue.edu

2008; Strickland and Krishnan, 2005). Although neural adaptation has also been proposed as a basis for the temporal effect in simultaneous masking, it cannot explain all aspects of the temporal effect (Bacon and Healy, 2000). The temporal window model does not predict the temporal effect in simultaneous masking.

The MOCR hypothesis for forward masking would be consistent with a decrease in gain, but at the level of the cochlea. A decrease in the gain of the active process would be expected to produce effects that would be distinguishable from the effects of a temporal window or of neural adaptation. Specifically, a decrease in the gain of the active process should produce a decrease in the gain of the input-output function of the cochlea.

This hypothesis may be studied using a paradigm called additivity of masking. In this paradigm, the masking produced by two maskers is compared to the masking produced by each masker alone. Previous research on the effects of two maskers has shown that it can be assumed that the effects of the maskers add linearly, if cochlear compression is taken into account (e.g., Penner and Shiffrin, 1980; Oxenham and Moore, 1994). The premise of the present study is that there are probably at least two types of forward masking. Forward masking by short maskers seems unlikely to be due to the efferent feedback to the cochlea. The most rapid MOCR effects fall in the range of 60–80 ms (Backus and Guinan, 2006). Therefore, a short forward masker may be used that should not activate the MOCR. Growth-of-masking (GOM) functions have been used in forward masking to obtain estimates of the cochlear input-output function (Oxenham and Plack, 1997). A masker approximately an octave below the signal frequency is used to mask a short signal, as a function of signal level. If the masker response is linear, the thresholds give an estimate of the input-output function at the signal frequency place. Thus, a GOM function measured with a short masker and signal should give an estimate of gain at the signal frequency place.

In the remainder of the paper, the longer masker will be referred to as the "precursor." For longer precursor durations, efferent feedback could play a role in forward masking. Because the MOCR is frequency specific, a long-duration precursor will be presented at the signal frequency, and the GOM function measured. If input-output functions estimated from the two GOM functions show a decrease in gain following a precursor, this would be consistent with MOCR activation. The signal threshold will also be measured with the precursor but no short forward masker, so that the results may be analyzed in terms of additivity of masking.

II. METHODS

A. Stimuli

The signal was a 6-ms, 4.0-kHz sinusoid, with 3-ms cosine-squared onset and offset ramps (no steady state). This frequency was chosen because large temporal effects in simultaneous masking have been found for this signal frequency. The off-frequency masker was a 2.8-kHz tone with a duration of 40 ms, including 5-ms cosine-squared gating. This masker duration was chosen to minimize activation of the MOCR, yet enable thresholds to be measured within the range of the equipment. There was no delay between masker offset and signal onset. In the precursor conditions, a 4.0-kHz precursor preceded the off-frequency masker with no delay between precursor offset and masker onset. Previous studies using a separate precursor in simultaneous masking have reported a temporal effect at the 40-ms precursor-signal delay used in this study (Bacon and Smith, 1991; Bacon and Healy, 2000).

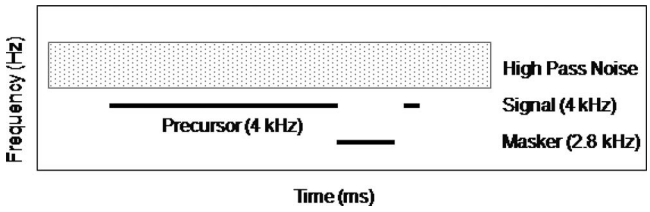


FIG. 1. Schematic showing the spectral (y-axis) and temporal (x-axis) characteristics of the (4 kHz, 6 ms) signal, (2.8 kHz, 40 ms) off-frequency masker, (4 kHz) on-frequency precursor with variable duration and level (bold lines), and high-pass noise (hatched rectangle).

The effects of precursor level and duration were examined. The precursor level was fixed at 50 or 60 dB SPL, based on pilot data that showed that these levels were effective. The duration was set at 160 ms. This duration was based on previous data in simultaneous masking that showed that the temporal effect plateaued for a signal delay of about 200 ms from masker onset (Zwicker, 1965a). The combined duration of the precursor and masker was 200 ms. The effect of precursor duration was also examined by using a 40-ms, 60-dB SPL precursor.

Throughout the experiment, high-pass noise was presented to prevent off-frequency listening. The lower cutoff frequency was $1.2f_s$, where f_s refers to the signal frequency. The spectrum level of this high-pass noise was set at 40 dB below the signal level. This level was used in a similar study (Rosengard *et al.*, 2005). The high-pass noise was turned on 50 ms before precursor onset and turned off 50 ms after the offset of the signal, to avoid confusion with the other stimuli. Conditions are shown schematically in Fig. 1.

The stimuli were created digitally and were routed through four separate D/A channels (TDT DA3-4). They were low pass filtered at 10 kHz (TDT FT5 and FT6-2). The levels of the stimuli were controlled by programmable attenuators (TDT PA-4), mixed (TDT-SM3) and routed to a headphone buffer (TDT HB6) prior to presentation through one of two ER-2A insert earphones to a listener seated in a sound-treated booth. These earphones have a flat frequency response from 250 to 8000 Hz.

B. Procedures

A three-interval forced choice task with a two-down, one-up stepping rule was used to determine thresholds. Subjects were asked to identify the interval containing the signal by pressing a key on a computer keyboard. Visual feedback was provided via a computer monitor. Within each trial, the signal level was fixed and the level of the off-frequency masker varied based on the response. The initial step size was 5 dB, and decreased to 2 dB after the second reversal.

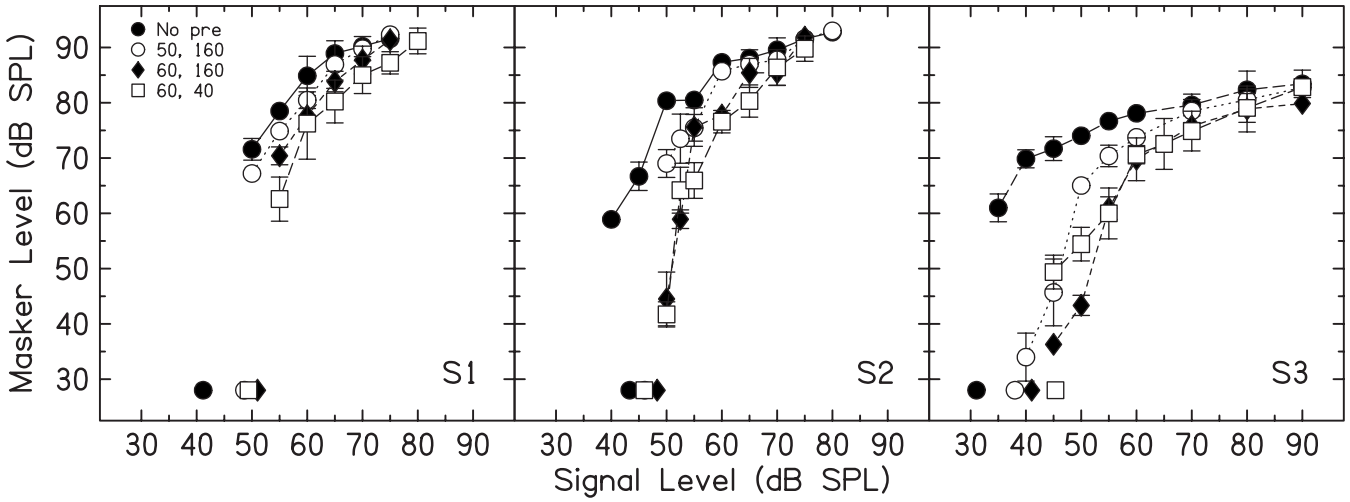


FIG. 2. Plots showing the masker level necessary to mask a signal, plotted as a function of signal level without a precursor (filled circle) and with the three different precursors (open circles: 50 dB SPL, 160 ms; filled diamonds: 60 dB SPL, 160 ms; open squares: 60 dB SPL, 40 ms). Symbols at bottom left of individual panels indicate signal thresholds obtained for each of the four conditions without the off-frequency masker.

Thresholds were taken as the mean of the last even number of reversals at the smaller step size in a set of 50 trials. This adaptive tracking procedure estimated the 70.7% correct point on the psychometric function (Levitt, 1971). The thresholds from at least two runs were averaged to obtain the final threshold. Runs were discarded if the standard deviation exceeded 5 dB or the threshold exceeded the limits of the equipment. Signal thresholds in quiet or following the precursor alone (with no masker) were also measured. When no masker was present, the delay between the precursor and signal was still fixed at 40 ms. Data were collected over several experimental sessions, each lasting 1–1.5 hours.

The listener without previous experience in psychoacoustic tasks was trained for 2–3 hours prior to data collection. In a given session, data were collected for all precursor conditions at one or two signal levels.

C. Subjects

Three subjects participated in this study. All subjects had air conduction thresholds within normal limits (<20 dB HL) and normal middle ear function bilaterally. There were two females and one male, with a median age of 28 years. Two of these subjects had prior experience in psychoacoustic tasks.

III. RESULTS

GOM functions for the three listeners are shown in Fig. 2. Filled circles are off-frequency masker thresholds with no precursor. Thresholds are also shown for the off-frequency masker following the 50-(open circles) and 60-dB SPL (filled diamonds) 160-ms precursors, and the 60-dB SPL, 40-ms precursor (open squares). The on-frequency precursor decreased masker thresholds for low signal levels, but had little effect at high signal levels. This effect increased with precursor level. For a 60-dB SPL precursor, there was little to no effect of decreasing the precursor duration from 160 to 40 ms.

Signal thresholds were also obtained in the presence of the on-frequency precursor with no off-frequency masker. These are indicated by symbols at the bottom of individual panels. The precursors cause a shift in threshold of approximately 2–15 dB. In a few cases, listeners were able to detect a signal that was slightly below quiet threshold, in the presence of the off-frequency masker. For example, for S2, the absolute threshold of the signal alone was 43 dB SPL (filled circle), but when the off-frequency masker was presented, the signal could be detected at 40 dB SPL. This slight improvement in signal threshold with a forward masker has been seen in several subjects, and has been observed previously (Zwislocki *et al.*, 1959).

Input-output functions were estimated for each of the precursor conditions using the lower two segments of a three-line function described by Plack *et al.* (2004).

$$L_{\text{out}} = L_{\text{in}} + G \quad (L_{\text{in}} \leq \text{BP}_1), \tag{1}$$

$$L_{\text{out}} = cL_{\text{in}} + k_1 + G \quad (\text{BP}_1 < L_{\text{in}} \leq \text{BP}_2), \tag{2}$$

where L_{in} was the level of the signal, L_{out} was the estimated output for a given signal level, and G, c, and BP_1 were free parameters. The two-line function had a slope of 1 below the lower breakpoint (BP_1), and a compressive slope (c) between the two BPs. The correction factor k_1, where $k_1 = \text{BP}_1(1-c)$, ensured that the two lines met at the breakpoint. As the GOM data did not show a BP at higher signal levels, BP_2 was assumed to be fixed at 100 dB SPL for all subjects. For some listeners, the presence of the precursor produced a slope greater than 1 on the lower leg of the GOM (e.g., S3). This has been seen before with a precursor (Strickland, 2008), and may be due to an effect near threshold which is not predicted by the model (see Plack and Skeels, 2007, for a discussion). These points were excluded from the fit. A least-squares minimization procedure was used to estimate the free parameters. The parameter estimates from the model are shown in Table I. The model fit the experimental data very well, with rms errors typically less than 3 dB.

In examining Table I, it can be seen that BP_1 increased with the addition of the on-frequency precursor for all subjects. The slope c also increased for S1 and S2. This change

TABLE I. Parameters from the two-line fit to the data using a technique derived from Plack *et al.* (2004). G =gain, c=slope of compressive function, BP_1=lower breakpoint, BP_2=upper breakpoint. BP_2 was fixed at 100 dB for all subjects and conditions.

Subject	Precursor level (dB)	Precursor duration (ms)	BP1	c	G	rms error
S1	No	···	65.54	0.27	23.47	0.98
	50	160	69.84	0.51	19.87	1.40
	60	160	82.22	0.85	17.20	1.19
	60	40	70.00	0.74	13.50	2.82
S2	No	···	60.39	0.32	26.22	2.22
	50	160	63.76	0.44	22.26	1.93
	60	160	75.51	0.88	16.25	4.76
	60	40	82.84	0.87	14.80	2.04
S3	No	···	43.67	0.39	27.93	0.98
	50	160	59.05	0.29	15.21	0.59
	60	160	67.61	0.40	6.39	1.90
	60	40	67.85	0.20	7.85	1.27

increased with precursor level. As a result, the maximum gain G decreased with the addition of an on-frequency precursor. A 10-dB increase in on-frequency precursor level resulted in a decrease in G of approximately 2.5–9 dB. The short duration (40 ms) precursor produced the same decrease in gain as a longer (160 ms) precursor of the same level.

IV. DISCUSSION

Although the results show a decrease in the gain of input-output functions following a precursor, it is possible that they could simply reflect additivity of the masking of the on-frequency precursor and the off-frequency masker. This possibility is explored below.

A. Additivity of masking

Many previous studies have examined the combined effects of two temporally nonoverlapping maskers on signal threshold, based on their individual effects. This has been called additivity of masking. The results in these studies may be explained if it is assumed that the effects of the two maskers add linearly. For example, suppose there are two maskers, and each is at the level at which it just masks a 70-dB SPL tone. Then when the two maskers are presented sequentially, the threshold for the tone should increase to 73 dB SPL, as if the intensity effects of the two maskers are added linearly. The threshold for the tone may increase more than 3 dB, and this has been attributed to compression.

Estimates for the individual effects for the precursor and masker were obtained by assuming that the GOM measured without a precursor estimated the input-output function at the signal place. The masker effect of the precursor alone was estimated by using the signal threshold in the presence of the precursor alone (symbols at the bottom of Fig. 2) and using the GOM function to estimate the output level. For example, for S3, the threshold for the signal following a 50-dB SPL precursor was 38 dB SPL (open circle). On the input-output function, using the function fitted by the equations, this signal would be masked by an off-frequency masker of 66 dB SPL. Thus, this fixed precursor has an effective level of 66 dB SPL. The precursor level was fixed for a given GOM function, so it was assumed that its effective level was constant. The level of the masker varied according to the signal level. The effective masker level was taken from the input-output function for each signal level. For example, for S3, the masker level needed to mask a 50-dB SPL signal was 74 dB SPL. Now, when the precursor is presented before the masker, what level should the masker be so that the combined effect of the two together is 74 dB SPL? The masker response is assumed to be linear. By subtracting the intensities, it can be determined that a masker level of 73.3 dB SPL would be needed. The actual masker level measured for a 50-dB SPL signal following a 50-dB SPL precursor was 65 dB SPL. This level is much lower than would be predicted by additivity of masking.

Figure 3 shows the masker levels predicted by additivity of masking for the 50- and 60-dB SPL precursors (symbols), along with the data from Fig. 2 replotted as lines. For S1 and S2, the predicted results for the two precursors overlie each other for most signal levels, while the actual data do not. It can be seen that the precursors cause a larger change in masker level than is predicted by additivity of masking.

B. Decrease in gain

Now consider the hypothesis that the gain of the GOM function does decrease following a long-duration precursor at the signal frequency. The interpretation would be that the on-frequency precursor turned down the gain in the cochlea, while the off-frequency masker produced some other type of forward masking. The decrease in gain is graded with precursor level. The decrease in gain observed here is similar to that reported in a simultaneous masking study by Strickland (2008). She reported a decrease in maximum gain of 4–6 dB for every 10 dB increase in precursor level, which is similar to the 2.5–9 dB decrease in the present data. That study also found a decrease in compression with a precursor for some listeners and not others, as in the present study. These data are consistent with Rosengard *et al.* (2005), who found that

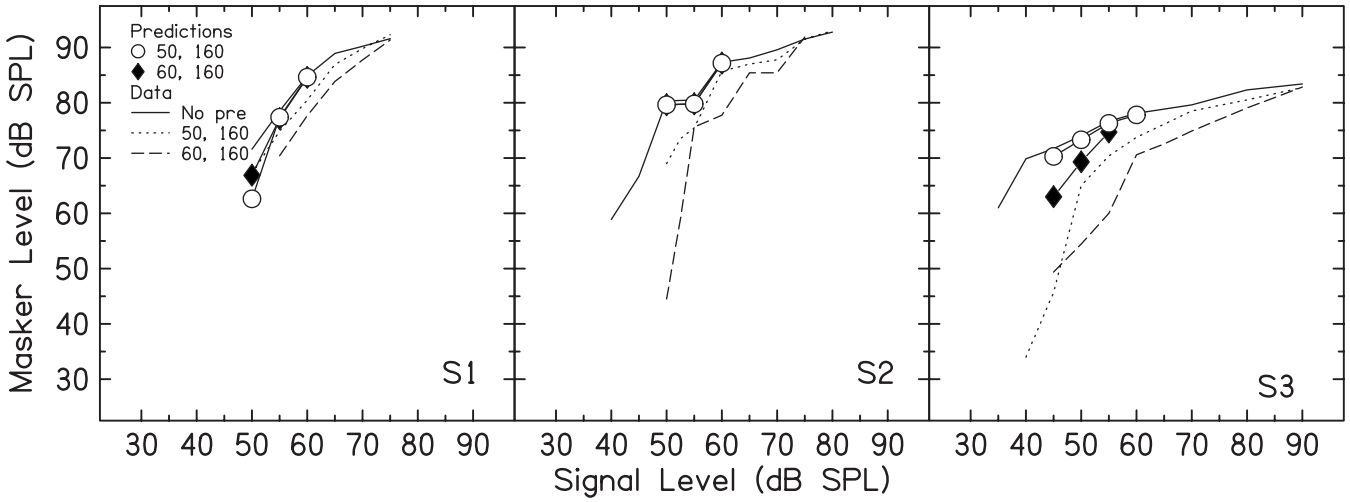


FIG. 3. Predictions using an additivity of masking analysis (symbols), along with data from Fig. 2 replotted as lines.

compression decreased with gain, but not with Plack *et al.* (2004), who found no correlation between compression and gain. In the present study, S3, who had the most data points on the compressive part of the GOM function, shows the least change in compression with a precursor. Thus, the apparent change in compression may be due to the limited number of data points to be fitted for the other listeners.

Data from both simultaneous and forward masking are consistent with the hypothesis that preceding stimulation decreases the gain of the basilar membrane input-output function. As noted in the Introduction, this would be consistent with the activation of the MOCR. The MOCR hypothesis fits in the context of adaptation of sensory systems in response to the changing environment. Forward masking would then just be a by-product of a system which is generally beneficial for listening in background noise in simultaneous masking.

C. Comparison to other aspects of temporal effects in simultaneous masking

In this study, the effect of the on-frequency precursor was characterized by a decrease in masker level, which may seem contrary to some temporal effects in simultaneous masking, where the masker level increases. The present data are consistent with some conditions in Strickland (2001), where the temporal effect was measured in simultaneous masking with a notched-noise masker. Adding a broadband noise precursor before the masker decreased masker levels for wide notch widths. Carlyon (1989) showed that a narrowband noise precursor at the signal frequency increased signal threshold. This is similar to the effect of the on-frequency precursor in our experiment, which decreases the masker level required to mask the signal. Both the previous simultaneous masking and the current forward masking results are consistent with a decrease in gain at the signal frequency following a precursor.

The effectiveness of the short-duration precursor was surprising. In a simultaneous masking study, Bacon and Smith (1991) showed that a short-duration precursor was no different from a longer precursor in its effectiveness, as long as it was long enough to activate the efferent system. In this study, it was expected that the 40-ms on-frequency precursor would be too short to activate the efferent system. However, in addition to precursor duration, it appears that the delay between precursor and signal onsets may play a role. For a 160-ms precursor, this delay is 200 ms, whereas for the shorter duration precursor, it is 80 ms. The time course of onset of the MOCR is on average about 70 ms (Backus and Guinan, 2006; Guinan, 2006). It is, therefore, possible that the 40-ms on-frequency precursor could have activated the MOCR.

Another interesting aspect of this study is the fact that a temporal effect is seen when the preceding stimulation is only at the signal frequency. Past studies have typically shown that in order to produce a temporal effect, the precursor and/or simultaneous masker must contain energy above the signal frequency (Zwicker, 1965b; McFadden, 1989; Bacon and Smith, 1991). One reason that this might be true is that it would eliminate the possibility of off-frequency listening. If listeners are able to attend to filters above the signal frequency, where the growth of excitation to the signal is more linear, a temporal effect might not be seen. Another reason could be that the temporal effect depends on suppression. The results of the present study are consistent with the off-frequency listening hypothesis. High-pass noise presented with the signal eliminated the possibility of off-frequency listening. This is consistent with temporal effects seen with tonal stimuli at high frequencies, where off-frequency listening would also not be effective due to the sharp increase in thresholds above the signal frequency (Schmidt and Zwicker, 1991; Carlyon and White, 1992). Thus, this shows that a temporal effect can be seen in a condition in which suppression is not playing a role. This is not to suggest that suppression never plays a role in the temporal effect, only that it is not a necessary condition.

ACKNOWLEDGMENTS

We thank Skyler Jennings for his valuable comments on the manuscript. This research was supported in part by funds from the Speech, Language, and Hearing Sciences Department at Purdue University and Grant No. R01-DC008327

from the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH).

Backus, B. C., and Guinan, Jr., J. J. (**2006**). "Time-course of the human medial olivocochlear reflex," J. Acoust. Soc. Am. **119**, 2889–2904.

Bacon, S. P., and Healy, E. W. (**2000**). "Effects of ipsilateral and contralateral precursors on the temporal effect in simultaneous masking with pure tones," J. Acoust. Soc. Am. **107**, 1589–1597.

Bacon, S. P., and Smith, M. A. (**1991**). "Spectral, intensive, and temporal factors influencing overshoot," Q. J. Exp. Psychol. A **43**, 373–399.

Carlyon, R. P. (**1989**). "Changes in the masked thresholds of brief tones produced by prior bursts of noise," Hear. Res. **41**, 223–235.

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.

Guinan, Jr., J. J. (**2006**). "Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans," Ear Hear. **27**, 589–607.

Harris, D. M., and Dallos, P. (**1979**). "Forward masking of auditory nerve fiber responses," J. Neurophysiol. **42**, 1083–1107.

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.

Liberman, M. C. (**1989**). "Rapid assessment of sound-evoked olivocochlear feedback: Suppression of compound action potentials by contralateral sound," Hear. Res. **38**, 47–56.

McFadden, D. (**1989**). "Spectral differences in the ability of temporal gaps to reset the mechanisms underlying overshoot," J. Acoust. Soc. Am. **85**, 254–261.

Meddis, R., and O'Mard, L. P. (**2005**). "A computer model of the auditory-nerve response to forward-masking stimuli," J. Acoust. Soc. Am. **117**, 3787–3798.

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.

Oxenham, A. J. (**2001**). "Forward Masking: Adaptation or integration?" J. Acoust. Soc. Am. **109**, 732–741.

Oxenham, A. J., and Moore, B. C. J. (**1994**). "Modeling the additivity of nonsimultaneous masking," Hear. Res. **80**, 105–118.

Oxenham, A. J., and Plack, C. J. (**1997**). "A behavioral measure of basilar-membrane non-linearity in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **101**, 3666–3675.

Penner, M. J., and Shiffrin, R. M. (**1980**). "Nonlinearities in the coding of intensity within the context of a temporal summation model," J. Acoust. Soc. Am. **67**, 617–627.

Plack, C. J., Drga, V., and Lopez-Poveda, E. A. (**2004**). "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss," J. Acoust. Soc. Am. **115**, 1684–1695.

Plack, C. J., and Skeels, V. (**2007**). "Temporal integration and compression near absolute threshold in normal and impaired ears," J. Acoust. Soc. Am. **122**, 2236–2244.

Relkin, E. M., and Turner, C. W. (**1988**). "A reexamination of forward masking in the auditory nerve," J. Acoust. Soc. Am. **84**, 584–591.

Rosengard, P. S., Oxenham, A. J., and Braida, L. D. (**2005**). "Comparing different estimates of cochlear compression in listeners with normal and impaired hearing," J. Acoust. Soc. Am. **117**, 3028–3041.

Schmidt, S., and Zwicker, E. (**1991**). "The effect of masker spectral asymmetry on overshoot in simultaneous masking." J. Acoust. Soc. Am. **89**, 1324–1330.

Shore, S. E. (**1998**). "Influence of centrifugal pathways on forward masking of ventral cochlear nucleus neurons," J. Acoust. Soc. Am. **104**, 378–389.

Smith, R. L. (**1977**). "Short-term adaptation in single auditory nerve fibers: Some poststimulatory effects," J. Neurophysiol. **40**, 1098–1112.

Smith, R. L. (**1979**). "Adaptation, saturation, and physiological masking in single auditory-nerve fibers," J. Acoust. Soc. Am. **65**, 166–178.

Strickland, E. A. (**2001**). "The relationship between frequency selectivity and overshoot," J. Acoust. Soc. Am. **109**, 2063–2073.

Strickland, E. A. (**2004**). "The temporal effect with notched-noise maskers: Analysis interms of input-output functions," J. Acoust. Soc. Am. **115**, 2234–2245.

Strickland, E. A. (**2008**). "The relationship between precursor level and the temporal effect," J. Acoust. Soc. Am. **123**, 946–954.

Strickland, E. A., and Krishnan, L. A. (**2005**). "The temporal effect in listeners with mild to moderate cochlear hearing impairment," J. Acoust. Soc. Am. **118**, 3211–3217.

Warr, W. B. (**1980**). "Efferent components of the auditory system," Ann. Otol. Rhinol. Laryngol. **90**, 114–190.

Warr, W. B., and Guinan, Jr., J. J. (**1979**). "Efferent innervation of the organ of Corti: Two separate systems," Brain Res. **173**, 152–155.

Zwicker, E. (**1965a**). "Temporal effects in simultaneous masking by white-noise bursts," J. Acoust. Soc. Am. **37**, 653–666.

Zwicker, E. (**1965b**). "Temporal effects in simultaneous masking and loudness," J. Acoust. Soc. Am. **38**, 132–141.

Zwislocki, J., Pirodda, E., and Rubin, H. (**1959**). "On some poststimulatory effects at the threshold of audibility," J. Acoust. Soc. Am. **31**, 9–14.

Unsupervised bird song syllable classification using evolving neural networks

Louis Ranjard[a] and Howard A. Ross[b]
Bioinformatics Institute, School of Biological Sciences, University of Auckland, Auckland 1142, New Zealand

(Received 4 December 2007; revised 5 May 2008; accepted 6 March 2008)

Evolution of bird vocalizations is subjected to selection pressure related to their functions. Passerine bird songs are also under a neutral model of evolution because of the learning process supporting their transmission; thus they contain signals of individual, population, and species relationships. In order to retrieve this information, large amounts of data need to be processed. From vocalization recordings, songs are extracted and encoded as sequences of syllables before being compared. Encoding songs in such a way can be done either by ear and spectrogram visual analysis or by specific algorithms permitting reproducible studies. Here, a specific automatic method is presented to compute a syllable distance measure allowing an unsupervised classification of song syllables. Results obtained from the encoding of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*) songs are compared to human-based analysis. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2903861]

PACS number(s): 43.60.Np, 43.60.Lq, 43.80.Ka [JAS] Pages: 4358–4368

I. INTRODUCTION

Song birds (Passeriformes) have learned, rather than innate, songs.[1] The learning occurs when the bird is immature and in some cases, learning continues throughout a bird's life. This learning process is divided into two steps; first, a bird hears and memorizes a song from another bird, and second, it tries to imitate the song as accurately as possible. This imitation is not perfect, and so songs evolve through generations as small changes occur. From one generation to the next, a particular component of a song can be kept or disappear because no bird would have copied it. The more frequently this component occurs, the greater chance it has to be copied. Therefore, it is possible to define a neutral model of evolution of songs.[2] If two populations with a common origin are isolated, one can expect that the songs of each will accumulate modifications independently and then, with a sufficient number of generations, become significantly different. Being able to detect those changes can help us to infer population histories and relationships.

Here, bioacoustic methods developed for encoding bird songs as sequences of discrete syllables are presented. First, syllables are extracted from segmented song recordings and their spectrograms are encoded. They are then classified using a new kind of artificial neural network which utilizes dynamic time warping for the learning stage. One way to perform sequence comparison is to use alignment algorithms that minimize a distance function between a pair of sequences.[3] This approach is popular in the field of molecular biology, in speech processing where it is commonly named dynamic time warping and also in bioacoustics.[4,5] This technique has been used because it allows us to perform unsupervised classification. Here, a pairwise syllable distance measure, calculated on the basis of mel-cepstrum coefficients dynamic time warping, will be introduced. This measure is proportional to the number of operations required to transform one sequence into another, while producing an alignment. Then, this distance is used for classifying syllables into different clusters. Self-organizing maps are a kind of neural network designed for data representation and classification.[6] During a learning stage, the network is trained with the data samples. However, this process can be very time consuming and the size of the map has to be arbitrarily chosen; therefore, modifications to the shape of the self-organizing map neural network have been developed. Evolving tree neural networks[7] are not only faster to train but also allow unsupervised clustering as they grow through a learning process, reaching a size proportional to the number of clusters required. The syllable distance introduced above is incorporated in such networks and weighted average sequences are derived from the syllable sequence alignments, supporting the learning process. Syllable classification obtained from the encoding of a small data set of songs belonging to a subspecies of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*) shows how the classification from this approach corresponds to those defined by classic approaches. This analysis suggests that such a method can be useful in analyzing larger data sets and so permit large scale studies of bird vocalizations.

II. METHODS

Classic approaches in bioacoustics use basic signal features and/or human knowledge for both detection and segregation of bioacoustic signals from noise.[8] Although former technical limitations are no longer restrictive,[8] only a few song analysis tools have been inspired by automatic speech processing[9,10] and are fully automatic.[11] When humans apply expert knowledge to song analysis, it can introduce subjec-

[a]Electronic mail: l.ranjard@auckland.ac.nz
[b]Electronic mail: h.ross@auckland.ac.nz

 0001-4966/2008/123(6)/4358/11/$23.00

tivity which can impair the reproducibility of bird song studies. Furthermore, this approach becomes very time consuming and thus is not suitable for large data sets. However, modern speech processing techniques offer high accuracy in sound analysis for speech recognition or speaker verification.[12,13] Different measures of word distance have been developed[14] and they can be extended to provide syllable distance measures. On the one hand, animal vocalizations seem to be less complex than human speech but, on the other hand, available recordings are often collected in poor acoustic conditions. Therefore, the signal-to-noise ratio can be low. Moreover, the meanings of the vocalizations remain unknown so the significant features of the song may remain undetected.

A. Song segmentation and syllable encoding

No common definition of a song note or a syllable has been set in the literature. Here, it has been considered that a syllable is a part of a song characterized by a high value of autocorrelation of the signal and with a continuity in the fundamental frequency. Syllable boundaries are found by segregating the syllables from the background noise using the autocorrelation function of the signal throughout a song recording. In a second step, the boundaries at the beginning and end of the syllables are readjusted. The spectral roll-off, defined as the frequency below which a specific percentage of the energy occurs, is analyzed to complete this step. A minimum threshold in the length of a syllable is set at 50 ms. Dooling[15] showed that birds can distinguish shorter sounds but a minimum number of samples is required to perform accurate feature extraction.

a. Autocorrelation. The cross correlation is a standard method for estimating how two signals are linearly related. In the field of music analysis, it is useful for instrument recognition.[16] For each 50% overlapping window of 128 samples under 44.1 kHz sampling rate, the maximum $c_{\max}$ of the cross correlation of the signal $s_l(n)$ of the lth analysis window is calculated as

$$c_{\max}(l) = \max \sum_{n=0}^{N-m-1} s_l(n+m)s_l^*(n), \quad \forall\, 1 \le m \le N, \quad (1)$$

where s_l^* is the complex conjugate of s_l and N is the number of samples in s_l. This function is then smoothed by calculating a moving average, thus eliminating the slight local variations, in analysis windows of 512 samples with 50% overlap. Then, the potential syllables are characterized as the sections of this function greater than a specific threshold. This threshold is set as the median value of the function.

b. Frequency roll-off 60%. The spectral roll-off point ro is defined as the frequency value at which 60% of the signal energy is contained below in the spectrum.[17] Windows of 512 samples with 50% overlap are used to compute this value through the signal. The spectral roll-off point is higher during a syllable emission than during noise because bird songs are generally high-pitched and field recordings contain low frequency background noise. Slabbekoorn *et al.*[18] showed that birds can adjust the frequency of their vocalizations depending on the environmental noise. Therefore, it can be helpful for segregating the signal from the noise.

$$\mathrm{ro}(l) = f_{\mathrm{ro}}, \quad (2)$$

where l defines a window and f_{ro} is computed from the spectrum X of this window as

$$\sum_{k=1}^{f_{\mathrm{ro}}} X(k) = 0.60 \sum_{k=1}^{f_{\mathrm{Nyquist}}} X(k), \quad (3)$$

where f_{Nyquist} is half the sampling frequency. Each boundary is replaced by the local minima of the roll-off function, in a window starting 50 ms before and ending 50 ms after the time point found after analyzing the autocorrelation function.

In order to represent the syllables, it is possible to use frequency band filtering and then calculate the mel-cepstrum coefficients as well as the first-order and second-order delta features. This feature extraction approach has a great success in human speaker recognition applications.[19] The mel-frequency bands have been defined for human ears and therefore there is little support for their use with bird vocalizations. It has been decided to switch the frequency band filters in order to get more sensitivity in high frequencies where most bird vocalization occurs. However, specific bandwidth filter banks should be defined as species specific. More precisely, the first of the 26 filterbank channels starts at 1000 Hz and the last one terminates at 22.05 kHz. Under a sampling frequency of 44.1 kHz, a Hamming window of 128 samples with 50% overlap has been used for computing the spectra and the signal had first-order pre-emphasis applied using a coefficient of 0.97. Twenty coefficients were calculated and the $C0$th cepstral parameter was used as the energy component. This number of coefficients has been chosen based on empirical results (not shown). Two frames before and two frames after the current one were used to estimate the first- and second-order temporal derivatives. Energy normalization was implemented by subtracting the maximum value of the energy and adding 1.0. The cepstral coefficients were rescaled by liftering the cepstra using a coefficient of 22 so that they have similar magnitudes. This window size is much smaller than that used in previous work, for example, Trawicki,[20] and therefore provides sufficient precision to analyze the changes in the frequency content of each syllable. Therefore, each syllable is represented by a sequence of vectors, where each vector is composed of 63 coefficients, the mel-cepstrum coefficients, and the delta features, of consecutive overlapping windows.

B. Syllable distance measure D_s

Different approaches exist to measure a distance between two sounds; some use feature extraction, others spectral cross correlation.[21] There are also different approaches[3,14,22] for performing sequence comparisons and dynamic time warping. An interesting advantage of dynamic time warping is that it allows the computation of an average from the alignment of a pair of variable-length sequences. Moreover, slight changes in the rate of sound emissions are tolerated while computing the alignment. These techniques consist of finding the optimal alignment between the two sequences of vectors using dynamic programming. This in-

volves computing an edit distance that is proportional to the minimum of the sum of operation costs required to transform one sequence into the other one. The values of the operation costs are proportional to a distance measure between mel-cepstrum coefficient vectors. Many different distances have been implemented for speech processing.[12] Here, the Euclidean distance has been used for each pair of vectors in the sequences to be aligned. Let $X=x_1 \ldots x_N$ and $Y=y_1 \ldots y_M$ be two vector sequences to be compared. Five edit operations are considered:

(a) Substitution $S(v,w)$ defines the cost associated with the substitution of the vector v for the vector w. This cost is the vector distance described above.
(b) Insertion $I(v)$ defines the cost associated with the insertion of the vector v. This cost is set as half the average of the substitution cost.
(c) Deletion $D(v)$ defines the cost associated with the deletion of the vector v. This cost is set as half the average of the substitution cost.
(d) Compression $C(vw,x)$ defines the cost associated with the compression of the vectors vw into the vector x. It is defined as the mean of the substitution cost of the vector v for x and the substitution cost of the vector w for x.
(e) Expansion $E(v,wx)$ defines the cost associated with the expansion of the vector v into the vectors wx. It is defined as the mean of the substitution cost of the vectors w for v and the substitution cost of the vector x for v.

Then, a graph of size $G(a,b)$ with $1 \leqq a \leqq N$ and $1 \leqq b \leqq M$ is computed as

$$G(a,b) = \min \begin{cases} G(a-1,b) + I(x_a) \\ G(a,b-1) + D(y_b) \\ G(a-1,b-1) + S(x_a,y_b) \\ G(a-1,b-2) + C(x_a,y_{b-1}y_b) \\ G(a-2,b-1) + E(x_{a-1}x_a,y_b). \end{cases} \quad (4)$$

This graph is used to calculate the distance between a pair of syllables $D_s=G(N,M)/N+M$. Then, tracing the warping path back allows us to find the weighted average sequence of vectors as defined in Refs. 3 and 23 performing a time interpolation. In each step in the warping path, a weighted vector is computed

$$c_k = qx_a + (1-q)y_b \quad (5)$$

and

$$c_k = 0.5qx_a + 0.5qx_{a-1} + (1-q)y_b, \quad (6)$$

in the case of a compression or expansion. For deletions and insertions, the vector used in the average sequence is simply the one conserved in the alignment. The corresponding time point is the average between those of both sequences

$$t_k = |qt_a + (1-q)t_b|. \quad (7)$$

In a last step, all vectors belonging to the same time point are averaged in order to obtain a continuous time sequence.

C. Self-organizing neural network

Evolving trees are a variant of self-organizing maps[6,7] which allow the treatment of large amounts of data. An advantage of this approach is that it produces a lower dimensional representation of the data set hierarchically ordered, and this hierarchy can be used for the identification of clusters in the data set. Therefore, it is a suitable approach for unsupervised classification problems. Moreover, the learning time of the network is considerably smaller than for a self-organizing map as it is not necessary to compare each input vector with each network's weight matrix. Furthermore, the network is able to grow in order to reach a size suitable for the classification of a specific data set.

The syllable distance measure D_s defined above is used in order to hierarchically find the best matching unit in the evolving tree. Starting from the root of the network, a data sample is aligned with every child neuron weight matrix, choosing the closest one to go to the next level until a leaf neuron is reached. In this case, this neuron is defined as the best matching unit. The learning process is carried out in two stages. First, the data set is explored as the network grows quickly, and second the network is fine-tuned by pulling the neurons closer to the data samples which they aim to classify. Therefore, it is important to slow down the growing and learning rate of the network as it is trained on the sample data set. During an epoch, every data sample is used just once. For each syllable in the data set, the closest neuron, or best matching unit, is found in the network. Then, its weight matrix is updated by aligning the vector sequences and computing a weighted average sequence. Each neuron has a hit counter which is incremented every time it is chosen as a best matching unit. If this counter goes beyond the splitting threshold, the neuron is subdivided and child neurons are created. The number of new neurons created depends on the number of leaves specified. A link measure distance has been used to compute the distance between neurons in the network. This distance is computed as the number of edges separating two neurons in the network, using an implementation of the depth-first search algorithm. The neighborhood function of the neural network is defined as

$$h(c(t),i) = \alpha(t)\exp\left(\frac{-d(c(t),i)^2}{2\sigma(t)^2}\right), \quad (8)$$

where $d(c(t),i)$ is the distance between the neuron weight matrix $c(t)$ and the sample i, $\alpha(t)$ is the learning rate defined as

$$\alpha(t) = \max \begin{cases} \alpha(0)\exp\left(\frac{-t^2}{(0.75T)^2}\right) \\ \alpha_{\min}, \end{cases} \quad (9)$$

where T is the total number of epochs. $\sigma(t)$ is the neighborhood size at epoch t, defined as

$$\sigma(t) = \sigma(0)\exp\left(\frac{-t}{0.5T}\right). \quad (10)$$

The neighborhood function $h(c(t),i)$ determines the weight q used in the calculation of weighted average sequences. The

TABLE I. Values for the parameter of the network training. The left column shows the values used by default and the right column shows the selected ones, after estimation.

	Default values	Selected values
Number of epoch	5	5
Splitting threshold	50	20% of data set size
Neighborhood strength	3	1
Initial leaf number	2	3
Final leaf number	2	2
Gamma	0.95	0.90
Initial learning rate	0.90	0.90
Final learning rate	0.01	0.01

number of children can also be set to decrease linearly after each epoch, through the learning process

$$n(t+1) = \max\begin{cases} n(t) - \dfrac{n(t)}{T} \\ n_{\min}. \end{cases} \tag{11}$$

Moreover, a factor γ affects the counter of each neuron in the network, slowing down the expansion of the network[24]

$$\text{count}(c(t+1)) = \gamma \text{count}(c(t)). \tag{12}$$

D. Parameter selection

Two different features of the final neural network are used to assess the quality of the training process. First, the mapping precision is defined as the average distance between a data sample and the best matching unit in the tree. The second one is the tree size, which means the number of neurons of the neural network at the end of the learning process. The goal of unsupervised clustering is to find the best dimensionality reduction of a given data set which is directly linked to this value. In their experiments, Pakkanen *et al.*[24] used a splitting threshold of 60, implying the creation of four new neurons, to analyze a data set of 1000 vectors during ten epochs. In another experiment, they used a splitting threshold of 50, the creation of three new neurons after each split and Ref. 25 used different values of those parameters. In this study, several values of these parameters were tested in order to understand better how they can affect the final characteristics of the network. For those assessments, a data set of 200 White-crowned Sparrow syllables was employed. The default parameters values used are shown in Table I. To assess the quality of the neural network, the size of the tree and the mapping precision were evaluated for different values of the parameters.

The longer the network is trained the greater the number of neurons which will reside in the final network. The mapping precision will be better as each neuron would have been averaged with more data samples, summarizing fewer samples. As expected, the size of the final network linearly increased with the number of epochs, Fig. 1(a). However, the mapping precision did not behave in a similar way and decreased in two different stages. Mapping precision rapidly decreased at first, but in a second stage, it decreased more slowly. Therefore, a small number of epochs is sufficient to train a network. The best value for this parameter, given a specific data set, is not proportional to the number of data samples. It is actually related to the number of clusters in the data set, and more precisely the number of examples of each given cluster. Indeed, a network trained on a data set comprising only few clusters but with a great number of samples in each of them will reach a high mapping precision quicker than if it is trained with a more heterogeneous data set. The number of expected clusters is not known; therefore, it is not obvious how to choose a correct value for the number of epochs. In most of the experiments, a value of 5 was used.

The value of the splitting threshold will determine the rate at which the network grows. With increasing values for the splitting threshold, the size of the tree underwent an exponential decay, Fig. 1(b). At the same time, the mapping precision increased exponentially, corresponding to a higher average distance between a data sample and its best matching unit. After a stage of rapid decrease in the tree size and increase in the mapping precision lasting until a threshold of around 50, those values changed more slowly. The point of change occurs when the splitting threshold is approximately 20% of the data sample size.

The neighborhood strength defines the distance from the best matching unit, expressed as number of neurons, until the point at which the update will have an effect in the network. A high value means that an important region of the tree will be updated at each learning step. A value of 0 indicates that the classification is similar to *kmeans* clustering as pointed out by Ref. 24. The tree size was not sensitive to this parameter (data not shown). Considering the mapping precision, small values, a distance between 1 and 3, returned the best results. A value of 1 means that only the mother neuron in the tree network will be updated. Nevertheless, this could be different with larger networks. This parameter could also be expressed as a function of the size of the network. However, a value of 1 has been chosen in most of the experiments.

The number of leaves created at each learning step is expected to have a direct effect on the size of the final tree. It is less obvious what effect it will have on the mapping precision. Assessments, performed using a constant leaf number throughout the learning process, showed that the best results were obtained with three new leaves created at each growing step, Fig. 1(c). With higher values, assessments did not show any improvement in the mapping precision, even if the learning process required more calculation. In fact, to find the best matching unit for a given data sample, its distance to every neuron weight matrix is computed at each level in the network. Therefore, it is effective to use small values for this parameter.

Another set of assessments was performed using a decreasing number of leaves. Different initial values of the number of leaves were tested with a constant final number of leaves (data not shown). With a greater initial number of leaves, it is expected that the network will first organize itself to explore the data set while growing quickly, but at the end of the learning process each neuron will only need to be pulled closer to the data samples with little requirement for growth. This resulted in smaller networks but with similar mapping precision and thus potentially a smaller number of

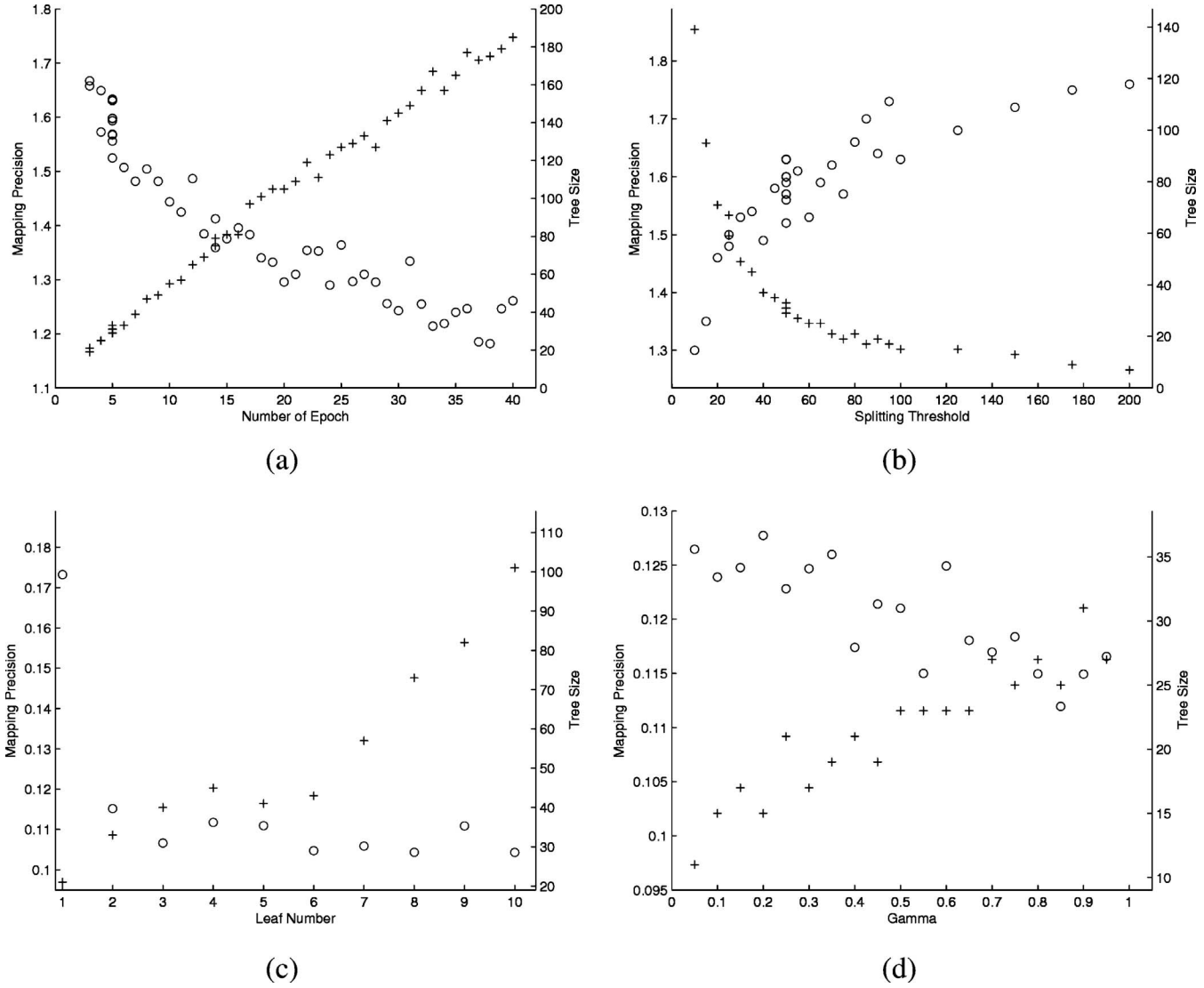


FIG. 1. Neural network size (+) and precision (○) are affected by varying values of the learning parameters: the number of epochs (a), the splitting threshold (b), the number of leaves (c), and gamma (d). The mapping precision is the average distance between the data samples and the best matching units in the network; therefore, lower values correspond to higher precision.

cluster centers. All assessments returned a better precision than the one obtained with a constant leaf number of 2, except for an initial value of 4. Surprisingly, with values greater than 8, the precision worsened while the final network was larger. More assessments were performed with decreasing number of leaves throughout the learning process. Different values for the initial and the final number of leaves were tested (data not shown). It appeared that initial numbers between 7 and 10 return the best mapping precision with an increasing network size. A small final number of leaves at the final stage of the learning process is shown to give the best results. Overall, the best results were obtained with a value decreasing from 7 to 3, 7 to 1, 8 to 3, and 8 to 1. A ratio around 2.5 between the initial and final numbers produced the best precision. More computing will be required for training the network, with higher initial number of leaves, because of the way the search for best matching units is performed. Therefore, these values have to be minimized and an initial value of 3 was selected.

Another way to slow the growth of the tree is to multiply the hit counter of each neuron by a factor gamma after the end of every epoch. In assessments using a constant number of leaves (2) at each split, larger values of the gamma factor reduced the mapping precision and increased the network size, Fig. 1(d). However, a better mapping precision can be obtained by varying the number of leaves, but with final trees of greater size.

The learning rate will define how strongly the neurons are pulled toward the data sample at every network update. This rate also decreased through the learning process in order to allow the network to first explore the data set and then assess it accurately. Although different initial values of the learning rate did not affect the final result significantly, better precision is returned with small final learning rates (data not shown). As a result of these assessments, the set of selected parameter values is presented in Table I and were used for other analysis.

It could be asked how well these results can be extrapolated to data sets of greater size. In particular, the first stages of learning, while the network is small, require the algorithm

TABLE II. Dialects are defined in Ref. 26 on the basis of the different types of simple syllables. The corresponding cluster numbers obtained with neural network classification are given for simple and complex syllables. Clusters in bold uniquely define a dialect. Cluster 27 contains one occurrence of a syllable belonging to the dialect B and the four complex syllables of the dialect E.

	Complex syllable		Simple syllable	
Dialect	Syllable type	Cluster	Syllable type	Cluster
A	22	11, 16, **30**, 31	1	**10**, 21, **23**, 24, **25**
B	12, 13	11, **12**, **27**, 31	2a	**4**, **19**, 22, 24, **26**
C	21	**6**, **8**, **20**, 31	7	**13**, 21, 22
D	n/a	**1**	8, 9	**3**, **32**
E	24	**27**	Absent	Absent

to be able to conquer the data set and thus to grow at a rate proportional to its size. Consequently, the parameters affecting this section, the initial leaf number and the neighborhood strength, should be expressed as a function of the data set size.

The final step of the syllable classification is performed by matching the data samples extracted from recordings to the neurons of the network. Each neuron defines a cluster center. An extra cluster merging step can be added to the classification. This step involves computing the pairwise distance between neurons and then merging the corresponding clusters if this distance falls below a specific threshold. It is also possible to apply a limit on the distance between a data sample and its closest neuron. Neurons of different depths define different levels of classification, basically obtaining clusters of different sizes and precision.

E. Validation with White-crowned Sparrow songs

To test whether an automated approach could reproduce a human classification, this method was applied to a set of songs from a species with well-documented dialects. A set of 17 songs, recorded between 1998 and 2000, was chosen to represent the diversity of song types sung by a subspecies of White-crowned Sparrow (*Zonotrichia leucophrys pugetensis*), in the west of the United States of America. Previous studies[26,27] have characterized dialects in this region by manually grouping the syllables of the same type. Human judges chose features to define each dialect and classified songs in accordance with these features. In Refs. 26 and 27, the visual classification is performed in two steps. First, syllables are classified into four main types: whistle, buzz, complex syllables, and simple syllables, depending on their spectrogram main shape and the position at which they occur in a song. Second, syllables belonging to the complex syllable and simple syllable types are classified into subtypes. It is this second classification which yields the definition of dialects. These 17 songs were visually classified into five dialects using the catalog of syllables of Nelson,[26] Table II. The complex syllable of dialect E was not found but it is visually different enough to constitute a new complex syllable type. To classify the syllables using the method described here, the rules introduced above for parameter selection were used. All syllables of this data set were classified by training a network, with the parameter values given above. After a neuron merging step, the network obtained was composed of 71 neurons with a mapping precision of 0.086. All data samples were classified and a threshold of 0.05 was used on the distance between weight matrices for cluster merging.

III. RESULTS

Using the method described here, the set of White-crowned Sparrow songs was found to contain 170 syllables in 32 clusters. The segmentation of the songs was performed by analyzing the autocorrelation and the roll-off of the songs as explained before. However, the threshold of the autocorrelation function has been manually modified in some cases, for the purpose of obtaining a consistent segmentation of the songs. Figure 2 shows an example of the spectrogram derived from one of those songs as well as the limits between syllables after song segmentation. When human experts classified the data set by visual inspection, 128 syllables were identified. The reason for this difference is that whistles and complex syllables have sometimes been subdivided by the segmentation algorithm, as can be noticed in Fig. 2. The resulting segments contain a continuous trace in the spectrogram separated from each other by short gaps of low frequency roll-off. By visual inspection, 28 of the 32 clusters are consistent with the four main syllable types (buzz, whistle, complex, and simple syllables). Figure 3 shows the spectrograms of the different data samples being part of clusters 4, 5, and 8. These three clusters illustrate characteristic cases. The presence of strong background noise, Fig. 3(a), has an influence on the clustering. Syllables are clustered in accordance with their dominant frequency, Fig. 3(b), and the variation in the shape of their spectrograms, Fig. 3(c).

This study differs from earlier studies in that all syllables are computationally classified at the same time rather than being classified in two steps. The neural network approach identified a larger number of clusters than had been

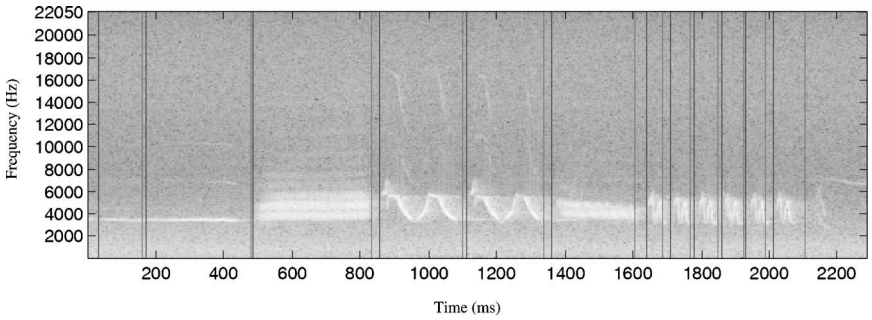


FIG. 2. Spectrogram of a song (Oak Harbor, WA, 1998, dialect D), showing the boundaries between syllables. The vertical axis is frequency in Hz and the horizontal axis is time in ms. This song consecutively contains two whistle syllables, a buzz syllable, two complex syllables, another buzz syllable, and finally six simple syllables which constitute an end trill.

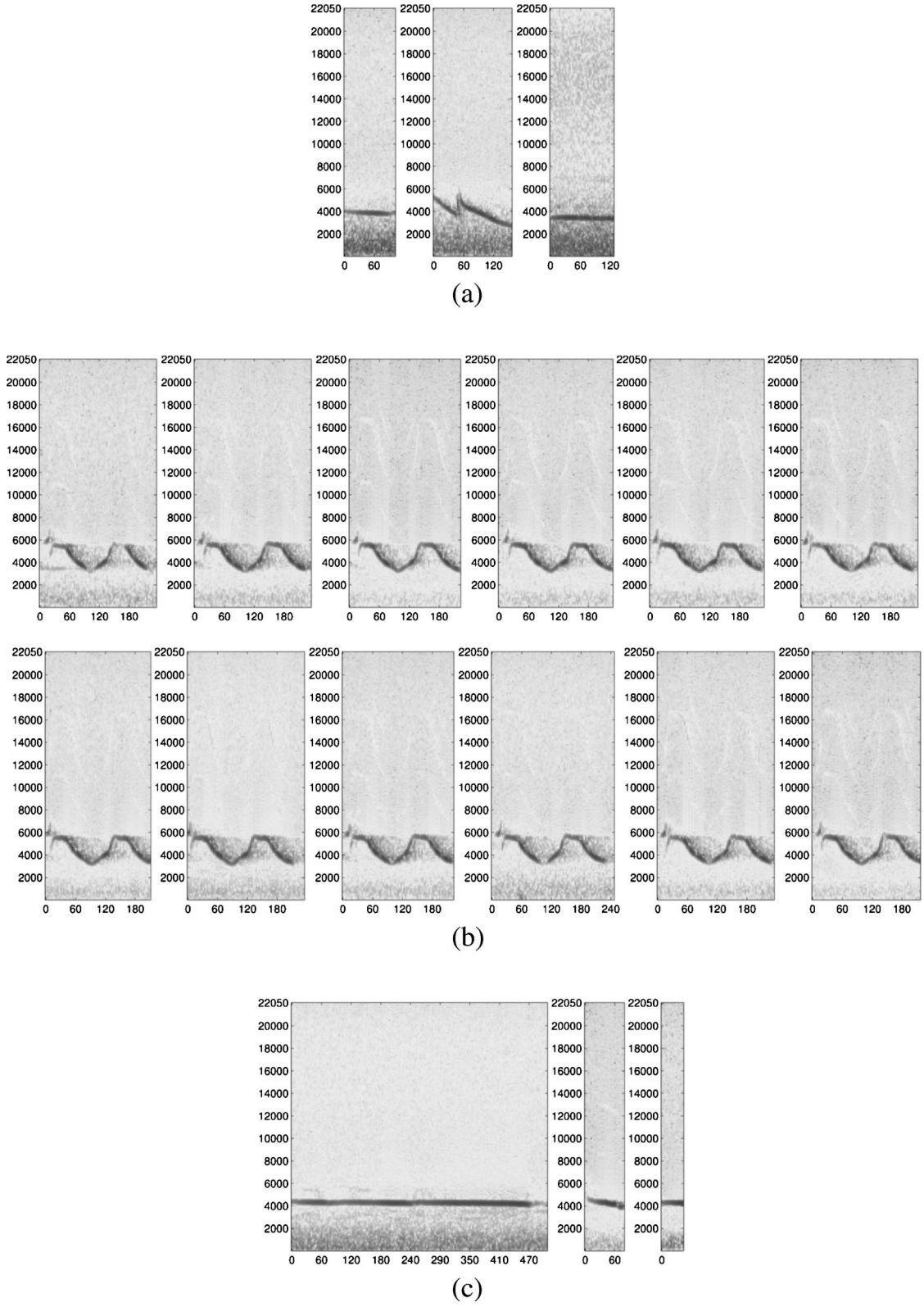


FIG. 3. Syllable spectrograms of the members of three different clusters. The first cluster (a), number 4, is affected by the presence of strong noise. Cluster 5 (b) appears to be consistent and Cluster 8 (c) comprises syllables sharing similar dominant frequencies. Axes are the same as in Fig. 2.

previously recognized so that a given syllable type corresponded to several clusters. For example, the simple syllable type 9 corresponds to the clusters 3 and 32, Table II. Moreover, several clusters contain only one representative. For two cases, the presence of strong noise has influenced the classification. This was particularly true when syllables had a short duration, Fig. 3. When syllables are very short, there is less information after sound encoding for performing the comparisons and the syllables are classified according to their dominant frequency rather than their spectral shape. Whistle syllables were classified according to their dominant frequency and buzz syllables according to their spectral

TABLE III. Occurrence of syllable types per localities, i.e., the number of times each syllable type is found in each different location. For each syllable cluster, the first level of visual classification is given, B for buzz, W for whistle, SS for simple syllable, and CS for complex syllable. The neural network cluster number is shown as well. The locations are ordered from north to south. Oak Harbor being the furthest north.

Neural net	18	16	30	10	25	23	20	13	7	12	19	26	2	3	9	5	32
Visual	W	CS	CS	SS	SS	SS	CS	SS	CS	CS	SS	SS	B	SS	W	CS	SS
Location[a]																	
OH													6	27	6	12	8
PC												1					
N										1	5						
Y									1								
F							1	3									
B		2	2	2	6	1											
PO	1																
Neural net	1	17	6	29	8	21	23	31	14	11	15	27	24	28	4		
Visual	W	W	B-CS	W	W-CS	SS	SS	CS	W	CS-W	B	CS	SS	B	W-SS		
OH														6	1		
PC								1	1	1	1	1	2		2		
N					1		1			2	1						
Y			3	1	1	9		3	1								
F	1	2	3	1			4	4									
B	2	1	3	1	1	2		4	2	1			2				
PO				2					1	1	1	4		3			

See Ref. 26 for location. OH, Oak Harbor; PC, Pacific City; N, Newport; Y, Yachats; F, Florence; B, Bandon; PO, Port Orford.

shape. Complex syllables, when they had been extracted in a similar way as in Refs. 26 and 27, were classified according to their geographical origin, for example, cluster 5, Fig. 3(b). However, when the segmentation process divided those syllables in short different parts, the alignment algorithm clustered the syllables with similar dominant frequency.

A unique set of clusters was identified for each of the visually specified syllable types used to define a dialect (Table II). The only exception was cluster 27 where a complex syllable, unique in the data set, was clustered with the four complex syllables of the dialect E. Similarity among dialects A, B, and C is shown by the shared possession of clusters from both the complex and simple syllables. Dialects D and E, on the other hand, had only unique syllable clusters.

Table III shows that each location possessed particular syllable types, which can be a consequence of the presence of dialects. However, some syllable types were present at different locations, showing regional patterns, for example, clusters 1, 17, 6, 29, and 8, while others did not, Table III. To test for regional patterns, the average geographic pairwise distance between the locations of every occurrence were calculated for each cluster. Then, a test statistic was defined as the average pairwise distance over all clusters. In the presence of dialect, this distance is expected to be smaller than if syllable clusters were randomly distributed among locations. Indeed, the number of shared syllables between the songs of any two locations should decrease with the geographical distance separating them. Over all syllables, the observed test statistic was less than that obtained for a random distribution ($N=100\,000$ replicates, data not shown). When each main group of syllables is examined individually, the strongest differences were for simple and complex syllables. A slight difference was apparent for whistle syllables, but in the case of buzz syllables, no significant difference between the statistic calculated with observed occurrences and that derived from random occurrences of syllables was noticeable.

Moreover, the alignment distances between the simple syllables reflect this geographic pattern, Fig. 4. The pairwise distance matrix between the most frequent syllables produced at the end of the songs for each location was used to build a hierarchical nearest neighbor tree. These syllables are repeated during the end trill of the songs, except for the songs of Port Orford which lack this part. This tree groups the locations sharing the same dialects, both Pacific City and Newport, and Yachats and Florence, together. However, it also suggests that some parts of the syllables are shared between dialects. The beginning of the syllables produced at Yachats and Florence looks similar to the terminal syllables of Bandon, and the end part resembles the syllables of Pacific City and Newport. Therefore, some syllables could be hybrids, created by the union or concatenation of syllables. Another explanation could be that these syllables have evolved as the consequence of the deletion of a part in the original syllables. The structure of the spectrogram of the terminal syllable at Oak Harbor are very different from the others, as can be seen in the tree.

IV. DISCUSSION

The segmentation of bird songs into syllables and the subsequent classification of those syllables have largely been based on human perception and aesthetics.[28] How birds perceive and parse songs is not known. The aim of this study has been to apply analytic methods to this problem, to extract reproducible classifications for large data sets.

The characteristics of evolving neural networks can be modified by using different values of the parameters direct-

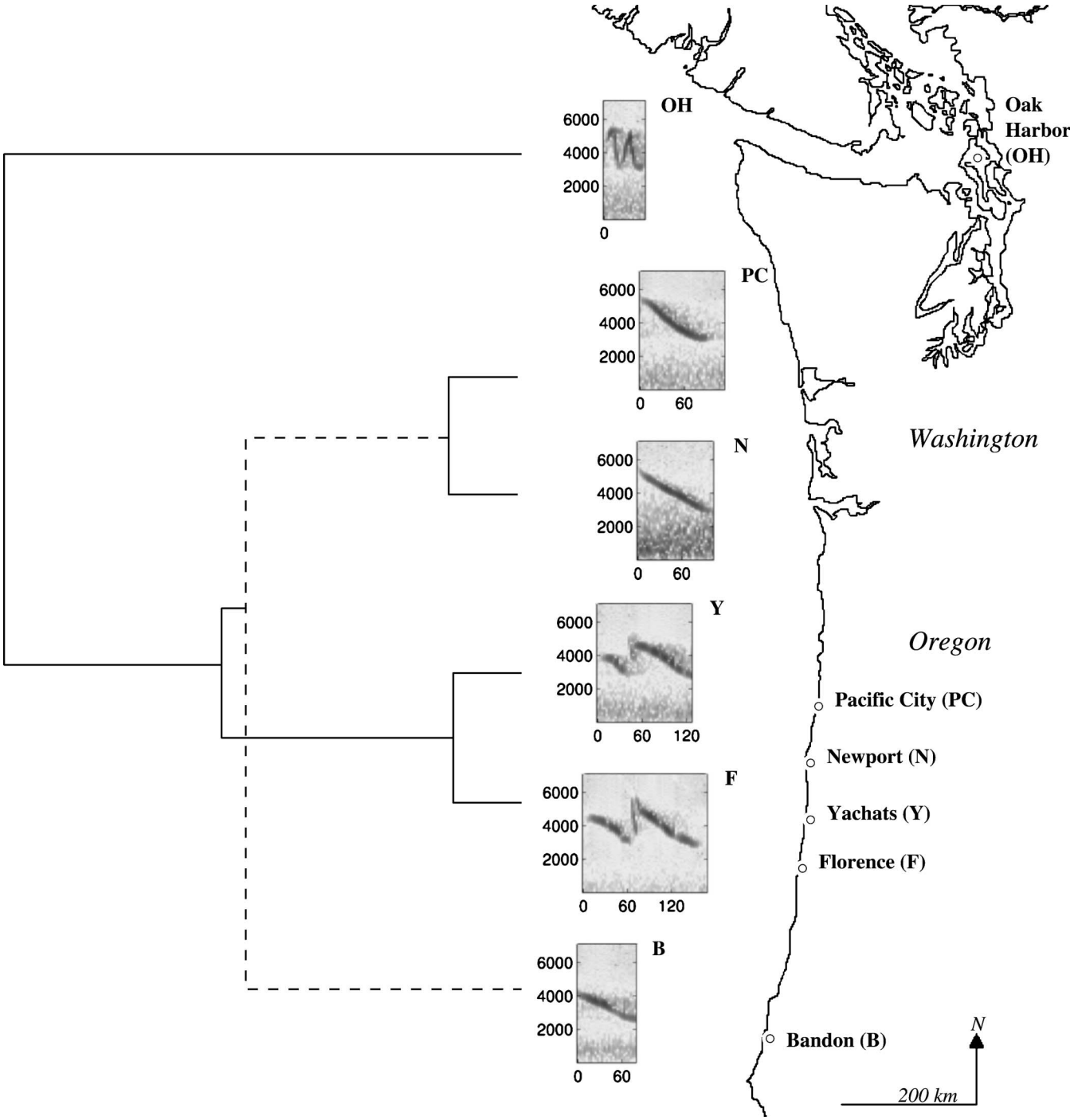


FIG. 4. Spectrograms of the most common clusters occurring at the end part of the songs. The spectrograms are ordered according to the geographic location where they were recorded. The spectrograms are positioned so that similar parts are vertically aligned in the time dimension. A nearest neighbor tree is shown on the left. Axes of the spectrograms are the same as in Fig. 2.

ing its growth. Those parameters are not independent but computer power limitations forbid an exhaustive parameter assessment. For example, the splitting threshold needs to be adjusted with the size of the tree. With a large initial number of leaves, the size of the network will increase rapidly and therefore the splitting threshold should be increased if one wants to limit the size of the final network. Single parameter assessments were conducted in order to get a better insight into their implications. The investigators may fine-tune the parameters for each specific data set depending on whether they wish to optimize the mapping precision or the size of the final network. Indeed, prior knowledge can indicate which aspect of the classification is more important. For example, the size of the network can be adjusted according to a range in the number of expected clusters.

Two main reasons can explain why the neural network classification conflicts with the visual classification in a few cases. First, the presence of strong background noise generates clusters which have been built principally on this criterion. With a visual approach, the noise is not taken into account for grouping syllables. In this work, frequencies below 1 kHz have been ignored, but the threshold could have been

increased, reducing the chance of creating spurious clusters. One could also focus on frequencies where birds have higher hearing sensitivity by excluding high frequencies and reducing the frequency bandwidth used for mel-cepstrum computation. In their work, Anderson *et al.*[4] limited the signal to 10 kHz. In this analysis, the recordings were not filtered before feature extraction and encoding of syllables. Ideally, the cepstrum coefficients should be computed with the help of a frequency scale specific to the bird species studied. Second, conflicts occur when long syllables are split into short parts during the segmentation of the songs. In this case, the information about frequency shifts throughout the syllable production is lost. The algorithm will cluster syllables differently depending on their duration. This shows the importance of the segmentation algorithm to this method. Appropriate threshold values must be chosen to obtain robust song segmentation.

The alignment distance calculation also has a strong effect. In particular, the insertion/deletion cost set for the alignment algorithm affects the importance of the syllables' duration in the distance calculation. Diminishing this cost will allow syllables of different lengths to be clustered together because of the resulting small pairwise distance. Consequently, the frequency similarities between the syllables will have greater importance. Weighted gap ends or weighted Euclidean distance between vectors are conceivable techniques for improving this distance measure. As noticed before, better results would be obtained by applying a limit to the distance between a sample and its best matching unit before affecting it to the cluster. In this way, potentially more centered clusters will be returned but at the expense of having some unclassified syllables. With larger data sets, each syllable type may occur more frequently so that each neuron would receive more hits of similar syllables in a constant number of epochs. This should improve the computation of the neuron weight matrices. An interesting aspect of this method is that it is able to retrieve small clusters containing rare syllables. This shows that the main cluster centers can be found in the data set. Therefore, it is possible to classify new recordings in order to examine how related they are to previously processed songs.

The distribution of the White-crowned Sparrow syllables, extracted from songs recorded in the west of America, strongly supports the presence of dialect. It is apparent that such culturally transmitted traits are continuously varying in space and time. The simple and complex syllables present the strongest geographical pattern but the whistle syllables seem to follow this trend too. Heterogeneity in the geographical patterns confirms that distinct positions in the White-crowned Sparrow songs evolve differently. Unique syllable clusters were found for most, but not all, dialects. A larger data set of songs and improved segmentation parameters would result in better resolution of syllable clusters and a closer match to human-based song classification. In some cases, the syllable's spectral structure seems to be a combination of different syllables, as seen in the end trill of the songs. Birds could potentially mix syllables together to produce new types. Thus, it is important to define relevant distance metric for performing comparisons independently from human biases.

V. CONCLUSION

Bird song syllable classification is a difficult task and the development of automatic methods acknowledges that the true classification is unknown. Indeed, bird song syllables are continuously varying characters. Retrieving cluster centers from syllable data sets can be achieved using evolving neural networks and a distance measure based on dynamic time warping. This method allows the processing of large data sets and reproducibility. Nevertheless, particular care must be given to the segmentation of the songs into syllables, to the encoding of the syllables, and to the choice of the distance measure and the parameters of the neural network learning process. It has been shown that this method is useful for the analysis of bird song evolution at different levels. First, the geographical distribution of syllables offers the opportunity to study dialects and potential population structure. Second, the distance measure gives a better insight into the fine spectrogram structure relationships between syllables.

ACKNOWLEDGMENTS

This work has been feasible thanks to the song recordings provided by the Borror Laboratory of Bioacoustics, Department of Evolution, Ecology, and Organismal Biology, Ohio State University, Columbus, OH and it has been supported by the Marsden Fund Council from Government funding, administrated by the Royal Society of New Zealand.

[1] P. Marler and M. Tamura, "Culturally transmitted patterns of vocal behavior in sparrows," Science **146**, 1483–1486 (1964).
[2] A. Lynch and A. J. Baker, "A population memetics approach to cultural evolution in chaffinch song: Meme diversity within populations," Am. Nat. **141**, 597–620 (1993).
[3] J. B. Kruskal and M. Liberman, "The symmetric time-warping problem: From continuous to discrete," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (CSLI, Stanford, CA, 1999), Chap. 4.
[4] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," J. Acoust. Soc. Am. **100**, 1209–1219 (1996).
[5] J. C. Brown, A. Hodgins-Davis, and P. J. O. Miller, "Classification of vocalizations of killer whales using dynamic time warping," J. Acoust. Soc. Am. **119**, EL34-EL40 (2006).
[6] T. Kohonen, "The self-organizing map," Proc. IEEE **78**, 1464–1480 (1990).
[7] J. Pakkanen, J. Iivarinen, and E. Oja, "The evolving tree—A novel self-organizing network for data analysis," Neural Processing Letters **20**, 199–211 (2004).
[8] D. W. Bradley and R. A. Bradley, "Application of sequence comparison to the study of bird songs," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, edited by D. Sankoff and J. B. Kruskal (CSLI, Stanford, CA, 1999), Chap. 6, pp. 189–209.
[9] M. D. Skowronski and J. G. Harris, "Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition," J. Acoust. Soc. Am. **119**, 1817–1833 (2006).
[10] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196 (1998).
[11] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," Anim. Behav. **59**, 1167–1176 (2000).

[12]S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd ed. (Dekker, New York, 2000).
[13]B. Gold and N. Morgan, *Speech and Audio Signal Processing, Processing and Perception of Speech and Music* (Wiley, New York, 2000).
[14]G. Kondrak, "Phonetic alignment and similarity," Computers and the Humanities **37**, 273–291 (2003).
[15]R. J. Dooling, M. R. Leek, O. Gleich, and M. L. Dent, "Auditory temporal resolution in birds: Discrimination of harmonic complexes," J. Acoust. Soc. Am. **112**, 748–759 (2002).
[16]S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," IEEE Trans. Audio, Speech, Lang. Process. **14**(4), pp. 1401–1412 (2006).
[17]M. Vacher, D. Istrate, L. Besacier, J. Serignat, and E. Castelli, "Life sounds extraction and classification in noisy environment," in SIP 2003: Fifth IASTED (The International Association of Science and Technology for Development) International Conference on Signal and Image Processing, Honolulu, HA, edited by M. Hamza, August 13–15, 2003.
[18]H. Slabbekoorn and A. den Boer-Visser, "Cities change the songs of birds," Curr. Biol. **16**(23), 2326–2331 (2006).
[19]S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Process. **ASSP–29**, 254–272 (1981).
[20]M. B. Trawicki, M. T. Johnson, and T. S. Osiejuk, "Automatic song-type classification and speaker identification of norwegian ortolan bunting (*Emberiza hortulana*) vocalizations," in 2005 IEEE Workshop on Machine Learning for Signal Processing (2005), pp. 277–282.
[21]S. Sharp and B. Hatchwell, "Individuality in the contact calls of cooperatively breeding long-tailed tits (*Aegithalos caudatus*)," Behaviour **142**, 1559–1575 (2005).
[22]B. J. Oommen, "String alignment with substitution, insertion, deletion, squashing, and expansion operations," Inf. Sci. (N.Y.) **83**, 89–107 (1995).
[23]P. Somervuo and T. Kohonen, "Self-organizing maps and learning vector quantization for feature sequences," Neural Processing Letters **10**, 151–159 (1999).
[24]J. Pakkanen, J. Iivarinen, and E. Oja, "The evolving tree—Analysis and applications," IEEE Trans. Neural Netw. **17**, 591–603 (2006).
[25]J. Vesanto, "Neural network tool for data mining: SOM toolbox," Technical Report, Proceedings of TOOLMET 2000 — 3rd International Symposium on Tool Environments and Development Methods for Intelligent Systems, April 13–14, University of Oulu, Oulu, Finland. http://citeseer.ist.psu.edu/388267.html.
[26]D. A. Nelson, K. I. Hallberg, and J. A. Soha, "Cultural evolution of Puget sound white-crowned sparrow song dialects," Ethology **110**, 879–908 (2004).
[27]L. F. Baptista, "Geographical variation in song and dialects of the Puget sound white-crowned sparrows," Condor **79**, 356–370 (1977).
[28]P. S. Warren, "Geographic variation and dialects in songs of the bronzed cowbird (*Molothrus aeneus*)," Auk **119**, 349–361 (2002).

Spatial release from energetic and informational masking in a selective speech identification task[a)]

Antje Ihlefeld and Barbara Shinn-Cunningham[b)]
*Auditory Neuroscience Laboratory, Boston University Hearing Research Center, 677 Beacon St., Boston, Massachusetts 02215, USA*1

(Received 10 November 2006; revised 12 March 2008; accepted 13 March 2008)

A masker can reduce target intelligibility both by interfering with the target's peripheral representation ("energetic masking") and/or by causing more central interference ("informational masking"). Intelligibility generally improves with increasing spatial separation between two sources, an effect known as spatial release from masking (SRM). Here, SRM was measured using two concurrent sine-vocoded talkers. Target and masker were each composed of eight different narrowbands of speech (with little spectral overlap). The broadband target-to-masker energy ratio (TMR) was varied, and response errors were used to assess the relative importance of energetic and informational masking. Performance improved with increasing TMR. SRM occurred at all TMRs; however, the pattern of errors suggests that spatial separation affected performance differently, depending on the dominant type of masking. Detailed error analysis suggests that informational masking occurred due to failures in either across-time linkage of target segments (streaming) or top-down selection of the target. Specifically, differences in the spatial cues in target and masker improved streaming and target selection. In contrast, level differences helped listeners select the target, but had little influence on streaming. These results demonstrate that at least two mechanisms (differentially affected by spatial and level cues) influence informational masking. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2904826]

PACS number(s): 43.66.Dc, 43.66.Pn, 43.66.Qp [RLF] Pages: 4369–4379

I. INTRODUCTION

When listening selectively for target speech in a background of competing talkers, at least two forms of masking can interfere with performance: energetic and informational masking (see, e.g., Freyman *et al.*, 1999; Brungart *et al.*, 2001; Arbogast *et al.*, 2002; Brungart *et al.*, 2005; Kidd *et al.*, 2005). Spatially separating the target from concurrent masker(s) can improve performance, an effect known as spatial release from masking (SRM; e.g., see Hirsh, 1948; Cherry, 1953; Arbogast *et al.*, 2002). While traditional binaural models can account for spatial release from energetic masking (e.g., Zurek, 1993), the mechanisms underlying spatial release from informational masking are poorly understood.

Two stimulus characteristics are often said to contribute to informational masking: (1) similarity between target and masker with respect to perceptual (e.g., Freyman *et al.*, 1999; Darwin and Hukin, 2000; Brungart, 2001a) or linguistic attributes (Hawley *et al.*, 2004; Van Engen and Bradlow, 2007), and (2) uncertainty about either target or masker (e.g., Lutfi, 1993; Kidd *et al.*, 2005a; Freyman *et al.*, 2007). Past work suggests that at least some of these effects of informational masking are due to failures in segregation and/or attention (e.g., Brungart *et al.*, 2005; Edmonds and Culling, 2006). However, there is no clear consensus on how these two processes contribute to spatial release from informational masking.

Previous studies comparing energetic and informational masking indicate that analysis of response errors can dissociate effects caused by energetic and informational masking. For instance, in selective speech identification tasks with interference from informational masking, subjects often report words from the masker rather than the target message(s) (Brungart, 2001b; Kidd *et al.*, 2005a; Wightman and Kistler, 2005). In contrast, for selective-listening tasks that are dominated by energetic masking, errors are more randomly distributed. The current study attempts to tease apart the mechanisms underlying informational masking through a more detailed analysis of response patterns than has been undertaken in previous studies. The analyses are driven by the hypothesis that similarity between target and masker can interfere with (1) extracting local time-frequency segments from the acoustic mixture, (2) connecting segments from the same source across time (streaming) and/or (3) selecting the correct target segments (or stream) even if they are properly segmented and streamed.

We explored how spatial separation between target and masker influences the pattern of responses, and how these patterns are affected by the level difference between target and masker. Listeners were asked to report a phrase from a variable-level target message that was presented simultaneously with a fixed-level masker message. The locations of target and masker were simulated over headphones to be either co-located or spatially separated by 90°. In addition,

[a)]Portions of this work were presented at the 2005 Mid-Winter meeting of the Association for Research in Otolaryngology.
[b)]Author to whom correspondence should be addressed. Electronic mail: shinn@cns.bu.edu.

target level varied over a wide range so that differences in level between target and masker could provide listeners with a cue to select the target and/or better link target keywords across time into a coherent target stream.

Analysis of response errors revealed systematic changes with spatial configuration and target level in the likelihoods of reporting all target keywords, all keywords from the masker, or a mixture of keywords from target and masker. The pattern of results suggests that the relative contributions of energetic and informational masking change systematically with the target-to-masker broadband energy ratio (TMR). At near-zero-dB TMRs, when informational masking occurs, space and intensity cues may help listeners track keywords across time to form a proper stream, as well as enable listeners to select the proper keywords or streams out of the mixture.

II. METHODS

A. Subjects

Four subjects (ages 21–24) were paid for their participation in the experiments. All subjects were native speakers of American English and had normal hearing, confirmed by an audiometric screening. All subjects gave written informed consent (as approved by the Boston University Charles River Campus Institutional Review Board) before participating in the study.

B. Stimuli

Raw speech stimuli were taken from the Coordinate Response Measure corpus (CRM, see Bolia *et al.*, 2000), which consists of highly predictable sentences of the form "Ready ⟨*call sign*⟩, go to ⟨*color*⟩ ⟨*number*⟩ now." The call sign was one of the set ["Baron," "Eagle," "Tiger," and "Arrow"]; the color was one of the set ["white," "red," "blue," "green"]; and the number was one of the digits between one and eight, excluding the number seven (as it is the only two-syllable digit and is therefore relatively easy to identify). For each session, one of the four call signs was randomly selected as the target call sign.

In each trial, two different sentences were used as sources. The call signs, numbers, and colors in the two utterances were randomly chosen, but constrained to differ from each other in each trial (with one sentence always containing the target call sign). In order to minimize differences between competing messages, talker 0 was used for both sentences.

Each speech signal was processed to produce intelligible, spectrally sparse speech signals (e.g., see Shannon *et al.*, 1995; Dorman *et al.*, 1997; Arbogast *et al.*, 2002; Brungart *et al.*, 2005). All processing was implemented in MATLAB 6.5 [see Fig. 1(a) for a diagram of the processing scheme]. Each target and masker source signal was bandpass filtered into 16 fixed frequency bands of 1/3 octave width, with center frequencies spaced evenly on a logarithmic scale between 175 Hz and 5.6 kHz (every one-third octave). The envelope of each band was extracted using the Hilbert transform. Subsequently, each envelope was multiplied by a pure tone carrier at the center frequency of that band. Figure 1(b) shows an example stimulus spectrum with all 16 bands (consecutive bands are shown in alternating shades). In contrast to many previous experiments using amplitude-modulated sine-wave carrier speech (e.g., Arbogast *et al.*, 2002; Gallun *et al.*, 2005; Brungart *et al.*, 2005; Kidd *et al.*, 2005b), the frequency bands of the current stimuli were not equalized to have similar spectral amplitudes, so that the high-frequency bands have less energy than the low-frequency bands.

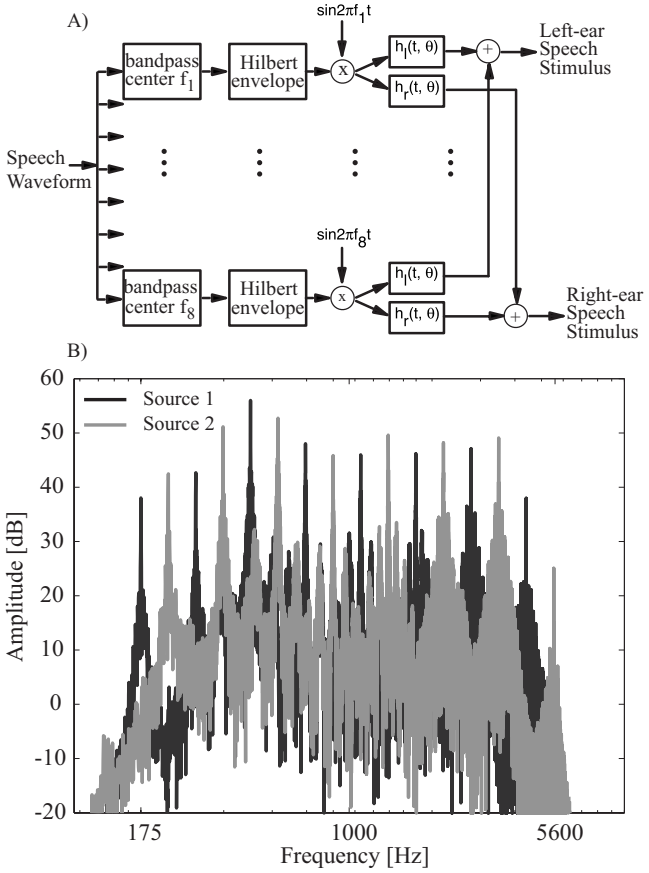


FIG. 1. (a) Flow chart showing how the stimuli were generated. (b) An example of the resulting interleaved spectra of two competing spectrally sparse messages, one in gray and one in black.

On each individual trial, eight of the 16 bands were chosen randomly while ensuring that four of these bands were selected from the lower eight bands (175–882 Hz) and four were selected from the upper eight bands (1.1–5.6 kHz). This resulted in a set of $(8!/(4!4!))^2$ or 4900 different possible spectral combinations. The eight bands were then summed to create the raw waveform for one source. The remaining eight bands were used to construct the other source using otherwise identical processing. As a result, the two raw sources had identical statistics over the course of the experiment, but differed within a trial in their timbre, call sign, and keywords (and, on most trials, level).

The raw source waveforms were scaled to have the same fixed, broadband root-mean-square (RMS) energy prior to spatial processing (described below). When target and masker were set to the same level of broadband RMS energy, the within-band energy ratio of one utterance to another was on the order of 20 dB at all frequencies [cf. Fig. 1(b)]. In fact, a model of the auditory periphery shows that for these

stimuli, energetic masking is likely to play a significant role within each target band only when the selected broadband TMR is −20 dB and less (see Shinn-Cunningham *et al.*, 2005a). Within each trial, the two spatially processed sentences were closely time aligned (this is an inherent property of the CRM corpus; no steps were taken to alter the temporal synchrony present in the original CRM corpus). To determine how synchronized the colors and numbers are when two sentences from Talker 0 are begun simultaneously, the timing of the key words was measured using a routine programmed in MATLAB 7.0. The color and number words were 307 and 290 ms long, on average, with standard deviations of 160 and 148 ms (for colors and numbers, respectively). The onset times of the keywords were almost synchronous across utterances; the standard deviations for the onset times of the color and number words across all stimuli are 29 and 164 ms, respectively.

C. Spatial synthesis

The raw signals were simulated at a distance of 1 m in the horizontal plane containing the ears, either at azimuth 0° (straight ahead) or 90° (to the right of the listener), using head-related transfer functions (HRTFs). HRTFs were measured in a classroom using a Knowles Electronics Manikin for Acoustic Research. The first 10 ms of the HRTFs were time windowed and band limited between 200 Hz and 10 kHz to get pseudo-anechoic HRTFs (see Shinn-Cunningham *et al.*, 2005b for details about the HRTFs used in this spatial processing). The prefiltered utterances were processed with appropriate HRTFs to simulate the desired configuration on a given trial.

D. Procedures

At the start of each session, a random call sign was selected to serve as the target call sign. The target call sign was always in the variable-level talker message. Listeners were instructed to report the color and number of the message with the target call sign, ignoring the message of the fixed-level talker, which will be referred to as the masker. A trial was scored as correct and subjects were given feedback that they were correct if and only if they reported both target keywords.

In each trial, the masker was presented at the same RMS level (which was approximately 70 dB sound pressure level SPL, when presented through the hardware). The target level was varied relative to that of the masker by an amount that was random from trial to trial, chosen from one of six levels (−40, −30, −20, −10, 0, and +10 dB, relative to the level of the masker prior to spatial processing). Subsequently, the binaural signals for the two talkers were summed to produce the two-talker stimulus.

There were four possible spatial configurations, two in which the target and masker were co-located (at either 0 or 90°) and two in which the talkers were spatially separated (target at 0° and masker at 90°, or target at 90° and masker at 0°).

In each run, the spatial configuration of the two talkers was fixed (i.e., the target and masker were played from constant locations throughout the run). In half of the sessions, subjects were told the call sign and location of the target message prior to each run; in the other half of the sessions only the call sign was identified *a priori* (although the location for the target was fixed throughout a block of trials). However, this difference in instructions had no consistent effect on results, so the data were collapsed across the different instructions.[1]

To ensure that subjects could understand the highly processed speech stimuli, subjects went through an initial screening in which they had to report the color and number of one message presented in quiet (processed by 0° HRTF) with at least 90% accuracy over 50 trials. None of the subjects failed this initial screening. Following the screening, each subject performed a training session consisting of 300 trials (at least one run of 50 trials for each spatial configuration, and an additional run of 50 trials for each of two randomly picked spatial configurations).

Following training, subjects performed four sessions of the experiment (one session per day). Additional data were collected in four additional sessions discussed in the companion paper (Ihlefeld and Shinn-Cunningham, submitted), in which listeners were asked to report both sets of keywords from both of the concurrent messages. Each session consisted of 12 runs (three runs for each of the four spatial configurations). The order of the runs in a session was random, but constrained so that each spatial configuration was performed once before any were repeated. A run consisted of eight repetitions of each of the six TMRs (48 trials per run). The orders of the runs within each session were separately randomized for each subject. Given that each subject performed four sessions of this experiment, each subject performed 96 repetitions of each specific configuration (8 repetitions/run × 3 runs/session × 4 sessions).

III. RESULTS

A. Percent correct

Given the four possible colors and seven possible numbers, the probability of responding correctly by chance is 1/28 or about 4%; however, if subjects understood only the masker color and number, realized they had heard the masker (not the target), and eliminated these possible responses, the likelihood of responding correctly by chance is 1/18 (6%).

The top panel of Fig. 2 shows the across-subject mean percent correct as a function of TMR for each spatial configuration; error bars show the standard errors of the mean across subjects. Results for all subjects show similar trends, so only across-subject averages are shown. In all configurations, performance improves with increasing TMR. When target and masker were spatially separated, performance was always better than for the co-located cases (dashed lines fall above solid lines). For both co-located and spatially separated conditions, performance at low TMRs was near chance levels. For spatially separated configurations, performance improved to near 100% for near-zero TMRs; however, in the co-located conditions, performance only reached about 80% at the highest TMRs. Moreover, for the co-located configurations, performance at TMR=0 dB, when the target and

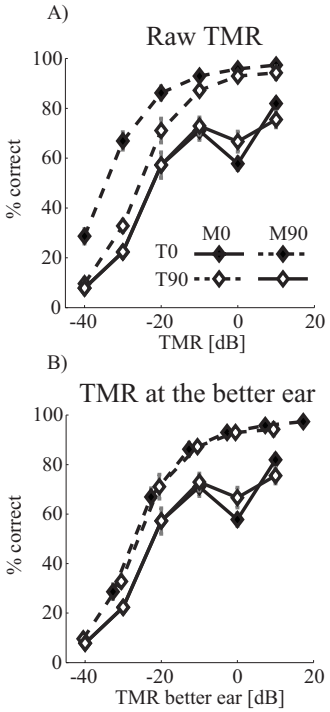


FIG. 2. Overall percent correct as a function of the TMR for the four tested spatial configurations, averaged across subjects. In general, performance improves with TMR, and is better for spatially separated than co-located sources. Error bars show the across-subject standard error of the mean. Filled symbols show results for the target at 0°, open symbols for the target at 90°. Results for spatially separated target and masker are shown with dashed lines. Results for co-located sources are shown with solid lines. (a) Results plotted as a function of the broadband target to masker ratio (TMR). (b) The same results re-plotted as a function of TMR_{be} (correcting for differences in the acoustic TMR at the better ear).

masker were at the same level, was actually worse than at TMR=−10 dB. This is similar to a plateau effect that has been observed in previous studies (Brungart, 2001b; Shinn-Cunningham *et al.*, 2005a).[2]

For co-located configurations, results are essentially identical when the target and masker play from in front of or from the side of the listener. In the separated conditions, results were best when the target originated in front of the listener and the masker from the side. Previous work suggests that the difference in performance for the two spatially separated configurations can be accounted for by considering the differences in the broadband TMR at the acoustically better ear (Shinn-Cunningham *et al.*, 2005a). Prior to the spatial simulation, RMS values of the processed speech waveforms were normalized to have the same broadband RMS, and then the overall level of the target was adjusted downward to produce the desired TMR. However, the spatial processing also altered the target and masker levels, so that the TMR in the signals reaching the ears varied with spatial configuration and could differ at the two ears (depending on the spatial configuration). In the condition where the target was at 0° and the masker at 90°, the TMR at the left ear was, on average, 7 dB greater than the nominal TMR, because the masker energy reaching the left ear was significantly reduced by the acoustic head shadow (note, however, that due to the random selection of frequency bands making up the target and masker on a given trial, the actual TMR at the ears varied somewhat from trial to trial). When the target was at 90° and the masker was in front, the TMR at the right ear was almost equal to the nominal TMR, averaging 1 dB lower than the nominal TMR (of course, for this configuration, the TMR at the left ear is 7 dB lower than the nominal TMR). Note that on average, the TMR for co-located target and masker equaled the nominal TMR.

TABLE I. Possible response types and chance performance for each category. On each trial, subjects responded by reporting a color (C) and a number (N). The subscripts denote whether a keyword was part of the target (T) or masker (M) message; X denotes that the reported word was not present in either the target or the masker.

Response type	Chance	Responses
Masker error	3.6%	$[C_MN_M]$
Mix error	7.1%	$[C_TN_M]$ or $[C_MN_T]$
Drop-1 error	25.0%	$[C_TN_X]$ or $[C_XN_T]$
Drop-2 error	35.7%;	$[C_XN_X]$
Combination error	25.0%.	$[C_MN_X]$ or $[C_XN_M]$
Correct	3.6%	$[C_TN_T]$

To take into account these acoustic effects, data were re-plotted as a function of the TMR at the acoustically better ear (denoted by TMR_{be}) by shifting the raw data in the top panels of Fig. 2 by the appropriate amounts in dB for each spatial configuration (for discussion see Shinn-Cunningham *et al.*, 2005a). As seen in the bottom panels of Fig. 2, this adjustment accounts for differences in performance for the two spatially separated configurations (results for the two separated configurations are indistinguishable following this correction).

B. Analysis of response errors

All incorrect responses were categorized into one of five mutually exclusive error types whose definitions are listed in Table I. Figure 3 plots each kind of error. Errors generally decreased with increasing TMR_{be}. However, the relative likelihoods of the different types of errors depended on TMR_{be} and spatial configuration. At low $TMRs_{be}$ (−20 dB and below), *drop errors* (reporting keywords not in either the target or the masker messages) were the most common errors. For TMR_{be} of −10 dB and greater, *masker* (reporting both masker keywords) and *mix error*s (reporting a mix of target and masker keywords) were the most common errors when sources are co-located, but both types of error were rare for spatially separated sources.

*Drop error*s are shown in panels (A) and (B). The proportion of trials with *drop error*s decreased steeply with increasing TMR_{be} in all spatial configurations, consistent with a decrease in the amount of energetic masking with increasing TMR_{be}. For TMR_{be} of −10 dB and greater very few *drop error*s occurred. For $TMRs_{be}$ between −30 dB and −10 dB (where floor and ceiling effects can be ignored), the number of drop errors was larger for co-located than for separated configurations (dashed lines fall below solid lines), although this difference was small.

Panel (C) displays *mix error*s, where subjects report one target and one masker keyword. At low $TMRs_{be}$ (−40 dB to −20 dB) *mix error*s increased as TMR_{be} in-

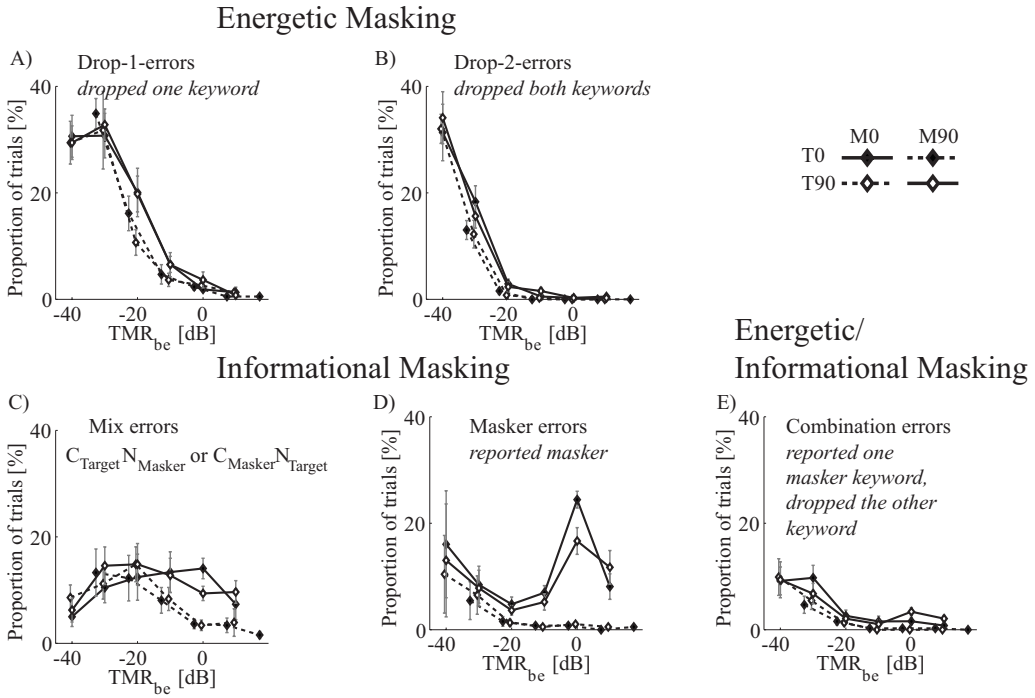


FIG. 3. Analysis of the response errors as a function of TMR_{be} for each kind of error, averaged across subjects. The dependence of each type of response error on TMR_{be} differs, indicating differences in the kind of masking responsible for the errors. Error bars show the across-subject standard error of the mean. Filled symbols show results for the target at 0°, open symbols for the target at 90°. Results for spatially separated target and masker are shown with dashed lines. Results for co-located sources are shown with solid lines. Labels above the panels indicate the posited kind of perceptual interference thought to contribute to the different types of response errors. (a) *Drop-1* errors, where listeners report one target word and one word not present in either the target or masker. (b) *Drop-2* errors, where listeners report two words not present in either the target or masker. (c) *Mix error*s, where listeners report one word from the target and one from the masker. (d) *Masker error*s, where listeners report both words from the masker. (e) *Combination* errors, which encompass all incorrect responses that are not in the above categories (e.g., reporting one masker word and guessing one word).

creased, with no significant differences between co-located and separated spatial configurations. For TMR_{be} −10 dB and greater, the rate of *mix error*s decreased with increasing TMR_{be} in the spatially separated configurations (dashed lines), but was nearly constant in the co-located configurations (given the across-subject variability; see solid lines). In other words, spatial separation between target and masker reduced the likelihood that listeners reported one target and one masker keyword, but only when the TMR_{be} was larger than −20 dB.

Masker errors are shown in panel (D). When sources were spatially separated (dashed lines), *masker* errors decreased monotonically with increasing TMR_{be}. Essentially no *masker* errors occurred for $TMRs_{be}$ of −10 dB or greater when target and masker were spatially separated. At all TMR_{be}, more *masker* errors occurred when the two sources were at the same location than when they were spatially separated (solid lines are above the dashed lines). *Masker* errors in the co-located configurations were nonmonotonic, decreasing as TMR_{be} grew from −40 dB to −20 dB, increasing as TMR_{be} grew from −20 dB to 0 dB, and then decreasing again for TMR_{be} of 10 dB.

Panels (C) and (D) show different trends in the likelihoods of obtaining *mix* versus *masker* errors. In particular, the ratio of the percentages of *mix* errors and *masker* errors varied nonmonotonically as a function of TMR_{be} (ratio not shown) and depended on whether or not target and masker were spatially separated or co-located. At −40 dB TMR_{be}, there were roughly half as many *mix* errors as *masker* errors. This ratio increased monotonically until around −20 dB TMR_{be}, where *mix* errors occurred at least four times more often than *masker* errors. For −10 dB TMR_{be} and greater, in the spatially separated configurations, *masker* errors were essentially nonexistent and *mix* errors decreased monotonically. In contrast, in the co-located configurations, *masker* errors peaked at 0 dB TMR_{be}, while *mix* errors were roughly constant for TMR_{be} greater than −10 dB. Together, these trends show that the spatial configuration of the sources and TMR_{be} affected *mix* and *masker* errors in different ways. This suggests that the relative importance of different forms of interference depends on the relative levels and locations of target and masker.

Combination errors, shown in panel (E), are relatively uncommon. These errors decreased as a function of TMR_{be}; almost no *combination* errors occurred for $TMRs_{be}$ of −20 dB and greater, and there were no significant differences between the spatially separated and co-located configurations.

C. Spatial gains

For each individual subject, logistic fits for the two co-located configurations were derived and averaged, as were the logistic fits for the two spatially separated configurations (after accounting for the TMR at the better ear; see Appendix for details). Between 30 and −20 dB TMR_{be} the *vertical difference between these averaged spatially separated and averaged co-located logistic fits* (the percent spatial gain) was

approximately 20% for subjects S1 and S4, and 7% for subjects S2 and S3. At the greatest $TMRs_{be}$, the percent spatial gain was between 20% (S1 and S4) and 25% (S2 and S3). The *horizontal shift between the co-located and separated configurations* (the dB spatial gain) was approximately 6 dB for subjects S1 and S4 and 2 dB for subjects S2 and S3 (in the performance range between 40% correct and 60% correct).

IV. DISCUSSION

In this selective attention task, listeners were asked to report the content of the message that contained a particular call sign (Brungart, 2001a). This target message was usually presented at a lower level than the fixed-level masker. To perform well in this selective task, listeners had to be able to segregate the (monosyllabic) target keywords from the acoustic mixture and report the proper keywords. Both energetic masking and informational masking interfered with performance in this task. Two factors emphasize the role of informational masking in the current study, at least at TMRs of −10 dB and above. First, target and masker messages were presented in nonoverlapping spectral bands (for a detailed analysis of simulated auditory nerve firing patterns for these types of stimuli see Shinn-Cunningham *et al.*, 2005a). Second, the target and masker were designed to be perceptually similar (e.g., acoustically, semantically, linguistically, etc.).

A. Cues for distinguishing target and masker

Target and masker keywords were nearly synchronous, and no semantic cues aided in distinguishing the target from the masker message. Because of the way the stimuli were processed, the messages did not have a strong pitch. The timbres of the target and masker signals varied unpredictably from trial to trial and these timbre differences were not very salient, so that it is unlikely that listeners relied on timbre to select the target. Thus, relatively few cues were available to help listeners differentiate the target from the masker. When target and masker were spatially separated, both spatial location and level differences between target and masker could be used to select target segments or the target stream from the mixture. However, when target and masker were co-located, the main cue enabling target selection was the level difference between target and masker (in those trials where target and masker had different levels).

B. Energetic masking and informational masking change systematically with TMR_{be}

Energetic masking is reduced and informational masking further emphasized when target and masker speech are presented in spectrally interleaved narrowbands (Arbogast *et al.*, 2005). In addition, several studies suggest that the amount of interference from informational masking is large in selective listening tasks that use the Coordinate Response Measure (CRM) corpus (Bolia *et al.*, 2000; Brungart, 2001a; Brungart *et al.*, 2005; Kidd *et al.*, 2005b; Wightman *et al.*, 2006). To emphasize the effects of informational masking, the current study employed spectrally interleaved bandpass filtered target and masker speech that were both derived from the CRM corpus.

We expected energetic masking to dominate at low $TMRs_{be}$ and to decrease with increasing TMR_{be}. At low $TMRs_{be}$ (−40 dB to −20 dB), the most common response errors were *drop error*s (nearly 60% of all trials), suggesting that indeed, at these low target levels, energetic masking was the dominant form of masking. In addition, *masker error*s occurred at a rate well above chance. In some ways, it is surprising that *masker error*s existed at all at these low $TMRs_{be}$: listeners were reporting the content of the more intense talker, which they should have realized was the masker and thus excluded from their response. This kind of result, where listeners seem unable to ignore a talker that they should know is the masker, has been observed in other studies of informational masking (e.g., Brungart and Simpson, 2004; Kidd *et al.*, 2005a). Such errors may occur because listeners are not completely certain that the message they heard was from the masker, or because it is too confusing to switch to a strategy of guessing a word not heard in these trials while still reporting the heard words in the other trials. Regardless, these *masker error*s at low TMR_{be} likely reflect a failure to hear the target (energetic masking).

Both *drop* and *masker error*s decreased as TMR_{be} increased from −40 dB to −20 dB, consistent with a decrease in energetic masking. Moreover, spatial separation yielded a slightly lower rate of both of these energetic-masking-caused errors, supporting the idea that binaural decorrelation processing provides a small release from energetic masking for the low-frequency portions of the target within 10–15 dB of masked threshold (Zurek, 1993; Durlach, 1972; Shinn-Cunningham *et al.*, 2005a). This release from energetic masking due to binaural decorrelation may not have required *perceived* spatial differences between target and masker (Colburn and Durlach 1965; Edmonds and Culling, 2005a; Edmonds and Culling, 2005b; Culling *et al.*, 2006), but simply a change in interaural correlation caused by the addition of the near-threshold signal (when the masker energy falling within the target bands can rival the target energy).

The pattern of errors was very different for $TMRs_{be}$ of −10 dB and greater. As the TMR_{be} increased from −20 dB, *drop* and *combination* errors disappeared, suggesting that energetic masking became negligible at the mid- to high-range $TMRs_{be}$. In this range, target-masker similarity of level and location (see Secs. IV D and IV E, below) determined how well listeners could extract the target from the mixture. *Masker* and *mix error*s were more likely when target and masker were co-located. This shows that once the target was audible and properly segmented from the acoustic mixture, informational masking dominated. Moreover, this pattern suggests that when the listener had trouble selecting the target from the properly segmented mixture, perceived spatial location was a salient, robust cue for identifying the target segments and/or target stream.[3]

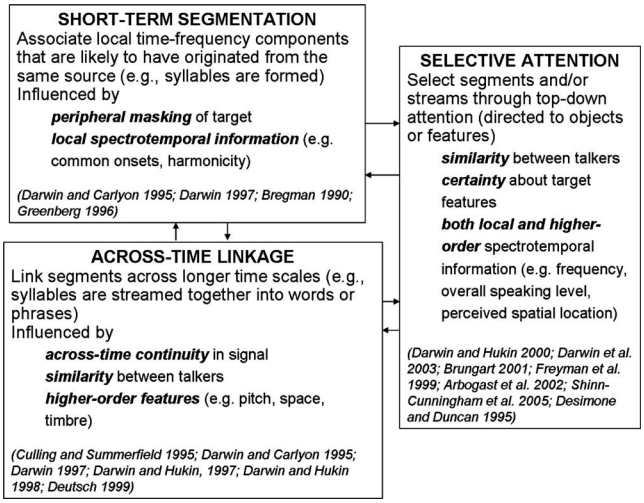


FIG. 4. Flow chart of the proposed conceptual framework of masking.

C. Conceptual framework explaining different forms of masking

At least three intricately linked mechanisms should affect masking in this kind of speech identification task: (1) short-term segmentation, (2) across-time linkage, and (3) selective attention (see definitions below and Fig. 4).

(1) Short-term segmentation is defined as the process by which all or part of the acoustic mixture is automatically segregated into local time-frequency components that are likely to have originated from the same source (e.g., see Bregman, 1990). Although segmentation may be influenced by attention, it is primarily based on the primitive spectro-temporal structure of the sound sources (e.g., see Darwin and Carlyon, 1995; Darwin, 1997). We assume that in the current task, when energetic masking is low, short-term segmentation can robustly extract speech segments on the time scale of syllables (Greenberg, 1996). Conversely, when energetic masking is high enough, target segmentation should fail, forcing listeners to guess the target keywords (or to adopt a different response strategy such as reporting the masker message).

(2) Temporal discontinuities (e.g., stop consonants, silent gaps) limit the duration of segments. Proper stream formation depends on "across-time linkage," the process of binding short-term segments across such discontinuities. Previous studies suggest that continuity of higher-order features such as timbre, perceived location, and overall intensity are important for across-time linkage (e.g., see Culling and Summerfield, 1995; Darwin and Carlyon, 1995; Darwin, 1997; Darwin and Hukin, 1997; Darwin and Hukin, 1998; Deutsch, 1999).

(3) Even if short-term segmentation and across-time linkage are performed flawlessly, listeners still must choose the correct stream from a sound mixture using "selective attention," a mechanism that may be directed both to local and higher-order spectro-temporal information (Freyman *et al.*, 1999; Darwin and Hukin 2000; Arbogast *et al.*, 2002; Darwin *et al.*, 2003; Shinn-Cunningham *et al.*, 2005a). Depending on the degree of similarity between target and interferers and the amount of certainty the listener has about the target features, selective attention can enhance sounds with desirable features and suppress others by biasing the sensory representation (e.g., Desimone and Duncan, 1995), thereby bringing a selected object into the perceptual foreground.

In performing the current selective listening task, listeners may use one of two response strategies. First, listeners may select all segments or the stream with a desired feature. In that case, the listener either must know these target features ahead of time, or they must estimate features of the target call sign when it occurs and selectively attend to the streams or segments with the estimated features of the target call sign. Alternatively, if listeners link segments into proper streams using across-time continuities inherent in each message, they may solve the task by attending to the stream that contains the target call sign. In fact, in conditions where higher-order acoustic features do not disambiguate which segments or which stream to attend to (e.g., when two perceptually similar messages are presented, and only a call sign within the target utterance defines which message is the target), listeners may have to use this strategy to solve the task. Specifically, given that timbre differences are relatively weak and unreliable in this task, listening for the target call sign and then properly linking it to the subsequent target words may be the only way to perform the task in the co-located configuration when the TMR_{be} is 0 dB. How do the current results support the conceptual framework? The following two sections shed light on this question.

D. Different mechanisms underlie masker errors and mix errors

Even though listeners were asked to report only the target message, listeners could almost always understand both the target and the masker messages when the TMR was −10 dB or higher and the target could be segmented properly.[4] Based on the conceptual framework (Sec. IV C), this observation suggests that *masker* errors occur either because the listener independently selects the wrong keywords for both the color and the number, or because the listener selects the wrong one of two properly formed streams. In contrast, *mix* errors may occur when the listener independently selects one correct keyword and one masker keyword, or when the listener selects a perceptual stream that is a mix of target and masker keywords. Both of these possibilities occur only when the listener fails to link keywords properly across time.[5]

In the current study, TMR affects performance in two ways. In general, listeners perform better with increasing target level. However, because *differences* in the levels of target and masker provide a cue to select the target segments or streams, performance does not improve monotonically with TMR_{be}. When target and masker are equally intense and co-located, subjects perform noticeably worse than when the masker is 10 dB more intense than the target. Thus, level must provide a cue that aids the target selection and/or improves the proper across-time linkage of the target keywords.

Looking in even more detail at the errors in the co-located configurations, the percentage of *masker* errors shows a pronounced peak at 0 dB TMR_{be}, while, in stark contrast, the percentage of *mix* errors does not. This differ-

ence in the patterns for *masker* and *mix* errors suggests that at least two different mechanistic failures contribute to informational masking: failures in across-time linkage of segments and failures in the selection of target segments and/or the target stream. In particular, mix errors are no more common at 0 dB TMR_{be} than at −10 or 10 dB TMR_{be}. Therefore, the across-time linkage of the target keywords is unaffected by TMR_{be}, suggesting that level cues do not provide a strong cue for streaming the keywords together. In contrast, masker errors are more common at 0 dB TMR_{be} than at −10 dB and 10 dB TMR_{be}. Thus, attention can be directed to a source based on a level difference between target and masker (explaining the jump in masker errors at 0 dB TMR_{be}).

Previous studies found that the usefulness of level cues depends on the types of stimuli used. The utility of intensity differences between the target and masking talkers decreases in importance as the number of maskers increases (Brungart *et al.*, 2001a; Freyman *et al.*, 2004), and is reduced when the masker is very different from the target in perceptual quality (Brungart 2001c). These results are consistent with the idea that level helps in selecting the target from the sound mixture (a process that should get more and more challenging the greater the number of competing talkers), but is redundant when other cues differentiate target and masker. None of these results suggest a role of level in automatic streaming of utterances.

E. Spatial release from masking

At the lowest $TMRs_{be}$, there is no difference in the rate of *mix error*s for co-located and spatially separated sources. This is consistent with previous studies that suggest that spatial release from informational masking depends on perceiving competing streams from different locations (Freyman *et al.*, 2001; Arbogast *et al.*, 2002; Shinn-Cunningham *et al.*, 2005a; Gallun *et al.*, 2005). Evidence suggests that objects start to be perceptually separated before they are heard at different locations (Woods and Colburn, 1992; Litovsky and Shinn-Cunningham, 2001; Best *et al.*, 2007). The current results are consistent with the idea that spatial release from informational masking only occurs when syllables are properly segmented and syllables are perceived at different locations, and that perception of the segmented target at the correct location only occurs at high $TMRs_{be}$, above the $TMRs_{be}$ that first allow the target to be segmented from the mixture.

In principle, spatial similarity could also cause difficulty in segmenting the two messages, by increasing uncertainty about which time-frequency components belong to which message. However, past studies show that spatial cues have little influence on short-term segmentation (Darwin, 1997).

At the mid- to high-range $TMRs_{be}$, when the target is intense enough to be segmented from the masker, the spatial gains in performance are markedly greater than at the low $TMRs_{be}$. The spatial release from masking at these $TMRs_{be}$ is primarily caused by a reduction in *masker* errors and *mix* errors in the spatially separated configurations compared to the co-located configurations. Together with the interpretation of how *masker* errors and *mix* errors are affected by across-time linkage and selective attention (see Sec. IV C and IV D), the spatial differences in the pattern of *masker* errors suggest that spatial cues improve the ability to select either independent target segments or a properly formed target stream.

The rate of *mix* errors decreases as TMR_{be} increases from −20 dB for spatially separated sources, but is essentially constant for co-located sources. At first glance, this reduction in *mix* errors appears to show that listeners can use spatial location as a cue for linking segments across time. However, because mix errors can also occur when listeners independently select one correct and one wrong segment, this spatial improvement in mix errors may merely reflect an improvement in the likelihood that listeners independently select the correct target keywords in the spatially separated configurations compared to the co-located configurations. Overall, these results are consistent with previous studies suggesting that spatial location can help listeners selectively attend to already-formed syllables (e.g., see Darwin and Hukin, 1999).

V. SUMMARY AND CONCLUSIONS

The results of this selective listening task give strong evidence that the relative influences of energetic masking and informational masking change systematically as a function of TMR_{be}. The pattern of results is consistent with the idea that different attributes of two competing signals can be used to select a target out of the mixture and to link short-term segments across time, including level differences between target and masker and the spatial cues that were the main focus of this study.

The pattern of errors as a function of the level difference between target and masker suggests that distinct mechanisms contribute to the types of errors in this selective speech identification task. In particular, *drop* errors appear to be caused predominantly by energetic masking; *masker* errors are most likely caused by energetic masking at low $TMRs_{be}$ and failures in selective attention at higher $TMRs_{be}$; and *mix* errors are most likely to occur when both across-time linkage and selective attention fail.

Spatial separation improves performance at all $TMRs_{be}$; however, the improvements come from different mechanisms at different $TMRs_{be}$. At the lowest TMR_{be}, binaural processing reduces energetic masking of the target, which, in turn, makes the target easier to segment from the mixture. There is no evidence that spatial cues improve selection or across-time linkage at these low $TMRs_{be}$. At higher $TMRs_{be}$, spatial release occurs by increasing the likelihood that the listener selects the correct keywords or the correct stream out of the mixture. The data hint that spatial differences between target and masker may also improve across-time linkage of syllables, but this conclusion is confounded by the possibility that selective attention alone may reduce the probability of selecting the color and number (independently), which would in turn reduced *masker* as well as *mix* errors. Finally, level differences between target and masker allow a listener to select the proper keywords from a mixture, but do not improve the perceptual linkage of the adjacent keywords into a single stream.

TABLE II. Mean parameters of the psychometric function fits for the different spatial configurations, averaged across subjects (across-subject standard error of the mean is shown in round brackets). The upper asymptote of performance is higher for spatially separated sources than for co-located sources, but no other differences are significant. (A) Estimates of α, the TMR at the midpoint of the dynamic range in the psychometric function. (B) Estimates of $1/\beta$, the slope of the psychometric function at the midpoint of the dynamic range. (C) Estimates of $1-\lambda$, the upper asymptote of the functions.

	T0M0	T90M90	T0M90	T90M90
(A) Midpoint of dynamic range α [dB]	27.4 (1.4)	24.5 (1.6)	23.5 (1.8)	24.2 (1.0)
(B) Slope at the midpoint of dynamic range $1/\beta$ [% correct / dB]	27.8 (11.0)	24.8 (16.3)	16.7 (4.5)	19.0 (3.4)
(C) Upper asymptote of performance $1-\lambda$ [% correct]	71.3 (4.6)[a]	72.4 (5.2)[a]	96.5 (2.0)	93.9 (4.2)

[a]Statistically significant difference between co-located and spatially separated configurations.

ACKNOWLEDGMENTS

This work was supported in part by grants from AFOSR and NIDCD. The authors are grateful to Gerald Kidd, Virginia Best, Christine Mason, Frederick Gallun, Steve Colburn, Richard Freyman, and three anonymous reviewers for their helpful comments.

APPENDIX

For each of the four spatial configurations, and separately for each subject, percent correct performance as a function of TMR_{be} was fitted by a logistic function using a maximum-likelihood method with bootstrapping, *psignifit* version 2.5.6 (http://www.bootstrap-software.com/psignifit/, see Wichmann and Hill, 2001a). The probability of responding correctly at a given TMR, $\hat{P}(x)$, equals

$$\hat{P}(x) = \gamma + (1-\lambda-\gamma)\frac{1}{1+e^{\alpha-x/\beta}}, \tag{A1}$$

where γ is the lower bound on performance, $1-\lambda$ is the upper bound on performance at the largest TMR, α is the energy ratio at which percent correct performance is exactly halfway between chance and the best observed performance, and $1/\beta$ is the slope of the psychometric function evaluated at $x=\alpha$. Note that this fitting algorithm places a higher emphasis on the steep portion of the psychometric function than a minimum least square fitting constraint would have.

The lower bound on performance γ was set to 6%, chance level assuming that listeners hear and rule out the masker keywords. Although there is ample evidence that listeners do not always do this, the fits were quantitatively better when setting chance performance to 6% rather than the 4% that would occur if listeners chose randomly among all keywords.

The goodness of fit of the psychometric functions was evaluated using Efron's bootstrap technique (Wichmann and Hill, 2001a, Wichmann and Hill, 2001b). Residual differences between the predictions from the fits and data were compared to the error residuals between the predictions and 10 000 runs of Monte Carlo simulated data sets (whose statistics equaled the estimated distribution of the data). A deviance measure was calculated as described in Eq. (A2) (for detailed discussion see Wichmann and Hill, 2001a):

$$D = \sum_{i=1}^{K}\left[n_i y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + n_i(1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right], \tag{A2}$$

where y_i denotes performance (either measured or simulated) at the TMR denoted by i, $\hat{p}_i$ is the percent correct predicted by Eq. (A1) for the corresponding TMR, n_i is the number of trials at each TMR ($n_i=96$ for all i), and K equals the number of TMRs tested ($K=6$). The deviance D_{measured} between predictions and measured data was calculated. Similarly, the deviances were calculated for each of the 10 000 sets of simulated data, yielding a set of D^*. Based on the Monte Carlo generated distribution of these 10 000 values of D^*, 95% confidence intervals for D_{measured} were then estimated.

For each measured data set for which D_{measured} falls within the 95% confidence interval, the fitting function was considered to provide a good description of the underlying data. Twelve of the 16 fits (four functions for each of four listeners) fell within the 95% confidence intervals of the distribution fits. In the four cases that did not meet this criterion, the largest errors in the predictions were consistently due to performance dips in the measured data at 0 dB TMR_{be} for co-located configurations. However, those data points cannot be fit by any monotonically increasing function. Other than missing these dips, the fits were deemed adequate descriptions of the data.

The resulting parameters, averaged across subjects, are listed in Table II. Unlike the patterns of the raw percent correct performance, the normalized midpoint parameters α of the psychometric functions (left panel) and the normalized slopes at the midpoints of the psychometric functions, $1/\beta$ (center panel) do not vary significantly with spatial configuration [T-test; $p>0.01$]. The upper bounds $1-\lambda$ are lower in the co-located than in the separated configurations, reflecting the lower level of performance for co-located configurations at the greatest TMRs (T-test; $p<0.01$).

At first glance, the fact that the upper bounds $1-\lambda$ are the only parameters that differ significantly across co-located and spatially separated configurations may appear counterintuitive. While the raw performance data differ for spatially

separated and co-located configurations, when normalized so that they range between 0% and 100%, the logistic fits have similar midpoints and slopes.

The parameters α and $1/\beta$ capture differences in midpoint and slope *relative* to the lower and upper limits of performance. Specifically, the midpoint parameter α is the TMR at which performance is halfway between the lower and upper bounds on performance for a given psychometric function, which will be different absolute levels of performance if the upper bound (parameter $1-\lambda$) varies with condition. Similarly, the slope parameter quantifies the percent of change in performance between lower and upper bounds of the psychometric function per dB, not the change in percent correct per dB, and so will translate to different absolute %/dB slopes if the upper bound varies with condition.

To the extent that logistic fits are adequate descriptions of the underlying data, this may suggest that a difference in the upper limits of the logistic fits between spatially separated and co-located configurations is sufficient to account for performance differences. Indeed, this is consistent with the idea that for the co-located configurations, the listener's attention may have been misdirected more often than in the spatially separated configurations, causing a decrease in asymptotic performance (cf. Lutfi *et al.*, 2003). However, fitting the nonmonotic performance function of our raw data with a monotonic logistic function conceals systematic differences in the midpoints of the raw performance data. In other words, this way of analyzing the data hides the fact that at a given TMR, spatial cues lead to absolute improvements in the ability to select the target keywords from the mixture.

[1]The lack of differences between these two sets of instructions suggests that, across all spatial configurations, listeners did not benefit from *a priori* knowledge of the target location. However, listeners could have computed the spatial location of the target call sign in the first few trials of a run, directed attention to that estimated location on subsequent trials, and then selected the keywords based on their perceived locations. Thus, this lack of any effect of instructions may simply reflect the fact that listeners may have adopted a strategy in which they directed attention to the target location, independent of the instructions.

[2]In each block, the target is softer than the masker in 67% of trials, allowing listeners to perform relatively well simply by focusing on the less-intense talker. The dip in performance at 0 dB TMR_{be} has been attributed to the loss of this relative level cue for selecting target words from the mixture. Not all studies show a drop in performance at 0 dB TMR (e.g., see Arbogast *et al.*, 2002). Similarly, not all of them show an upper bound of 80% (rather than 100%) correct. However, those studies that do not show a drop in performance for equal intensity target and masker (and that had higher high-TMR performance) generally used more speech bands for the target than for the masker, or they used full speech. This may have made the target more salient, even when it was the same broadband level as the masker, and thus easier to understand than the masker. This is in line with findings by Brungart and colleagues, who show that the amount of across-ear interference in a dichotic masking paradigm increases with the number of masker bands for amplitude-modulated sine-wave carrier speech as well as modulated-noise-band speech (Brungart *et al.*, 2005), a result that suggests that performance decreases as the intelligibility of an informational masker increases.

[3]Unlike the processed speech used in the current study, in ordinary conversational speech within-stream continuity cues are much stronger, and pitch, semantic, and linguistic information help listeners to link syllables and words across time. The stimuli used here are likely to make it more difficult to properly track keywords from a target message across time compared to normal everyday discourse. This may cause listeners to rely more heavily on other cues in the stimuli (level, location) important for auditory scene analysis. These stimuli allow us to tease apart whether level and location contribute to across-time linkage and/or selection of keywords and/or streams.

[4]The listeners who participated in this experiment also participated in a companion study with identical stimuli in which they were asked to report keywords from both utterances (Ihlefeld and Shinn-Cunningham, submitted). Listeners could report all four keywords (of both target and masker) nearly as well as they could report the keywords of the target message. Within each of the spatial configurations, the percent correct performance in this selective task (reporting both target keywords correctly) is nearly equal to percent correct performance in the divided task in the companion study (reporting all four keywords correctly), never differing by more than 10%.

[5]Failures in short-term segmentation, across-time linkage, and/or selection can occur at any time during the presentation of the target and masker utterances. However, we only measured performance for color and number keywords. Therefore, based solely on our results, we cannot determine how across-time linkage between (for instance) the call sign and color depends on stimulus manipulations.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2005**). "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **117**, 2169–2180.

Best, V., Gallun, F. J., Carlile, S., and Shinn-Cunningham, B. G. (**2007**). "Binaural interference and auditory grouping," J. Acoust. Soc. Am. **121**, 1070–1076.

Bolia, R. S., Nelson, W. T., and Ericson, M. A. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).

Brungart, D. (**2001a**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D. S. (**2001b**). "Evaluation of speech intelligibility with the coordinate response measure," J. Acoust. Soc. Am. **109**, 2276–2279.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001a**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Brungart, D. S., and Simpson, B. D. (**2004**) "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," J. Acoust. Soc. Am. **115**, 301–310.

Brungart, D., Simpson, B., Darwin, C., Arbogast, T., and Kidd, G. J. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Colburn, H. S., and Durlach, N. I. (**1965**) "Time-intensity relations in binaural unmasking," J. Acoust. Soc. Am. **38**, 93–103.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785–797.

Culling, J. F., Edmonds, B. A., and Hodder, K. I. (**2006**). "Speech perception from monaural and binaural information," J. Acoust. Soc. Am. **119**, 559–565.

Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *Hearing*, edited by B. C. J. Moore (Academic, San Diego), pp. 387–424.

Darwin, C. J., and Hukin, R. W. (**1997**). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," J. Acoust. Soc. Am. **102**, 2316–2324.

Darwin, C. J., and Hukin, R. W. (**1998**). "Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony," J. Acoust. Soc. Am. **103**, 1080–1084.

Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: The role of interaural time differences," J. Exp. Psychol. Hum. Percept. Perform. **25**, 617–629.

Darwin, C. J., and Hukin, R. W. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of

fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.
Desimone, R., and Duncan, J. (**1995**). "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222.
Deutsch, D. (**1999**). "Grouping mechanisms in music," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch (Academic, San Diego).
Dorman, M., Loizou, P., 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.
Durlach, N. I. (**1972**). "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, edited by J. Tobias (Academic, New York), pp. 369–463.
Edmonds, B., and Culling, J. (**2005a**). "The role of head-related time and level cues in the unmasking of speech in noise and competing speech," Acta Acust. **91**, 546–553.
Edmonds, B. A., and Culling, J. F. (**2005b**). "The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay," J. Acoust. Soc. Am. **117**, 3069–3078.
Edmonds, B. A., and Culling, J. F. (**2006**). "The spatial unmasking of speech: Evidence for better-ear listening," J. Acoust. Soc. Am. **120**, 1539–1545.
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.
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., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**, 1040–1046.
Gallun, F., Mason, C., and Kidd, G. J. (**2005**). "Binaural release from informational masking in a speech identification task," J. Acoust. Soc. Am. **118**, 1614–1625.
Greenberg, S. (**1996**). "Understanding speech understanding: Towards a unified theory of speech perception," in *Proceedings of the ESCA Workshop on the "Auditory basis of speech perception,"* Keele University, Keele, England, pp. 1–8.
Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.
Hirsh, I. J. (**1948**). "The influence of interaural phase on interaural summation and inhibition," J. Acoust. Soc. Am. **20**, 536–544.
Ihlefeld, A., and Shinn-Cunningham, B. G.. "Spatial release from energetic and informational masking in a divided speech identification task." J. Acoust. Soc. Am. **123**, 4380–4392.
Kidd, G. J., Arbogast, T., Mason, C., and Gallun, F. (**2005a**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.
Kidd, G. J., Mason, C., and Gallun, F. (**2005b**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**, 982–992.
Litovsky, R. Y., and Shinn-Cunningham, B. G. (**2001**). "Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression," J. Acoust. Soc. Am. **109**, 346–358.
Lutfi, R. A. (**1993**). "A model of auditory pattern analysis based on component-relative entropy," J. Acoust. Soc. Am. **94**, 748–758.
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.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, and Larson, E. (**2005a**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. **91**, 967–979.
Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005b**). "Localizing nearby sound sources in a classroom: Binaural room impulse resonses," J. Acoust. Soc. Am. **117**, 3100–3115.
Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.
Wichmann, F., and Hill, N. (**2001a**). "The psychometric function: I. Fitting, sampling and goodness-of-fit," Percept. Psychophys. **63**, 1293–1313.
Wichmann, F., and Hill, N. (**2001b**). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," Percept. Psychophys. **63**, 1314–1329.
Wightman, F. L., and Kistler, D. J. (**2005**). "Informational masking of speech in children: Effects of ipsilateral and contralateral distractors," J. Acoust. Soc. Am. **118**, 3164–3176.
Wightman, F. L., Kistler, D. J., and Brungart, D. (**2006**). "Informational masking of speech in children: Auditory-visual integration," J. Acoust. Soc. Am. **119**, 3940–3949.
Woods, W. S., and Colburn, H. S. (**1992**). "Test of a model of auditory object formation using intensity and interaural time difference discrimination," J. Acoust. Soc. Am. **91**, 2894–2902.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (College-Hill Press, Boston, MA).

Spatial release from energetic and informational masking in a divided speech identification task[a)]

Antje Ihlefeld and Barbara Shinn-Cunningham[b)]
Auditory Neuroscience Laboratory, Boston University Hearing Research Center, 677 Beacon St., Boston, Massachusetts 02215, USA

(Received 3 January 2007; revised 12 March 2008; accepted 13 March 2008)

When listening selectively to one talker in a two-talker environment, performance generally improves with spatial separation of the sources. The current study explores the role of spatial separation in divided listening, when listeners reported both of two simultaneous messages processed to have little spectral overlap (limiting "energetic masking" between the messages). One message was presented at a fixed level, while the other message level varied from equal to 40 dB less than that of the fixed-level message. Results demonstrate that spatial separation of the competing messages improved divided-listening performance. Most errors occurred because listeners failed to report the content of the less-intense talker. Moreover, performance generally improved as the broadband energy ratio of the variable-level to the fixed-level talker increased. The error patterns suggest that spatial separation improves the intelligibility of the less-intense talker by improving the ability to (1) hear portions of the signal that would otherwise be masked, (2) segregate the two talkers properly into separate perceptual streams, and (3) selectively focus attention on the less-intense talker. Spatial configuration did not noticeably affect the ability to report the more-intense talker, suggesting that it was processed differently than the less-intense talker, which was actively attended. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2904825]

PACS number(s): 43.66.Dc, 43.66.Pn, 43.66.Qp [RLF] Pages: 4380–4392

I. INTRODUCTION

Previous studies on selective listening show that listeners are generally good at retrieving information from a source at a location they are attending, but perform poorly when asked to recall messages from unexpected locations (Cherry, 1953; Yost, 1997; Arbogast and Kidd, 2000). Nonetheless, in a recent study, listeners did remarkably well when asked to report two simultaneous messages, and overall performance was only weakly influenced by the amount of spatial separation between two concurrent speech sources (Best *et al.*, 2006). Other studies confirm that while listeners typically can attend to only one auditory message at a time (Cherry, 1953; Broadbent, 1954), they have some capacity to process semantic information from messages outside the immediate focus of attention (e.g., see Moray, 1959; Lawson, 1966; Cowan, 1995; Conway *et al.*, 2001; Rivenez *et al.*, 2006).

Previous work indicates that when trying to understand several sources at the same time, listeners may actively attend to one during the presentation of the stimulus and then selectively read out information about other source(s) from temporary buffers, after the stimulus ended (Conway *et al.*, 2001; Best *et al.*, 2006). This suggests that when asked to report two concurrent sources, listeners may exploit spatial cues to selectively attend to one source during the presentation and then process the other source from memory.

It is not clear whether selective spatial attention operates on a buffered representation of the other source like it does on an ongoing stimulus. If the listener cannot access presegregated objects in the buffered representation, competition between sources may play out differently for a recalled message than for a message that is selectively attended during the stimulus presentation. Therefore, it is difficult to predict how the spatial configuration of the sources will influence the ability to extract information about a recalled message. This paper attempts to disentangle how the location of the attended message in a two-talker setting influences the ability to extract information from two simultaneously presented messages.

Studies of selective listening identify numerous forms of interference that can limit performance in speech identification tasks (cf. Brungart, 2001; Brungart *et al.*, 2005; Freyman *et al.*, 2005; Kidd *et al.*, 2005; Ihlefeld and Shinn-Cunningham, 2008). Energetic masking occurs when the masker interferes with the peripheral representation of the target (Cherry, 1953; Spieth *et al.*, 1953; Moray, 1959; Ebata *et al.*, 1968). Informational masking occurs either because listeners (1) cannot segregate the target from the masker, and/or (2) cannot select the target out of a mixture of similar, properly segregated maskers, possibly because they are uncertain as to which sound features constitute the target (Arbogast *et al.*, 2002; Brungart and Simpson, 2002; Durlach *et al.*, 2003; Lutfi *et al.*, 2003; Brungart *et al.*, 2005; Watson, 2005; Best *et al.*, 2005; Shinn-Cunningham *et al.*, 2005).

[a)]Portions of this work were presented at the 2005 Mid-Winter meeting of the Association for Research in Otolaryngology.
[b)]Author to whom correspondence should be addressed. Electronic mail: shinn@cns.bu.edu.

 0001-4966/2008/123(6)/4380/13/$23.00

Both energetic masking and informational masking are likely to affect performance when listeners try to report multiple simultaneous target messages (see also Best *et al.*, 2006). In selective listening, the spatial separation of competing sources influences performance by improving the audibility of the target (reducing energetic masking, e.g., see Zurek, 1993), improving perceptual segregation of the sources (reducing one form of informational masking, e.g., see Freyman *et al.*, 1999), and decreasing confusions between target and masker (reducing the other form of informational masking, e.g., see Brungart, 2001).

The current study has two aims. The first aim is to gain a better understanding of how energetic masking and informational masking interfere with the ability to report the less-intense target when listeners are asked to report two simultaneous targets. The second aim is to examine the roles of spatial factors on performance in a divided listening task, and determine how they change as the relative contributions of energetic masking and informational masking vary.

Listeners were asked to report the content of two concurrent utterances. The spatial separation between the talkers was varied from block to block, and the broadband energy ratio between the talkers was varied within each block to systematically change the relative contributions of energetic masking versus informational masking (see companion paper about a selective attention experiment with identical stimuli; Ihlefeld and Shinn-Cunningham, 2008). To the extent that divided listening consists of first selectively attending to one message and then reporting the other message, spatial separation should improve the ability to report the actively attended message through a combination of acoustic effects at the better ear for the attended message, binaural processing, and spatial release from informational masking through spatially directed attention (e.g., see Best *et al.*, 2006; Ihlefeld and Shinn-Cunningham, 2008). However, spatial separation is also likely to influence the ability to process the other message, either negatively (because listeners attend the location of the initially processed target message, which impairs performance for sources from other locations), or positively (because the two targets are perceptually more distinct). Evidence for both effects was found by Best and colleagues (Best *et al.*, 2006). In the current study, *post-hoc* analysis of response patterns supports the idea that listeners actively attended to the less-intense target and recalled the more-intense target through a different mechanism. We find that spatial separation of the concurrent messages (1) improved the ability to report the actively attended source in ways comparable to improvements in selective listening (Ihlefeld and Shinn-Cunningham, 2008), but (2) neither helped nor hindered the ability to report the other message.

II. METHODS

The methods used in this study are essentially identical to the methods used in the companion paper which tested selective attention (Ihlefeld and Shinn-Cunningham, 2008). The same subjects participated in both studies. Stimuli and procedures were identical, except for the instructions (to report one of the two sources in the selective task described in the companion paper, or to report both messages in the current divided task). Methods are summarized briefly here (see Ihlefeld and Shinn-Cunningham, 2008, for more details).

A. Subjects

Four subjects (ages 21–24) were paid for their participation in the experiments. All subjects were native speakers of American English and had normal hearing, confirmed by an audiometric screening. All subjects gave written informed consent (as approved by the Boston University Charles River Campus Institutional Review Board) before participating in the study.

B. Stimuli

Raw speech stimuli were taken from the Coordinate Response Measure corpus (CRM, see Bolia *et al.*, 2000). Sentences were processed to produce intelligible, spectrally sparse speech signals (e.g., see Shannon *et al.*, 1995; Dorman *et al.*, 1997; Arbogast *et al.*, 2002; Brungart *et al.*, 2005). Each target and masker source signal was bandpass filtered into 16 logarithmically spaced, adjacent frequency bands of 1/3 octave width (center frequencies 175 Hz–5.6 kHz). The envelope of each band was extracted using the Hilbert transform. Subsequently, each envelope was multiplied by a pure-tone carrier at the center frequency of that band.

On each individual trial, eight of the 16 bands were chosen randomly (four from the lower eight frequency bands and four from the upper eight frequency bands) to create the raw waveform for one source. The remaining eight bands were used to construct the other source using otherwise identical processing. The raw source waveforms were scaled to have the same fixed, broadband root mean square (RMS) energy reference level prior to spatial processing (described below).

C. Spatial synthesis

Raw signals were processed with head-related transfer functions (HRTFs) of an acoustic manikin to simulate sources from 0° (in front) or 90° (to the side) azimuth, at a distance of 1 m in the horizontal plane (see Ihlefeld and Shinn-Cunningham, 2008, for details).

D. Procedures

One talker, referred to as the fixed-level talker (target_F) was always presented at the same reference RMS level (set to approximately 70 dB sound pressure level prior to spatial processing). The level of the other, variable-level talker (target_V) was attenuated relative to target_F by an amount that varied randomly from trial to trial, chosen from one of five levels (−40, −30, −20, −10, and 0 dB). Subsequently, the binaural signals for the two talkers were summed to produce the two-talker stimulus. As a result of this manipulation of target_V, the nominal energy ratio between the two talkers varied (without taking into account spatial processing ef-

fects). The broadband energy ratio between $target_V$ and $target_F$ will be denoted by T_VT_FR. In this study, T_VT_FR ranged from −40 dB to 0 dB.

There were four possible spatial configurations, two in which the two talkers were co-located (at either 0 or 90°) and two in which the talkers were spatially separated ($target_V$ at 0° and $target_F$ at 90°, or $target_V$ at 90° and $target_F$ at 0°). In each run, the spatial configuration of the two talkers was fixed (i.e., the talkers were played from the same location throughout the run).

Stimuli were digital-to-analog converted, amplified using Tucker-Davis System 3 hardware, and presented over Sennheiser HD 580 headphones to subjects seated in a sound-treated booth. Following each trial, subjects indicated perceived target keywords using a graphical user interface (GUI), after which the GUI indicated the correct response.

At the start of each session, a random call sign was selected to serve as the call sign of $target_V$, matching the procedures used in the companion study of selective attention (Ihlefeld and Shinn-Cunningham, 2008). In contrast, the call sign of $target_F$ varied randomly from trial to trial. $Target_V$ and $targe_F$ always had different call signs. Listeners were instructed to report the colors and numbers of both talkers. They were not explicitly instructed to report these keywords in proper pairs corresponding to the two physical sources, nor were they made aware of the fact that $target_V$ had a fixed call sign throughout the session. A trial was scored as correct and subjects were given feedback to that effect only if they reported all of the four keywords in any order.

Prior to testing, all subjects went through an initial screening in which they had to report the color and number of one talker of processed speech presented in quiet (processed by a 0° azimuth HRTF). In order to proceed with the experiment, they had to achieve at least 90% correct over the course of 50 such trials. None of the subjects failed this initial screening. Following the screening, each subject performed a training session consisting of 300 trials (at least one run of 50 trials for each spatial configuration, and an additional run of 50 trials for each of two randomly picked spatial configurations).

Following training, subjects performed four sessions of the experiment (one session per day). In the other four sessions, subjects performed a selective-attention task (reported in Ihlefeld and Shinn-Cunningham, 2008). Each session consisted of 12 runs (three runs for each of the four spatial configurations) of either the selective or the divided task. The order of the sessions and the order of the runs within each session were separately randomized for each subject, but constrained so that each spatial configuration and each of the two tasks was performed once before any were repeated. A run consisted of eight repetitions of each of the five T_VT_FR s (40 trials per run). The orders of the runs within each session were separately randomized for each subject. Given that each subject performed four sessions of this experiment, each subject performed 96 repetitions of each specific configuration (8 repetitions/run × 3 runs/session × 4 sessions).

E. Hypotheses

In the current task, the listener was asked to report the content of both of two simultaneous messages. On each trial, subjects responded by first reporting one color-number pair and then reporting a second color-number pair. The color and number from $target_V$ will be denoted by C_V and N_V, respectively. Similarly, the color and number from $target_F$ will be denoted by C_F and N_F. Color and number responses that are not keywords in either message will be signified by C_X and N_X, respectively. The order and pairing in which keywords were reported was not important for the score that listeners received. Specifically, for stimulus $[C_VN_VC_FN_F]$, the following four responses were scored as correct: $[C_VN_VC_FN_F]$, $[C_FN_FC_VN_V]$, $[C_VN_FC_FN_V]$, and $[C_FN_VC_VN_F]$, where order of report corresponds to the pair order within the brackets.

The ability to correctly report what both talkers were saying depends on whether the listener can hear and segregate the target words. In addition, listeners need to divide their processing resources between the two competing talkers. As in selective listening tasks (Brungart, 2001; Kidd *et al.*, 2005; Ihlefeld and Shinn-Cunningham, 2008), analyzing response errors made in divided listening tasks may illuminate the underlying response strategies that listeners use. Several factors can contribute to a failure to hear a target. Importantly, in the current experiment, listeners may not hear a target message because (1) it was energetically masked by the other source (energetic masking), or because (2) listeners failed to hear out and remember that target, even though it was well above the threshold of audibility (informational masking).

The relative influence of energetic masking compared to informational masking is likely to depend on the energy ratio between the two talkers (Ihlefeld and Shinn-Cunningham, 2008). If listeners truly selectively attend to one target message and then recall the other message, then the pattern of errors should depend on the kind of interference present for the attended target (Brungart *et al.*, 2001), while the ability to report the recalled target will depend on how well it is represented in memory. When $target_V$ is at least 20 dB less intense than $target_F$, $target_V$ may be difficult to hear (energetic masking; Ihlefeld and Shinn-Cunningham, 2008). In such trials, $target_F$ should have a clean representation both in the direct sensory input and in any temporary buffer and should therefore be easy to recall, regardless of the spatial configuration of the talkers. In such conditions, the intelligibility of the less-intense talker should be the main factor limiting divided-attention performance. Thus, performance should improve as the relative level of the less-intense talker increases (much as performance in selective listening improves with increasing target-to-masker ratio; e.g., see Arbogast *et al.*, 2002; Shinn-Cunningham *et al.*, 2005). When the two competing talkers are spatially separated, the overall energy ratio of the less-intense talker relative to the more-intense talker will improve at one ear. Furthermore, binaural cues will increase the audibility of the less-intense talker by a modest amount when it is near detection threshold (Zurek,

1993). Therefore, to the extent that the less-intense target determines performance, divided performance should improve with spatial separation.

If attention can be focused on only one location at a time, increasing spatial separation between the two concurrent messages may also increase the number of *drop* errors for the recalled message (e.g., responding $[C_V N_V C_X N_X]$ if target_V was attended, or $[C_F N_F C_X N_X]$ if target_F was attended, where C_x and N_x denote a color and number not present in either utterance; see Best *et al.*, 2005). Note that while in selective listening a failure to hear the single target can cause listeners to erroneously report the content of the masker message, in the current divided task, listeners will end up guessing the content of the source they tried to attend while still reporting the message of the other target that they heard. Here, in the majority of trials, target_F is relatively intense and salient, whereas target_V is usually much harder to hear than target_F. If listeners therefore attend to target_V at its location, spatial separation may increase the number of drop errors for target_F.

Finally, while listeners were not explicitly instructed to report the keywords of both talkers in proper pairs, they may have a natural tendency to do so. It should be difficult for listeners to report both messages without confusions when the talkers are similar in level and have the same perceived location. Specifically, when both targets are clearly audible but perceptually similar, listeners may have difficulty segregating the talkers; or they may be able to segregate the words and recall keywords from both messages, but may confuse which talker spoke which words (informational masking). Although there was no penalty for responding this way, listeners reporting a mix of target_F and target_V keywords (i.e., $[C_V N_F C_F N_V]$ or $[C_F N_V C_V N_F]$), henceforth *mix responses*, may reflect less complete perceptual segregation and streaming of the two sources compared to trials in which they report the keywords in proper pairs (i.e., $[C_V N_V C_F N_F]$ or $[C_F N_F C_V N_V]$), which will be called *fully correct* responses. Any systematic patterns in the relative likelihood of mix versus fully correct response likely reflect differences in the degree of perceptual segregation of the target and the masker.

III. RESULTS

Section III A analyzes the probability of reporting all four keywords correctly, independent of their pairing and response ordering. More detailed analysis of the kinds of response errors and order of responses are given in subsequent sections.

A. Percent correct

On each trial, subjects responded by reporting two color—number pairs. After each trial, subjects received feedback that they were correct if and only if they reported all four keywords of both utterances, regardless of how they paired keywords from the talkers. Therefore, the likelihood of responding correctly by chance equals $4\times1/4\times1/7\times1/3\times1/6$ or 0.8%. However, if subjects heard target_F but could not hear and had to randomly guess the keywords for target_V, the likelihood of being correct by chance would equal $1/3\times1/6$ or 6%.

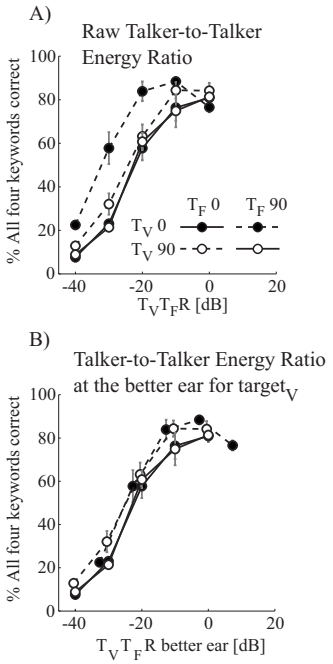


FIG. 1. Percent correct as a function of energy ratio between target_V and target_F ($T_V T_F R$). Error bars show the across-subject standard error of the mean. In general, performance improves with $T_V T_F R$, and is better for spatially separated than co-located sources. Filled symbols show results for target_V at 0° and open symbols show results for target_V at 90°. Results for spatially separated targets are shown with dashed lines. Results for co-located sources are shown with solid lines. (A) Results plotted as a function of the broadband target to target broadband energy ratio ($T_V T_F R$). (B) The same results re-plotted as a function of $T_V T_F R_{be\text{-}V}$ (correcting for differences in the acoustic $T_V T_F R$ at the better ear for target_V).

1. Overall percent correct

Figure 1(a) shows percent correct as a function of $T_V T_F R$, averaged across subjects (error bars show the across-subject standard error). All subjects showed relatively similar results, so only the across-subject average results are shown. As the intensity of target_V increased, performance improved. Performance was better in the spatially separated configurations than in the co-located configurations (dashed lines fall above solid lines). At the lowest $T_V T_F R$, performance was near 6% for the co-located configurations, 18% for T_V at 90° and T_F at 0°, and 22% for T_V at 0° and T_F at 90°. For all spatial configurations, performance improved with increasing intensity of target_V until it reached an upper bound of roughly 70%.

Performance was essentially identical for co-located sources whether they played from in front or from the side of the listener. For the spatially separated configurations, performance was better when target_V was playing from in front and target_F from the side of the listener than when their positions were reversed. As shown in the companion paper (Ihlefeld and Shinn-Cunningham, 2008; see also Shinn-Cunningham *et al.*, 2005), differences in the broadband acoustic target-to-masker energy ratio at the better acoustic ear accounted for differences in selective listening perfor-

mance for different spatially separated spatial configurations. The RMS energy of the two messages was equated prior to spatial processing; the level of $target_V$ was then adjusted to produce the desired T_VT_FR. However, spatial processing also altered the levels of the talkers at each ear. When the two talkers were spatially separated, there was always one ear (the acoustically better ear for $target_V$) that received a higher broadband T_VT_FR than the other ear. When $target_V$ was in front of the listener and $target_F$ was to the right side of the listener, the T_VT_FR at the left ear was on average 7 dB greater than the nominal T_VT_FR prior to spatial processing for the stimuli used in this study (see analysis in Shinn-Cunningham *et al.*, 2005). Conversely, when $target_F$ came from in front and $target_V$ was to the right of the listener, the right ear was the better ear for $target_V$, with a T_VT_FR that was on average 1 dB lower than the nominal T_VT_FR prior to spatial processing. Note that in the co-located configurations, T_VT_FR at both ears equaled the nominal T_VT_FR (on average).

Figure 1(b) shows the data from Fig. 1(a) re-plotted as a function of the T_VT_FR at the better ear for $target_V$ ($T_VT_FR_{be\text{-}V}$) by shifting the raw data horizontally by the appropriate dB amount for each spatial configuration. This adjustment completely accounts for performance differences between the two spatially separated configurations [dashed lines in Fig. 1(b) are virtually identical], just as in the companion study of selective listening (Ihlefeld and Shinn-Cunningham, 2008).

2. Spatial gains

For each subject and spatial configuration, percent correct performance as a function of $T_VT_FR_{be\text{-}V}$ was fitted by logistic functions (see Appendix B). For each individual subject, the psychometric function fits for the two co-located configurations were averaged, as were the psychometric function fits for the two spatially separated configurations (after accounting for the acoustic advantage at the better ear for $target_V$). Between −30 and −20 dB $T_VT_FRs_{be\text{-}V}$, the vertical difference of these averaged spatially separated and co-located psychometric function fits (the percent spatial gain) was 6% for subjects S1 and S3, 10% for subject S2, and 13% for subject S4. At the greatest $T_VT_FRs_{be\text{-}V}$ the percent spatial gain, equal to the difference in upper bounds, was between approximately 5% (subjects S1, S2, and S3) and 11% (S4). The horizontal shift between the logistic fit (the dB spatial gain) was approximately 2–4 dB (for all subjects).

3. Analysis of response pairing

In general, despite the fact that listeners were not instructed to report the messages in correct pairings, they tended to do so. Was there a positive effect of spatial separation on the likelihood of reporting keywords in pairings that correspond to the target messages? To examine this question, all trials where subjects responded correctly were analyzed in more detail. In the majority of the trials in which all four keywords were reported, they were reported in proper pairings (i.e., in 91% of all fully correct trials subjects either reported $[C_VN_VC_FN_F]$ or $[C_FN_FC_VN_V]$; fully correct).

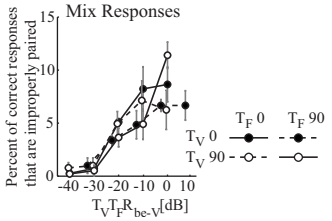


FIG. 2. Mix responses as a function of $T_VT_FR_{be\text{-}V}$ for each spatial configuration, averaged across subjects. Error bars show the across-subject standard error of the mean. Mix responses increase with increasing $T_VT_FR_{be\text{-}V}$. Filled symbols show results for $target_V$ at 0°, while open symbols show results for $target_V$ at 90°. Results for spatially separated sources are denoted with dashed lines and for co-located sources with solid lines.

The pattern of fully correct responses was very similar to the pattern for overall correct responses (shown in Fig. 1) and was not analyzed in more detail. However, analysis of the order in which the proper pairs were reported revealed interesting patterns, considered further in the Appendix A.

The less-common trials in which subjects responded with all four keywords correct, but in improper pairings (i.e., in which they reported $[C_VN_FC_FN_V]$ or $[C_FN_VC_VN_F]$) are denoted as mix responses (even though listeners were given feedback in that these responses counted as correct) to reflect the fact that subjects mixed the keywords from the two target streams in their responses.

Chance performance for mix responses was $2\times1/4\times1/7\times1/3\times1/6$ or 0.4%. Overall, the rate of mix responses (Fig. 2) was a small subset of the correct responses (shown in Fig. 1). Figure 2 shows the pattern of mix responses as a function of $T_VT_FR_{be\text{-}V}$. Mix responses increased with increasing $T_VT_FR_{be\text{-}V}$. In other words, the more similar the two targets became in level, the more likely listeners were to mix up keywords from the two sources. There are no clear differences in how often listeners made mix responses across different spatial configurations, except near 0 dB $T_VT_FR_{be\text{-}V}$, where slightly more mix responses occurred when sources were co-located compared to when they were separated (dashed lines fall below solid lines near 0 dB $T_VT_FR_{be\text{-}V}$). In other words, subjects were most likely to mix the streams when the sources were both at the same intensity (0 dB $T_VT_FR_{be\text{-}V}$) and at the same location in space. While this effect does not reach statistical significance in this study ($F(1,3)=7.741$, $p=0.069$), it is consistent with results from our companion selective listening study, which showed the greatest number of confusions between target and masker when the sources were co-located and at nearly the same level (Ihlefeld and Shinn-Cunningham, 2008).[1]

In selective listening, differences in both level and location can improve a listener's ability to selectively attend to the target source. To a lesser extent than in the selective listening task, these same factors reduced confusions between the competing talkers in this divided task. This is consistent with the idea that listeners first selectively attended to $target_V$ and then recalled $target_F$.

B. Spatial effects on reporting the second message

In Sec. III A, only those trials in which all four keywords were reported were analyzed. However, this analysis

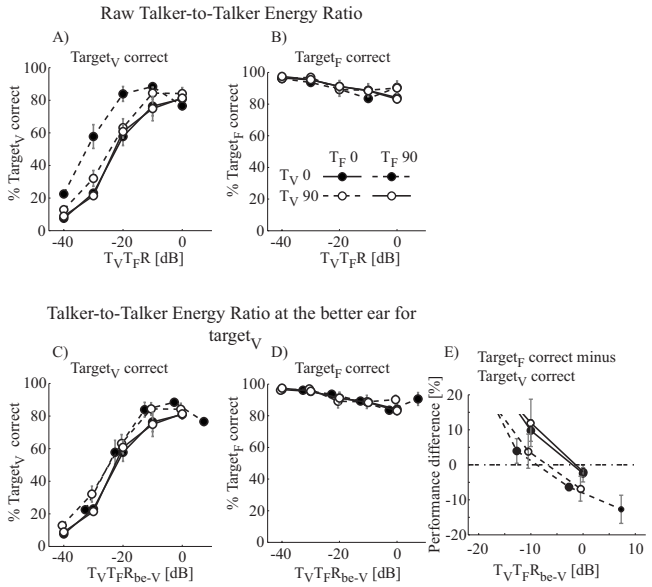


FIG. 3. Probability of reporting each target correctly in a proper pairing as a function of $T_VT_FR_{be\text{-}V}$, averaged across subjects. Error bars show the across-subject standard error of the mean. Spatial separation improves performance for target$_V$ but has no significant effect on target$_F$. Filled symbols show results for target$_V$ at 0° and open symbols for target$_V$ at 90°. Results for spatially separated targets are shown with dashed lines and for co-located sources with solid lines. (A) Results for target$_V$ correct as a function of T_VT_FR). (C) The same target$_V$ results re-plotted as a function of $T_VT_FR_{be\text{-}V}$ (correcting for differences in the acoustic T_VT_FR at the better ear for target$_V$). (B) Results for target$_F$ correct as a function of T_VT_FR. (D) Target$_F$ results re-plotted as a function of $T_VT_FR_{be\text{-}V}$. (E) Target$_F$ correct minus target$_V$ correct (difference between curves in panels D and C) when the two messages are similar in level (i.e., around $T_VT_FR_{be\text{-}V}$ near 0 dB).

ignored those trials in which part but not all of the response was correct. As discussed in the Introduction, we hypothesized that spatial separation negatively affects the ability to process both messages. In particular, if listeners attend to the location of the initially processed target message, it may impair performance for the second target when the second target comes from a different location than the initially processed target. To examine the effect of spatial separation on reporting the second message, here, all trials when listeners succeeded in reporting one of the two messages are analyzed separately for target$_V$ and target$_F$. When subjects reported both keywords of target$_V$ as a pair (i.e., either responded either [$C_VN_VC\_N\_$] or [$C\_N\_C_VN_V$], where "_" denotes a target$_F$ keyword or a keyword not present in either message), a trial was scored as target$_V$ correct. Analogously, a trial was scored as target$_F$ correct when the response contained the color and number of target$_F$ in one pair (i.e., subjects responded either [$C_FN_FC\_N\_$] or [$C\_N\_C_FN_F$]. Note that although it was not explicitly pointed out to them, in principle, listeners could differentiate between target$_V$ and target$_F$, because target$_V$ (the target that was usually softer and therefore harder to hear) had a call sign that was fixed throughout the course of the session. In contrast, the call sign of target$_F$ varied randomly from trial to trial (but never equaled the call sign of target$_V$; see also Sec. II D).

Figure 3 plots the percentage of trials in which a message was reported in correct pairing for target$_V$ [Fig. 3(a)] and target$_F$ [Fig. 3(b)] as a function of T_VT_FR and as a function of $T_VT_FR_{be\text{-}V}$ [target$_V$ and target$_F$ in Figs. 3(c) and 3(d), respectively]. As a function of T_VT_FR, performance for target$_V$ [Fig. 3(a)] was better when the two talkers are spatially separated than when they are co-located, and was best when target$_V$ is at 0° and target$_F$ is at 90°. When plotted as a function of $T_VT_FR_{be\text{-}V}$, results for target$_V$ were similar for the two spatially separated configurations [Fig. 3(c)]. Plotted this way, there was a small spatial gain of about 2–3 dB for both target$_V$ in front and target$_F$ to the side (dashed lines fall above solid lines). In contrast, performance for target$_F$ [Fig. 3(b)] did not depend strongly on spatial configuration (the four lines are overlapping). Normalization to take into account the better-ear ratio for target$_F$ [plotting as a function of $T_VT_FR_{be\text{-}V}$; Fig. 3(d)] had little effect on the psychometric curve describing the probability of correctly reporting target$_F$ and did not reveal any systematic effect of spatial configuration on performance.[2] However, ceiling effects may account for the lack of spatial effects on performance for target$_F$.

Finally, while for the majority of trials performance for the more-intense target$_F$ was better than for target$_V$, one might expect that the two talkers were equally hard to understand near 0 dB $T_VT_FR_{be\text{-}V}$, when they were near the same intensity. However, subjects performed better for target$_V$ than for target$_F$. Given the scales of panels 3(C) and 3(D), it is difficult to make direct comparisons and see this difference. To make this pattern clearer, for each spatial configuration, the differences between the percentages correct for target$_V$ correct and target$_F$ are shown in Fig. 3(e) for points near 0 dB $T_VT_FR_{be\text{-}V}$ (error bars show across-subject standard errors). For trials with $T_VT_FR_{be}$ near 0 dB, the differences in performance between target$_V$ and target$_F$ are consistently negative for all spatial configurations [i.e., performance is better for target$_V$ than for target$_F$; Fig. 3(e); coding of line style and symbols is the same as in the other panels].

Further analysis of report order in Appendix A shows that there are systematic differences in how listeners prioritize the two messages. These results suggest that both (1) when listeners selectively attended to target$_V$ and (2) when sources were spatially separated in this divided task, the source from in front of the listener was inherently more salient than the source to the side (see Appendix A).

In summary, spatial separation improved the ability to hear out target$_V$, but had no significant effect on performance for target$_F$. Moreover, subjects appeared to devote more processing resources to target$_V$, as evidenced by the fact that they were better at reporting target$_V$ than target$_F$ when both were equally intense (or even when target$_V$ is slightly less intense than target$_F$), even when the two messages were co-located. In other words, listeners appear to have attended to the location of the initially processed target message, but this did not impair performance for the second target when the second target came from a different location than the initially processed target.

C. Incorrect responses

Performance for all trials in which responses were not counted correct was analyzed in more detail to see if there

TABLE I. Definitions of and chance probabilities of different error types

Response type	Chance	Responses
$Target_V$ drop error	6.7%	$[C_FN_F\ C_XN_V]$, $[C_FN_F\ C_VN_X]$, $[C_FN_F\ C_XN_X]$, $[C_XN_V\ C_FN_F]$, $[C_VN_X\ C_FN_F]$, $[C_XN_X\ C_FN_F]$
$Target_F$ drop error	6.7%	$[C_VN_V\ C_XN_F]$, $[C_VN_V\ C_FN_X]$, $[C_VN_V\ C_XN_X]$, $[C_XN_F\ C_VN_V]$, $[C_FN_X\ C_VN_V]$, or $[C_XN_X\ C_VN_V]$
$Target_V$ combination error	6.7%	$[C_VN_X\ C_XN_V]$, $[C_VN_X\ C_FN_V]$, $[C_VN_F\ C_XN_V]$, $[C_XN_V\ C_VN_X]$, $[C_FN_V\ C_VN_X]$, or $[C_XN_V\ C_VN_F]$
$Target_F$ combination error	6.7%	$[C_FN_X\ C_XN_F]$, $[C_FN_X\ C_VN_F]$, $[C_FN_V\ C_XN_F]$, $[C_XN_F\ C_FN_X]$, $[C_VN_F\ C_FN_X]$, or $[C_XN_F\ C_FN_V]$
Mix response	0.4%	$[C_VN_F\ C_FN_V]$ or $[C_FN_V\ C_VN_F]$
Fully correct	0.4%	$[C_VN_V\ C_FN_F]$ or $[C_FN_F\ C_VN_V]$
Other	72.2%	All trials that were not fully correct, mix responses, drop errors, or combination errors were scored as other.

was any evidence for how listeners prioritized the messages. Table I shows the definitions of the different response types and the distributions of the different incorrect responses [Tables I and II, respectively; Note that the total sum of the percentages of incorrect responses, fully correct responses, and mix responses equals 100%].

Across all $T_VT_FR_{be\text{-}V}$, drop errors (where subjects reported both keywords of one target but failed to report both

TABLE II. Distribution of types of incorrect responses in percent as a function of $T_VT_FR_{be\text{-}V}$ averaged across subjects (standard error in parentheses). The top half of the table shows results when the targets are coming from the same location; the bottom half shows results when the targets are spatially separated.

T_VT_FR	−40 dB		−30 dB		−20 dB		−10 dB		0 dB	
					Co-located configurations					
Incorrect Response [%]	$T_V0°$ $T_F0°$	$T_V90°$ $T_F90°$	$T_V0°$ $T_F0°$	$T_V90°$ $T_F90°$	$T_V0°$ $T_F0°$	$T_V90°$ $T_F90°$	$T_V0°$ $T_F0°$	$T_V90°$ $T_F90°$	$T_V0°$ $T_F0°$	$T_V90°$ $T_F90°$
$Target_V$ drop error	88 (3)	87 (2)	72 (3)	74 (3)	35 (9)	34 (8)	16 (8)	19 (13)	10 (6)	12 (4)
$Target_F$ drop error	0 (1)	0 (1)	2 (1)	2 (1)	4 (2)	5 (4)	7 (2)	7 (2)	13 (2)	14 (3)
$Target_V$ combination error	1 (1)	2 (1)	3 (1)	3 (2)	4 (2)	3 (2)	5 (3)	3 (4)	3 (2)	3 (2)
$Target_F$ combination error	0 (0)	0 (0)	1 (1)	1 (2)	1 (1)	1 (1)	2 (1)	2 (1)	4 (2)	3 (2)
Other	3 (1)	2 (1)	2 (3)	2 (1)	3 (1)	3 (3)	2 (2)	3 (1)	3 (2)	2 (2)
					Spatially separated configurations					
	$T_V0°$ $T_F90°$	$T_V90°$ $T_F0°$	$T_V0°$ $T_F90°$	$T_V90°$ $T_F0°$	$T_V0°$ $T_F90°$	$T_V90°$ $T_F0°$	$T_V0°$ $T_F90°$	$T_V90°$ $T_F0°$	$T_V0°$ $T_F90°$	$T_V90°$ $T_F0°$
$Target_V$ drop error	74 (2)	83 (4)	36 (15)	63 (11)	11 (7)	30 (9)	8 (1)	12 (5)	7 (7)	7 (3)
$Target_F$ drop error	1 (0)	0 (1)	2 (1)	1 (0)	7 (2)	6 (6)	14 (2)	8 (8)	19 (4)	14 (6)
$Target_V$ combination error	2 (2)	2 (2)	3 (2)	3 (2)	4 (3)	4 (3)	2 (1)	2 (1)	2 (1)	2 (2)
$Target_F$ combination error	1 (1)	0 (0)	1 (1)	1 (1)	2 (2)	2 (2)	1 (1)	2 (2)	3 (2)	1 (1)
Other	2 (1)	2 (4)	3 (4)	1 (2)	1 (2)	3 (2)	2 (2)	2 (2)	1 (1)	1 (1)

of the keywords of the other target) were the dominant kind of response error. In all spatial configurations, the relative likelihood of $target_V$ drop errors decreased with increasing $T_VT_FR_{be\text{-}V}$, while $target_F$ drop errors increased. $Target_V$ drop errors were less common when the targets were spatially separated than when they were co-located ($F(1,3)=92.549$, $p=0.002$). In contrast, the percentage of $target_F$ drop errors did not vary significantly with the spatial configuration of the talkers ($F(1,3)=2.131$, $p=0.24$). Combination errors, which occur when listeners succeed in segregating one of the targets out of the acoustic mixture but fail to properly stream it across time, tended to increase with increasing $T_VT_FR_{be\text{-}V}$. However, while the relative number of $target_F$ combination errors increased monotonically with increasing $T_VT_FR_{be\text{-}V}$ ($F(4,12)=4.871$, $p=0.014$), the percentage of $target_V$ combination errors increased between −40 and −20 dB $T_VT_FR_{be\text{-}V}$, and then either decreased or remained constant as the two targets became more similar in level (no significant effect of $T_VT_FR_{be\text{-}V}$, $F(4,12)=7.921$, $p=0.341$). Although this is not a strong trend, it was consistent across all spatial configurations. This result hints that level cues influence the segregation and streaming of $target_V$ more than $target_F$. Other errors are very uncommon and do not depend consistently on $T_VT_FR_{be\text{-}V}$ or spatial configuration.

IV. DISCUSSION

A previous divided-listening study found that spatial separation between concurrent messages improves performance slightly, but that the dominant benefit of spatial separation was from purely acoustic effects (Best *et al.*, 2006). However, that study presented two messages of equal intensity, making it difficult to assess the full impact of other spatial factors. That study also found evidence for two opposing effects of spatial separation in divided listening: spatial separation leads to an improvement in perceptual segregation of the concurrent sources, but a degradation in the ability to process both of the two simultaneous sources. In one experiment in that study, the two competing sources were equated such that they were equally intelligible in a selective listening task, but listeners were instructed to report the target message that was relatively more to the left before the target message that was relatively more to the right. These instructions caused listeners to devote more attentional resources to the left source that they had to report first. As a result, listeners made more errors for the lower-priority, right source. Moreover, the effects of spatial separation on the two sources differed. Best *et al.* concluded that listeners actively attended the higher-priority, left source and then recalled the lower-priority, right source, and that spatial separation had very different effects on the ability to report the two sources.

The current results support and extend these findings. Here, intelligibility of two spectrally degraded competing targets was investigated as a function of their broadband energy ratio for different spatial configurations. In this divided task, the ability to understand the less-intense talker dominated the pattern of performance, and performance improved as the energy ratio of the less-intense talker to the more-intense talker increased at the ear that had the more favorable energy ratio for the less-intense talker. Although listeners in the current study were not explicitly instructed as to which source to give higher priority, results suggest that listeners actively attended to the less-intense talker, which was usually harder to hear. As in the study by Best *et al.* (2006), this prioritization caused different effects of spatial separation on the lower- and higher-priority messages. Specifically, we found that spatial separation of the messages improved the ability to report the higher-priority message, but had little effect on the lower-priority message.

In the visual literature, three main models of spatial attention have been proposed. When extended to auditory tasks, these models predict that spatial separation will impair performance in a divided-listening task (cf. McMains and Somers, 2005). In the "zoom lens" model, the tuning of a single, spatial attentional filter widens in order to encompass spatially dispersed targets of interest, causing a trade-off between response accuracy and the size of the attentional field. The "multiple spotlights" model proposes simultaneous sampling of the auditory scene at several target locations, predicting a trade-off between processing efficiency and the total spatial extent of the attended regions. The "rapidly moving spotlight" model assumes that a single spotlight switches between spatially separated talkers, predicting that performance should degrade with increasing spatial separation of the targets.

The current results show that performance was better when the targets are spatially separated compared to when they are co-located, suggesting that these models of visual spatial attention cannot readily be applied to the current auditory divided-attention task. Of course, there are a number of differences in the demands of our auditory task and those of the visual tasks that typically are used to test models of dividing visual attention. For instance, by their very nature, auditory messages evolve over time, requiring listeners to sustain attention on a target message in order to analyze it and extract its meaning. The need to sustain attention on a message over time may make a strategy in which listeners switch attention between targets ineffective. Instead, the current results are consistent with listeners prioritizing the two targets differently, and processing them through different mechanisms.

While performance for the keywords of $target_F$ was essentially unaffected by the spatial configuration of the concurrent sources, performance for the actively attended $target_V$ was better in the spatially separated than in the co-located configurations. Most of the effects of spatial separation on performance for $target_V$ are consistent with the effects of spatial separation in selective listening (Arbogast *et al.* 2002; Ihlefeld and Shinn-Cunningham, submitted). However, no such effects occurred for $target_F$ (e.g., there was no reduction of $target_F$ drop errors when sources were spatially separated). Moreover, even at 0 dB $T_VT_FR_{be\text{-}V}$ where the two talkers should have been equally salient (albeit somewhat difficult to keep segregated), performance was slightly better for $target_V$ than for $target_F$ [cf. Fig. 3(e)]. We infer that listeners actively attended to $target_V$, and did so in

part based on its fixed call sign (which was the only cue distinguishing $target_V$ from $target_F$ when sources were co-located and at the same level).

Current results show that as in selective listening, spatially separating the two targets improved the intelligibility of the actively attended message ($target_V$), presumably through some combination of acoustic improvements at the better ear for $target_V$, binaural processing benefits that improved the audibility of $target_V$ (e.g., see Zurek, 1993) and spatial attention benefits that allowed listeners to selectively attend to $target_V$ by directing attention to its location. In this task, where the ability to report $target_V$ determined overall performance, a strategy of actively attending to $target_V$ may have been near optimal, at least if listeners could not actively attend to both messages simultaneously. After the better-ear advantage for $target_V$ was taken into account, the dominant remaining spatial effect (ignoring report order; see Appendix) was that $target_V$ drop errors were less common for spatially separated than for co-located sources.

In contrast, with the exception of performance at 0 dB $T_VT_FR_{be\text{-}V}$, mix responses and combination errors did not vary with spatial separation for either $target_V$ or $target_F$. When both talkers were relatively easy to hear, spatial separation did not influence the ability to segregate the competing messages, except when spatial cues were the sole reliable feature for differentiating the two talkers. Informal listening suggested that for −20 dB T_VT_FR and greater, two distinct sources could be heard. However, we did not measure whether listeners heard the two target messages from two distinct locations. Therefore it is difficult to assess the extent to which listeners used spatial attention to perform the current task.

In order to perform this task, listeners needed to properly identify the two messages; it was not necessary to link each keyword to the proper source in order to have a trial scored as correct. However, percent correct performance in this divided listening task was nearly as good as performance in the companion selective listening task in which listeners were asked to report only one of the two messages (Ihlefeld and Shinn-Cunningham, 2008). This suggests that listeners were indeed able to link the keywords to distinct sources, but further studies are needed to gain a better understanding of how the ability to identify keywords and the ability to correctly pair a message with its source influence divided listening.

The relatively high incidence of drop errors at high $T_VT_FR_{be\text{-}V}$ suggests that the ability to track two simultaneous talkers was limited. However, overall performance in the divided task was surprisingly high compared to performance in many previous studies. Many researchers (Cherry, 1953; Broadbent, 1954; Moray, 1959; Treisman and Geffen, 1967; for a review see Stifelman, 1994) suggest that listeners are limited in their ability to report two or more simultaneous messages. For instance, although listeners can recall basic properties of a channel that is not actively attended (such as the sex of the talker), most of the target words from that channel cannot be reported correctly (e.g., Cherry, 1953; Treisman and Geffen, 1967). However, these previous studies investigated identification tasks with a relatively high processing load, such as asking listeners to shadow sustained messages (i.e., "Repeat what you hear in the right ear"). In a study that examined a detection task with a lighter processing load (using tones instead of word targets), listeners could detect targets equally well in attended and rejected channels (Lawson, 1966). The processing and memory load required for the highly predictable, relatively short CRM messages used in the current task may have been low enough that listeners could process and/or temporarily store the contents of both of the two simultaneous utterances.

At 0 dB $T_VT_FR_{be\text{-}V}$, subjects performed better for the keywords from $target_V$ than for the keywords from $target_F$, even though both talkers were equally intense and should have been equally intelligible. This suggests that listeners assigned higher processing priority to $target_V$. At least one previous study shows that the order of responses in a divided attention task reflects the priority that listeners give each target (Bonnel and Hafter, 1998). Examination of response order in Appendix A shows that on those trials where subjects reported all four keywords correctly, as $T_VT_FR_{be\text{-}V}$ increased subjects were increasingly likely to report keywords from $target_V$ first. In contrast, the percentage of responses in which listeners reported one target keyword from the variable-level talker and guessed at least one other word did not change systematically with $T_VT_FR_{be\text{-}V}$. In other words, response order did not just depend on $T_VT_FR_{be\text{-}V}$, but depended on whether listeners got all keywords correct, i.e., how well they extracted each of the two messages on a particular trial. In addition, when talkers were spatially separated, the report order was biased towards reporting the message from in front of the listener before the message from the side.

Overall, these results support the idea that response order depended on the relative certainty that the listener had about the two messages, with the listeners first reporting the message about which they were most sure. The relative certainty of the messages appears to depend on both the relative saliency of the two targets as well as the amount of attention that the listener devoted to a target. In turn, the inherent salience of the messages depended on (1) the audibility of the messages, (2) the relative intensities of the messages, and (3) the spatial locations of the messages (where messages from in front were inherently more salient). In summary, the results support the idea that subjects gave higher priority (and selectively attended) to $target_V$. However, when listeners tried but failed to understand $target_V$, they resorted to reporting $target_F$ first, and then reporting their best-guess response for $target_V$.

Together, these results suggest that listeners used two different processing strategies in monitoring the two concurrent targets. Spatial separation improved the ability to understand keywords from $target_V$, presumably because listeners actively tried to attend to $target_V$ and were more successful in performing this selective attention task when $target_V$ came from a different location than $target_F$. In contrast, performance for $target_F$ showed little effect of spatial separation, consistent with the idea that $target_F$ was recalled from a temporary storage that was at best weakly affected by the spatial configuration of the sources or by spatially directed attention.

V. CONCLUSIONS

In this divided listening task with two concurrent target messages, performance improved as the ratio of the broadband energy of a less-intense talker to the energy of a simultaneous fixed-level talker increased. Overall, listeners were relatively good at reporting the fixed-level talker, which was generally easy to hear.

Results are consistent with listeners actively attending to the harder-to-hear source ($target_V$), and then recalling $target_F$.

Overall performance (the probability of reporting all four keywords) improved with increasing spatial separation.

- After taking into account better ear effects for the high-priority $target_V$, overall performance depended primarily on whether the sources were co-located or separated.
- Improvements with spatial separation of the competing messages came about primarily through spatial gains in performance for the less-intense, high-priority $target_V$. Effects of spatial configuration on the low-priority $target_F$ were negligible.

Listeners naturally tended to report messages in proper pairings, even though they were not instructed to do so. Spatial separation of sources reduced the likelihood of confusing the two messages and reporting the keywords in inconsistent pairings. However, this benefit was very small and was only observed near 0 dB $T_VT_FR_{be\text{-}V}$, where listeners had few other cues to segregate the mixture.

ACKNOWLEDGMENTS

This work was supported in part by ONR and AFOSR. We thank Virginia Best, Gerald Kidd, Steve Colburn, and Chris Mason, Richard Freyman and two anonymous reviewers for helpful comments on earlier versions of this manuscript.

APPENDIX A: REPORT ORDER

In a companion study, when asked to report only the keywords from $target_V$ (ignoring the message from $target_F$) subjects performed nearly as well as they did here, when asked to report both messages. Together with the current results, this finding suggests that listeners had little difficulty reporting the usually more-intense $target_F$ in addition to $target_V$, and that the ability to report $target_V$ was the main factor limiting performance. Therefore, both saliency (i.e., the inherent, bottom-up strength of $target_V$ relative to $target_F$) and attention (i.e., the listener's ability to select $target_V$ from the mixture) should have influenced how well listeners perform in this task. The order in which subjects naturally choose to report the target keywords can reflect how they prioritize each target (Bonnel and Hafter, 1998). Therefore, results were analyzed *post-hoc* to examine whether there was a consistent pattern in the order in which listeners chose to report the color-number pairs.

Figure 4(a) shows the percentage of trials in which the first color-number pair corresponded to the keywords from either one of the two messages (i.e., where subjects correctly

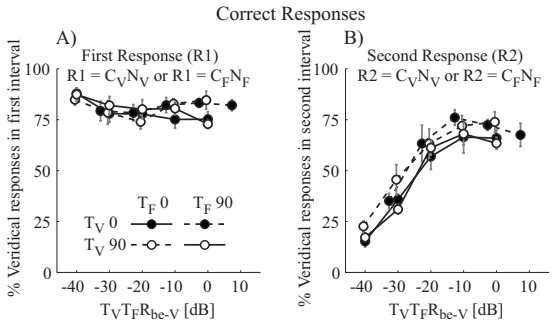


FIG. 4. Probability of reporting the first and second response pairs correctly in a proper pairing as a function of $T_VT_FR_{be\text{-}V}$, averaged across subjects. Error bars show the across-subject standard error of the mean. The probability of correct first responses is always greater than that of correct second responses. Subjects are more likely to respond without error in the second interval when the targets are spatially separated than when they are co-located. Filled symbols show results for $target_V$ at 0° and open symbols show results for $target_V$ at 90°. Results for spatially separated sources are shown with dashed lines and co-located sources are shown with solid lines. (A) First response interval. (B) Second response interval.

reported either $[C_VN_V]$ or $[C_FN_F]$ as the first pair, ignoring the second response, which could be correct or wrong), as a function of $T_VT_FR_{be\text{-}V}$. Figure 4(b) shows the corresponding probability of a correct color-number pair being reported in the second pair (i.e., either $[C_VN_V]$ or $[C_FN_F]$), ignoring responses in the first color-number pair).

Figure 4(a) shows that for all spatial configurations, subjects responded without error in the first interval in 80% or more of the trials. The likelihood that the first pair was correct was very similar for all spatial configurations. However, in the spatially co-located configurations, the percentage of those first-pair responses that were correct decreased slightly with increasing $T_VT_FR_{be\text{-}V}$ (consistent with subjects confusing the two target messages when they were both at the same level and from the same location), whereas this probability did not change with $T_VT_FR_{be\text{-}V}$ in the two spatially separated configurations.

Figure 4(b) shows that for all spatial configurations, the percentage of correct second-pair responses increased with increasing $T_VT_FR_{be\text{-}V}$. Moreover, subjects were more likely to respond without error in the second interval when the two talkers were spatially separated than when they are co-located [dashed lines are above solid lines in Fig. 4(b)].

Comparing results of Figs. 4(a) and 4(b), the probability of a correct first-pair response was much greater than the probability of a correct second-pair response for all conditions, indicating that subjects tended to respond first with a color-number pair that they were more sure was correct (though this was not the only criterion, as $target_V$ influenced the report order too; see analysis below).

All of the first-pair responses (both correct and incorrect) were broken down into six possible response types, depending on whether subjects reported both the color and number of $target_V$ ($[C_VN_V]$, a correct response on the first pair), both color and number of $target_F$ ($[C_FN_F]$; another form of correct response on the first pair), a mix of keywords from both targets ($[C_VN_F]$ or $[C_FN_V]$; a mix response), one keyword from $target_V$ and one word that was not from either talker ($[C_VN_X]$ or $[C_XN_V]$; a form of drop error), or two

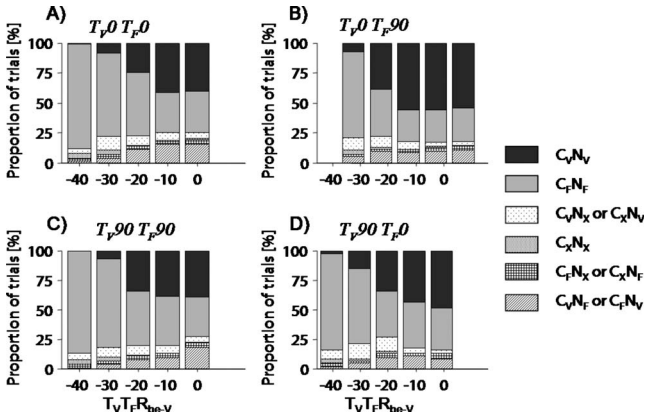


FIG. 5. Analysis of the first-pair responses as a function of $T_VT_FR_{be\text{-}V}$ for each spatial configuration, averaged across subjects. As $target_V$ becomes increasingly more audible, subjects become more likely to report it first. Different fill patterns denote different errors. $[C_VN_V]$ responses are solid black, $[C_FN_F]$ solid gray. $target_V$ guesses $([C_VN_X]$ and $[C_XN_V])$ are represented by sparsely dotted fill and completely random guesses $([C_XN_X])$ by densely dotted fill. $[C_FN_X]$ and $[C_XN_F]$ are denoted by square-grid fill. $[C_VN_F]$ and $[C_FN_V]$ are represented by rightward diagonal hatches. Each panel shows one spatial configuration. The left two panels (A, C) show results when the targets are coming from the same location. The right two panels (B, D) show results when the targets are spatially separated. The top row (A, B) shows results for the two configurations with $target_V$ in front of the listener. The bottom row (C, D) shows results when $target_V$ is to the side of the listener.

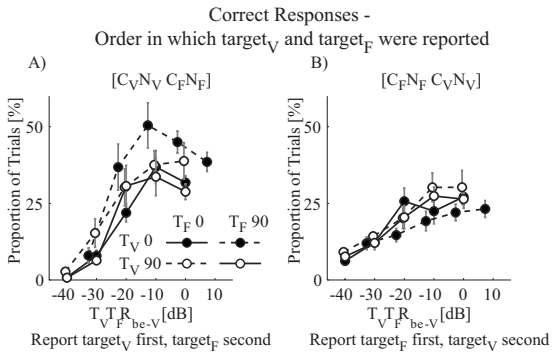


FIG. 6. Order in which listeners reported properly streamed keywords of $target_V$ and $target_F$ in the fully correct trials as a function of $T_VT_FR_{be\text{-}V}$. Error bars show the across-subject standard error of the mean. Results suggest that report order is a measure of the relative certainty the listener has about the content of the two messages. The absolute spatial configuration affected report order, suggesting that the source in front of the listeners was inherently more salient than source to the side of the listener. Filled symbols show results for the target at 0° and open symbols for the target at 90°. Results for spatially separated sources are shown with dashed lines and for co-located sources with solid lines. (A) First response is $target_V$ and second response is $target_F$. (B) First response is $target_F$, and second response is $target_V$.

words that were not part of either target ($[C_XN_X]$; another form of drop error). All first responses that did not fit any of these criteria were scored as other responses ($[C_FN_X]$ or $[C_XN_F]$), but such responses were rare. Note that the probabilities of responding ($[C_VN_V]$), ($[C_FN_F]$), ($[C_VN_X]$ or $[C_XN_V]$), $[C_XN_X]$, ($[C_VN_F]$ or $[C_FN_V]$), and ($[C_FN_X]$ or $[C_XN_F]$) sum to 1.0.

Figure 5 shows the distribution of the first responses as a function of $T_VT_FR_{be\text{-}V}$ for each spatial configuration. For $T_VT_FR_{be\text{-}V}$ of −20 dB and below, $[C_FN_F]$ was the dominant response type (gray solid fill). For $T_VT_FR_{be\text{-}V}$ greater than −20 dB, the most common first-pair response was $[C_VN_V]$ (black solid fill). This shows that as $target_V$ became louder and easier to hear, subjects became more and more likely to report it first. The proportion of trials in which subjects heard only part of $target_V$ (i.e., reported one keyword from $target_V$ and guessed the other word, $[C_VN_X]$ or $[C_XN_V]$, shown as sparsely dotted fill) was small and did not change systematically with $T_VT_FR_{be\text{-}V}$ (compare size of sparsely dotted-fill areas from left to right in each panel). This suggests that when listeners were not sure of the content of $target_V$, they tended to report it second, rather than first.

The percentage of trials in which listeners intermingled keywords from both talkers (reported $[C_VN_F]$ or $[C_FN_V]$) increased with increasing $T_VT_FR_{be\text{-}V}$ (see rightward diagonal hatch areas in Fig. 5), especially for the co-located spatial configurations (panels A and C). This increase in mix responses in the first responses was consistent with the overall pattern of mix responses (cf. Sec. III A 3). Completely random drop errors in the first response (reporting $[C_XN_X]$) only occurred at the lowest $T_VT_FR_{be\text{-}V}$ (densely dotted fill), and were very unlikely compared to the other responses. Similarly, other errors did not occur often and did not change consistently with either $T_VT_FR_{be\text{-}V}$ or spatial configuration (square-grid fill).

The ways in which subjects ordered and paired responses on the subset of trials when they were fully correct was analyzed to see how listeners naturally grouped the keywords, conditioned on them being fully correct. Figure 6(a) shows the percentage of correct trials in which subjects first reported $target_V$ and then $target_F$ ($[C_VN_VC_FN_F]$). Figure 6(b) shows the percentage of correct trials in which subjects first reported $target_F$ and then $target_V$ ($[C_FN_FC_VN_V]$). In both panels, performance is plotted as a function of $T_VT_FR_{be\text{-}V}$ for the four different spatial configurations.

For both report orders, the percentage of fully correct trials increased with increasing $T_VT_FR_{be\text{-}V}$. For $T_VT_FR_{be\text{-}V}$ less than −20 dB, subjects were more likely to report $target_F$ before $target_V$ [plotted percentages are higher in Fig. 6(b) than in Fig. 6(a)]. For $T_VT_FR_{be\text{-}V}$ of −20 dB and greater, subjects were most likely to report $target_V$ first [plotted percentages are higher in Fig. 6(a) than in Fig. 6(b)]. Interestingly, there were differences in these likelihoods that depended on the absolute locations of the talkers: Subjects were more likely to report keywords from $target_V$ first when $target_V$ was in front and $target_F$ was to the side than when $target_V$ was to the side and $target_F$ was in front [in Fig. 6(a), the dashed line with filled symbols is above the other lines], even after taking into account the talker energy ratios at the better ear for $target_V$ (or $target_F$, see[2]). This trend reverses for trials in which listeners first reported $target_F$ and then reported $target_V$. In those trials, the percentage of correct reports was greater when $target_F$ was from in front of the listener and $target_V$ was to the side than in the reverse configuration [in Fig. 6(b), the dashed line with filled symbols is below the other lines].

Overall, these results suggest that the listeners were actively attending to $target_V$, but tended to report the message

TABLE III. Mean parameters of the psychometric function fits for the different spatial configurations, averaged across subjects (across-subject standard error of the mean is shown in round brackets). The midpoint parameters are greater for co-located than for spatially separated sources; the upper bound of performance is higher for spatially separated sources than for co-located sources; no other differences are significant. (A) Estimates of α, the TMR at the midpoint of the dynamic range in the psychometric function. (B) Estimates of $1/\beta$, the slope of the psychometric function at the midpoint of the dynamic range. (C) Estimates of $1-\lambda$, the upper asymptote of the functions.

	T_v0T_f0	T_v90T_f90	T_v0T_f90	T_v90T_f90
(A) Midpoint of dynamic range α [dB]	27.3 (2.2)	27.2 (2.9)	25.2 (3.1)	25.1 (3.2)
(B) Slope at the midpoint of dynamic range $1/\beta$ [% correct/dB]	19.2 (5.4)	20.4 (7.6)	19.5 (5.0)	16.0 (1.6)
(C) Upper asymptote of performance $1-\lambda$ [% correct]	85.6 (6.2)	85.4 (5.8)	91.6 (6.0)	92.5 (4.3)

that they were most sure of first. The effect of the absolute locations of the talkers on report order suggests that a message from in front of the listener was more salient (and that listeners were therefore more sure of its content) than a message from the side of the listener. Note that this was the only aspect of performance for which the absolute locations of the talkers mattered (after accounting for the acoustic effects of the better ear for $target_V$); all other effects of spatial configuration depended only on whether the talkers were spatially separated or co-located.

We conclude that at least three factors affected the relative certainty listeners had about the content of the competing messages: listeners were actively trying to attend to $target_V$, which enhanced the neural representation of $target_V$ (when listeners were successful at hearing $target_V$). However, the ability to hear $target_V$ depended directly on $T_VT_FR_{be\text{-}V}$. On top of both of these effects, the source from in front of the listener appeared to be inherently more salient than the other source, which caused an asymmetry in report orders for the two spatially separated configurations.

In this task, listeners were not instructed to report the two messages in any particular order, and were not penalized if they incorrectly paired keywords from the two competing messages. Despite this, report order depended systematically on $T_VT_FR_{be\text{-}V}$, on whether the messages were spatially separated or co-located, and on the absolute spatial configuration of the sources. The consistency of these effects, even without any explicit instruction to the subjects, suggests that listeners naturally adopted a strategy in this divided attention task in which they gave top priority to the usually harder-to-hear variable-level target over the fixed-level target.

APPENDIX B: FITS TO PSYCHOMETRIC FUNCTIONS

Psychometric functions were fit to the percent correct scores as a function of T_VT_FR for each subject and condition (Wichmann and Hill, 2001a; see also Ihlefeld and Shinn-Cunningham, 2008). The estimated probability of responding correctly at a given T_VT_FR, $\hat{P}(x)$ was fit as

$$\hat{P}(x) = \gamma + (1 - \lambda - \gamma)\frac{1}{1 + e^{\alpha - x/\beta}}, \quad \text{(B1)}$$

where γ is the lower bound on performance (chance performance, set to 6%), $1-\lambda$ is the upper bound on performance at the largest T_VT_FR, α is the energy ratio at which percent correct performance is halfway between chance and asymptotic performance, and $1/\beta$ is the slope of the psychometric function evaluated at $x=\alpha$.

The goodness of fit of the psychometric functions was evaluated with a deviance criterion that was derived using Efron's bootstrap technique (Wichmann and Hill 2001a, Wichmann and Hill, 2001b). Fourteen of the 16 fits (four functions for each of four listeners) meet the 95% confidence interval deviance criterion. The relatively poor data fit in the other two cases was not caused by outliers (subjectively, even these fits summarized the results adequately).

The upper bound parameter $1-\lambda$ and the slope parameter $1/\beta$ were fitted to maximize the likelihood of observing the actual data given the psychometric function, using the psignifit toolbox in MATLAB 6.5. The resulting parameters, averaged across subjects, are shown in Table III. *T*-tests were employed to test for differences between the within-subject averages of the two spatially co-located configurations and the two spatially separated configurations. The midpoint parameter α of the psychometric function was significantly larger in the spatially co-located than in the spatially separated configurations (*t*-test; $p<0.01$). The slopes at the midpoints of the psychometric functions, $1/\beta$, did not vary significantly with spatial configuration (*t*-test; $p<0.01$). The upper bounds $1-\lambda$ were significantly lower in the co-located than in the separated configurations (*t*-test; $p<0.01$), reflecting the lower level of performance for co-located configurations at the greatest $T_VT_FR_{be\text{-}V}$.

[1]Considering that (1) listeners were not instructed to report keywords in proper pairing and that (2) listeners also received correct feedback for mix responses, mix responses could have resulted from a response strategy whereby listeners did not attempt to report keywords in proper pairings. However, given that listeners had a strong natural tendency to report the keywords in proper pairings, this is not a very likely explanation for the occurrence of mix responses (see also Appendix A).

[2]Responses were also analyzed as a function of the better ear for $target_F$ ($T_VT_FR_{be\text{-}F}$; results not shown here). However, this analysis does not reveal any consistent pattern in the data that could help in explaining any of the effects of spatial separation in the data. In fact, when plotted as a function of $T_VT_FR_{be\text{-}F}$, the performance curves for the spatially separated configurations end up being shifted away from each other, seemingly increasing the difference between spatially separated configurations.

Arbogast, T. L., and Kidd, Jr., G. (**2000**). "Evidence for spatial tuning in informational masking using the probe-signal method," J. Acoust. Soc. Am. **108**, 1803–1810.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (**2006**). "The influence of spatial separation on divided listening," J. Acoust. Soc. Am. **120**, 1506–1516.

Best, V., Ozmeral, E., Gallun, F. J., Sen, K., and Shinn-Cunningham, B. G. (**2005**). "Spatial unmasking of birdsong in human listeners: Energetic and informational factors," J. Acoust. Soc. Am. **118**, 3766–3733.

Bolia, R. S., Nelson, W. T., and Ericson, M. A. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bonnel, A., and Hafter, E. (**1998**). "Divided attention between simultaneous auditory and visual signals," Percept. Psychophys. **60**, 179–190.

Broadbent, D. (**1954**). "The role of auditory localization in attention and memory span," J. Exp. Psychol. **47**, 191–196.

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.

Brungart, D. S., and Simpson, B. D. (**2002**). "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal," J. Acoust. Soc. Am. **112**, 664–676.

Brungart, D., Simpson, B., Darwin, C., Arbogast, T., and Kidd, G. J. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Conway, A. R., Cowan, N., and Bunting, M. F. (**2001**). "The cocktail party phenomenon revisited: The importance of working memory capacity," Psychon. Bull. Rev. **8**, 331–335.

Cowan, N. (**1995**). *Attention and Memory: An Integrated Framework* (Oxford University Press).

Dorman, M., Loizou, P., 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.

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.

Ebata, M., Sone, T., and Nimura, T. (**1968**). "Improvement of hearing ability by directional information," J. Acoust. Soc. Am. **43**, 289–297.

Freyman, R., Helfer, K., and Balakrishnan, U. (**2005**). "Spatial and spectral factors in release from informational masking in speech recognition," Acta Acust. **91**, 537–545.

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.

Ihlefeld, A., and Shinn-Cunningham, B. G. (**2008**). "Spatial release from energetic and informational masking in a selective speech identification task." J. Acoust. Soc. Am. **123**, 4369–4379.

Kidd, G. J., Arbogast, T., Mason, C., and Gallun, F. (**2005**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.

Lawson, E. A. (**1966**). "Decisions concerning the rejected channel," Q. J. Exp. Psychol. **18**, 260–265.

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.

McMains, S., and Somers, C. (**2005**). "Processing efficiency of divided spatial attention mechanisms in human visual cortex," J. Neurosci. **25**, 9444–9448.

Moray, N. (**1959**). "Attention in Dichotic Listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**, 56–60.

Rivenez, M., Darwin, C. J., and Guillaume, A. (**2006**). "Processing unattended speech," J. Acoust. Soc. Am. **119**, 4027–4040.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, and Larson, E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. **91**, 967–979.

Spieth, W., Curtis, J., and Webster, J. (**1953**). "Responding to one of two simultaneous messages," J. Acoust. Soc. Am. **26**, 391–396.

Stifelman, L. (**1994**). "The cocktail party effect in auditory interfaces: A study of simultaneous presentation," MIT Media Laboratory Technical Report.

Treisman, A., and Geffen, G. (**1967**). "Selective attention: Perception or response?" Q. J. Exp. Psychol. **19**, 1–17.

Watson, C. (**2005**). "Some comments on informational masking," Acta Acust. **91**, 502–512.

Wichmann, F., and Hill, N. (**2001a**). "The psychometric function: I. Fitting, sampling and goodness-of-fit," Percept. Psychophys. **63**, 1293–1313.

Wichmann, F., and Hill, N. (**2001b**). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," Percept. Psychophys. **63**, 1314–1329.

Yost, W. A. (**1997**). "The cocktail party problem: Forty years later," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. anderson (Erlbaum, New York), pp. 329–348.

Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (College-Hill Press, Boston, MA).

Frequency discrimination learning in children[a)]

Lorna F. Halliday,[b)] Jenny L. Taylor, A. Mark Edmondson-Jones, and David R. Moore
MRC Institute of Hearing Research, University Park, Nottingham, NG7 2RD, United Kingdom

(Received 14 June 2007; revised 12 February 2008; accepted 13 February 2008)

Psychoacoustic thresholds of pure tone frequency discrimination (FD) in children are elevated relative to those of adults. It has been shown that it is possible to improve FD thresholds in adults, following a single (subhour) training session. To determine whether FD thresholds in children may be improved by training and, consequently, reduced to adult levels, 100 normally hearing 6- to 11-year-old children and adults received ~1 h of training on a FD task at 1 kHz. At the start of training, a quarter of all child participants had FD thresholds that resembled those of naïve adults (adult-like subgroup). Another quarter achieved thresholds that were adult-like at some point during training (trainable subgroup). For the remainder (nonadult-like subgroup), thresholds did not reach those of naïve adult listeners at any point in the training session. Subgroup membership was linked to the influence of three factors—age, nonverbal IQ, and attention. However, across subgroups, learning was found not to generalize to either a different standard frequency (4 kHz) or a variable (roving) presentation paradigm. The results indicate that it is possible for some children to achieve FD thresholds comparable to those of naïve adults, either natively or after limited training.
 [DOI: 10.1121/1.2890749]

PACS number(s): 43.66.Fe, 43.66.Ba [BLM] Pages: 4393–4402

I. INTRODUCTION

In recent years there has been an upsurge in studies of auditory perceptual learning in children (see Moore and Amitay, 2007, for review). To a large extent, the motivation behind these studies has come from evidence that childhood deficits in discriminating brief, or rapidly presented sounds, may be linked to problems in processing the sound structure of speech, with subsequent problems for language learning (e.g., Tallal, 2004). Consequently, many of these studies have focused on training a variety of tasks (both speech and nonspeech) incorporating a temporal processing component, for several hours a day over a period of several weeks, and have typically used children with language-learning impairments (LLI) such as specific language (SLI) and reading impairment (SRI), and "higher-level" (nonauditory) outcome measures such as performance on standardized tests of language and literacy (Habib *et al.*, 1999, 2002; Merzenich *et al.*, 1996; Schäffler *et al.*, 2004; Tallal *et al.*, 1996; Temple *et al.*, 2003; but see Moore *et al.*, 2005). Research in adults has, however, shown that it is possible to induce auditory learning in nonclinical groups, using a single type of sound stimulus, and a (comparatively) less intensive training regime. Here we examine whether it is possible to induce that same kind of learning in children.

Frequency discrimination (FD) skills are of interest for a number of reasons. First, there is evidence to suggest that FD thresholds are poorer in children and adults with LLI relative to typically developing controls (Ahissar *et al.*, 2006; Baldeweg *et al.*, 1999; Banai and Ahissar, 2004; Cacace *et al.*, 2000; Fischer and Hartnegg, 2004; Halliday and Bishop, 2006; Hill *et al.*, 2005; McArthur and Bishop, 2004a, b; Ors, 2002; Schäffler *et al.*, 2004; cf. Walker *et al.*, 2002). If such auditory processing deficits play a role in the genesis of LLI, as has been suggested (e.g., Tallal, 2004), then an understanding of how these abilities may be trained may have implications for remediation. Second, psychophysical studies have shown that FD thresholds are poorer in typically developing young children relative to adults and have a long developmental trajectory (e.g., Jensen and Neff, 1993; Thompson *et al.*, 1999), with some reports suggesting that thresholds continue to improve beyond 12 years of age (Maxon and Hochberg, 1982). Finally, research using adult listeners has shown that it is possible to obtain robust learning effects on a FD task (Amitay *et al.*, 2005; 2006a; Campbell and Small, 1963; Delhommeau *et al.*, 2002; 2005; Demany, 1985; Demany and Semal, 2002; Grimault *et al.*, 2003; Hawkey *et al.*, 2004; Irvine *et al.*, 2000; Moore, 1973; Wright and Sabin, 2007). This raises the possibility that young children might be able to achieve adult-like levels of performance on a FD task, if they are given sufficient training.

There is now a growing body of research into FD learning in adults, which has led to a number of principle findings. FD learning in adults is prolonged, with improvements in thresholds being observed even after multiple hours of training (Amitay *et al.*, 2005; Campbell and Small, 1963; Delhommeau *et al.*, 2002; 2005; Demany and Semal, 2002; Grimault *et al.*, 2003; Irvine *et al.*, 2000; Moore, 1973; Wright and Sabin, 2007). However, it is also rapid, occurring during even a single (<1000 trials) training session (Amitay *et al.*,

[a)] Portions of this work were presented in "Auditory frequency discrimination learning in children," BSA Short Papers Meeting, Cambridge, UK, 2006 (Abstract No. 8) and in "Frequency discrimination learning in children: Effects of age and intelligence," ARO Midwinter Meeting, Denver, CO, 2007 (Abstract No. 926).

[b)] Current affiliation: Developmental Cognitive Neuroscience Unit, UCL Institute of Child Health, 30 Guilford Street, London, WC1N 1EH, United Kingdom. Electronic mail: l.halliday@ich.ucl.ac.uk

2006a; Hawkey *et al.*, 2004). FD learning has also been shown to transfer to frequencies within the range mediated by the same mechanism (i.e., temporal coding at low frequencies, tonotopic coding at high frequencies) (Demany, 1985), although more recent findings indicate that at least a small component of learning is frequency- (Demany and Semal, 2002; Delhommeau *et al.*, 2005; Irvine *et al.*, 2000), duration-, and ear-specific (Delhommeau *et al.*, 2002; Demany and Semal, 2002). Finally, FD thresholds show considerable individual variability between listeners, even after training (Amitay *et al.*, 2005). To our knowledge, none of these findings have been tested in children.

Only two studies to date have examined the effects of training on FD abilities in typically developing (Soderquist and Moore, 1970), and reading-impaired (Schäffler *et al.*, 2004) children. In Schäffler *et al.* (2004) 140 participants (aged 7 to 21 years) with SRI received training on a battery of five auditory psychophysical tests, one of which was FD. Listeners were presented with an adaptive two-alternative forced-choice paradigm, whereby they were required to state which of two tones (initially 1000 vs 1100 Hz) was higher in frequency. Over a minimum period of 10 days, participants played one task per day, with each daily session lasting approximately 10–15 min. Although prior to training, 72% of these children fell below the 16th percentile on FD, this had fallen to 13% following training. Schäffler *et al.* (2004) therefore concluded that it was possible to improve the FD skills of children with SRI with daily training.

A more detailed approach was taken by Soderquist and Moore (1970), who recruited 54, 5-, 7-, and 9-year-old children from mainstream education, half of whom received training on a FD task using the method of constant stimuli. The remaining, control children, participated in normal classroom activities during this time. On a given training trial, children were presented with one standard (300 or 350 Hz) tone, and a comparison tone of a different frequency. Children were required to state whether the comparison tone was higher or lower in frequency than the standard tone, and verbal feedback was provided stating the correctness of each response. Children received 600 trials of training (100 trial blocks over 6 sessions, which were, where possible, on consecutive days). Consistent with findings of a prolonged developmental trajectory for FD, Soderquist and Moore (1970) found that 5 year olds showed significantly higher (poorer) FD thresholds than their 7- and 9-year-old peers at the start of training. Moreover, they reported a nonsignificant trend for the trained group to show lower (better) FD thresholds relative to controls following training, with the most dramatic effects being observed for the youngest (5-year-old) group. These trends remained apparent when a subset of participants was retested 11 months later. The findings of Soderquist and Moore (1970) therefore suggest that it may be possible to train FD abilities in children, and that this training may have lasting effects.

The goal of the present study was to determine whether, like adults, children show learning on a FD task even after limited (<1000 trials) training. We investigated the effects of training on FD thresholds in four groups of typically developing listeners: 6 to 7 year olds, 8 to 9 year olds, 10 to 11 year olds, and adults. We asked whether (a) FD thresholds in children could be trained to reach a similar level to those of naïve (nontrained) adults, and (b) improvements in FD would generalize to a different frequency, and to randomly presented roving frequencies, following a single session of training.

II. METHODS

A. Participants

Participants were subdivided into four groups according to age: 6 to 7 year olds, 8 to 9 year olds, 10 to 11 year olds, and adults (18 to 40 year olds). Participants were native English speakers who had no prior experience of psychoacoustic testing. Prior to inclusion in this study, participants were required to pass an audiometric hearing test (pure-tone thresholds ≤25 dB HL bilaterally, at 0.5, 1, 2, and 4 kHz) administered in accordance with the British Society of Audiology's recommended procedure (British Society of Audiology, 2004), and to be able to reliably discriminate between a 1- and 1.5-kHz tone. Eight children (three 6 to 7 year olds, one 8 to 9 year old, and four 10 to 11 year olds) and three adults were excluded from the study on the basis of elevated audiometric thresholds; twenty-two children (thirteen 6 to 7 year olds, six 8 to 9 year olds, and three 10 to 11 year olds) were excluded on the basis of being unable to reliably discriminate between 1 and 1.5 kHz. This resulted in 100 participants being included in this study (see Table I). Children were recruited via two primary schools in Nottinghamshire, UK, and received vouchers that could be redeemed for gifts for their participation. Adults were recruited via adver-

TABLE I. Descriptive statistics and comparisons of the four age groups.

	6 to 7 years	8 to 9 years	10 to 11 years	Adults	*p* value
N	17	29	24	30	
Age (years; $M \pm$ s.d.)	7.42 ± 0.45	9.17 ± 0.67	10.83 ± 0.55	26.24 ± 5.76	
Age range	6.33–7.85	8.01–9.98	10.03–11.73	19.20–40.60	
Gender (males/females)	6/11	13/16	7/17	15/15	0.424[b]
Nonverbal IQ[a]	93.76 ± 9.79	101.69 ± 12.33	98.17 ± 13.54	108.10 ± 11.10	0.001[c]
Nonverbal IQ range	79–114	76–128	71–123	86–127	

[a]Standard scores derived from the combined nonverbal subscales of the Weschler Abbreviated Scale of Intelligence (WASI; The Psychological Corporation, 1999).
[b]Pearson chi-square test.
[c]One-way ANOVA.

tisements posted on public notice boards, and were paid for their participation. Informed consent was obtained from all adult participants, and from the parents of all child participants. This study was approved by the Nottingham NHS Research Ethics Committee.

B. Stimuli and apparatus

The stimuli were 200-ms tone bursts with 10-ms raised-cosine ramps, separated by an interstimulus interval of 500 ms. Stimuli were digitally generated on each presentation, using custom software running on a PC. The signal waveforms were generated at a sampling rate of 44.1 kHz, and were output as 16-bit samples using a sound card (Darla Echo; Echo Digital Audio Corporation, Carpinteria, CA). Stimuli were presented diotically via Sennheiser HD-25 headphones at a level of 70 dB SPL (reference 20 μPa). Presentation levels were calibrated using a Brüel and Kjær measuring amplifier (type 2660), microphone (type 4192), and artificial ear (type 4153). Participant responses were recorded via a touch screen.

C. Design and procedure

Child participants were tested over two sessions. The first session took place in a quiet room in the child's school, and consisted of an audiometric screen, followed by two demo trials designed to ensure that children could reliably discriminate between 1 and 1.5 kHz. Children who passed both tests were then invited to attend a second session, lasting approximately 3 h, which was conducted in a laboratory sound-attenuating chamber. This session consisted of a full audiometric hearing test, further demo trials, a training phase, and two generalization posttraining tests (see Sec. II C 1 and 2). In an effort to introduce novelty into the training phase, cognitive tests were administered between training blocks (see Table I). Adult participants attended the second testing session only. Other than that, the training and testing procedures for the adult and child participants were identical. Participants were tested individually during both sessions.

1. Training phase

Training was delivered via eight blocks of 75 trials, comprising three interleaved tracks running concurrently. For each track, FD thresholds were estimated using an adaptive, three-interval, three-alternative (oddball) forced-choice paradigm. In each trial, two intervals contained a standard, 1-kHz tone, and the third, randomly determined interval contained a higher-frequency target tone, the frequency of which varied adaptively from trial to trial. The listener's task was to detect the interval that contained the target tone. The frequency of the target tone was adaptively varied on each trial using a staircase procedure. The initial ΔF was 50% of the standard frequency (target tone $F=1500$ Hz). An initial "lead-in" one-down one-up rule was used to speed up approach to the ΔF region of interest. During this phase, ΔF was halved after each step, until the first error occurred. The staircase then followed a three-down one-up rule targeting the 79% correct point on the psychometric function (Levitt, 1971). During this phase, ΔF was multiplied or divided by a factor of $\sqrt{2}$.

Training was delivered via child-friendly computer games in which each auditory interval corresponded to a visual event on the computer screen. Participants were given an unlimited time to respond, and the initiation of each new trial was self-paced. Positive trial-by-trial visual feedback was provided for all correct responses, and tokens accumulated at the bottom of the touch screen as a measure of past performance success.

2. Posttraining phase

Following the training phase, two additional tests were administered, in counterbalanced order. These tests were designed to examine whether learning generalized (a) across frequency ("4-kHz test"), and (b) to randomly roving frequencies ("roving test"). Both tests followed the same procedure as a single block (3×25 trials) of the training phase (see Sec. II C 1), with one difference: in the 4-kHz test, thresholds were estimated using a standard tone of 4-kHz; in the roving test, the standard tone was randomly roved on a trial-by-trial basis between frequencies outside the 1-kHz critical band (570, 1170, and 2150 Hz).

D. Parameter estimation

Although threshold estimates derived from psychometric function fitting and from traditional (reversal averaging) methods have been shown, in adults, to yield equivalent values for a FD task (Amitay *et al.*, 2006b), the former has the distinct advantage over the latter in providing an estimate of performance at the upper asymptote in addition to threshold. This provides an estimate of lapse rate (i.e., the extent to which an individual would fail to reach 100% correct at the highest ΔF on any given run). Lapse rate can, therefore, be used to provide an estimate of inattentiveness, which has been proposed as a possible mechanism underlying the poorer performance of young children relative to adults on auditory psychophysical tasks (e.g., Lutfi and Wightman, 1996; Wightman and Allen, 1992; cf. Wightman *et al.*, 2003). Consequently, psychometric functions were fitted to the data of each participant for each block using the fitting technique described by Wichmann and Hill (2001a; b; http://bootstrap-software.org/psignifit/). This technique has been described in detail elsewhere (see Amitay *et al.*, 2006b; Hill, 2001; Wichmann and Hill, 2001a; b) so only an outline will be given here. A two-parameter logistic psychometric function (percentage of correct responses as a function of ΔF) was fitted to the pooled data from the three concurrent adaptive tracks for each block for each participant. The probability of responding correctly when the adaptive parameter value is ΔF, in % Hz, is

$$\psi(\Delta F) = \gamma + \frac{1-\gamma-\lambda}{1+e^{-(\log(\Delta F)-\alpha)/\beta}} \quad (1)$$

where γ is the guessing rate (chance performance level), and $1-\lambda$ is the performance level at the greatest ΔF levels (λ is also termed the lapse rate), such that higher estimates of λ correspond to poorer performance. With the exception of α

and β which were unconstrained, γ was fixed at 0.33, and λ was constrained to have a maximum of 30%. An upper limit of 30%—rather than the 6% default advocated by Wichmann and Hill (2001a, b)—was used in recognition of the possibility that children would show a greater propensity to attentional lapses than adult listeners. Note that while the model generates estimates of threshold (the 79% point on the psychometric function), lapse rate, and slope, only threshold and lapse rate will be reported here owing to the variability of slope estimates in datasets (like those here) where few points are available (Kaernbach, 2001; Leek *et al.*, 1992).

Goodness of fit was evaluated by running Monte Carlo simulations to obtain confidence intervals for model deviance (Hill, 2001; Wichmann and Hill, 2001a). Approximately 3% of the psychometric function fits for the training blocks, 7% for the 4-kHz tests, and 4% for the roving tests, fell outside the 95% confidence intervals. To avoid any bias toward underdispersion, data that failed the goodness of fit were included in the analysis. However, we have separately confirmed that excluding these cases did not essentially change the pattern of results shown here.

E. Missing data and statistical analyses

Time constraints resulted in the following missing data: Three participants were unable to complete the training blocks, two of whom (one 6 to 7 year old, and one 8 to 9 year old) completed six out of eight and one of whom (10 to 11 year old) completed seven out of eight of the training blocks; Five participants (two 6 to 7 year olds, one 8 to 9 year old, and two 10 to 11 year olds) were unable to complete the 4-kHz test, and three participants (one 6 to 7 year old and two 10 to 11 year olds) were unable to complete the roving test. The available data for these participants were included in the analyses.

Data were excluded from the analysis if the model returned threshold estimates that either hit ceiling (exceeded the maximum stimulus presentation ΔF of 50%) or were deemed biologically implausible (fell below 0.1%). The justification for excluding ceiling data was threefold. First, as these thresholds were estimated on the basis of extrapolation beyond the stimulus range used in this experiment, confidence intervals for these threshold estimates were wide, and therefore their reliability was low. Second, participants had, prior to being included in the study, shown that they were able to reliably discriminate a ΔF of 50% against a 1-kHz standard. Failure to discriminate this same difference on a given block was therefore taken as an indication of noncompliance to the task. Finally, despite training, many of the children in the youngest age groups reported that they found the posttraining tests difficult, a finding we believe to be of interest. Consequently, for the 4-kHz and roving tests, the proportion of children per age group who exceeded ceiling is included in a separate analysis. Approximately 3% of the estimated thresholds for the training blocks, and 14% and 21% of those for the 4-kHz and roving tests, respectively, exceeded ceiling. Just 0.4% of threshold estimates for the training blocks were classified as biologically implausible. Excluding these data from the main analyses will tend to understate the thresholds for these tests. This makes the subsequent findings regarding the 4-kHz and roving tests conservative. Following data exclusion, a total of 770 of a possible 800 (100 participants × 8 blocks) training blocks, and 155 of a possible 200 posttest blocks were included in the analyses.

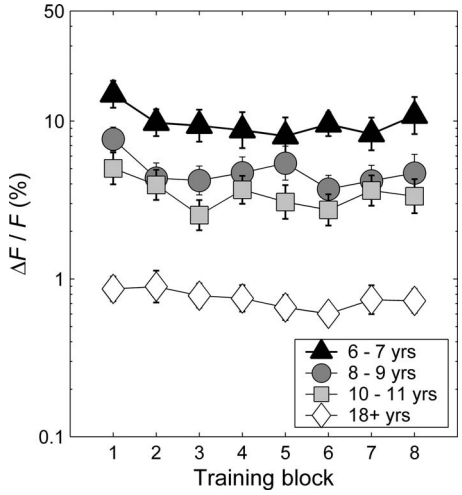


FIG. 1. Geometric mean thresholds, expressed as $\Delta F/F$ (%), over the eight training blocks for the 6- to 7-year-old group (triangles), the 8- to 9-year-old group (circles), the 10- to 11-year-old group (squares), and the adult group (diamonds). Error bars indicate ±1 standard error.

Estimates of threshold (in % Hz) for each of the four groups for each block violated the normality assumption owing to a positive skew. All statistical analyses were therefore conducted on log-transformed thresholds, which showed a normal distribution (Kolmogorov-Smirnov test; $p > 0.05$). Estimates of lapse rates were also non-normally distributed, owing primarily to floor effects. These data were therefore analyzed using nonparametric statistics. Inspection of the data revealed that performance within groups was highly variable, with thresholds spanning over two orders of magnitude (base 10). Consequently, data were analyzed first in terms of group means, and second on the basis of individual (subgroup) differences.

III. RESULTS

A. Training

1. Group analysis

Mean estimates of threshold for each age group (Fig. 1) suggest three principle findings. First, FD thresholds showed a consistent improvement (reduction) with increasing age. Second, thresholds improved with training—lower thresholds were obtained in later than in earlier training blocks. Third, the learning curves were characterized by an initial rapid improvement, followed by a leveling-off of performance.

The effects of age (Group) and training (Block) on FD thresholds were investigated using a linear mixed models procedure (SPSS version 14.0, SPSS Inc., Chicago, IL, 2005). As nonverbal IQ differed significantly between age groups (see Table I) and is known to be associated with FD (Amitay *et al.*, 2005), this was entered as a covariate in the

model. Linear mixed models allow the correlation structure within the data to be modeled, while also permitting a high degree of flexibility regarding the models that can be considered. In particular, it permits the inclusion of threshold data for subjects where other thresholds are missing. All valid models incorporating some or all of the main effects and interactions of factors Group and Block and the covariate nonverbal IQ (IQ) were fit, and the model minimizing Akaike's information criterion (AIC) was selected on the grounds of providing the best data fit subject to an allowance for model parsimony. Other information criteria [Hurvich and Tsai's Criterion (AICC), Bozdogan's Criterion (CAIC), and Schwarz's Bayesian criterion (BIC)] were also inspected and found to be consistent with the AIC optimizing model. These criteria consistently identified the main effects model comprising Group, Block, and IQ as being the best model according to the stated model selection criteria. All models assume a random intercept-only model structure.

Thresholds improved generally with age, resulting in a highly significant main effect of Group, $F(3,95.13)=24.49$, $p<0.001$. Pairwise comparisons using Fisher's least significant difference (LSD) indicated that all child groups had significantly ($p<0.001$) higher (poorer) thresholds than those of adults, and that thresholds for the 10 to 11 year olds were significantly lower than those for the 6- to 7-year-old group ($p<0.001$). Thresholds for the 8 to 9 year olds did not differ significantly from those of either the 6- to 7-year-old ($p=0.091$) or 10- to 11-year-old groups ($p=0.056$).

Inspecting the training effect, we also found a highly significant effect of Block, $F(7,663.61)=7.28$, $p<0.001$. Pairwise comparisons (Fisher's LSD) confirmed an initial reduction in thresholds between blocks 1 and 2 and blocks 2 and 3 ($p=0.003$, and $p=0.01$, respectively), with little, if any, subsequent learning (blocks 3–8: $p>0.05$). This was explored in more detail by a reparametrization of the training block factor into seven dummy variables which suggested significant incremental differences between blocks 1 and 2 ($p=0.004$), with no significant difference between subsequent blocks ($p>0.05$). The absence of a significant training Block × Group interaction suggests no evidence of this training pattern being different in the four age groups.

Finally, we observed that nonverbal IQ also had a significant effect on estimated thresholds, $F(1,94.88)=17.53$, $p<0.001$, with higher IQs being associated with better (lower) thresholds. Again, the absence of a significant Group × IQ interaction suggests no evidence that this effect varied by age group.

As outlined earlier, one of the advantages of psychometric function fitting is that, in addition to thresholds, the procedure also yields estimates of the extent to which an individual would fail to reach 100% correct at the highest ΔF levels. The estimated lapse rates were greater than 5% (never greater than 30% owing to model constraints) on 27%, 24%, 20%, and 6% of all blocks for the 6 to 7 year olds, 8 to 9 year olds, 10 to 11 year olds, and adults, respectively. Estimates of lapse rate were found not to vary significantly (Kruskal–Wallis tests) as a function of block in any age group, suggesting that attention did not wane with repeated testing. Consequently, data were combined across all blocks

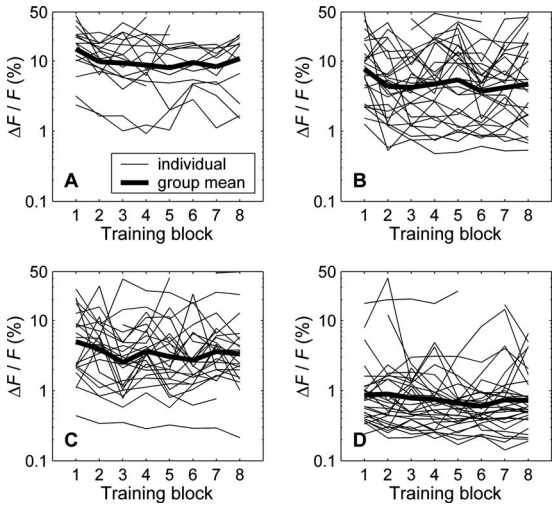


FIG. 2. Individual (thin lines) and group geometric mean (thick lines) thresholds over the eight training blocks for (A) the 6- to 7-year-old group, (B) the 8- to 9-year-old group, (C) the 10- to 11-year-old group, and (D) the adult group.

for further analyses. A Kruskal–Wallis test found a significant effect of group on lapse rates, $H(3)=70.83$, $p<0.001$. Post-hoc comparisons (Mann–Whitney) confirmed that estimates of lapse rate decreased significantly as a function of age. Estimates of lapse rate for the 6 to 7 year olds (7%) were higher (poorer) than those of the 8 to 9 year olds (3%; $p=0.004$) and the 10 to 11 year olds (2%; $p=0.004$) which, in turn, were higher than those of adults (0.4%; $p<0.001$).

2. Subgroup analysis

Figure 2 presents both individual and group mean thresholds as a function of block and age group. As indicated earlier, group means masked considerable variability in performance both within groups and across blocks. In an effort to assess this variability, child participants were divided into subgroups on the basis of their performance across the eight training blocks. Participants were classified as "adult-like" if they obtained a threshold that was within 1 s.d. of the log adult mean for block 1 (-0.062 ± 0.422) on their first training block. Participants were classified as "trainable" if they obtained a threshold that was within 1 s.d. of the log adult mean for block 1 on any training block other than block 1. Finally, participants were classified as "nonadult-like" if they failed to obtain a threshold that was within 1 s.d. of the log adult mean for block 1 for any of the eight training blocks.

Figure 3 presents the individual and subgroup mean thresholds as a function of block and subgroup. Prospective subgroupings were independently verified using a linear mixed models analysis of variance (ANOVA) (Subgroup × Block, on log thresholds), which confirmed main effects of Subgroup, $F(2,66.73)=90.66$, $p<0.001$, and Block, $F(7,443.03)=6.60$, $p<0.001$, as well as a significant Subgroup × Block interaction, $F(14,443.13)=3.27$, $p<0.001$. Post-hoc analyses (linear mixed models ANOVA with Fisher's LSD) showed that whereas thresholds for both the nonadult-like and the trainable subgroups varied significantly as a function of block, $F(7,215.30)=2.57$, $p=0.014$,

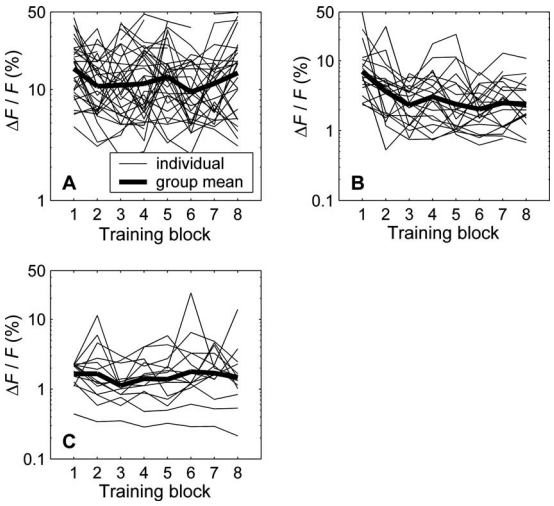


FIG. 3. Individual (thin lines) and group geometric mean (thick lines) thresholds over the eight training blocks for (A) the nonadult-like subgroup, (B) the trainable subgroup, and (C) the adult-like subgroup.

and $F(7,138.09)=8.19$, $p<0.001$, respectively, thresholds for the adult-like subgroup did not, $F(7,90.04)=1.18$, $p=0.323$. The trainable subgroup showed a significant learning effect in early (blocks 1 to 2: $p=0.001$; blocks 2 to 3: $p=0.023$), but not later (3–8) blocks ($p>0.05$). Thresholds for the nonadult-like subgroup showed an atypical learning pattern, with thresholds showing a significant reduction between blocks 1 and 2 ($p=0.019$) and blocks 5 and 6 only ($p=0.036$).

Characteristics of the three subgroups were assessed in terms of age, nonverbal IQ, and attention (Table II). Half of the children overall could be classified as adult-like or trainable and this proportion increased with age, χ^2 (4)=15.09, $p=0.005$. The effect of age was confirmed with a one-way ANOVA, $F(2,69)=8.65$, $p<0.001$. Post-hoc analyses using Scheffé tests showed that children classified as either adult-like or trainable were on average older than those classified as nonadult-like, $p=0.001$, and $p=0.027$, respectively. The adult-like and trainable groups did not differ significantly in age, $p=0.430$. The three subgroups also differed in their nonverbal IQ, $F(2,69)=3.98$, $p=0.023$. Post-hoc analyses (Scheffé) showed that the adult-like group had significantly higher nonverbal IQ than the nonadult-like group, $p=0.027$. The trainable group did not differ significantly in nonverbal IQ from either the nonadult-like or the adult-like groups, $p=0.366$ and $p=0.384$, respectively. Finally, the three subgroups differed in their lapse rates [Kruskal–Wallis: $H(2)=305.85$, $p<0.001$]. The nonadult-like group had higher lapse rates than the trainable group (Mann–Whitney: $p<0.001$) which, in turn, had higher lapse rates than the adult-like group ($p<0.001$).

B. Generalization of learning across frequency

1. Group analysis

As reported earlier, despite training, a relatively high proportion (14%) of threshold estimates for the 4-kHz test exceeded ceiling. This proportion varied significantly as a function of age (approximately 33% of 6 to 7 year olds; 21% of 8 to 9 year olds; 5% of 10 to 11 year olds; and 3% of adults), as confirmed by a χ^2 analysis, $\chi^2(3)=10.60$, $p=0.013$.

Figure 4(A) presents the mean estimates of threshold for each age group, excluding those participants whose thresholds for the 4-kHz test exceeded ceiling. For ease of comparison, thresholds for the 4-kHz test are presented alongside those obtained from blocks 1 and 8 of the training phase. The data presented in Fig. 4(A) provide no evidence for generalization of learning across frequency. To assess the significance of this trend, a linear mixed models ANOVA [Group $\times$ Condition (block 1 versus block 8 versus 4-kHz test] was conducted on log-transformed thresholds. The main effect of Group was significant, $F(3,76.35)=24.07$, $p<0.001$. Pairwise comparisons (Fisher's LSD) showed that all three child groups had significantly higher thresholds than adults ($p<0.001$). Thresholds for the 6 to 7 year olds were significantly higher than those of the 10- to 11-year-old group ($p=0.03$). No other group comparisons reached significance. Thresholds also differed significantly as a function of Condition, $F(2,149.02)=9.68$, $p<0.01$. Pairwise comparisons indicated that thresholds on block 8 of the training session were significantly lower than for both block 1 ($p=0.001$) and the 4-kHz test ($p<0.001$). Thresholds on the 4-kHz test did

TABLE II. Descriptive statistics and comparisons of the three training subgroups of children.

	Adult-like	Trainable	Nonadult-like	p value
N 6 to 7 year olds	0	3	14	
N 8 to 9 year olds	5	9	15	
N 10 to 11 year olds	9	9	6	0.005[b]
Age (years; $M \pm$ s.d.)	10.27 ± 1.42	9.69 ± 1.40	8.71 ± 1.32	0.001[c]
Gender (males/females)	5/9	5/16	16/19	0.258[b]
Mean lapse rate	0.001 ± 0.05	0.011 ± 0.03	0.065 ± 0.10	0.001[d]
Nonverbal IQ[a]	105.50 ± 10.32	99.76 ± 14.39	95.06 ± 10.88	0.023[c]
Nonverbal IQ range	91–123	71–128	76–123	

[a]Standard scores derived from the combined nonverbal subscales of the WASI (The Psychological Corporation, 1999).
[b]Pearson chi-square test.
[c]One-way ANOVA.
[d]Kruskal–Wallis test.

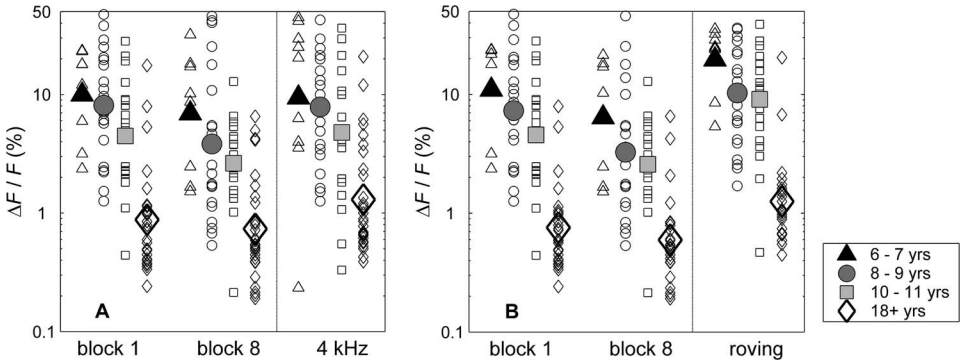


FIG. 4. Individual (small symbols) and group geometric mean (large symbols) thresholds for blocks 1 and 8 of the training phase respectively, and (A) the 4-kHz test (4 kHz), and (B) the roving test (roving), for the 6- to 7-year-old group (triangles), the 8- to 9-year-old group (circles), the 10- to 11-year-old group (squares), and the adult group (diamonds). Data are displayed only for those individuals who had estimated thresholds on (A) the 4-kHz or (B) the roving tests, respectively.

not differ significantly from those of block 1 ($p=0.522$). This pattern was consistent across groups (Group × Condition interaction was not significant).

2. Subgroup analysis

To assess individual differences in performance on the 4-kHz test, participants were again divided into the same three subgroups as outlined earlier (adult-like, trainable, and nonadult-like) and the analysis repeated [see Fig. 5(A)]. A mixed models analysis (Subgroup × Condition) showed significant main effects of Subgroup, $F(2,48.60)=46.27$, $p<0.001$, and Condition, $F(2,94.94)=8.30$, $p<0.001$. Pairwise comparisons (Fisher's LSD) confirmed that the adult-like group had significantly lower thresholds than the trainable group ($p<0.001$) which, in turn, had significantly lower thresholds than the nonadult-like group ($p=0.003$). Across subgroups, thresholds for the 4-kHz test were significantly higher than for block 8 ($p<0.001$) but not for block 1 ($p=0.785$) of training. The interaction between Subgroup and Condition was (just) nonsignificant, $F(4,94.90)=2.26$, $p=0.069$.

C. Transfer of learning across stimulus presentation paradigm

1. Group analysis

The proportion of participants hitting ceiling on the roving test decreased significantly with age [$\chi^2(3)=14.84$, $p=0.002$]. Approximately 50% of 6 to 7 year olds, 28% of 8 to 9 year olds, 5% of 10 to 11 year olds, and 10% of adults hit ceiling on this test.

Mean threshold estimates for the roving test are shown in Fig. 4(B). A mixed models analysis confirmed a main effect of Group, $F(3,71.14)=36.34$, $p<0.001$. All child participant groups obtained higher thresholds than adults (Fisher's LSD: $p<0.001$), and the 6- to 7-year-old group obtained higher thresholds than the 10- to 11-year-old group ($p=0.02$). The main effect of Condition was also significant, $F(2,137.13)=34.52$, $p<0.001$. Across groups (Group × Condition interaction nonsignificant), thresholds for the roving test were significantly higher than for both block 1 ($p<0.001$) and block 8 ($p<0.001$) of the training session.

2. Subgroup analysis

Mean estimates of threshold for the roving test for each subgroup are shown in Fig. 5(B). Mixed models analysis showed significant main effects of Subgroup, $F(2,46.81)=39.66$, $p<0.001$, and Condition, $F(2,89.71)=32.16$, $p<0.001$, and a significant Subgroup × Condition interaction, $F(4,89.72)=3.62$, $p=0.009$. Pairwise comparisons (Fisher's LSD) confirmed that the adult-like group had significantly lower thresholds than both the trainable and the nonadult-like groups ($p<0.001$), and that the trainable group, in turn, had significantly lower thresholds than the nonadult-like group ($p<0.001$). Across subgroups, thresholds on the roving test were significantly higher than on block 1 or block 8 of training ($p<0.001$). Post-hoc analyses (linear mixed models ANOVA with Fisher's LSD) showed that whereas roving thresholds for both the nonadult-like and the trainable subgroups did not differ significantly from those of training block 1 ($p>0.05$), thresholds for the adult-like group were significantly *higher* on the roving test compared to block 1 ($p<0.001$).

IV. GENERAL DISCUSSION

Our results demonstrate four main findings. First, FD thresholds across all (6- to 7-year-old, 8- to 9-year-old,

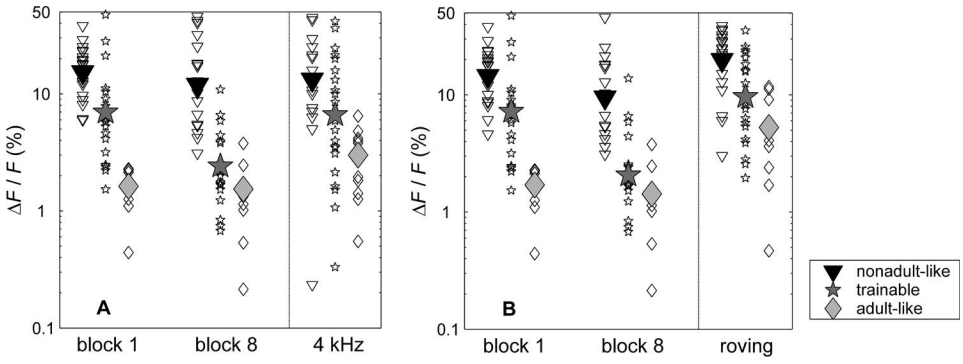


FIG. 5. Individual (small symbols) and group geometric mean (large symbols) thresholds for blocks 1 and 8 of the training phase (block 1 and block 8, respectively) and (A) the 4-kHz test (4 kHz), and (B) the roving test (roving), for the nonadult-like (inverted triangles), trainable (stars), and adult-like (diamonds) subgroups. Data are displayed only for those individuals who had estimated thresholds on (A) the 4-kHz or (B) the roving tests, respectively.

10- to 11-year-old, and adult) groups improved with training, thus confirming that it is possible to induce auditory learning in children after limited (600 trials) practice. However, across age groups, learning was confined to early training blocks ($<\sim 225$ trials). Second, FD thresholds improved with age. Nevertheless, the degree of learning did not vary with age. Third, there was marked interindividual and intra-age-group variability in FD thresholds, both at the start and during the course of training. Finally, across age groups, FD learning did not generalize to a different frequency or a variable (roving) stimulus presentation paradigm.

A. Effects of age

The results of our study confirm those of previous reports of poorer FD abilities in children relative to adults (Jensen and Neff, 1993; Maxon and Hogberg, 1982; Thompson *et al.*, 1999). In general, younger (6- to 7-year-old) children performed more poorly on our FD task than older (10- to 11-year-old) children, and the performance of all child groups was poorer than that of adults, even after training. Our findings therefore suggest that FD abilities continue to develop late into childhood, and do not in general reach adult levels until after 11 years. However, it is important to note that we only included children who were able to reliably discriminate between the highest ΔF levels used in this study (50%). There were a number of children, particularly in the youngest (6- to 7-year-old) age group, who we had to exclude on this basis. This is consistent with the results of Thompson *et al.* (1999), who found that two-thirds of their youngest (5-year-old) children were unable to learn the experimental task. We do not know whether the children excluded from our study would have shown learning on our FD task with training. However, it is likely that the performance of the 6- to 7-year-old group in particular would have been poorer still if these children had been included in our study.

We also found that while children did show evidence of learning, they did not, on average, achieve FD thresholds that were commensurate with those of our naive adult listeners, even after training. A question that is still outstanding is whether it would have been possible to achieve this with additional training. We know that we did not train FD abilities in this study to those of optimal levels of human performance: FD thresholds for the majority of our participants were higher than those that have been reported previously at 1 kHz, in studies involving highly trained adult listeners (e.g., Delhommeau *et al.*, 2002; Moore, 1973). It is therefore likely that FD thresholds for the children could have been improved had we administered further training. Nevertheless, the leveling-off of learning that we saw in later blocks suggests that auditory learning may be more efficiently achieved in children if training is delivered in "bite-sized" pieces (multiple, short training sessions, across different days), a possibility that we are currently investigating.

B. Individual differences

Despite the fact that mean FD thresholds for the child groups did not reach those of adult levels, there was marked interindividual variability in FD thresholds, with performance in all age groups ranging over almost two orders of magnitude. Inspection of these individual differences led to the identification of three subgroups of children. Around 20% of naïve children achieved similar FD thresholds to naïve adults (adult-like subgroup). Thirty percent achieved thresholds comparable to those of naïve adult listeners at some point during the training session (trainable subgroup). The remaining 50% of children did not achieve FD thresholds comparable to those of naïve adults, even after training (nonadult-like subgroup). Our findings suggest that for some children at least, it is possible to achieve adult-like thresholds on a FD task, either natively, or after a limited amount of training.

What was responsible for these individual differences? We did not ask parents about their child's musical training, so it is possible that those children who obtained adult-like FD thresholds at the beginning of training were the ones who were learning to play a musical instrument (Kishon-Rabin *et al.*, 2001; Micheyl *et al.*, 2006; Spiegel and Watson, 1984). However, we did find evidence that a number of other factors may have contributed to both the initial performance, and the susceptibility of children to training, on our FD task. In general, the adult-like group was characterized by being older, having slightly above average nonverbal IQ, and showing fewer attentional lapses. The trainable group was similar in age and IQ to the adult-like group, but showed a greater number of attentional lapses. Finally, the nonadult-like group was younger, and had poorer attention than both the adult-like and trainable groups, but had lower IQ than the adult-like group only.

A relation between intelligence and FD is perhaps not surprising in the context of previous reports (Amitay *et al.*, 2005; Deary, 1994; Talcott *et al.*, 2002), and supports a link between "higher-level" cognitive abilities and performance on a FD task. Our results shed further light on this issue by indicating that nonverbal IQ is likely to be associated with training outcome—the higher the IQ, the more likely a particular child will be to achieve adult-like performance at some point on a FD task. The association between attention and FD thresholds is also intuitive, given the attentional demands that were inherent in our task. To succeed on a given trial, listeners were required to attend three intervals, and to compare at least two of them in frequency. Successful performance on a given block required listeners to sustain attention across 75 of these trials, and across eight of these blocks in a single session. This was clearly difficult for some children, and it may be that the strong associations that we saw among FD thresholds, IQ, and attention were exacerbated by the demands of our testing and training procedure. Nevertheless, that FD performance (and thus subgroup membership) was strongly linked to age was a consistent finding in this study. Our findings therefore suggest that in order to achieve adult-like performance on a psychophysical FD task, a child must first have reached a certain level of cognitive maturation.

C. Generalization of learning

Finally, we also asked whether the learning effects observed in this study would generalize to a FD task that used

a different standard frequency (4-kHz test), or would transfer to one that used a stimulus presentation paradigm in which the standard frequency roved from trial to trial (roving test). Perhaps our most striking finding here was that, despite training, many children were unable to get a reliable threshold on these tests. This was particularly evident for the youngest (6- to 7-year-old) children, and for the roving test, where 50% of 6 to 7 year olds obtained threshold estimates that exceeded the maximum stimulus presentation ΔF of 50%. One possible explanation is that this was due to fatigue, as these posttraining tests came at the end of a relatively long test session. However, fatigue cannot account for why more children hit ceiling on the roving than on the 4-kHz test, as the two tests were counterbalanced across listeners. It is also unlikely that procedural difficulties were responsible, as the posttraining tests were not administered until participants had completed 600 trials of training on the fixed 1-kHz FD task, which had identical response demands. Finally, we do not believe that more reliable threshold estimates would have been achieved if we had allowed ΔF to increase further: For both the 4-kHz and the roving tests, the majority of children were able to discriminate a ΔF of 50% on a given trial. The difficulty they had, it seemed, was in switching their listening strategy to either a different standard frequency or to different roving frequencies. Our tentative interpretation, therefore, is that these ceiling effects may reflect the difficulty that young children had in switching their attention to a different standard frequency and/or presentation paradigm following training on a fixed, 1-kHz standard tone. This interpretation suggests that young children may experience interference, rather than generalization, on an auditory discrimination task following training on a different task.

There was also no evidence for generalization of learning in those listeners who did not reach ceiling on the posttraining tests. Across age and subgroups, thresholds were either not significantly different from (in the case of the 4-kHz test) or significantly *higher* than (in the case of the roving test) thresholds for block 1 of the training. Indeed, mean thresholds for these tests are likely to have been even higher, had we included threshold estimates which exceeded a ΔF of 50%. For the roving task, these results are consistent with those of Amitay *et al.* (2005), who found little evidence for transfer of learning from a fixed to a roving stimulus presentation paradigm, using tasks that were very similar to those employed here. Our finding of limited generalization across frequencies is less consistent with the literature, in which reports of at least partial generalization of learning to different frequencies abound (Demany, 1985; Delhommeau *et al.*, 2005; Grimault *et al.*, 2003; Irvine *et al.*, 2000). However, two explanations may account for our findings. First, our finding that thresholds on the posttraining tests were not significantly different from those of naïve listeners at 1 kHz does not necessarily provide firm evidence against generalization of learning. We do not know what thresholds would have been achieved by *naïve* listeners on these two tests, although there are some reports of a worsening of FD with increasing frequency (Delhommeau *et al.*, 2005; Irvine *et al.*, 2000). It is therefore possible, although unlikely, that naïve listeners would have obtained higher thresholds on the 4-kHz and roving tests had they not received training and that, coincidentally, this training reduced thresholds to levels similar to (or higher than) those of naïve listeners for a fixed task at 1 kHz. Second, there are differences between the experimental procedures employed here and in those of previous studies. In particular, our listeners received considerably less training (~1 h) relative to those of former studies (e.g., 12 and 24 h, in the cases of Delhommeau *et al.*, 2005; and Grimault *et al.*, 2003, respectively). It is therefore possible that the extent to which auditory learning generalizes along the trained dimension, or to a different paradigm, may depend on listeners having received a critical amount of exposure to the trained stimulus and task. This explanation is consistent with recent reports that generalization of learning on a duration discrimination task to an untrained standard frequency may depend not on the total amount of training that listeners have received, but rather on the number of trials listeners have received *per session* (Wright and Sabin, 2007).

D. Conclusion

In sum, the data reported here provide evidence that primary school-age children can show learning on a psychoacoustic FD task following limited (600 trials) training. For half of these children, FD thresholds were comparable to those of naïve adult listeners, either at the start of, or during the course of learning, although this was more likely in older children. However, three caveats deserve mention. First, although at the start of training our adult-like subgroup achieved thresholds that were comparable to those of naïve adult listeners, unlike adults they did not show significant learning as the session progressed. In this way our term 'adult-like' may be a misnomer. This finding suggests that naïve performance and the ability to show learning on a FD task may be governed by different mechanisms. Second, we do not know how long the training effects we observed here might persist, or how much additional learning we could have seen with additional training. Further research is needed to see what best describes the time course of auditory learning in child listeners. Third, the remaining half of the children we tested did not achieve adult-like thresholds even with training. Our results suggest that whether or not a child can achieve adult-like FD thresholds may depend on that child having reached a particular level of cognitive maturation.

ACKNOWLEDGMENTS

The authors would like to thank Emilie Vavasour, Emma Booker, and Katie Dangerfield for assistance with data collection, and Tim Folkard for the software package used for psychophysical testing. Many thanks to Sygal Amitay and to two anonymous reviewers for helpful comments on an earlier version of this manuscript. This research was entirely supported by the Medical Research Council, United Kingdom.

Ahissar, M., Lubin, Y., Putter-Katz, H., and Banai, K. (**2006**). "Dyslexia and the failure to form a perceptual anchor," Nat. Neurosci. **9**, 1558–1564.
Amitay, S., Hawkey, D. J. C., and Moore, D. R. (**2005**). "Auditory fre-

quency discrimination learning is affected by stimulus variability," Percept. Psychophys. **67**, 691–698.
Amitay, S., Irwin, A., and Moore, D. R. (**2006a**). "Discrimination learning induced by training with identical stimuli," Nat. Neurosci. **9**, 1446–1448.
Amitay, S., Nelson, A., Hawkey, D. J. C., Cowan, J., and Moore, D. R. (**2006b**). "A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners," J. Acoust. Soc. Am. **119**, 1616–1625.
Baldeweg, T., Richardson, A., Watkins, S., Foale, C., and Gruzelier, J. (**1999**). "Impaired auditory frequency discrimination in dyslexia detected with mismatch evoked potentials," Ann. Neurol. **45**, 495–503.
Banai, K., and Ahissar, M. (**2004**). "Poor frequency discrimination probes dyslexics with particularly impaired working memory," Audiol. Neuro-Otol. **9**, 328–340.
British Society of Audiology. (**2004**). Recommended Practice "Pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels," http://www.thebsa.org.uk/ (accessed 1 June 2007).
Cacace, A. T., McFarland, D. J., Ouimet, J. R., Schrieber, E. J., and Marro, P. (**2000**). "Temporal processing deficits in remediation-resistant reading-impaired children," Audiol. Neuro-Otol. **5**, 83–97.
Campbell, R. A., and Small, A. M. (**1963**). "Effect of practice and feedback on frequency discrimination," J. Acoust. Soc. Am. **35**, 1511–1514.
Deary, I. J. (**1994**). "Intelligence and auditory-discrimination: Separating processing speed and fidelity of stimulus representation," Intelligence **18**, 189–213.
Delhommeau, K., Micheyl, C., and Jouvent, R. (**2005**). "Generalization of frequency discrimination learning across frequencies and ears: Implications for underlying neural mechanisms in humans," J. Assoc. Res. Otolaryngol. **6**, 171–179.
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.
Fischer, B., and Hartnegg, K. (**2004**). "On the development of low-level auditory discrimination and deficits in dyslexia," Dyslexia **10**, 105–118.
Grimault, N., Micheyl, C., Carlyon, R. P., Bacon, S. P., and Collet, L. (**2003**). "Learning in discrimination of frequency or modulation rate: Generalization to fundamental frequency discrimination," Hear. Res. **184**, 41–50.
Habib, M., Espesser, R., Rey, V., Giraud, K., Bruas, P., and Gres, C. (**1999**). "Training dyslexics with acoustically modified speech: Evidence of improved phonological performance," Brain Cogn **40**, 143–146.
Habib, M., Rey, V., Daffaure, V., Camps, R., Espesser, R., Joly-Pottuz, B., and Demonet, J. F. (**2002**). "Phonological training in children with dyslexia using temporally modified speech: A three-step pilot investigation," Int. J. Lang Commun. Disord. **37**, 289–308.
Halliday, L. F., and Bishop, D. V. M. (**2006**). "Auditory frequency discrimination in children with dyslexia," J. Res. Read. **29**, 213–228.
Hawkey, D. J., Amitay, S., and Moore, D. R. (**2004**). "Early and rapid perceptual learning," Nat. Neurosci. **7**, 1055–1056.
Hill, N. J. (**2001**). "Testing hypotheses about psychometric functions: An investigation of some confidence interval methods, their validity, and their use in assessing optimal sampling strategies," Doctoral dissertation, Oxford University, Oxford.
Hill, P. R., Hogben, J. H., and Bishop, D. V. M. (**2005**). "Auditory frequency discrimination in children with specific language impairment: A longitudinal study," J. Speech Lang. Hear. Res. **48**, 1136–1146.
Irvine, D. R. F., Martin, R. L., Klimkeit, E., and Smith, R. (**2000**). "Specificity of perceptual learning in a frequency discrimination task," J. Acoust. Soc. Am. **108**, 2964–2968.
Jensen, J. K., and Neff, D. L. (**1993**). "Development of basic auditory discrimination in preschool children," Psychol. Sci. **4**, 104–107.
Kaernbach, C. (**2001**). "Slope bias of psychometric functions derived from adaptive data," Percept. Psychophys. **63**, 1389–1398.
Kishon-Rabin, L., Amir, O., Vexler, Y., and Zaltz, Y. (**2001**). "Pitch discrimination: Are professional musicians better than non-musicians?," J. Basic Clin. Physiol. Pharmacol. **12**, 125–143.
Leek, M. R., Hanna, T. E., and Marshall, L. (**1992**). "Estimation of psychometric functions from adaptive tracking procedures," Percept. Psychophys. **51**, 247–256.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Lutfi, R. A., and Wightman, F. (**1996**). "Guessing or confusion? Analytic predictions for two models of target-distractor interference in children," Assoc. Res. Otolaryngol. Abstr. **19**, 142.
Maxon, A. B., and Hochberg, I. (**1982**). "Development of psychoacoustic behavior—sensitivity and discrimination," Ear Hear. **3**, 301–308.
McArthur, G. M., and Bishop, D. V. M. (**2004a**). "Frequency discrimination deficits in people with specific language impairment: Reliability, validity, and linguistic correlates," J. Speech Lang. Hear. Res. **47**, 527–541.
McArthur, G. M., and Bishop, D. V. M. (**2004b**). "Which people with specific language impairment have auditory processing deficits?," Cogn. Neuropsychol. **21**, 79–94.
Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., and Tallal, P. (**1996**). "Temporal processing deficits of language-learning impaired children ameliorated by training," Science **271**, 77–80.
Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A. J. (**2006**). "Influence of musical and psychoacoustical training on pitch discrimination," Hear. Res. **219**, 36–47.
Moore, B. C. J. (**1973**). "Frequency difference limens for short-duration tones," J. Acoust. Soc. Am. **54**, 610–619.
Moore, D. R., and Amitay, S. (**2007**). "Auditory training: Rules and applications," Semin. Hear. **28**, 99–110.
Moore, D. R., Rosenberg, J. F., and Coleman, J. S. (**2005**). "Discrimination training of phonemic contrasts enhances phonological processing in mainstream school children," Brain Lang **94**, 72–85.
Ors, M. (**2002**). "Time to drop 'specific,' in specific language impairment," Acta Paediatr. **91**, 1025–1030.
Psychological Corporation. (**1999**). *WASI: Wechsler Abbreviated Scale of Intelligence* (Psychological Corporation, San Antonio).
Schäffler, T., Sonntag, J., Hartnegg, K., and Fischer, B. (**2004**). "The effect of practice on low-level auditory discrimination, phonological skills, and spelling in dyslexia," Dyslexia **10**, 119–130.
Soderquist, D. R., and Moore, M. J. (**1970**). "Effect of training on frequency discrimination in primary school children," J. Aud. Res. **10**, 185–192.
Spiegel, M. F., and Watson, C. S. (**1984**). "Performance on frequency-discrimination tasks by musicians and nonmusicians," J. Acoust. Soc. Am. **76**, 1690–1695.
Talcott, J. B., Witton, C., Hebb, G. S., Stoodley, C. J., Westwood, E. A., France, S. J., Hansen, P. C., and Stein, J. F. (**2002**). "On the relationship between dynamic visual and auditory processing and literacy skills; results from a large primary-school study," Dyslexia **8**, 204–225.
Tallal, P. (**2004**). "Opinion—Improving language and literacy is a matter of time," Nat. Neurosci. **5**, 721–728.
Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., Schreiner, C., Jenkins, W. M., and Merzenich, M. M. (**1996**). "Language comprehension in language-learning impaired children improved with acoustically modified speech," Science **271**, 81–84.
Temple, E., Deutsch, G. K., Poldrack, R. A., Miller, S. L., Tallal, P., Merzenich, M. M., and Gabrieli, J. D. E. (**2003**). "Neural deficits in children with dyslexia ameliorated by behavioural remediation: Evidence from functional MRI," Proc. Natl. Acad. Sci. U.S.A. **100**, 2860–2865.
Thompson, N. C., Cranford, J. L., and Hoyer, E. (**1999**). "Brief-tone frequency discrimination by children," J. Speech Lang. Hear. Res. **42**, 1061–1068.
Walker, M. M., Shinn, J. B., Cranford, J. L., Givens, G. D., and Holbert, D. (**2002**). "Auditory temporal processing performance of young adults with reading disorders," J. Speech Lang. Hear. Res. **45**, 598–605.
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. (**2001**). "The psychometric function. II. Bootstrap-based confidence intervals and sampling," Percept. Psychophys. **63**, 1314–1329.
Wightman, F., and Allen, P. (**1992**). "Individual differences in auditory capability among preschool children," in *Developmental Psychoacoustics*, edited by L. A. Werner, and E. W. Rubel (American Psychological Association, Washington, DC, **1992**), pp. 113–133.
Wightman, F., 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.
Wright, B. A., and Sabin, A. (**2007**). "Perceptual learning: how much daily training is enough?," Exp. Brain Res. **180**, 727–36.

Estimation of the detection probability for Yangtze finless porpoises (*Neophocaena phocaenoides asiaeorientalis*) with a passive acoustic method

T. Akamatsu
NRIFE, Fisheries Research Agency, Hasaki, Kamisu, Ibaraki 314-0408, Japan

D. Wang, K. Wang, S. Li, S. Dong, and X. Zhao
Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, People's Republic of China

J. Barlow
NOAA Fisheries, Southwest Fisheries Science Center, La Jolla, California 92037

B. S. Stewart
Hubbs-SeaWorld Research Institute, 2595 Ingraham Street, San Diego, California 92109

M. Richlen
Department of Zoology, University of Hawai'i, Edmondson 152, Honolulu, Hawaii 96822

(Received 3 November 2007; revised 10 March 2008; accepted 1 April 2008)

Yangtze finless porpoises were surveyed by using simultaneous visual and acoustical methods from 6 November to 13 December 2006. Two research vessels towed stereo acoustic data loggers, which were used to store the intensity and sound source direction of the high frequency sonar signals produced by finless porpoises at detection ranges up to 300 m on each side of the vessel. Simple stereo beam forming allowed the separation of distinct biosonar sound source, which enabled us to count the number of vocalizing porpoises. Acoustically, 204 porpoises were detected from one vessel and 199 from the other vessel in the same section of the Yangtze River. Visually, 163 and 162 porpoises were detected from two vessels within 300 m of the vessel track. The calculated detection probability using acoustic method was approximately twice that for visual detection for each vessel. The difference in detection probabilities between the two methods was caused by the large number of single individuals that were missed by visual observers. However, the sizes of large groups were underestimated by using the acoustic methods. Acoustic and visual observations complemented each other in the accurate detection of porpoises. The use of simple, relatively inexpensive acoustic monitoring systems should enhance population surveys of free-ranging, echolocating odontocetes. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912449]

PACS number(s): 43.66.Gf, 43.80.Ev, 43.80.Ka [WWA] Pages: 4403–4411

I. INTRODUCTION

Estimating abundance of marine mammals is essential for their conservation and management. Visual observation is the most commonly used method to estimate abundance of aquatic mammals. These animals must surface to breathe and then are visible to ship-based or airborne observers. However, on ship or aerial surveys, not all aquatic mammals surface within the visual range of observers due to relatively long dive times for some species (e.g., Okamura *et al.*, 2006) and avoidance of ships (Richardson *et al.*, 1995). Consequently, an unknown proportion of animals near or on the survey track line are not detected.

Strip or line transect survey methods allow the estimation of total population size based on the incomplete detection of local abundance (Buckland *et al.*, 1993). A key assumption of this method is that all animals within the strip width or on the transect line are detected. This condition is generally not satisfied. However, the detection probability can be calculated by using independent visual observers, which is often based on the same observation platform. Observation events of an individual animal, or a group of animals, by two independent observers are then matched. Based on an assumption of independent sampling, the detection probability of the primary observer can be calculated as the number of matched events over the total number of events observed by the secondary observer (Buckland *et al.*, 1993). In the present study, we employed the strip transect method to compare independent visual and acoustical detections of finless porpoises.

Detection probability is the key to estimate the number of animals. Once the detection probability within a specific distance of the survey track line has been determined, the total number of animals can be estimated from this probability (Buckland *et al.*, 1993). This simple but well established method has been widely applied to assess abundance of marine mammals including blue whales (Calambokidis and Barlow, 2004), humpback whales (Calambokidis *et al.*, 2004), sperm whales (Lewis *et al.*, 2007; Barlow and Taylor, 2005), killer whales (Zerbini *et al.*, 2007), dugongs (Shirakihara *et al.*, 2007), spotted seals (Mizuno *et al.*, 2002), and several species of dolphins and porpoises (de Segura *et al.*, 2006;

Mullin and Fulling, 2004; Hammond *et al.*, 2002) including finless porpoises (Yoshida *et al.*, 1997).

Small odontocetes, such as dolphins and porpoises, are relatively difficult to detect. Group size is difficult to estimate because of the brief periods that animals appear at the surface when breathing and close interanimal distances. Odontocetes swim at speeds of 1.2–5 m/s (Akamatsu *et al.*, 2002; Hanson and Baird, 1998) with a dive duration of 1–3 min and perhaps longer when feeding. This means that dolphins and porpoises can travel several hundred meters underwater without being observed visually at the surface.

Acoustical method can be used as an independent observation to compare to the primary visual observer. Vocalizations of marine mammals stand out from ambient noise (Wartzok and Ketten, 1999; Richardson *et al.*, 1995). They can be detected remotely with passive acoustic methods. Low frequency vocalizations of baleen whales are good candidates for passive acoustic surveys as they propagate relatively great distance underwater. Vocalizations (i.e., song patterns) are often unique to species. Extensive acoustic studies of blue whales (Oleson *et al.*, 2007), right whales (Wade *et al.*, 2006), minke whales (Rankin *et al.*, 2007), and humpback whales (Tiemann *et al.*, 2006) have been conducted.

Not only the presence of the specific species but also additional information could be monitored by acoustical observations. Vocalizations of sperm whales have been helpful in estimating abundance (Barlow and Taylor, 2005) and even documenting dive patterns (Thode, 2004). Passive acoustic methods have applied the identification of multiple species by the characteristics of whistles (Oswald *et al.*, 2007).

High frequency sonar signals of odontocetes have been also used for the observation of odontocetes. For example, a monaural acoustic detection system (T-POD) has been developed that is now commercially available for detecting high frequency sonar pulses of a few species (Philpott *et al.*, 2007; Verfuss *et al.*, 2007). Jefferson *et al.* (2002) applied line transect methods for the survey of finless porpoises simultaneously with T-PODs.

Acoustic detection methods have several advantages over simple visual observations. Detection performance of hardware and software systems can be standardized independent from the observers' abilities. Moreover, they prevent cueing of observers to sightings allowing independent observations between the methods. A hydrophone array determines the distance and direction to a vocalizing animal, which can be directly compared to visual estimates. Because acoustic detection methods do not require human observers, they are useful as an independent detection method during visual transect surveys.

Acoustic detection methods are not without limitations. The probability of detecting animals with passive acoustic methods is affected by the signal-to-noise ratio and by the production rate and the temporal pattern of vocalizations. Vocalizations with low source levels can only be detected when the receiver is close to the source in noisy environments. Animals that are silent for long periods will evade detection.

The echolocation signals of odontocetes are a primary target of passive acoustic detection methods. Source levels of those sounds are up to 170 dB for small porpoises (Akamatsu *et al.*, 2002) and over 220 dB for other species (e.g., bottlenose dolphins; Au, 1993). Yangtze finless porpoises (*Neophocaena phocaenoides asiaeorientalis*) in the semi-natural reserve at Shishou, Hubei, China, produce series of ultrasonic echolocation pulses (i.e., click trains) every 5.1 s on average (Akamatsu *et al.*, 2005a). They rarely travel more than 20 m without vocalizing. Harbor porpoises also produce click trains relatively often (i.e., every 12.3 s; Akamatsu *et al.*, 2007), suggesting that porpoises do not usually travel far without producing detectable sounds. Frequent sound production is essential for effective detection using passive acoustic monitoring systems. Because of these characteristics, porpoises appear to be good candidates for applying passive acoustic monitoring systems while avoiding the few limitations of the method.

II. MATERIALS AND METHODS

A. Acoustic observation

We used acoustic data loggers (i.e., A-tags; ML200-AS2, Marine Micro Technology, Saitama, Japan) to make passive acoustic observations of Yangtze finless porpoises during surveys on the Yangtze River between Yichang and Shanghai, China. The survey was conducted between 6 November and 13 December 2006. The hydrophone sensitivity of the data logger was −201 dB/V at 120 kHz (100–160 kHz within −5 dB band), which is close to the dominant sonar signal frequency of finless porpoises. Each data logger had two hydrophones 11 cm apart to record the difference in the arrival time of each pulse with a resolution of 271 ns. Every 0.5 ms, the logger stored the intensity of the received pulse in the dynamic range of 136.1–160.7 dB peak to peak, which is referred to a 1 μPa reference. The A-tag that we used during this survey had an identical signal processing to the earlier model (W20-ASII; Little Leonardo, Tokyo, Japan; Akamatsu *et al.*, 2005b) but also had a digital detection threshold setting. These A-tags record the difference in time of arrival between sounds received by each hydrophone, which can be used to estimate the conical bearing angle to a sound source.

We made a round-trip survey in two research vessels (*Kekao* and *Honghu*) simultaneously, between Yichang and Shanghai covering the entire habitat of the baiji and Yangtze finless porpoise (Turvey *et al.*, 2007) between 6 November and 13 December 2006. Here, we report acoustic survey data only for the downriver survey (1669 km) because water flow noise contamination was lower when traveling with the river current. As depicted in Fig. 1, one vessel (Kekao) towed two data loggers; the distal one was 2 m ahead of a monaural hydrophone (C54XRS; Cetacean Research Technology, Seattle, WA, USA) on an 87 m cable that included a 7 m proximal extension to a preamplifier (VP2000, Reson, Denmark) that was onboard. This hydrophone was used to monitor low frequency whistles of baiji (*Lipotes vexillifer*) during the survey (Turvey *et al.*, 2007). The second data logger was 17 m ahead of the distal one. We calculated the spatial locations of the porpoises by simple geometric determination when sonar signals were received by both A-tags. Each A-tag stored the

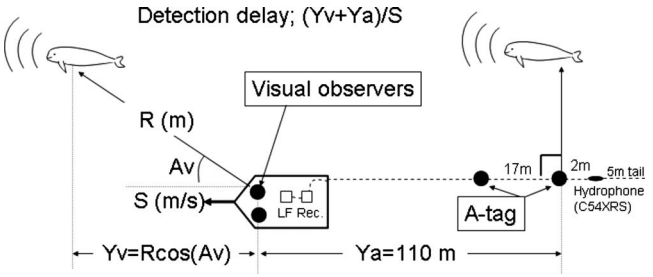


FIG. 1. Two A-tags were towed 110 m behind the visual observers on the vessel (Kelao). Supplemental hydrophone for the low frequency monitoring is placed 2 m behind the distal A-tag. Pictorial representation of parameters used to calculate the expected delay time between visual and acoustic detections. Black circles are the locations of the visual observers and acoustic recorders. Visual detection occurred before the acoustic detection. The delay lag was calculated by using the difference between the distances of visual and acoustic detection along the cruise line. The time of acoustic detection is the zero crossing point at the rear A-tag, which means that the animal was almost perpendicular to the cruise line and abeam of the data logger.

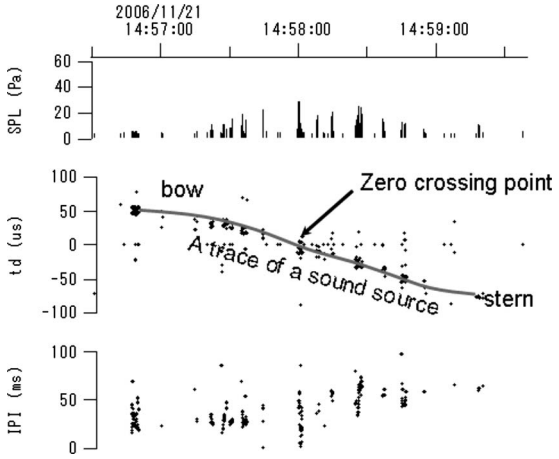


FIG. 2. Echolocation signals from single porpoise passing by the data logger. Top panel: The received sound pressure level (SPL) in Pa; middle panel: the time arrival difference of sonar sounds (T_d) in μs. A trace of the time difference (T_d), indicated as a gray line, changing from positive to negative corresponds to an individual passing from bow to stern relative to the data logger. Lower panel: Interpulse interval in ms. Note that the SPL has a maximum value near the zero crossing point of T_d, which suggests that the porpoise was closest to the data logger at that time.

sound source direction calculated from the time arrival difference of sound between the stereo hydrophone of A-tag. By using two independent angles from the separated two data loggers, location of the sound source could be calculated. The other vessel (Honghu) towed one data logger on a rope 80 m behind the ship.

We added a 5 m length of 5-mm-diameter kremona rope behind the distal data logger on each vessel to stabilize the position of data loggers and to prevent them from swinging. We placed floats at 5 m intervals on the tow cable or rope to prevent the data loggers from dragging on the river bottom. A 2 kg lead weight was fixed 1 m in front of each data logger to keep it approximately 50 cm underwater and prevent surface splashing that would result in broadband noise contamination.

B. Acoustic counting of animals

Biosonar signals from porpoises were identified by their regular interpulse intervals of approximately 30–70 ms (Fig. 2), which is typical of free-ranging finless porpoises (Akamatsu *et al.*, 1998). The source of noise we recorded came mostly from passing cargo vessels and had randomly changing interpulse intervals and intensities unlike the biosonar signals from porpoises. The time arrival difference that corresponded to the bearing angle of porpoise vocalizations always changed from positive to negative (Fig. 2), which meant that the porpoise was passing by the vessel from bow to stern. Because the survey vessels were moving faster (i.e., at 15 km/h) than the average swimming speed of porpoises (4.3 km/h; Akamatsu *et al.*, 2002), none of the animals could catch up with the vessel. The detection time of the animal was defined as the point at which the signal arrival time difference was nearly equal to zero (i.e., the zero crossing point). At that moment, the animal was adjacent to the data logger on a line perpendicular to the cruise line. When two or more porpoises could be discriminated in a group, the time at each zero crossing point was used for the analysis. In the present study, we used the data obtained by the distal A-tag for the Kekao vessel. If sound was detected away from the zero crossing point and the animal was not vocalizing near the data logger, the time of the sonar signal detection was used. To avoid double counting for short traces that were temporally close, we conservatively assumed that traces within 3 min of each other were from the same porpoise. The 3 min duration corresponds to the 750 m distance the vessel proceeds. This is similar to ±300 m, which is the detection distance of the A-tag presented in Sec. III.

C. Visual observation

We made continuous visual observations during daylight hours from the top decks of both vessels. The primary observation team on each vessel consisted of two observers (left and right obsevers) who continually searched for porpoises using 7×50 Fujinon binoculars and occasionally with unaided eyes (Turvey *et al.*, 2007). A data recorder in the middle of the visual observers recorded sighting time, latitude and longitude position, estimated radial distance and bearing to the animals by using an angle board, observer number, group size, distance from the sighting to the nearest river bank, and a code for habitat type (Turvey *et al.*, 2007). Six or seven observers rotated among these positions every half hour and rested for 90–120 min between shifts.

There was one independent observer on watch continuously during daylight hours on each vessel to look for porpoises that may have been missed by the primary observers. The independent observer focused on the area directly ahead of the vessel to guard the track line (Buckland *et al.*, 1993). Two very experienced observers alternated in the independent observer's position every 60 min. Independent observers searched with 25× binoculars on Kekao and with 7× binoculars on Honghu.

For the analysis hereafter, visual detection made within 300 m of the vessel track was used to match the maximum detection distance of acoustic recording system for the comparison of two types of observations.

D. Matching of multimodal detections

To calculate the probability of detection of porpoises, we matched detections made by the primary visual observers with those from the acoustic data loggers to determine if they referred to the same porpoise or porpoise group. Matched detections are defined as the detection of the same animals by both visual and acoustical modes during a particular time window. We could not directly compare time of visual observation (T_v) with the time of acoustic detection (T_a) because porpoises were visually observed only abeam of or ahead of the vessels, whereas they were acoustically detected behind the vessel (Fig. 1). This resulted in a time difference between the two independent detections. The time lag can be estimated based on the distance along the cruise line between the visual detection and the data logger (Y_v+Y_a in Fig. 1) divided by the vessel speed (S, 15 km/s). The standard clocks of visual observers and the acoustic system were set to GPS time. The distance to an animal ahead of the cruise line from the visual observer (Y_v) was calculated from the visually observed distance (R) and the relative angle to the animal (A_v) as

$$Y_v = R\cos(A_v). \tag{1}$$

For both vessels, the distance from the visual observer to the data logger (Y_a) was 110 m, including the cable length (80 m) and the distance from the visual observer to the stern (30 m) of the vessel. The delay time between the visual and acoustic detection of identical animals (T_d) was calculated by using the survey vessel speed (S) as

$$T_d = (Y_v + Y_a)/S. \tag{2}$$

For this, we assumed that the animal did not move far (at 1.2 m/s) compared to the distance traveled by the vessel (at 4.2 m/s) during the period.

We used an arbitrary time window (T_w) for matching detections from the independent methods. Each detection time was assigned to a time bin of T_w s and the number of animals detected in each time bin was summed. If the number of animals in any time bin was one or more for both visual and acoustic detections, the detection was defined as being matched. The matched detection should satisfy the following condition:

$$\text{integer}(T_a/T_w) = \text{integer}[(T_v + T_d)/T_w], \tag{3}$$

where T_v and T_a are the times of visual and acoustic detection and T_d is the expected delay time of the acoustic detection compared to the visual detection of the identical porpoise or porpoise group. We considered any porpoises detected during the time window to belong to the same group. Although this working definition is different from the biological definition of a group, we think that it is most useful for comparing multimodal detections from a moving platform.

E. Estimate of detection probability

The comparison of data obtained by primary observers and that obtained by an independent observer allows the estimation of the probability of detection of porpoises by visual observers (Buckland *et al.* 1993). The total number of animals in the strip transect (N) within 300 m of the vessel track can be calculated by using the number of visual detections (N_v; Fig. 3), the number of acoustic detections (N_a), and the number of detections matched by both methods (m). The number of detections matched by both methods is calculated according to the procedure in the previous section. All these numbers are observable. The total number of animals in the strip transect (N) as well as the detection probability by visual (P_v) and acoustic (P_a) methods are unknown. The number of groups detected acoustically is the total number of animals in the strip transect times the acoustic detection probability,

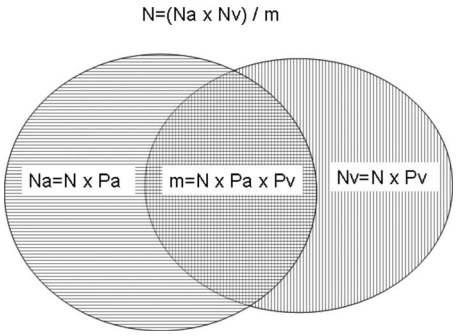


FIG. 3. Simple detection model of two independent observation methods in a strip transect. Here, we assume that the acoustic and visual observers have detected N_a and N_v individuals in the strip transect during the entire survey. The number of matched detections is m. All these parameters are observable, whereas the total number of the target animals in the strip transect (N) and the detection probabilities of two independent methods (P_a and P_v) are possible to calculate.

$$N_a = NP_a. \tag{4}$$

Further, the number of groups detected visually (N_v) is

$$N_v = NP_v. \tag{5}$$

As long as the two observation methods are independent, the number of matched groups (m) is

$$m = NP_aP_v = N(N_v/N)(N_a/N). \tag{6}$$

Here, the total number of animals in the strip transect (N) and each of the detection probabilities using known parameters is

$$N = (N_vN_a)/m, \tag{7}$$

$$P_a = m/N_v, \tag{8}$$

and

$$P_v = m/N_a. \tag{9}$$

III. RESULTS

We counted 204 porpoises from Kekao and 199 from Honghu, by using acoustic methods. In comparison, primary observers detected 163 porpoises from Kekao and 162 from Honghu within 300 m of the vessel track. An example of acoustical and visual detections is depicted in Fig. 4. Many single porpoises were detected acoustically and large group size was detected visually. On November 20 and 21, 2006,

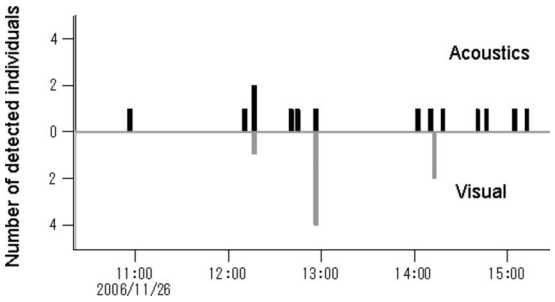


FIG. 4. An example of detection events by acoustics and visual observations from Kekao. Visual detections farther out than the 300 m strip width perpendicular to the cruise line are not included.

the vessels went into Poyang Lake, where some populations of finless porpoises were found. We excluded this period because the towed hydrophone array system was not used due to heavy ship traffic.

A. Observable distance by acoustic systems and an appropriate transect width

We calculated spatial locations of porpoises from a simple geometric determination of the angle of the acoustic signals from each of the two data loggers towed 17 m apart behind Kekao. Simultaneous recording of direction with two data loggers matched the visual detection 49 of the 204 sightings. The maximum detection distance of porpoises using data loggers was 329 m though most porpoises were detected only within 250 m (Fig. 5). The visual observation distances were up to 400 m from the vessels.

We assumed that the observable transect width for our acoustic system was 300 m, which included 95% of the acoustic detections. Consequently, porpoises that were beyond 300 m could not be detected acoustically. Appropriate truncation of distant sightings can reduce the bias of density estimation (Barlow, 1995).

B. Time window

The number of the matched events is related to the duration of the time window. A longer time window results in a greater number of matched events. Increasing the window length too much could potentially result in incorrectly matched groups. A shorter time window produces a lower

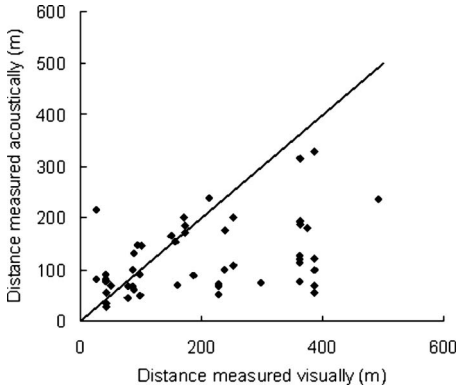


FIG. 5. Comparison of visual and acoustic detection distances for groups that were linked by a detection time window of 120 s.

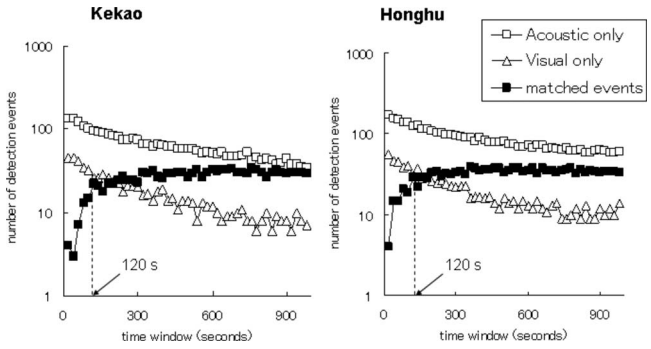


FIG. 6. Number of matched and unmatched detection events. As the time window increases, the number of matched events (black squares) increases while the number of unmatched acoustic and visual events decreases. The number of matched events becomes saturated for a time window of 120 s.

number of matched events and reduces false matching. However, this has the potential of missing matches even if the two independent detections were the same animal.

To determine an appropriate length of the time window, detected numbers of matched events were calculated according to time window lengths ranging from 20 to 1000 s (Fig. 6). For both vessels, the number of matched events (black squares) was quite low when the time window was short (e.g., 20 or 40 s). In this case, many matches are expected to be missed. Matched events increased quickly as the duration of the time window increased up to 120 s, indicating that the number of matching events for identical groups increased. However, the number of matched events became saturated for time window over 140 s. As the time window lengthened, the total number of time bins decreases, whereas the number of matched events including false matches rises. Therefore, the total number of matched events is stable and independent of the time window, even for two detection events that are random and uncorrelated. This means that visual and acoustic detection events are correlated with each other for time windows shorter than 120 s.

C. Detection probability

The calculated probability of acoustic detections was approximately twice that of visual detections for any time window less than 1000 s duration for both vessels (Fig. 7). Acoustic detection probability was consistently greater than that of visual observations regardless of the time window's

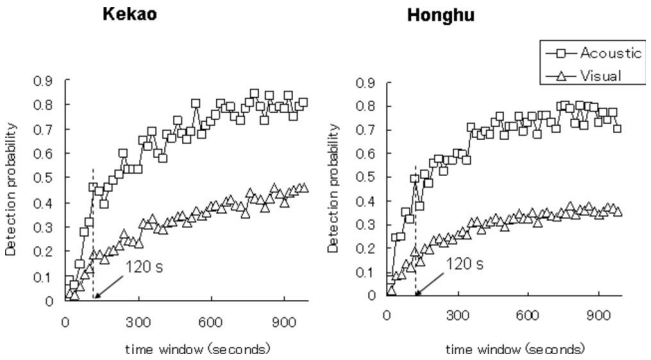


FIG. 7. Detection probability of acoustic (squares) and visual (triangles) observations as a function of the time window width.

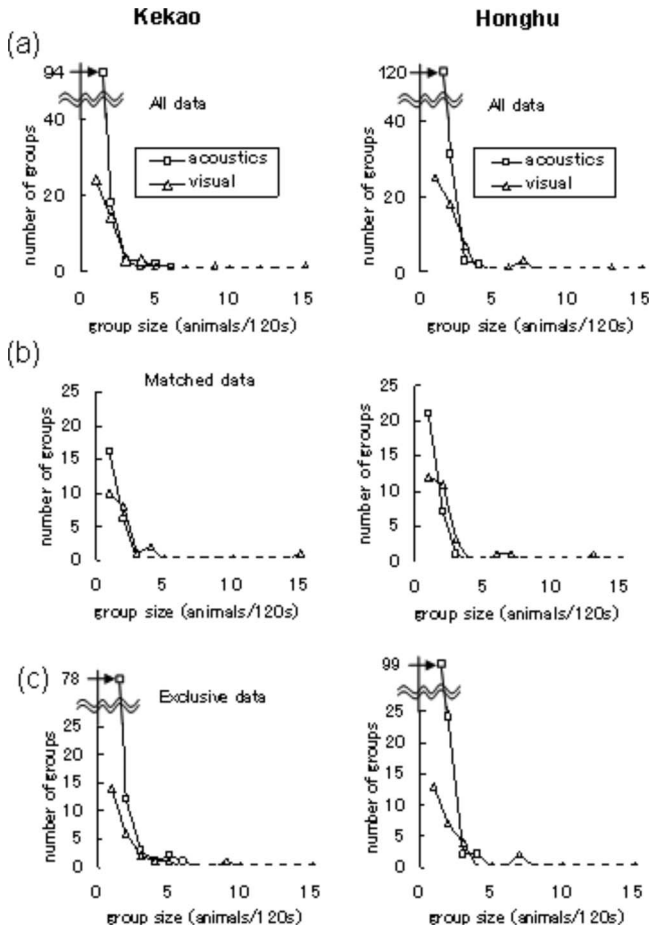


FIG. 8. Comparisons of group size detected acoustically and visually. (a) Accumulated data of all detections show a large difference in the detected number of isolated animals. (b) Matched detection of the two methods that is linked by a 120 s time window. (c) Exclusive data indicate a large difference in the number of detections of single animals depending on method.

duration, though the detection probability for time window lasting more than 140 s may have included false matches of acoustic and visual detections.

D. Group size

We defined group size as the number of porpoises detected during a particular time window. We chose a 120 s time window to compare the estimated group size from the methods. There was a large difference between the acoustic and visual observations for groups of one or two animals [Fig. 8(a)]. This pooled distribution can be resolved into a matched component and an exclusive component [Fig. 8(b) and 8(c)]. The exclusive component is the animals that are detected only by the acoustic or visual method. The matched component shows 50% more acoustic detections than visual detections of single porpoises. For exclusive data, the number of acoustic detections of single porpoise was five times the number from visual observations [Fig. 8(c)].

The large difference in the detection of isolated animals shown in Fig. 8 is the probable cause of the differences in detection probabilities between visual and acoustic methods shown in Fig. 7. To examine this effect, we recalculated the matched number of detections by using only two or more individuals observed in the time window. That is, all of the single animals detected visually and acoustically within a specific time window were ignored. The results are shown in Fig. 9. In this case, the detection probabilities for the two methods were almost the same at any time window. The number of matched events (black squares) gradually increased compared to Fig. 6.

IV. DISCUSSION

The stereo passive acoustic system using A-tag data loggers was successful in detecting and counting finless porpoises in the Yangtze River. The numbers of porpoises counted by the two survey vessels were similar for acoustic (204 vs 199) and visual (163 vs 162) methods. The maximum acoustic detection range of 300 m is approximately double the effective detection distance for finless porpoises using acoustic data loggers reported by Wang *et al.* (2005). In that study, the distance was 150 m with a correct detection level of 77.6% and a false alarm level of 5.8%. Correct acoustic detections did not occur at distances greater than 250 m in their study, possibly because of the less sensitive hydrophone (−210 dB/V). The strip width of 300 m was chosen for the present analysis of acoustic and visual detection based on the maximum acoustical detection distance collected in this study.

The probability of detecting finless porpoises using passive acoustic methods was twice that for visual observations for both vessels during any time window (Fig. 7). This was due to a large difference between the two methods in detecting single porpoises. The acoustic system detected five times more porpoises than did the visual observations [Fig. 8(c)].

Finless porpoises are known to be among the most difficult aquatic animals to detect visually because they are small, lack a dorsal fin, do not jump or porpoise above the water surface, are only slightly darker than the turbid waters of the Yangtze River, and are undetectable from a ship when submerged. Small groups of dolphins and porpoises are generally more difficult to detect compared to large whales. Baleen whales and large toothed whales usually produce large and visible respirations that may linger for several seconds or more and are visible over relatively great distances. Only large schools of dolphins and porpoises are easier to detect visually though estimates of group size may be more difficult.

When we considered detections of two or more porpoises and ignored sightings of solitary porpoises, the detection probabilities were similar for visual and acoustic methods on both vessels (Fig. 9). This indicates that the differences between the estimates of porpoises seen from the two methods was due to the difficulty in visually detecting solitary porpoises.

A. Performance of acoustic detection system

Although the acoustic method was better than the visual method in detecting soliltary animals, acoustic detection of porpoises was limited by its inability to count more than five porpoises simultaneously during a 120 s time window. For animals congregated in a small area, it is difficult to differentiate the sources of sounds from individual porpoises es-

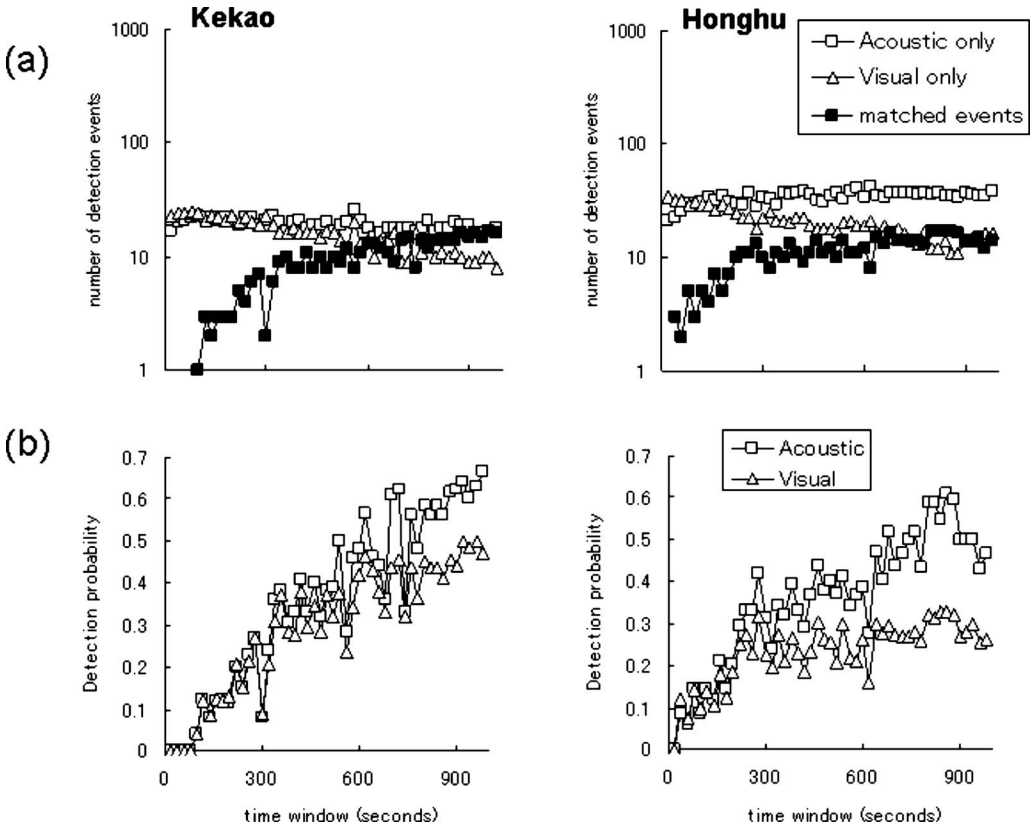


FIG. 9. (a) Number of matched detections using only the data for group size of two or more porpoises. (b) The detection probabilities of visual and acoustic methods for multiple individual groups were similar during all time windows shorter than 300 s. Note that longer time window tends to include two separated detection of animals in single time bin. This caused the total number of detection of multiple individuals occasionally increased according to the time window.

pecially for larger groups. This is evidently due to the short 11 cm base line of the A-tag data logger. The 271 ns time of arrival difference resolution corresponds to a 3.7 MHz sampling frequency, which is sufficient to measure the trigger within a wavelength. However, the level of received pulses was not the same for the two hydrophones on a single data logger. Even within a wavelength, the trigger point of each hydrophone changed. This robust resolution will be improved by using systems with longer base lines.

The acoustic detection distance is influenced by the source level, source directionality, and sound propagation. The source level of finless porpoises is estimated at approximately 163.7–185.6 dB re 1 μPa at 1 m for the on-axis direction (Li *et al.*, 2006), and the sound pressure level for the off-axis beam is 162 dB peak to peak for a 1 μPa reference (Akamatsu *et al.*, 2005c). The detection threshold level of the A-tag data logger was 136.1 dB, which is around 30–50 dB lower than the source levels in different directions from a porpoise. Sound propagation in shallow water systems, such as the Yangtze River, is complex. Our vessel traveled mostly within the shipping lane, which is around 20 m deep, though porpoises could travel in shallow waters along the river banks.

B. Future works

Once the detection probability within the strip transect is obtained, the number of animals within the strip transect can be calculated as the observed number of animals divided by the detection probability. Abundance is the density of the animal within the strip width times the area of the focal sites as long as the density is able to be used outside of the transect width. However, several parameters should be examined before conducting this calculation.

First, the rate of sound production by an animal strongly affects the detection probability. Biologging observation of phonation behavior will help to understand this parameter (Akamatsu *et al.*, 2005a). Second, animal behavior affects the detection probability. For the precise matching between visual and acoustical detections, extrapolation of animal movement during two detections will be needed. Ship avoidance behavior can be observed by the hydrophone array system to identify the sound source. Third, the detection probability is also influenced by the heterogeneity of the independent observers, for example, one may be much better at locating animals than the other. If some of the animals are easy to spot but others are difficult, this also contributes to a heterogeneity bias in the calculated detection probability. To solve this issue, independent double acoustical monitoring will work. When identical systems are operated simultaneously, no heterogeneity is expected. Comparing detections by primarily and secondary acoustical system will provide less biased detection probability.

In conclusion, simple and relatively inexpensive acoustic logging systems like the one we used in this survey of finless porpoises in the Yangtze River should enhance population surveys of cetaceans that vocalize often and travel in relatively small groups. The passive acoustic survey system

worked well for detecting solitary porpoises, which are hard to detect otherwise by visual methods, though it did not work as well for counting porpoises in large groups. Consequently, combining stereo passive acoustic methods with traditional visual observation methods should provide more accurate estimates of population abundance for dolphins and porpoises.

ACKNOWLEDGMENTS

The Yangtze Freshwater Dolphin Expedition 2006 was organized cooperatively by the Institute of Hydrobiology of the Chinese Academy of Sciences, the Administrative Committee of Changjiang Fisheries Resource of the Ministry of Agriculture of China, and the baiji.org Foundation. The acoustic research portion of the expedition was supported by the Program for the Promotion of Basic Research Activities for Innovative Biosciences (ProBRAIN), the Research and Development Program for New Bioindustry Initiatives, Grant-in-Aid for Scientific Research (B) 19405005, National Natural Science Foundation of China (30730018) and the Fisheries Research Agency of Japan. Sponsorship was provided by the SeaWorld-Busch Conservation Fund, Budweiser Wuhan, Anheuser-Busch Inc., SGS, DEZA, BAFU, the Manfred Hermsen Stiftung Foundation, the Hubbs-SeaWorld Research Institute, the Hubbs Society, and the Ocean Park Conservation Foundation of Hong Kong. We specially thank Samuel T. Turvey, Robert L. Pitman, Barbara L. Taylor, Kotoe Sasamori, Leigh A. Barrett, Randall R. Reeves, Zhuo Wei, Xianfeng Zhang, L. T. Pusser, and John R. Brandon for their work on the Yangtze River survey.

Akamatsu, T., Teilmann, J., Miller, L. A., Tougaard, J., Dietz, R., Wang, D., Wang, K., Siebert, U., and Naito, Y. (**2007**). "Comparison of echolocation behaviour between coastal and riverine porpoises," Deep-Sea Res., Part II **54**, 290–297.

Akamatsu, T., Wang, D., Wang, K., and Naito, Y. (**2005a**). "Biosonar behaviour of free-ranging porpoises," Proc. R. Soc. London, Ser. B **272**, 797–801.

Akamatsu, T., Matsuda, A., Suzuki, S., Wang, D., Wang, K., Suzuki, M., Muramoto, H., Sugiyama, N., and Oota, K. (**2005b**). "New stereo acoustic data logger for tagging on free-ranging dolphins and porpoises," Mar. Technol. Soc. J. **39**, 3–9.

Akamatsu, T., Wang, D., and Wang, K. (**2005c**). "Off-axis sonar beam pattern of free-ranging finless porpoises measured by a stereo pulse event data logger," J. Acoust. Soc. Am. **117**, 3325–3330.

Akamatsu, T., Ding, W., Kexiong, W., Zhuo, W., Qingzhong, Z., and Yasuhiko, N. (**2002**). "Diving behavior of freshwater finless porpoises (*Neophocaena phocaenoides*) in an oxbow of the Yangtze River, China," ICES J. Mar. Sci. **59**, 483–443.

Akamatsu, T., Wang, D., Nakamura, K., and Wang, K. (**1998**). "Echolocation range of captive and free-ranging baiji (*Lipotes vexillifer*), finless porpoise (*Neophocaena phocaenoides*) and bottlenose dolphin (*Tursiops truncatus*)," J. Acoust. Soc. Am. **104**, 2511–2516.

Au, W. W. L. (**1993**). *The Sonar of Dolphins* (Springer-Verlag, New York), pp. 133–137.

Barlow, J., and Taylor, B. L. (**2005**). "Estimates of sperm whale abundance in the northeastern temperate Pacific from a combined acoustic and visual survey," Marine Mammal Sci. **21**, 429–445.

Barlow, J. (**1995**). "The abundance of cetaceans in California waters. Part I: Ship surveys in summer and fall 1991," Fish. Bull. **93**, 1–14.

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (**1993**). *Distance Sampling: Estimating Abundance of Biological Populations* (Chapman and Hall, London), pp. 200–217.

Calambokidis, J., and Barlow, J. (**2004**). "Abundance of blue and humpback whales in the eastern North Pacific estimated by capture-recapture and line-transect methods," Marine Mammal Sci. **20**, 63–85.

Calambokidis, J., Steiger, G. H., Ellifrit, D. K., Troutman, B. L., and Bowlby, C. E. (**2004**). "Distribution and abundance of humpback whales (*Megaptera novaeangliae*) and other marine mammals off the northern Washington coast," Fish. Bull. **102**, 563–580.

de Segura, A. G., Crespo, E. A., Pedraza, S. N., Hammond, P. S., and Raga, J. A. (**2006**). "Abundance of small cetaceans in waters of the central Spanish Mediterranean," Mar. Biol. (Berlin) **150**, 149–160.

Hammond, P. S., Berggren, P., Benke, H., Borchers, D. L., Collet, A., Heide-Jorgensen, M. P., Heimlich, S., Hiby, A. R., Leopold, M. F., and Oien, N. (**2002**). "Abundance of harbour porpoise and other cetaceans in the North Sea and adjacent waters," J. Appl. Ecol. **39**, 361–376.

Hanson, M. B., and Baird, R. W. (**1998**). "Dall's porpoise reactions to tagging attempts using a remotely-deployed suction-cup tag," Mar. Technol. Soc. J. **32**, 18–23.

Jefferson, T. A., Hung, S. K., Law, L., Torey, M., and Tregenza, N. (**2002**). "Distribution and abundance of finless porpoises in Hong Kong and adjacent waters of China," Raffles Bulletin of Zoology, Supplement Series, **10**, 43–55.

Lewis, T., Gillespie, D., Matthews, L. J., Danbolt, M., Leaper, R., McLanaghan, R., and Moscrop, A. (**2007**). "Sperm whale abundance estimates from acoustic surveys of the Ionian Sea and Straits of Sicily in 2003," J. Mar. Biol. Assoc. U.K. **87**, 353–357.

Li, S., Wang, D., Wang, K., and Akamatsu, T. (**2006**). "Sonar gain control in echolocating finless porpoises (*Neophocaena phocaenoides*) in an open water," J. Acoust. Soc. Am. **120**, 1803–1806.

Mizuno, A. W., Wada, A., Ishinazaka, T., Hattori, K., Watanabe, Y., and Ohtaishi, N. (**2002**). "Distribution and abundance of spotted seals *Phoca largha* and ribbon seals *Phoca fasciata* in the southern Sea of Okhotsk," Evol. Ecol. Res. **17**, 79–96.

Mullin, K. D., and Fulling, G. L. (**2004**). "Abundance of cetaceans in the oceanic northern Gulf of Mexico, 1996–2001," Marine Mammal Sci. **20**, 787–807.

Okamura, H., Minamikawa, S., and Kitakado, T. (**2006**). "Effect of surfacing patterns on abundance estimates of long-diving animals," Fish. Sci. **72**, 631–638.

Oleson, E. M., Calambokidis, J., Barlow, J., and Hildebrand, J. A. (**2007**). "Blue whale visual and acoustic encounter rates in the southern California bight," Marine Mammal Sci. **23**, 574–597.

Oswald, J. N., Rankin, S., Barlow, J., and Lammers, M. O. (**2007**). "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **122**, 587–595.

Philpott, E., Englund, A., Ingram, S., and Rogan, E. (**2007**). "Using T-PODs to investigate the echolocation of coastal bottlenose dolphins," J. Mar. Biol. Assoc. U.K. **87**, 11–17.

Rankin, S., Norris, T. F., Smultea, M. A., Oedekoven, C., Zoidis, A. M., Silva, E., and Rivers, J. (**2007**). "A visual sighting and acoustic detections of Minke whales, *Balaenoptera acutoroshata* (Cetacea: Balaenopteridae), in nearshore Hawaiian waters," Pacific Science **61**, 395–398.

Richardson, W. J., Greene, Jr., C. R., Malme, C. I., and Thomson, D. H. (**1995**). *Marine Mammals and Noise* (Academic, San Diego), pp. 101–204.

Shirakihara, M., Yoshida, H., Yokochi, H., Ogawa, H., Hosokawa, T., Higashi, N., and Kasuya, T. (**2007**). "Current status and conservation needs of dugongs in Southern Japan," Marine Mammal Sci. **23**, 694–706.

Thode, A. (**2004**). "Tracking sperm whale (*Physeter macrocephalus*) dive profiles using a towed passive acoustic array," J. Acoust. Soc. Am. **116**, 245–253.

Tiemann, C. O., Martin, S. W., and Mobley, J. R. (**2006**). "Aerial and acoustic marine mammal detection and localization on navy ranges," IEEE J. Ocean. Eng. **31**, 107–119.

Turvey, S. T., Pitman, R. L., Taylor, B. L., Barlow, J., Akamatsu, T., Barrett, L. A., Zhao, X., Reeves, R. R., Stewart, B. S., Wang, K., Wei, Z., Zhang, X., Pusser, L. T., Richlen, M., Brandon, J. R., and Wang, D. (**2007**). "First human-caused extinction of a cetacean species?," Biology Letters **3**, 537–540.

Verfuss, U. K., Honnef, C. G., Meding, A., Dahne, M., Mundry, R., and Benke, H. (**2007**). "Geographical and seasonal variation of harbour porpoise (*Phocoena phocoena*) presence in the German Baltic Sea revealed by passive acoustic monitoring," J. Mar. Biol. Assoc. U.K. **87**, 165–176.

Wade, P., Heide-Jorgensen, M. P., Shelden, K., Barlow, J., Carretta, J., Durban, J., Leduc, R., Munger, L., Rankin, S., Sauter, A., and Stinchcomb, C. (**2006**). "Acoustic detection and satellite-tracking leads to discovery of rare concentration of endangered North Pacific right whales," Biology Letters **2**, 417–419.

Wang, K., Wang, D., Akamatsu, T., Li, S., and Xiao, J. (**2005**). "A passive acoustical monitoring method applied to observation and group size esti-

mation of finless porpoises," J. Acoust. Soc. Am. **118**, 1180–1185.
Wartzok, D., and Ketten, D. (**1999**) "Marine Mammal Sensory Systems," in *Biology of Marine Mammals*, edited by Reynolds, III, J. E., and Rommel, S. A. (Smithonian Institution, Washington), pp. 117–175.
Yoshida, H., Shirakihara, K., Kishino, H., and Shirakihara, M. (**1997**). "A population size estimate of the finless porpoise, *Neophocaena phocaenoides*, from aerial sighting surveys in Ariake Sound and Tachibana Bay Japan," Researches on Population Ecology **39**, 239–247.
Zerbini, A. N., Waite, J. M., Durban, J. W., LeDuc, R., Dahlheim, M. E., and Wade, P. R. (**2007**). "Estimating abundance of killer whales in the nearshore waters of the Gulf of Alaska and Aleutian Islands using line-transect sampling," Mar. Biol. (Berlin) **150**, 1033–1045.

Enhancement, adaptation, and the binaural system

Maja Šerman, Catherine Semal, and Laurent Demany[a)]
Laboratoire Mouvement, Adaptation, Cognition (UMR CNRS 5227), BP 63, Université Bordeaux 2, 146 Rue Leo Saignat, F-33076 Bordeaux, France

(Received 7 March 2007; revised 29 February 2008; accepted 3 March 2008)

In a test sound consisting of a burst of pink noise, an arbitrarily selected target frequency band can be "enhanced" by the previous presentation of a similar noise with a spectral notch in the target frequency region. As a result of the enhancement, the test sound evokes a pitch sensation corresponding to the pitch of the target band. Here, a pitch comparison task was used to assess enhancement. In the first experiment, a stronger enhancement effect was found when the test sound and its precursor had the same interaural time difference (ITD) than when they had opposite ITDs. Two subsequent experiments were concerned with the audibility of an instance of dichotic pitch in binaural test sounds preceded by precursors. They showed that it is possible to enhance a frequency region on the sole basis of ITD manipulations, using spectrally identical test sounds and precursors. However, the observed effects were small. A major goal of this study was to test the hypothesis that enhancement originates at least in part from neural adaptation processes taking place at a central level of the auditory system. The data failed to provide strong support for this hypothesis.
 [DOI: 10.1121/1.2902177]

PACS number(s): 43.66.Mk, 43.66.Pn [RYL] Pages: 4412–4420

I. INTRODUCTION

A. Monaural enhancement

An auditory "enhancement" phenomenon occurs when a sound with a given power spectrum (*A*) is followed on the same ear by a second sound with a power spectrum consisting of *A* plus some additional frequency content, *B*. The presentation of the first sound (the "precursor") appears to enhance the detectability, or the perceptual salience, of *B* in the second sound (the "test" sound). This is observable when, for instance, *A* is a sum of harmonics and *B* is another harmonic (Viemeister, 1980; Hartmann and Goupell, 2006), or when *A* is a wideband noise with spectral valleys filled by *B* (Wilson, 1970). Objective evidence for enhancement has been obtained in various experimental paradigms: simultaneous masking (e.g., Viemeister, 1980; Carlyon, 1989), forward masking (Viemeister and Bacon, 1982), pitch matching (Hartmann and Goupell, 2006), and vowel identification (Summerfield *et al.*, 1984, 1987).

What is the origin of these enhancement effects? It has often been hypothesized that they stem from neural adaptation at a relatively peripheral level of the auditory system (Viemeister, 1980; Summerfield *et al.*, 1987; Hicks and Bacon, 1992). The neural response to the components of the test sound that were already present in the precursor could be reduced following the presentation of the precursor, thus increasing the relative prominence of the novel part of the test sound. Palmer *et al.* (1995) found physiological support for this hypothesis in the auditory nerve of guinea pigs. However, this "simple" adaptation cannot account for the psychophysical fact that enhancing a tone increases its forward masking of a subsequent tone (Viemeister and Bacon, 1982). To account for the latter observation, it has been supposed that the precursor adapts (i.e., reduces) the ability of the corresponding part of the test sound to "suppress" (or inhibit) the novel part of the test sound (Viemeister, 1980; Viemeister and Bacon, 1982). However, the physiological study of Palmer *et al.* (1995) did not provide support for that idea at the auditory nerve level. Moreover, the adaptation-of-suppression hypothesis is at odds with psychophysical results reported by Wright *et al.* (1993). Currently, therefore, the contribution of neural adaptation to the perceptual enhancement effects described above is not clear. Carlyon (1989) has argued that their main source is not adaptation. His alternative hypothesis and other ones will be considered later in this paper.

B. Binaural enhancement?

No significant enhancement occurs when the test sound and its precursor are presented to opposite ears (Viemeister, 1980; Summerfield *et al.*, 1984, 1987; Carlyon, 1989; Kidd and Wright, 1994). As pointed out by Kidd and Wright (1994), this does not imply that the mechanism of enhancement is located at a peripheral level of the auditory system, below the level at which the two monaural pathways converge: It could be that the mechanism of enhancement has a central site and is sensitive to interaural relations. More specifically, enhancement phenomena might largely stem from a central form of neural adaptation (CNA). In the past few years, physiologists have uncovered previously unknown forms of adaptation in the auditory system. Ulanovsky *et al.* (2003, 2004) described a highly stimulus-specific form of CNA in the primary auditory cortex of cats. In addition, McAlpine *et al.* (2000), Malone *et al.* (2002), and Furukawa *et al.* (2005) suggested that in the inferior colliculus or the auditory cortex of mammals, where neurons are (broadly)

[a)]Author to whom correspondence should be addressed. Tel.: +33-55757-1651. Fax: +33-55690-1412. Electronic mail: laurent.demany@psyac.u-bordeaux2.fr

tuned to specific interaural phase differences (IPDs) in a given frequency region, the neural response to a given IPD is extremely dependent on the IPDs presented in the recent past. The authors of these investigations emphasized that the CNA revealed by their work increased the contrast between the neural representations of slightly different sounds heard in succession.

In the domain of binaural hearing, some psychophysical phenomena have already been interpreted as a consequence of CNA. A number of authors have reported "repulsive" aftereffects in sound localization or lateralization. Under certain conditions, after the presentation of a sound with some interaural time difference (ITD), the judged location of a second sound with the same spectrum but a different ITD is shifted in the direction opposite to the location of the previous sound (Thurlow and Jack, 1973; Kashino and Nishida, 1998; Phillips and Hall, 2005; Vigneault-MacLean *et al.*, 2007). This might be a genuinely sensory phenomenon due to CNA, although an alternative possibility is a change in response criterion (i.e., a nonsensory bias). In the same vein, Kashino (1998) and Getzmann (2004) reported that the ability to detect a difference between the spatial positions of two successive sounds *X* and *Y* can be improved by the presentation of a precursor identical to *X*. This might again be due to CNA, although other interpretations are possible.

In the study reported here, we attempted to support the hypothesis that one source of enhancement phenomena is CNA. Given that if CNA does exist and has perceptual consequences, it should logically manifest itself in enhancement phenomena, this study was more generally a search for psychophysical correlates of CNA.

II. EXPERIMENT 1

A. Rationale

The goal of this initial experiment was to determine if the binaural system can have an influence on the perceptual enhancement of spectral energy in a given frequency region by a precursor with no energy in that frequency region. In one experimental condition, the test sounds and the precursors were binaural stimuli with the same ITD (either +600 or −600 μs). In another experimental condition, the precursor and test sounds were also binaural stimuli but had opposite ITDs (+600 and −600 μs). There was no difference at all between these two conditions with respect to the *monaural* components of the stimuli. Thus, if enhancement is of a purely peripheral origin, one would have expected it to have the same magnitude in the two conditions. If, on the other hand, enhancement is partially caused by an *ITD-specific* (as well as frequency-specific) form of neural adaptation, then its magnitude ought to have been larger in the "same-ITD" condition than in the "opposite-ITD" condition. Our index of enhancement magnitude was the salience of the pitch sensation evoked by the frequency band which was missing in the precursor but present in the test sound. The audibility of the corresponding pitch was assessed by means of a pitch comparison task.

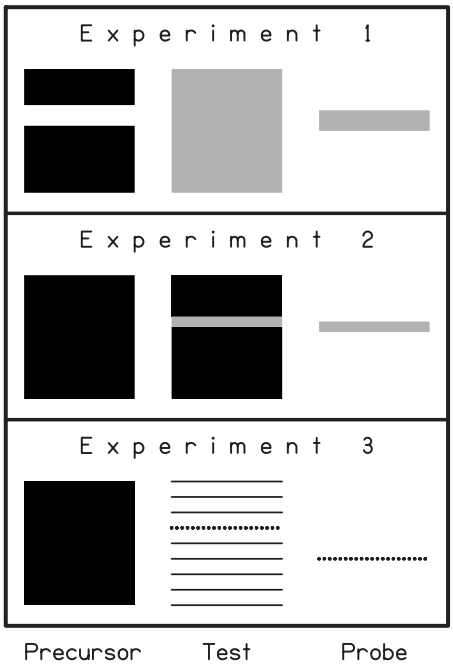


FIG. 1. Schematic illustration of the relationships between precursor, test, and probe sounds in each experiment. The vertical dimension of each panel represents frequency. Differences in binaural lateralization are denoted by a contrast between black and gray areas or between continuous and dotted lines. In experiment 1, the precursor sound was a two-octave band of noise with a 1/3-octave notch, the test sound was the same noise band without the notch, and the probe sound was a 1/3-octave noise band, slightly lower or higher in frequency than the previous notch; in two out of four conditions the precursor differed in lateralization from the two following sounds. In experiment 2, the precursor sound was a two-octave band of noise, the test sound was identical to the precursor except for an ITD change in a 1/6-octave target band, and the probe sound was a 1/6-octave noise band, slightly lower or higher in frequency than the target band; on a given trial, the precursor could be presented four times, presented only once, or not presented. In experiment 3, the precursor sound was again a two-octave band of noise, the test sound was a sum of nine pure tones 1/4-octave apart, and the probe sound was a single pure tone; eight components of the test sound had the same ITD as the precursor sound; the remaining component, with a different ITD, was the target stimulus; the probe sound was matched in frequency with either the target stimulus or some other component of the test sound; on a given trial, the precursor could be either long (1200 ms), short (300 ms), or not presented.

B. Method

1. Stimuli and task

On each trial, subjects were successively presented with (1) a precursor sound (total duration: 1000 ms), (2) a test sound (600 ms), and (3) a probe sound (400 ms). These three sounds—as well as all those used in our subsequent experiments—were gated on and off with 5 ms cosinusoidal amplitude ramps. There was no silent interval (ISI) between the offset ramp of the precursor and the onset ramp of the test sound. A 1200 ms ISI separated the test sound from the probe.

The spectral relations of the precursor, test, and probe sounds are depicted in the upper panel of Fig. 1. Each test sound was essentially a band of pink noise with a width of two octaves. It was generated by adding together 121 synchronous pure tones with equal amplitudes and a frequency spacing of 1/60 octave. The frequencies of the lowest and highest tones were 200 and 800 Hz. Each tone had a nominal

sound pressure level (SPL) of 49 dB, resulting in an overall level of 70 dB SPL. Given that their component tones were very close in frequency (and had random initial phases, see below), the test sounds were perceived as nothing but noise in the absence of a precursor.

The power spectrum of the precursor sounds was identical to that of the test sound, except for the omission of 21 adjacent tones, forming a 1/3-octave frequency band, among the 121 tones making up the test sounds. The spectral position of the omitted frequency band (called the "target band" hereafter) varied randomly from trial to trial without any constraint.

The probe sound presented on a given trial consisted of 21 tones 1/60-octave apart. These tones formed a 1/3-octave band which was identical to the target band except for an overall frequency shift of plus or minus 1/12 octave. The subject's task was to identify the direction of the frequency shift, which varied randomly from trial to trial. Due to the enhancement phenomenon, this pitch comparison task was relatively easy when, for instance, the precursor and test sounds were presented monaurally to the same ear. In contrast, we noted informally that the task became very hard when the test sound was removed.

The component tones of the precursor, test, and probe sounds had new random initial phases on each trial. *Within* a trial, moreover, the initial phase of a tone produced more than once varied randomly from presentation to presentation.

2. Conditions

Subjects were tested in four conditions. In the "monaural" condition, the test sound and its precursor were presented monaurally to the same ear. In the "contralateral" condition, the precursor and test sounds were also monaural stimuli but were presented to opposite ears. In the same-ITD condition, the precursor and test sounds were presented binaurally, with an ITD of 600 μs favoring the same ear; the ITD did not affect the onset and offsets of the stimuli, which were interaurally synchronous. In the opposite-ITD condition, finally, a 600 μs ongoing ITD was also present in the stimuli but favored opposite ears in the precursor and test sounds.

For each condition, the precursor sound was lateralized on the left or right in different subconditions. This created a total of eight subconditions, run in separate blocks of trials. With respect to lateralization, the probe sound was always identical to the test sound.

3. Procedure

The stimuli were digitally generated with a sampling rate of 20 kHz and a 24 bit amplitude quantization. They were presented to the subject through Sennheiser HD265 headphones in a double-walled sound-attenuating booth (Gisol, Bordeaux). Responses ("up" or "down") were given by means of mouse clicks on two labeled virtual buttons and were immediately followed by visual feedback. The formal experiment (following a small number of training sessions) consisted of eight sessions, each including eight blocks of trials—one block for each of the eight subconditions, run in a random order. We wished to process only the data collected on trials in which the target band had been located in the middle part of the test sound's spectrum, more precisely, the frequency region defined by the 61 most central tones (283–566 Hz). Thus, it was decided to terminate every block of trials after a fixed number of trials fulfilling this requirement; the number in question was 30. Overall, therefore, the analyzed data were collected in $30\times8\times2=480$ trials per condition (each condition including two subconditions) and subject.

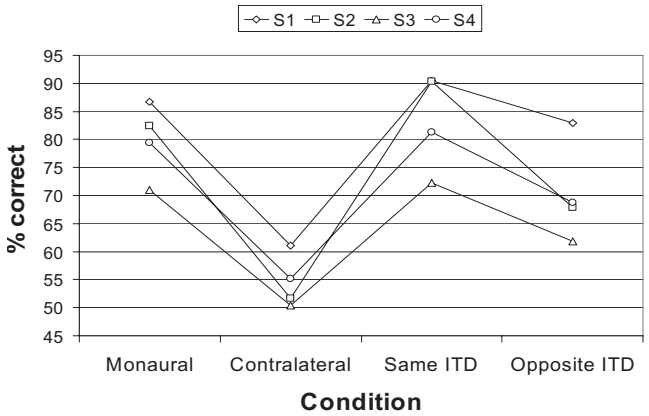


FIG. 2. Performance of each subject (S1, S2, S3, or S4) in the four conditions of experiment 1.

4. Subjects

Four listeners aged between 21 and 53 years (S1, S2, S3, and S4) were tested. All had normal pure-tone audiograms except for S1 who had a 25 dB dip in her left-ear audiogram between about 150 and 250 Hz. The results for this subject did not differ in trend from those of other subjects, neither in the present experiment nor in the subsequent experiments reported here. S1 and S2 were two of the authors.

Given that the experimental task consisted of ordinal pitch comparisons (up and down judgments), it was useful to check that the performance of each subject would not be limited by difficulty in making comparisons of that kind (Semal and Demany, 2006). This was checked by means of a preliminary test including 200 trials in which two successive sounds similar to the probe sounds used subsequently were presented. These two sounds differed by plus or minus (at random) 1/12 octave and the subject had to identify the direction of the corresponding change. S1, S2, S3, and S4 proved to be able to perform perfectly (100% correct) in this test.

C. Results and discussion

Figure 2 displays the percentage of correct responses obtained for each subject in each condition. It can be seen that the four subjects behaved similarly. As expected from previous research (e.g., Summerfield *et al.*, 1987), performance was much better in the monaural condition (mean score: 79.9% correct) than in the contralateral condition (54.6% correct). In the latter case, performance was close to chance. In the same-ITD condition (mean score: 83.6% correct), subjects were slightly more efficient than in the mon-

aural condition. In the opposite-ITD condition (mean score: 70.4% correct), performance was well above chance but definitely poorer than that in the same-ITD condition.

The latter finding is the main outcome. It is at odds with results obtained by Kidd and Wright (1994) in an experiment which had basically the same goal as ours but was quite different methodologically. Kidd and Wright (1994) used a simultaneous masking paradigm in which the signal was a monaural tone burst with a fixed frequency (1000 Hz) and a very short duration (4 ms); the masker, a notched noise, was presented either ipsilaterally, contralaterally, or binaurally with an interaural level difference favoring the contralateral ear. Kidd and Wright's (1994) data led them to suggest that enhancement is essentially a monaural, and thus a peripheral, phenomenon. Contrary to Kidd and Wright's (1994) conclusion, our data indicate that enhancement is significantly dependent on binaural processing. It is presently difficult to see precisely why the two experiments led to discrepant conclusions because their methodologies differed in many ways.

The fact that subjects were more successful in the same-ITD condition than in the opposite-ITD condition shows unambiguously that central factors play a role in enhancement. It is possible that the central influence observed here had something to do with "attention:" The ITD change occurring in the opposite-ITD condition might have "distracted" the subject, perhaps strongly enough to produce the obtained 13.2% drop in performance relative to that in the same-ITD condition. However, this does not seem very likely since, on each trial run in the opposite-ITD condition, the change in ITD was predictable and could be anticipated. A more appealing interpretation of the advantage obtained in the Same-ITD condition is that this advantage originates from an ITD-specific CNA.

The fact that performance was better in the opposite-ITD condition than that in the contralateral condition may also have a central origin since the change in subjective lateralization produced in the opposite-ITD condition was smaller than the change produced in the contralateral condition. However, on the other hand, this result was of course expected under the hypothesis that one source of enhancement is peripheral. The latter hypothesis is neither ruled out nor clearly supported by our data. Some support for it can be found in the slight increase of performance from the monaural condition to the same-ITD condition since two ears offer a statistical advantage over a single ear with regard to the production of a significant monaural enhancement effect.

III. EXPERIMENT 2

A. Rationale

If one source of enhancement phenomena is CNA and if CNA can be ITD specific (as suggested by the physiological studies cited above), then it should be possible to produce an enhancement effect based entirely on ITD manipulations, using spectrally identical precursor and test stimuli. We attempted to do so in experiment 2. The test sounds were similar to those employed in the same-ITD and opposite-ITD conditions of experiment 1 except that within each test sound, a small set of adjacent tones, forming a narrow frequency band (the target band), had an ITD that differed from the ITD of the other tones (the "background" tones). The target band was thus liable to evoke a pitch sensation corresponding approximately to the pitch of its center frequency; previous studies on this instance of "dichotic pitch" were reported by Dougherty *et al.* (1998) and Akeroyd and Summerfield (2000). In the absence of a precursor, however, this pitch was not easily heard. We attempted to enhance it by a precursor consisting of the same tones as those forming the test sound but in which all tones had the ITD of the subsequent background tones, as illustrated in the middle panel of Fig. 1. The precursor was intended to adapt specifically the background components of the test sound and to increase in this way the salience of the target band's pitch. We reasoned that if the benefit of the precursor was due to CNA, one would expect to obtain a larger benefit from repeating the precursor several times before the test sound than from presenting it only once. Indeed, the physiological studies of Ulanovsky *et al.* (2003, 2004) indicate that even though a single stimulus repetition is sufficient to observe a substantial amount of CNA, the effect increases with the number of repetitions; CNA appears to be a cumulative process. This led us to use three conditions in which the precursor was, respectively, presented four times, presented only once, and not presented.

B. Method

1. Stimuli and task

In each test sound, consisting again of 121 synchronous pure tones spaced by 1/60-octave intervals and ranging from 200 to 800 Hz, 11 adjacent tones formed the target band (width: 1/6 octave). These tones had an ITD (ITD_{target}) that differed from the ITD of the other tones ($ITD_{background}$). One of these two "ITDs" was in fact equal to zero; the other ITD favored the right ear and was subject dependent (more details in the next section); it did not affect the onsets and offsets of the stimuli. The spectral position of the target band varied randomly from trial to trial without any constraint. As before, each tone had a nominal SPL of 49 dB and a random initial phase (renewed from trial to trial) at a given ear. The test sounds had a total duration of 400 ms.

On each trial, after a 1200 ms ISI, the test sound was followed by a 400 ms probe sound which was, as in experiment 1, a transposition of the target band at a frequency distance of 1/12 octave; the ITD of this probe was equal to ITD_{target}. Again, the subject's task was to identify the direction of the frequency shift, this direction varying randomly from trial to trial, and visual feedback was provided following each response.

On a given trial, the test sound was preceded by either zero, one, or four presentations of a 200 ms precursor sound. The precursor did not differ from the test sound with respect to the power spectrum, but the ITD of *all* its component tones was equal to $ITD_{background}$. There was a 200 ms ISI between the final (or single) precursor presentation and the test sound. When the precursor was repeated, its presentations were also separated by 200 ms ISIs.

2. Procedure and subjects

The three conditions (zero, one, or four precursor presentations) were run in separate blocks of 50 trials. In contrast to experiment 1, we did not restrict data analysis to the trials on which the target band had been located in the middle part of the test sound's spectrum. Every block included 25 trials in which ITD_{target} was zero while $ITD_{background}$ was not zero and 25 trials in which the opposite was true. Any trial of one of these two types was always followed and/or preceded by a trial of the other type. This was done in order to avoid possible benefits of *across-trial* CNA in the blocks without precursor. In such blocks, had $ITD_{background}$ been fixed, the background components of the test sound presented on a given trial could have advantageously served as an adaptor to the background components of the test sound presented in the next trial.

In the experiment proper, each condition was run in eight blocks of trials (400 trials overall) per subject. The order of conditions was randomized within sessions. The experiment proper was preceded by preliminary training sessions during which, for each subject, the size of the nonzero ITD was varied in order to find an ITD value that avoided floor and ceiling effects. Except for this variation of the nonzero ITD, the preliminary training sessions did not differ from the following formal sessions.

Five listeners were tested. Three of them (S1, S2, and S3) also served as subjects in experiment 1. The two new subjects (S5 and S6) were in their 20s and had normal audiograms at both ears. Both of them performed without any error the preliminary pitch comparison test described in Sec. II B 4. The nonzero ITD value used in the experiment proper was 150 μs for S1, 500 μs for S2, and 200 μs for S3, S5, and S6.

C. Results and discussion

The scores obtained within blocks of trials were submitted to a three-way analysis of variance (ANOVA) [subject $\times$(number of precursors) $\times$ ITD_{target} (zero versus nonzero)]. This ANOVA indicated that the number of precursors had a highly significant main effect [$F(2,210)=23.5$, $P<0.001$] and did not interact significantly with the subject factor [$F(8,210)=1.2$, $P=0.32$] nor with ITD_{target} [$F(2,210)<1$]. The main effect of ITD_{target} was at the limit of statistical significance [$F(1,210)=4.1$, $P=0.04$] and there was a significant interaction between this factor and the subject factor [$F(4,210)=8.1$, $P<0.001$]. Figure 3 shows the effect of the number of precursors, with performance collapsed across the ITD_{target} variable. The mean percentage of correct responses was 72.3% when no precursor was presented, 79.3% for one precursor presentation, and 80.6% for four presentations. So, although presenting a precursor had a benefit, this benefit was essentially the same for one and four presentations. Overall, performance was slightly better when ITD_{target} differed from 0 (mean percentage of correct responses: 78.5%) than when ITD_{target} was equal to 0 (mean percentage: 76.4%); this trend is consistent with observations by Hartmann and Zhang (2003) concerning the audibility of the "Huggins pitch" (Cramer and Huggins, 1958) for low-frequency target bands.

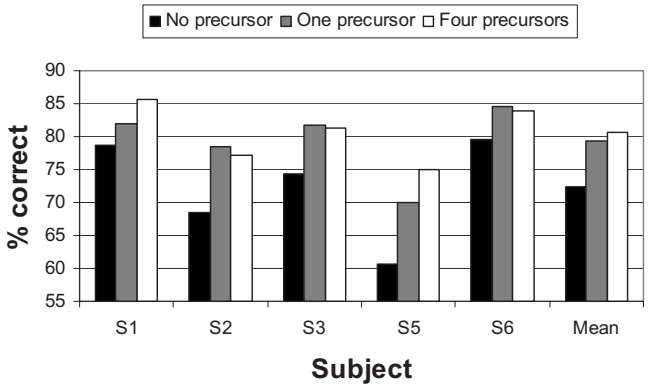


FIG. 3. Performance of each subject in the three conditions of experiment 2. For each condition, performance is collapsed across the ITD_{target} variable. The means across subjects are also displayed.

We can conclude from these results that it is possible to produce a perceptual enhancement effect based entirely on ITD manipulations, without adding new spectral energy to the precursor. The existence of this effect is consistent with the hypothesis that some aspects of human auditory perception are under the influence of CNA. However, CNA (like other forms of neural adaptation) is supposed to be a cumulative process (see, e.g., Ulanovsky *et al.*, 2003, 2004). Thus, the fact that we did not observe a significantly stronger enhancement after four presentations of the precursor than after a single one puts into question the role of CNA in the present experimental situation.

It must also be noted that the benefit of the precursor was only modest. In this respect, a natural conjecture is that a stronger benefit could have been found if the ISI separating the final (or single) precursor presentation from the test sound had been shorter than 200 ms. The reason why we chose this ISI rather than a much shorter one (or no ISI at all) is that, if the ISI had been very short, the benefit of the precursor would have been likely to arise in part from factors unrelated to CNA. Culling (2000) pointed out that, given the well-known "sluggishness" of the binaural system (e.g., Grantham and Wightman, 1978), a rapid ITD change in some frequency region is detectable not as an ITD change *per se* but rather as a decorrelation of the two monaural inputs in that frequency region, leading to the perception of the corresponding pitch. In the present experiment, we wanted to prevent subjects from using such a cue in order to identify the target band when a precursor preceded the test sound.

IV. EXPERIMENT 3

A. Rationale

In experiment 2, the precursor sounds had only a weak influence on performance even when they were repeated before the test sounds. This may have been due to their rather short duration (200 ms). We had thought that the repetition of a short precursor could be more efficient than only one presentation of a longer precursor because the binaural system is particularly sensitive to the onset of sounds (see, e.g.,

Hafter *et al.*, 1988). However, this may have been a wrong idea. In experiment 3, we used longer precursors with the hope of observing a stronger effect.

Experiment 3 was also motivated by the fact that, in experiment 2, the benefit of a precursor could, at least in theory, originate from factors unrelated to CNA. One such factor could be "timbral cuing:" The precursor indicated precisely to the subject, just before the test stimulus how the test stimulus, would sound if all its spectral components had the same ITD; the precursor thus provided a potentially useful timbral reference in auditory memory (see in this regard McFadden, 1966; Demany and Semal, 2008). Another possibility was sequential grouping. Carlyon (1989) suggested that this is the main source of enhancement. The idea is that when a sound (the precursor) is rapidly followed by an identical sound (the background of the test sound, i.e., the test sound minus the target), these two stimuli are automatically grouped into a single auditory stream; therefore, it is easier to perceive the target as a separate "auditory object" than when the precursor is absent or greatly differs from the test sound's background. In order to account for the results of experiment 2 with this scenario, one must assume that sequential grouping processes can produce a stream segregation effect on the basis of nothing but ITD cues. Contrary to this assumption, Darwin and Hukin (1999) have argued that "listeners do not explicitly track [spectral] components that share a common ITD." Nevertheless, it was desirable to determine if binaural enhancement can be produced with stimuli for which sequential grouping processes are unlikely to play a role (whereas CNA could have a large effect). The stimuli used in experiment 3 fulfilled this requirement. At the same time, they were such that enhancement could not be interpreted in terms of timbral cuing. The experiment was largely inspired by Summerfield *et al.* (1987), who (briefly) reported a comparable experiment on the monaural enhancement of newly arriving spectral energy.

B. Method

New test sounds were used, as well as a new task. As illustrated in the bottom panel of Fig. 1, each test sound was the sum of nine equal-amplitude pure tones, spaced by intervals of 1/4 octave and ranging in frequency from 200 to 800 Hz. The spacing of the tones was such that they were resolvable by the auditory system. The test sounds were thus perceived as tonal stimuli (chords). On each trial, eight components of the presented test sound had the same ITD ($ITD_{background}$) and the remaining component—the target—had a different ITD (ITD_{target}). The target tone was selected at random among the seven "inner" components (i.e., the components with frequencies differing from 200 and 800 Hz). The test sound was followed by a probe which was a single pure tone, equiprobably identical to the target tone or matched in frequency to some other inner component of the test sound (selected at random among the six candidates). The ITD of the probe was always equal to ITD_{target}. The task was to indicate if the probe tone differed from the target tone or not.

As in experiment 2, the precursor sounds consisted of 121 tones spaced by 1/60 octave and ranging from 200 to 800 Hz. All these tones had again an ITD equal to $ITD_{background}$. Since they were not resolvable by the auditory system, the precursor sounds were perceived as noise, whereas the test sounds were perceived as tonal chords. Therefore, when a test sound was preceded by a precursor, the background components of the test sound were not liable to be perceived as a repetition (or quasirepetition) of a portion of the precursor. In addition, since the test sound was separated from the precursor by a 100 ms silent ISI, the sequence was not liable to induce a "retrospective continuity illusion" (Carlyon *et al.*, 2005).

The precursor sounds and test sounds had the same overall level, 70 dB SPL. Thus, whereas the component tones of the precursors had (as before) a SPL of 49 dB, the SPL of the test sounds' component tones was 60.5 dB. Test sounds and probe tones had a duration of 200 ms. As mentioned above, the ISI between precursor and test sounds was 100 ms. The ISI between test and probe stimuli was 1200 ms.

On a given trial, the test sound was preceded by either a "long" precursor (1200 ms), a "short" precursor (300 ms), or no precursor at all. These three conditions were run in separate blocks of 40 trials (12 blocks for each condition in the experiment proper). Within each block, there were 20 trials in which $ITD_{background}=0$ and $ITD_{target}\neq 0$ (the right ear leading) and 20 trials in which the opposite was true. As in experiment 2 (and for the same reason), any trial of one of these two types was always followed and/or preceded by a trial of the other type. The order of conditions was randomized within sessions.

Five listeners were tested, among whom four (S1, S2, S3, and S4) had participated as subjects in at least one of the previous experiments. The fifth subject, S7, was in his 20s and had a normal audiogram at each ear. Following some preliminary sessions in which the nonzero ITD value was varied adaptively, this parameter was set to 110 μs for S1, 600 μs for S2, 200 μs for S3, 350 μs for S4, and 150 μs for S7.

C. Results

A three-way ANOVA of the scores measured within blocks of trials [subject×precursor duration (0,300, or 1200 ms)×ITD_{target} (zero versus nonzero)] revealed significant main effects of precursor duration [$F(2,330)=23.0$, $P<0.001$] and ITD_{target} [$F(1,330)=15.1$, $P<0.001$] but no significant interaction between these two factors [$F(2,330)<1$]. There was a significant interaction between the subject and precursor duration factors [$F(8,330)=2.2$, $P=0.03$] and a significant three-way interaction [$F(8,330)=2.0$, $P=0.05$]. As in experiment 2, overall, performance was better when ITD_{target} differed from 0 (mean percentage of correct responses: 78.6%) than when ITD_{target} was equal to 0 (mean percentage: 74.4%). The results concerning precursor duration, collapsed across the ITD_{target} variable, are displayed in Fig. 4. Comparisons between the three conditions with the Holm-Sidak method showed that the mean score obtained for each condition differed significantly ($P<0.05$) from the mean score obtained for each of the other two conditions.

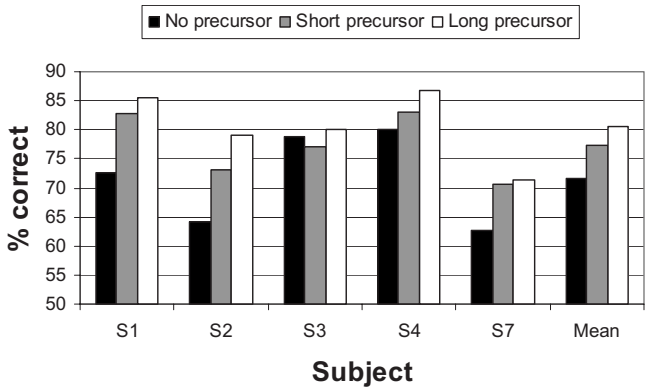


FIG. 4. Performance of each subject in the three conditions of experiment 3. For each condition, performance is collapsed across the ITD_{target} variable. The means across subjects are also displayed.

These differences were in the direction predicted by the CNA hypothesis. Adding long (1200 ms) precursors to the test sounds improved performance by 8.8% on average. Performance was also globally better in the long precursor condition than in the short precursor condition; however, this difference was quite small (3.2%).

V. GENERAL DISCUSSION

For any living organism, it is crucially important to detect *changes* in the environment. The existence of sensory mechanisms enhancing the internal representation of a new stimulus in a background of previously perceived stimuli is thus clearly profitable. The present research shows that, in humans, the binaural system—more specifically the processing of ITDs—can play a role in this enhancement. Experiment 1 indicated that it can modulate the enhancement of newly arriving spectral energy. Moreover, experiments 2 and 3 indicated that it is possible to enhance a pitch percept on the sole basis of ITD manipulations, without any spectral change. In the latter case, however, the observed effects were small.

It has been found in several previous studies (McFadden, 1966; Robinson and Trahiotis, 1972; Yost, 1985; Bernstein *et al.*, 2006) that the binaural detection of an interaurally antiphasic tonal burst ($S\pi$) added to a diotic burst of white noise ($N0$) can be easier when $N0$ and $S\pi$ are preceded by a "forward fringe" consisting of a sample of $N0$ alone than when $N0$ and $S\pi$ are simply pulsed synchronously in the absence of a fringe. One might think that this observation is closely related to what we found in experiments 2 and 3, but that is probably not the case. In the just-mentioned masking experiments, the signal ($S\pi$) had a fixed frequency across trials and it was superimposed on an independent noise, thus producing a *time-varying* ITD; it appeared that a forward fringe improved signal detection significantly only when the signal was very short. In contrast, the target sounds that we used in experiments 2 and 3 varied randomly from trial to trial and all the components of the test sounds had a *steady* ITD; moreover, we found informally that the duration of the test sounds was not a critical parameter with respect to enhancement. It should also be noted that CNA was never considered as a possible explanation for the benefit of forward fringes in the masking experiments using $N0S\pi$ configurations; Bernstein *et al.* (2006) proposed a quite different scenario.

Another previous study, by Kubovy and Howard (1976), is more closely related to the present work. Kubovy and Howard (1976) generated sequences of binaural chords in which each chord consisted of six synchronous pure tones with different IPDs. The tones had the same frequencies, ranging from 392 to 659 Hz and forming a diatonic musical scale, in all chords. In the initial chord, the IPDs were an arbitrary function of frequency. Each of the subsequent chords was identical to the first chord, except for a modification in the IPD of a single tone. The tone with the modified IPD changed from chord to chord in a sawtooth manner, going gradually from 392 to 659 Hz in some sequences and vice versa in other sequences. It appeared that, in such sequences, listeners were able to track the tone with the modified IPD from chord to chord; as a result, they perceived an ascending or descending cyclical melodic pattern. For a majority of the tested listeners, this task was feasible even when the chords were separated by an ISI as long as 1 s (but not much longer than that). An important point is that the modified IPDs had no special characteristic which would allow the listener to identify the target component of a chord in this chord alone; some kind of memory of the initial IPDs (generally repeated from chord to chord) was necessary in order to identify and track the target tones.

Using chord sequences which were similar but not identical to the sequences of Kubovy and Howard (1976), Culling (2000) obtained different results. He found that ascending and descending progressions of the target tone (with the modified IPD) were easy to discriminate from each other when the ISI was very short (0–20 ms), but that for most listeners, discrimination became nearly impossible for an ISI of 160 ms. The reason why discrepant results were obtained in these two experiments is not clear. In any case, the results of Kubovy and Howard (1976) are trustworthy. Just before the present study, we replicated their experiment with exactly the same stimuli. The outcome was consistent with their results. We tested five listeners, all of whom gave more than 70% of correct responses ("ascending" versus "descending") for a 200 ms ISI.

The perceptual phenomenon described by Kubovy and Howard (1976) is, we believe, a case of binaural enhancement (in which each chord, except for the very first one, serves both as a precursor sound and as a test sound). However, their experiment does not clearly demonstrate the existence of binaural enhancement. The mere fact that in their sequences of chord, it is possible to discriminate between ascending and descending progressions of the target tone does not prove that, within a chord, the target tone is more audible than the other tones. An alternative interpretation is that all tones can be heard out equally clearly and that the change in IPD simply serves as a pointer to the tone on which the listener's attention should be focused. In our experiments 2 and 3, by contrast, the audibility of the target sounds was assessed in the absence as well as the presence of precursor sounds, and a comparison was made between the two corresponding data sets.

In fact, Kubovy and Howard (1976) did not describe their finding as a case of enhancement and did not explicitly consider CNA as a possible explanation. Their conclusion was instead—or more vaguely—that there exists a form of auditory memory that is "in the service of perception, a *percept-forming* memory or in particular a *pitch-segregating* memory" (pp. 536–537). The present research, on the other hand, was primarily intended to test the idea that CNA is a source of enhancement and thus plays a role in human auditory perception. Our results provide some support to that idea but not a strong one; moreover, they suggest that, if it does exist, CNA has only small effects.

As already pointed out above, the presentation of a precursor sound before a target sound may improve the perception of the target sound for a number of reasons that have nothing to do with CNA. Generally speaking, the precursor may serve as an attentional cue. It may also be automatically grouped with the background of the target sound, thus isolating the target itself from irrelevant stimulation (see in this regard Best *et al.*, 2007). Yet another possibility, specific to binaural stimuli, is the interaural decorrelation process hypothesized by Culling (2000) and mentioned in Sec. III C. Our best evidence for a role of CNA is probably that obtained in experiment 3 because here, timbral cuing was ruled out, as well as decorrelation (due to the 100 ms ISI) and sequential grouping (the background components of the test sound could not be perceived as a repetition or an extension of part of the precursor). In that experiment, however, the effect of 1200 ms precursors was barely different from the effect of precursors which were four times shorter (300 ms), a problematic fact for the CNA hypothesis.

It is conceivable, of course, that we would have observed stronger binaural enhancement effects if we had used precursors lasting several seconds. Very long "adapter" (i.e., precursor) stimuli have been employed in the experiments concerning the effect of a sound on the subjective lateralization of a subsequent sound (e.g., Kashino and Nishida, 1998). However, very long precursors are not required in order to obtain monaural enhancement of newly arriving spectral energy (see, e.g., Viemeister, 1980). Moreover, if ITD-specific CNA is an extremely slow process, then this form of CNA is unlikely to play an important role in ordinary listening situations because people often move their head or their whole body.

ACKNOWLEDGMENTS

The research reported here was carried out while M.S. held a position of "chercheur associé étranger" at the CNRS; this author is now affiliated with Siemens Audiological Engineering Group (Erlangen, Germany). We thank Dr. Constantine Trahiotis for a very helpful discussion. Dr. Wesley Grantham and two anonymous reviewers made constructive comments on a previous version of this article.

Akeroyd, M. A., and Summerfield, Q. (**2000**). "The lateralization of simple dichotic pitches," J. Acoust. Soc. Am. **108**, 316–334.

Bernstein, L. R., Trahiotis, C., and Freyman, R. L. (**2006**). "Binaural detection of 500-Hz tones in broadband and in narrowband masking noise: Effects of signal/masker duration and forward masking fringes," J. Acoust. Soc. Am. **119**, 2981–2993.

Best, V., Gallun, F. J., Carlile, S., and Shinn-Cunningham, B. G. (**2007**). "Binaural interference and auditory grouping," J. Acoust. Soc. Am. **121**, 1070–1076.

Carlyon, R. P. (**1989**). "Changes in the masked thresholds of brief tones produced by prior bursts of noise," Hear. Res. **41**, 223–236.

Carlyon, R. P., Deeks, J. M., Grahn, J., Shtykov, Y., Hauk, F., and Pulvermuller, F. (**2005**). "The mysterious case of elephant noise: A retrospective continuity illusion?," *Poster presented at the Meeting of the Bristish Society of Audiology, Cardiff, UK.*

Cramer, E. M., and Huggins, W. H. (**1958**). "Creation of pitch through binaural interaction," J. Acoust. Soc. Am. **30**, 413–417.

Culling, J. F. (**2000**). "Auditory motion segregation: A limited analogy with vision," J. Exp. Psychol. Hum. Percept. Perform. **26**, 1760–1769.

Darwin, C. J., and Hukin, R. W. (**1999**). "Auditory objects of attention: The role of interaural time-differences in attention to speech," J. Exp. Psychol. Hum. Percept. Perform. **25**, 617–629.

Demany, L., and Semal, C. (**2008**). "The role of memory in auditory perception," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 77–113.

Dougherty, R. F., Cynader, M. S., Bjornson, B. H., Edgell, D., and Giaschi, D. E. (**1998**). "Dichotic pitch: A new stimulus distinguishes normal and dyslexic auditory function," NeuroReport **9**, 3001–3005.

Furukawa, S., Maki, K., Kashino, M., and Riquimaroux, H. (**2005**). "Dependency of the interaural phase difference sensitivities of inferior collicular neurons on a preceding tone and its implications in neural population coding," J. Neurophysiol. **93**, 3313–3326.

Getzmann, S. (**2004**). "Spatial discrimination of sound sources in the horizontal plane following an adapter sound," Hear. Res. **191**, 14–20.

Grantham, D. W., and Wightman, F. L. (**1978**). "Detectability of varying interaural temporal differences," J. Acoust. Soc. Am. **63**, 511–523.

Hafter, E. R., Buell, T. N., and Richards, V. M. (**1988**). "Onset-coding in lateralization: its form, site and function," in *Auditory Function: Neurological Basis of Hearing*, edited by G. Edelman, M. Cohen, and E. Gall (Wiley, New York), pp. 647–676.

Hartmann, W. M., and Goupell, M. J. (**2006**). "Enhancing and unmasking the harmonics of a complex tone," J. Acoust. Soc. Am. **120**, 2142–2157.

Hartmann, W. M., and Zhang, P. X. (**2003**). "Binaural models and the strength of dichotic pitches," J. Acoust. Soc. Am. **114**, 3317–3326.

Hicks, M. L., and Bacon, S. P. (**1992**). "Factors influencing temporal effects with notched-noise maskers," Hear. Res. **64**, 123–132.

Kashino, M. (**1998**). "Adaptation in sound localization revealed by auditory after-effects," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 322–328.

Kashino, M., and Nishida, S. (**1998**). "Adaptation in sound localization revealed by auditory after effects," J. Acoust. Soc. Am. **103**, 3597–3604.

Kidd, G. Jr., and Wright, B. A. (**1994**). "Improving the detectability of a brief tone in noise using forward and backward masker fringes: Monotic and dichotic presentations," J. Acoust. Soc. Am. **95**, 962–967.

Kubovy, M., and Howard, F. P. (**1976**). "Persistence of a pitch-segregating echoic memory," J. Exp. Psychol. Hum. Percept. Perform. **2**, 531–537.

Malone, B. J., Scott, B. H., and Semple, M. N. (**2002**). "Context-dependent adaptive coding of interaural phase disparity in the auditory cortex of awake macaques," J. Neurosci. **22**, 4625–4638.

McAlpine, D., Jiang, D., Shackleton, T. M., and Palmer, A. R. (**2000**). "Responses of neurons in the inferior colliculus to dynamic interaural phase cues: Evidence for a mechanism of binaural adaptation," J. Nondestruct. Eval. **83**, 1356–1365.

McFadden, D. (**1966**). "Masking-level differences with continuous and with burst masking noise," J. Acoust. Soc. Am. **40**, 1414–1419.

Palmer, A. R., Summerfield, Q., and Fantini, D. A. (**1995**). "Responses of auditory nerve fibres to stimuli producing psychophysical enhancement," J. Acoust. Soc. Am. **97**, 1786–1799.

Phillips, D. P., and Hall, S. E. (**2005**). "Psychophysical evidence for adaptation of central auditory processors for interaural differences in time and level," Hear. Res. **202**, 188–199.

Robinson, D. E., and Trahiotis, C. (**1972**). "Effects of signal duration and masker duration on detectability under diotic and dichotic listening conditions," Percept. Psychophys. **12**, 333–334.

Semal, C., and Demany, L. (**2006**). "Individual differences in the sensitivity to pitch direction," J. Acoust. Soc. Am. **120**, 3907–3915.

Summerfield, A. Q., Haggard, M. P., Foster, J. R., and Gray, S. (**1984**). "Perceiving vowels from uniform spectra: Phonetic exploration of an auditory after effect," Percept. Psychophys. **35**, 203–213.

Summerfield, A. Q., Sidwell, A., and Nelson, T. (**1987**). "Auditory enhancement of changes in spectral amplitude," J. Acoust. Soc. Am. **81**, 700–708.
Thurlow, W. R., and Jack, C. E. (**1973**). "Some determinants of localization-adaptation effects for successive auditory stimuli," J. Acoust. Soc. Am. **53**, 1573–1577.
Ulanovsky, N., Las, L., and Nelken, I. (**2003**). "Processing of low-probability sounds by cortical neurons," Nat. Neurosci. **6**, 391–398.
Ulanovsky, N., Las, L., Farkas, D., and Nelken, I. (**2004**). "Multiple time scales of adaptation in auditory cortex neurons," J. Neurosci. **17**, 10440–10453.
Viemeister, N. F. (**1980**). "Adaptation of masking," in *Psychophysical, Physiological and Behavioural Studies in Hearing*, edited by G. van den Brink and F. A. Bilsen (Delft University Press, Delft, The Netherlands), pp. 190–199.
Viemeister, N. F., and Bacon, S. P. (**1982**). "Forward masking by enhanced components in harmonic complexes," J. Acoust. Soc. Am. **71**, 1502–1507.
Vigneault-MacLean, B. K., Hall, S. E., and Phillips, D. P. (**2007**). "The effects of lateralized adaptors on lateral position judgements of tones within and across frequency channels," Hear. Res. **224**, 93–100.
Wilson, J. P. (**1970**). "An auditory afterimage," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden, The Netherlands), pp. 303–318.
Wright, B. A., McFadden, D., and Champlin, C. A. (**1993**). "Adaptation of suppression as an explanation of enhancement effects," J. Acoust. Soc. Am. **94**, 72–82.
Yost, W. A. (**1985**). "Prior stimulation and the masking-level difference," J. Acoust. Soc. Am. **78**, 901–907.

Monaural level discrimination under dichotic conditions[a)]

Daniel E. Shub[b)]
Speech and Hearing Bioscience and Technology Program, Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and Hearing Research Center, Biomedical Engineering Department, Boston University, Boston, Massachusetts 02215

Nathaniel I. Durlach
Hearing Research Center, Biomedical Engineering Department, Boston University, Boston, Massachusetts 02215 and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

H. Steven Colburn
Hearing Research Center, Biomedical Engineering Department, Boston University, Boston, Massachusetts 02215 and Speech and Hearing Bioscience and Technology Program, Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 25 May 2006; revised 19 March 2008; accepted 2 April 2008)

The ability to make judgments about the stimulus at one ear when a stimulus is simultaneously presented to the other ear was tested. Specifically, subjects discriminated the level of a 600 Hz target tone presented at the left ear while an identical-frequency distractor was simultaneously presented at the other ear. When there was no distractor, threshold was 0.7 dB. Threshold increased to 1.1 dB when a distractor with a fixed phase and level was introduced contra-aurally to the target. Further increases in threshold were observed when an across-presentation variability was introduced into the distractor phase (threshold of 1.6 dB) or level (threshold of 5.8 dB). When both the distractor level and phase varied, the largest threshold of 7.3 dB was obtained. These increases in threshold cannot be predicted by common binaural models, which assume that a target stimulus at one ear can be processed without interference from the stimulus at the nontarget ear. The measured thresholds are consistent with a model that utilizes two binaural dimensions that roughly correspond to the loudness and the position of a fused binaural image. The results show that, with binaurally fused tonal stimuli, subjects are unable to listen to one ear. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2912828]

PACS number(s): 43.66.Pn, 43.66.Ba, 43.66.Fe, 43.66.Rq [AK] Pages: 4421–4433

I. INTRODUCTION

Most models of binaural processing assume that one can voluntarily listen to the signal at one ear (right or left) even when the acoustic stimulus is binaural. More specifically, in the most frequently used models of binaural hearing [see the review by Colburn and Durlach (1978)], it is assumed that the input from each ear bifurcates into two pathways. One of these pathways proceeds to a network in which binaural interaction occurs and the other proceeds up the auditory system independent of the stimulus to the opposite ear (usually referred to as a monaural channel). Furthermore, these models assume that the listener is able to combine the information from these channels (e.g., switch between monaural and binaural listening). One consequence of these assumptions is that it should be impossible to degrade the performance of a task, where correct responses are completely determined by the stimulus at a single ear, by introducing signals into the other ear. In other words, there should be no "cross masking," or "contralateral masking," or "binaural disadvantage," or, using the term we prefer, "contra-aural interference."

In this study, the ability to access the monaural channel is measured by having subjects discriminate the level of a target tone presented at one ear while an identical-frequency distractor tone is simultaneously presented at the other ear. The objective task is based on the level of the stimulus at a single ear even though the natural perception of a binaural tone is a fused percept with a prominent loudness and position that are influenced by the stimuli at both ears. In addition to this dominant perception, however, there are also "secondary" percepts such as the image width or additional images (e.g., Hafter and Jeffress, 1968; Ruotolo *et al.*, 1979; Hartman and Constan, 2002). Even if there is no percept that corresponds to a monaural channel, the complexity of the binaural percepts makes modeling of the auditory system as having both monaural and binaural channels seem reasonable since the binaural percepts could provide information equivalent to that carried by the postulated monaural channels. Although data that are consistent with an accessible monaural channel (e.g., almost all of the data on the masking of tones by noise) certainly exist, it is by no means always the case.

[a)]Parts of this work were presented at the 27th Midwinter Meeting of the Association of Research in Otolaryngology [Shub, D. E., and Colburn, H. S. (**2004**). "Monaural Intensity Discrimination Under Dichotic Conditions," Assoc. Res. Otolaryngol. Abstr. 1521].

[b)]Present address: Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104. Electronic mail: dshub@sas.upenn.edu

There have been experiments of several types that have illustrated contra-aural interference, most of which involved broadband stimuli. For example, in studies of informational masking (e.g., Brungart and Simpson, 2002; Kidd *et al.*, 2003), monaural speech intelligibility was reduced by adding a speech masker to the other ear. In a precedence-effect experiment, Zurek (1979) demonstrated that detection of a lagging sound could be better under diotic conditions than under dichotic conditions; this experiment, together with the usual observation that diotic performance is the same as monotic performance, suggests that the dichotic case is generating contra-aural interference. In discrimination experiments, Bernstein and Oxenham (2003) showed that the ability to discriminate changes in the fundamental frequency of a harmonic tone complex was reduced under dichotic conditions, and Heller and Trahiotis (1995) reported that subjects could discriminate different noise tokens under monotic conditions, but that the subjects were unable to discriminate the tokens under some dichotic conditions. These studies interpreted the measured (contra-aural) interference by using the concepts of the central spectrum (Bilsen, 1977; Bilsen and Raatgever, 2000) and nonoptimal across-frequency processing.

Other cases of contra-aural interference cannot be easily interpreted with the concepts of the central spectrum and nonoptimal across-frequency processing. In particular, contra-aural interference has been reported in both masked (Taylor and Clarke, 1971; Taylor *et al.*, 1971a; Taylor *et al.*, 1971b; Yost *et al.*, 1972; Koehnke and Besing, 1992) and absolute (Zwislocki, 1972; Mills *et al.*, 1996) detection experiments. Small amounts of contra-aural interference have also been demonstrated in experiments in which subjects discriminate the level of a target tone, which is presented at one ear, in the presence of a simultaneously presented contra-aural distractor tone, with the same frequency and duration as the target (Rowland and Tobias, 1967; Yost, 1972; Bernstein, 2004). These level discrimination experiments were primarily investigations of binaural, specifically interaural level difference (ILD), processing. For some of these binaural conditions, when one headphone was removed (as was done in the monotic control conditions included in these studies), performance improved; it is these conditions that lead us to describe these experiments as showing contra-aural interference.

Although inconsistent with reports of contra-aural interference, the conceptualization and modeling of the auditory system including both monaural and binaural processing channels persist. One rationale for the inconsistency, which was suggested by Durlach and Colburn (1978), is that the "monaural channels" are difficult to access in some circumstances, particularly when the percepts are complicated. In these cases, one might expect that the paradigm, training, feedback, and instructions are important. In fact, the results of experiments in which secondary binaural cues are important can be heavily dependent on the experimental procedure (Hafter and Jeffress, 1968; Hafter and Carrier, 1970; Trahiotis, 1992). Therefore, tasks that use these secondary binaural cues to access the output of monaural processing channels might also be dependent on the experimental procedure.

Here, a number of steps were taken in an attempt to optimize the subjects' abilities to access possible monaural channel listening when discriminating the level of a target tone that was presented at one ear. The task was objectively defined with correct-answer feedback based on the level of the target. Subjects were instructed to optimize performance by using the correct-answer feedback to prevent the introduction of potential biases that might arise from the experimenter's description of useful percepts. A four-interval (two-cue), two-alternative forced-choice experimental paradigm was selected as it is particularly easy for subjects in complicated situations (Trahiotis, 1992). An adaptive paradigm, which is initiated with a difference between the two target levels that was large enough for consistently good performance, was used so that subjects would be able to experience a variety of possible subjective cues. Finally, performance was measured under a number of different dichotic conditions, with the monotic (monaural) condition as a reference, in which potential cues were systematically eliminated.

In a series of stimulus conditions specified at the nontarget ear, we systematically manipulated the amount of variability in the distractor level and phase to eliminate the reliability of the dominant binaural perceptions (loudness and position) for judgments about the level of the target. The most extreme condition was chosen, such that a model based on loudness and lateral position would predict substantial contra-aural interference, and thus, if little contra-aural interference was measured, this would provide evidence for the "monaural processing channels" described above and assumed in most binaural models. A number of intermediate conditions were also evaluated, such that if there was substantial contra-aural interference, one could determine whether performance is consistent with optimal use of only the outputs of the binaural processing channels. In the modeling portion of this work, we present predictions based on a model in which the decision device only has access to binaural processing channels that provide estimates of the lateral position and overall loudness.

The model is a two-dimensional decision-theoretic model with decision variables based on the overall loudness and lateral position. This model, like the experiment reported here, does not include variations over the frequency or the time dimensions of the stimulus; it is limited to a fused, stationary stimulus. When the distractor level and phase are fixed from interval to interval, both the loudness and lateral position are reliable cues for discriminating the level of the target. When the phase of the distractor is random from interval to interval, the reliability of the lateral position for discriminating the target is greatly reduced. When the level of the distractor is randomized, the reliabilities of both the loudness and the lateral position are reduced; however, the loudness and lateral position together still specify the target level. When the level and phase of the distractor are independently randomized, neither the loudness nor the lateral position nor their combination is adequate to specify the target level.

The rest of this document is organized as follows. Section II presents the experimental methods and explains the data analysis procedures. In Sec. III, models based on loud-

TABLE I. Distractor properties in the five conditions. In all of the conditions, the target has a frequency of 600 Hz, a duration of 300 ms, a phase of zero, and a reference level of 50 dB SPL. The distractor was simultaneously presented but contra-aurally to the target. Roving of the level and phase of the distractor was done on an interval-by-interval basis with values chosen from uniform distributions.

	Frequency (Hz)	Duration (ms)	Phase (rad)	Level (dB SPL)
No distractor	⋯	⋯	⋯	⋯
Fixed	600	300	0	50
Roving phase	600	300	Uniform $\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$	50
Roving level	600	300	0	Uniform (50,80)
Double rove	600	300	Uniform $\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$	Uniform (50,80)

ness and position decision variables are described and related to similar models in the literature. In Sec. IV, the psychophysical results are presented, along with the predictions of the models. The discussion in Sec. V includes comparisons between data and model predictions and explores the possibility that monaural processing channels exist. Finally, Sec. VI gives some concluding remarks.

II. EXPERIMENTAL METHODS

A. Subjects

Four subjects (S1–S4) completed the tasks. Subject S1 is the first author. With the exception of S1, subjects received an hourly wage for their participation. All subjects had pure tone thresholds below 20 dB HL at frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz in both ears. The subjects were between 19 and 31 years old. Subjects S1 and S2 had prior listening experience in similar tasks, while subjects S3 and S4 had no prior experience in psychoacoustic experiments.

B. Stimulus and procedures

The experimental task was designed to test the ability of subjects to discriminate the level of a 600 Hz target tone at the left ear in the presence of a 600 Hz distractor tone simultaneously presented at the right ear. A four-interval, two-alternative, forced-choice (4I-2AFC), adaptive paradigm was used and correct-answer feedback was given on every trial. Five different conditions were explored: the *no-distractor* condition and four conditions that differed in the presence/absence of interval-to-interval variation in the distractor phase and level. The four distractor conditions were *fixed* distractor, *roving-phase* distractor, *roving-level* distractor, and *double-rove* distractor (with both level and phase variation). Table I lists the properties of the distractor used in each condition.

The target and distractor tones had 300 ms durations and 25 ms rise/fall times, were simultaneously gated on and off, and had 500 ms of quiet between intervals. The target was presented at either the reference level of 50 dB SPL or the reference level plus an increment ΔL (in decibels). Tone phases were defined relative to the onset of the stimulus ramp; the target phase was always zero and the distractor phase was specified relative to this zero-phase target. When the distractor level was fixed, it was held at 50 dB SPL (the reference level of the target). When the level was roved (roving-level and double-rove conditions), the level of the distractor was randomly chosen on an interval-by-interval basis from a uniform distribution between 50 and 80 dB SPL. When the distractor phase was fixed, it was held at zero phase, and when it was roved (roving-phase and double-rove conditions), the phase of the distractor uniformly roved between $\pm\pi/2$ on an interval-by-interval basis.

In the 4I-2AFC paradigm used here (Bernstein and Trahiotis, 1982), listeners must distinguish the patterns *ABAA* from *AABA*, where *A* and *B* are defined by the level of the tone at the left ear (i.e., the target). Stimuli labeled *A* have the reference level at the left ear and stimuli labeled *B* have the level at the left ear incremented by ΔL. Subjects were informed about the two temporal patterns and instructed to maximize the percent correct by utilizing the trial-by-trial feedback. We did not explicitly tell the subjects to respond according to whether the level of the tone at the left ear was relatively higher on the second or third interval (even though such instructions would have been consistent with the objective task and with the feedback provided) because we thought it was better to avoid restricting the subjects' attention to specific subjective cues. In our pilot experiments, when subjects tried to focus their attention directly on the loudness at the left ear rather than on the loudness and lateralization of the fused image (and on how these subjective cues related to the feedback), they performed much worse. When considering the instructions used, it is also worth noting that subjects who knew that the key to the feedback was the level at the left ear (e.g., the first author who served as subject S1 as well as subject S2) and subjects who knew only that they should respond in a manner that, according to the feedback, led to a correct answer (subjects S3 and S4), produced roughly comparable data after the initial training period was completed.

A two-down one-up adaptive procedure, which was modeled after Levitt (1971), estimated the minimum change in the target level required to achieve a probability of a correct response of 0.7 in this paradigm. Each adaptive run consisted of 16 reversals and began with a random (large) initial value of ΔL (uniformly chosen between 15 and 25 dB). According to the two-down one-up adaptive rule, ΔL was initially adjusted by multiplying/dividing its current value in decibels by a scale factor of 1.8. After two reversals occurred, the increment was adjusted by multiplying/dividing by a scale factor of 1.4. The magnitude of the scale factor was further reduced when the fourth, sixth, and eighth reversals occurred to values of 1.2, 1.1, and 1.05, respectively. After the eighth reversal, the scale factor remained at 1.05 for eight additional reversals, at which point the adaptive run was concluded. The adaptive trials were self-paced and the subjects had an unlimited time to respond. The subjects received correct-answer feedback after every trial.

During each testing session, subjects completed four adaptive runs for each of the five different conditions. The ordering in which the conditions were presented was the same for each session: no distractor, fixed, roving phase, roving level, and double rove. Since the perceptual cues for each condition may have been different, subjects were alerted to the condition by using a unique letter for each condition. For each subject, data from the four adaptive runs for each distractor condition were collected in succession. The blocking of the adaptive runs in this manner was done to allow the subjects to refamiliarize themselves with the relevant perceptual cues. Subjects were given a minimum of 10 h of training before the reported data were collected. *Post hoc* analyses suggest that asymptotic performance was obtained early in the training period and that alerting the subjects to the condition before the block of runs began was sufficient to obtain the asymptotic performance. Further, in pilot listening, prolonged experience with the double-rove condition did not improve performance. For each condition, the reported results are based on 16 post-training adaptive runs.

C. Apparatus and materials

During the experiment, subjects sat in a sound treated room in front of a computer monitor. They responded "interval 2" or "interval 3" through a graphical interface via a computer mouse. On each trial, "lights" on a liquid crystal display monitor displayed the current interval number. The experiment was self-paced and listening sessions lasted for no more than 2 h with frequent rest breaks. A PC and Tucker–Davis Technology System II hardware (AP2, PD1, PA4, and HB6) generated the experimental stimuli at a sampling rate of 50 kHz. Stimuli were presented over Sennheiser HD 265 headphones.

D. Data analysis

Level discrimination thresholds have been reported in many forms and the analysis reported here follows the recommendations of Buus and Florentine (1991). In accordance with their recommendations, the performance metric used is ΔL, which is the decibel change of the target level. Also in accordance with the conclusions of Buus and Florentine (1991), the value ΔL is displayed on a logarithmic scale. The analysis and statistics are therefore based on the logarithm of ΔL. For example, the threshold was estimated from the geometric mean of the reversals of the adaptive runs and when averaging thresholds across conditions, the geometric mean was used. The results were also analyzed by considering ΔL on a linear display (i.e., using the arithmetic mean) and both analyses lead to similar conclusions.

The data from the experiment are reported in two ways. First, level discrimination thresholds were determined from the geometric mean of the values of ΔL that occurred on the last eight reversals of each adaptive run. Second, psychometric functions were fitted to the data from all the trials of the adaptive runs. The binary data for single trials (correct or incorrect) were used in fitting the psychometric functions instead of estimates of the probability of a correct response for each value of ΔL based on multiple trials since only a few trials were conducted with each value of ΔL. Even though there were hundreds of trials per condition, the experimental paradigm resulted in a large number of different values of ΔL being used; the adaptive runs began with a random initial value of ΔL and on each trial, ΔL was adjusted using the above-mentioned adaptive rule.

In accordance with Buus and Florentine (1991), d' is assumed to be proportional to ΔL. Converting the correct/incorrect data into d' is problematic since there are many values of ΔL for which there were only a small number of trials that the subjects got all correct, leading to a d' of infinity. By assuming that (1) d' is proportional to ΔL, (2) the observer is unbiased (which is reasonable in a 4I-2AFC task), and (3) performance is limited by Gaussian noise, one can relate d' to the probability of a correct response. Specifically, with these assumptions, the probability of a correct response depends on ΔL and can be expressed as

$$P_{\text{correct}}(\Delta L) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\Delta L} e^{-x^2/2\sigma^2} dx, \tag{1}$$

where σ is a fitting parameter and is related to the proportionality between d' and ΔL. Note that Eq. (1) implies that when ΔL is equal to zero (corresponding to no change in the target level) the probability of a correct response is 0.5, as is expected, and that σ is the only fitting parameter. When fitting psychometric functions to the subject data, the parameter σ was adjusted to minimize the root-mean-squared (rms) error between the predicted percent correct and all of the data for a given subject and condition (collapsed over adaptive runs).

For each subject and condition, analysis of the fitted psychometric function is based on the parameter σ and the rms error statistics. Confidence intervals for both σ and the rms error were calculated by randomly drawing, with replacement, the results of N trials (where N is the total number of trials for a given subject and condition) and fitting a psychometric function to these sampled data. For each subject and condition, 1000 random drawings were made. An additional estimate of the threshold is obtained by substituting the appropriate value of σ into Eq. (1) and then solving for the ΔL that yields a probability of a correct response of 0.7. These estimates of the threshold based on the psychometric functions agree with the measurements of the threshold based on the reversals of the adaptive runs and are not explicitly presented.

III. MODELING

The dominant perception of the stimulus when the target and distractor are present is a fused binaural image with a salient loudness and lateral position. It is unclear whether the subjects have access to the level of the monaural target through secondary cues, which are cues that are not necessarily direct perceptions of the stimulus at a single ear but that provide sufficient information to reconstruct the stimulus at a single ear. Our modeling is intended to provide evidence in regard to the existence (or absence) of the monaural processing channels that appear in most binaural models. A model that includes monaural processing channels would not

predict an effect of a contra-aural distractor, and the predictions of our experiments are trivial for that case. Instead, we investigate a model in which there are no separate monaural channels. Specifically, we consider a detection theoretic model based on a two-dimensional decision space that roughly corresponds to the loudness and lateral position of the fused binaural image. Other perceptual variables, such as the image width, are not included nor are temporal characteristics that would be expected to be important in some experiments (especially in studies of binaural masking level differences), such as distributions of interaural differences or lateral positions over time.

The two dimensions of the model denoted Λ and Θ are functions of the level at the left ear, L_{left} (in decibels), the level of the right ear, L_{right} (in decibels), the interaural time difference, T (in microseconds), and two internal noises N_Λ and N_Θ (both in decibels). The dimensions Λ and Θ are defined as

$$\Lambda = 10 \log_{10}(10^{L_{\text{left}}/10} + 10^{L_{\text{right}}/10}) + N_\Lambda \tag{2}$$

and

$$\Theta = L_{\text{left}} - L_{\text{right}} + kT + N_\Theta, \tag{3}$$

where k is the intensity-time trading ratio in dB/μs, which is related to the time-intensity trading ratios used by Hafter (1971) and Yost (1972). To aid comparisons between Λ and Θ, we have defined Θ in decibels even though it can be logically expressed in a number of other ways (e.g., interaural time difference or a dimensionless quantity related to a location along the interaural axis). Note that, under monotic conditions (i.e., either L_{left} or L_{right} is equal to negative infinity), Θ is equal to either positive or negative infinity and indicates an extreme position toward the stimulated ear.

Models based on position variables such as Θ have been used to predict the results of many binaural experiments (Hafter, 1971; Yost, 1972). Although models based on position variables have a strong predictive power for some types of experiments, they are insufficient to predict all binaural effects. Importantly, Bernstein (2004) reported that subjects could outperform an ideal observer of Θ when discriminating the ILD of tones with a random interaural time difference (ITD). The discrepancy between Bernstein's (2004) measurements of ILD discrimination and predictions of a simple position variable may be overcome by a variety of modifications, as discussed by Stern and Trahiotis (1997) in their review of position variable models, such as the image width or a comparison of separate monaural levels [as in the level meter model of Hartmann and Constan (2002)]. In this study, we consider predictions for monaural level discrimination based on the use of the overall (binaural) level Λ and the unmodified position variable Θ [as defined in Eqs. (2) and (3)].

Models based on the Λ dimension (the overall binaural level) have been evaluated as an aspect of central spectrum models (Bilsen, 1977; Bilsen and Raatgever, 2000). These studies of the central spectrum have focused on issues concerning across-frequency integration. Binaural loudness has been studied in a variety of ways (e.g., Zwicker and Zwicker, 1991; Sivonen and Ellermeier, 2006; Whilby *et al.*, 2006).

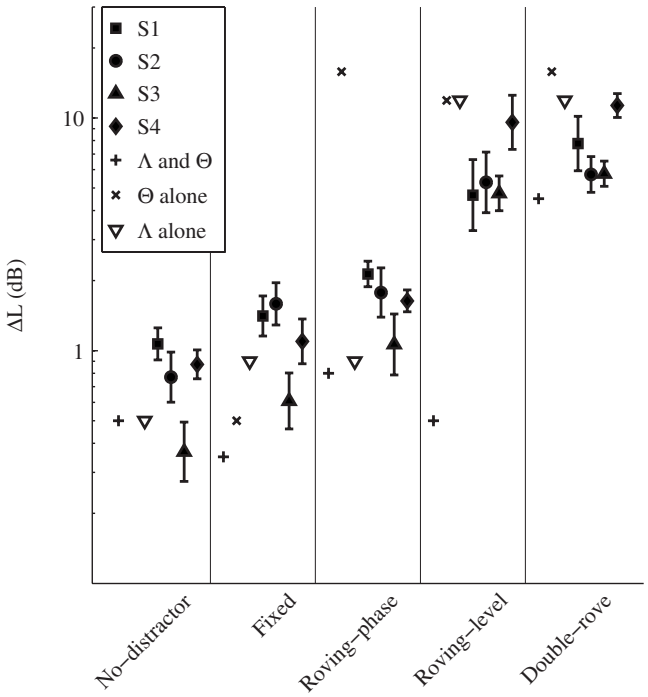


FIG. 1. Mean thresholds for the four subjects under the five different conditions. Error bars are two times the standard error of the mean. Note that the standard deviations are twice as large as two times the standard error of the mean based on 16 runs. The thresholds of the ideal observer of Λ and Θ, both alone and together, are also shown. In the no-distractor condition, the ideal observer of Θ never obtains threshold performance.

Sivonen and Ellermeier (2006) concluded that power summation (i.e., Λ) predicted binaural loudness best. Under monotic conditions, Λ mathematically reduces to the monaural level and models based on a monaural level detector have been explored for many monotic and diotic tasks.

Three models are explicitly considered here: the maximum likelihood observers of Λ alone, Θ alone, and Λ and Θ together. The dimensions Λ and Θ are specified with internal noises N_Λ and N_Θ, which are assumed to be zero-mean Gaussian random variables that are statistically independent across the dimensions, the observation intervals, and the trials. With these assumptions, there are three parameters, k, σ_Λ, and σ_Θ, which correspond to the intensity-time trading ratio and the standard deviations of the two internal noises. In order to predict generally accepted values for the just noticeable differences in the overall level, ILD, and ITD (cf. Viemeister 1988; Blauert, 1997), the intensity-time trading ratio k is fixed at 1 dB per 20 μs, and the standard deviations of the internal noises (σ_Λ and σ_Θ) are fixed at 0.5 dB throughout all the modeling. The performance of these three ideal observers in a 4I-2AFC task is derived in Appendix A.

IV. RESULTS

A. Psychophysical results

Figure 1 contains the geometric mean of the thresholds (calculated from the reversals of the adaptive runs) of the four subjects for each of the five different conditions. A repeated measure analysis of variance test found statistically significant effects of distractor condition and subject and a

statistically significant interaction between distractor condition and subject ($p<0.02$). The measured thresholds vary with the type of distractor and subject. Multiple planned paired t tests were used to test for statistically significant differences between conditions (all combinations). There are statistically significant differences ($p<0.05$) in performance between all pairs of conditions except between the roving-level and double-rove conditions ($p=0.09$). Performance in the no-distractor condition was the best with an average (across-subject geometric-mean) threshold value of ΔL of 0.7 dB. Performances with the roving-level and double-rove distractors were the worst with average thresholds of 5.8 and 7.3 dB, respectively. The fixed and roving-phase distractors showed small detrimental effects on performance, with average thresholds of 1.1 and 1.6 dB, respectively. Displaying the threshold value of ΔL (a decibel value) on a logarithmic scale, as opposed to a linear scale, expands the differences between the no-distractor condition and the fixed and roving-phase distractor conditions and compresses the differences between the no-distractor condition and the roving-level and double-rove distractor conditions. The subjects are clearly not basing their decisions on the level at the target ear; in other words, they are not "listening to one ear" since the addition of a distractor at the nontarget ear decreases performance.

Figure 2 shows example psychometric functions for a representative subject (S2) for all five conditions. The psychometric functions take into account all trials, whereas the threshold measurements take into account only trials at which reversals in performance occurred. Since the data were collected using an adaptive paradigm, most trials had values of ΔL near threshold. Traditionally, psychometric functions are not constructed from data collected with adaptive paradigms. We wished, however, to determine the extent to which performance was monotonic in ΔL. Visually, the data appear monotonic and generally consistent with the sigmoid function of Eq. (1). The values of σ (the single fit parameter for the psychometric function) that best fit the data are presented in the top panel of Fig. 3. Consistent with the threshold data presented in Fig. 1, the value of σ that best fits the data systematically varies across the conditions. Paired t tests show no statistically significant differences ($p>0.05$) in the values of σ among the no-distractor, fixed, and roving-phase conditions nor are there statistically significant differences between the roving-level and double-rove conditions ($p>0.05$). The differences in the values of σ for the no-distractor, fixed, and roving-phase conditions and for the roving-level or double-rove conditions, however, are statistically significant ($p<0.025$).

In addition to the changes in σ, there are also changes in the rms error between the fitted psychometric function and the data. The bottom panel of Fig. 3 shows the rms error. The rms error and the visual agreement between the fitted psychometric functions and the data are similar for all of the subjects. Paired t tests show no statistically significant differences ($p>0.05$) in the rms error for the no-distractor, fixed, and roving-phase conditions nor for the roving-level and double-rove conditions. The differences in the rms error for the no-distractor, fixed, and roving-phase conditions and for the roving-level and double-rove conditions, however, are statistically significant ($p<0.025$). These changes in the rms error are indicative of a change in how well the fitted psychometric functions fit the data.

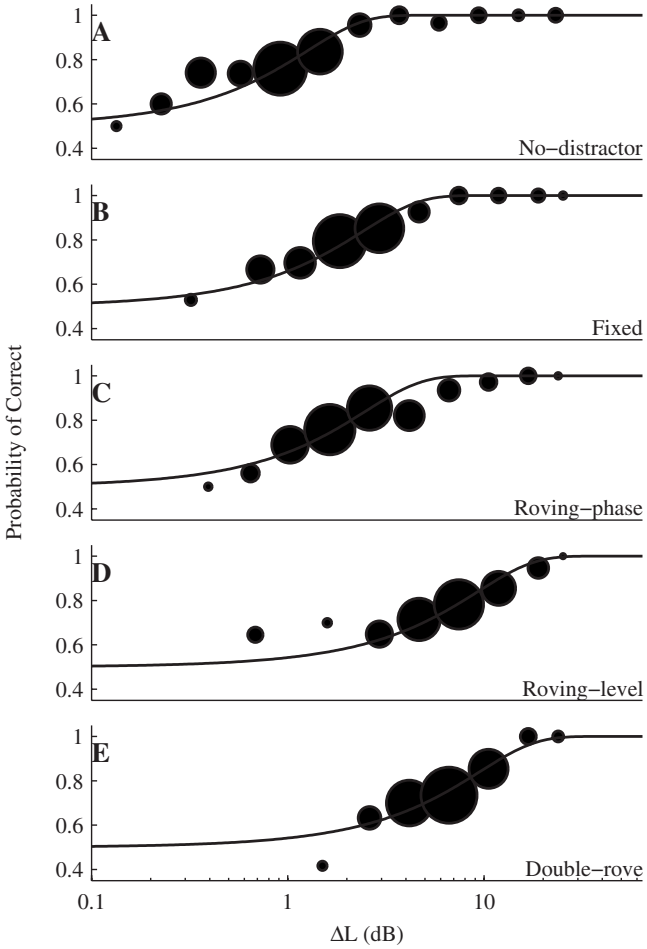


FIG. 2. Examples of the dependence of the probability of a correct response on the target increment for subject S2 for the five different conditions. Panels (A)–(E) correspond to the no-distractor, fixed, roving-phase, roving-level, and double-rove conditions, respectively. The data have been binned according to ΔL, and the size of the symbol is proportional to the number of trials that occurred within the bin. The best fitting psychometric functions by using the single parameter function in Eq. (1) are also shown for each condition.

B. Decision variable distributions

In order to understand the information about the level of the target, which is carried by the two decision variables Λ and Θ, their probability densities for a given L_{target} are examined. We consider the joint density of Λ and Θ ($f_{\Lambda,\Theta|L_{\text{target}}}$) as well as the densities of Λ ($f_{\Lambda|L_{\text{target}}}$) and Θ ($f_{\Theta|L_{\text{target}}}$) in isolation. The manner in which these probability densities were computed is described in Appendix B. For simplicity, we define L_{target} as being equal to the reference level L_0 plus an increment ΔL and plot the density functions for five different values of ΔL (0, 2, 4, 8, and 16 dB), such that, for the unincremented target, ΔL is equal to zero.

Figure 4 shows $f_{\Lambda|L_{\text{target}}}$ for the five different conditions. Since the distractor phase has no effect on Λ, $f_{\Lambda|L_{\text{target}}}$, is identical in the fixed and roving-phase conditions as well as

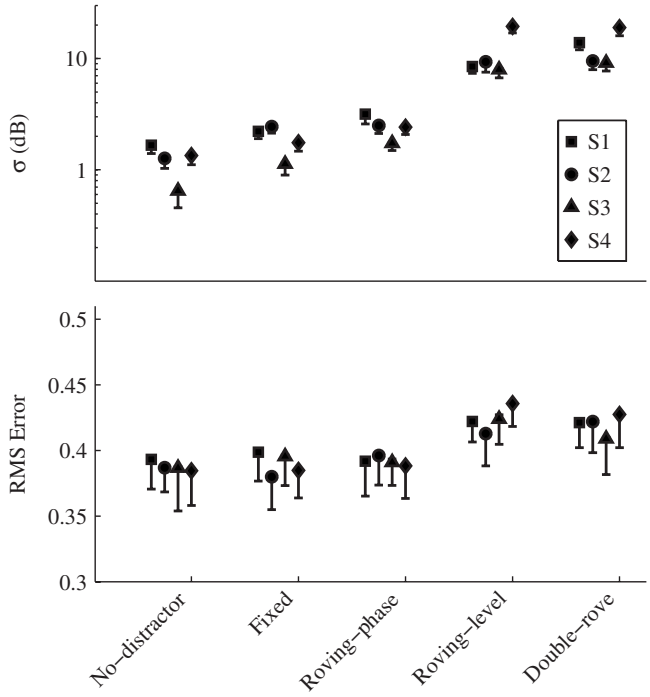


FIG. 3. Fit parameter σ (top panel) and the RMS error (bottom panel) for the psychometric functions that were fitted to the data of the four subjects under the five different conditions. Error bars are the 95% confidence intervals derived by resampling the data. When the error bars are absent, the confidence interval is on the order of the size of the symbol.

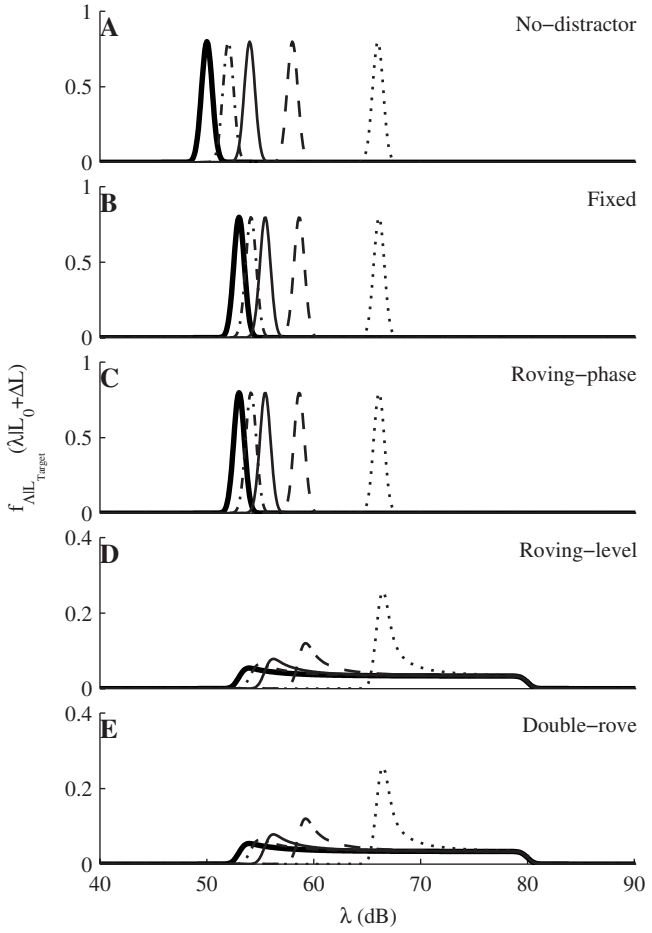


FIG. 4. Probability distributions of the overall level variable Λ. Panels (A)–(E) correspond to the no-distractor, fixed, roving-phase, roving-level, and double-rove conditions, respectively. Note that each panel plots the density function of Λ as a function of its argument λ for the target level L_{target} equal to the reference level L_0 plus the monaural level increment ΔL. Specifically, $f_{\Lambda|L_{\text{target}}}$ is plotted for five values of ΔL: 0 dB (thick solid), 2 dB (dot dash), 4 dB (thin solid), 8 dB (dashed), and 16 dB (dotted).

in the roving-level and double-rove conditions. In the no-distractor, fixed, and roving-phase conditions $f_{\Lambda|L_{\text{target}}}$ is Gaussian with standard deviation σ_Λ and a mean that depends on ΔL. The mean in the fixed and roving-phase conditions is higher than in the no-distractor condition due to the additional distractor energy; the effect of this additional energy decreases with increasing ΔL since Λ is calculated by adding the intensities (units of power per area) and not the decibel levels. In the roving-level and double-rove conditions, the random distractor level affects Λ, and therefore, $f_{\Lambda|L_{\text{target}}}$ is much broader. In these two conditions, changes to ΔL affect both the mean and the shape.

Figure 5 shows $f_{\Theta|L_{\text{target}}}$ in four of the conditions (the no-distractor condition is not included since Θ is undefined). Changes to ΔL again affect the mean in all conditions but do not affect the shape of $f_{\Theta|L_{\text{target}}}$ in any condition. In the fixed condition, $f_{\Theta|L_{\text{target}}}$ is Gaussian (with standard deviation σ_Θ). In the roving-phase and roving-level conditions, Θ depends on both the internal noise and a uniformly distributed random variable (either the distractor phase or level) and $f_{\Theta|L_{\text{target}}}$ is nearly uniform over a large range. In the double-rove condition, Θ is the sum of two uniformly distributed random variables (the distractor level and phase) and the internal noise and $f_{\Theta|L_{\text{target}}}$ is nearly trapezoidal in shape.

Figure 6 shows the region for which $f_{\Lambda,\Theta|L_{\text{target}}}$ is greater than 0.0001 for the roving-level and double-rove conditions, respectively. In the roving-level condition, the probability of Θ conditioned on Λ is narrow, and changes in ΔL substantially shift the distribution, making monaural level discrimination possible for the ideal observer (i.e., the distributions for the unincremented and incremented targets do not substantially overlap). However, the complexity of the distributions may lead a nonoptimal observer to have a substantially degraded performance. In the double-rove condition, the variables Λ and Θ together do not carry accurate information; the probability of Θ conditioned on Λ is broad (i.e., each value of Λ now corresponds to a range of Θ values) and changes in the increment size have only small effects. Therefore, the ideal observer of Λ and Θ together cannot discriminate between the unincremented and incremented target levels.

In summary, the Λ and Θ dimensions carry information both individually and jointly about the target level. Introducing variability into the distractor phase decreases the information in Θ but does not compromise the information in Λ. Introducing variability into the distractor level reduces the information in Λ and Θ individually but does not reduce the joint information. In order to substantially decrease the performance of the observer of both Λ and Θ together, and thereby making it advantageous to utilize a secondary perception such as the output of monaural processing channels (if it is available), variability must be introduced into both the distractor phase and level (i.e., the double-rove condition).

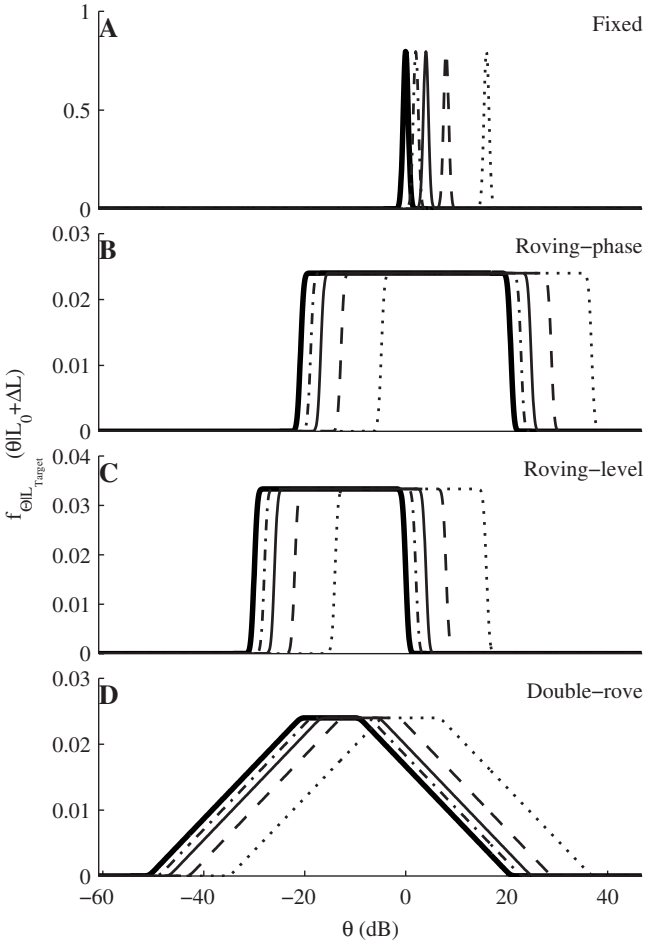


FIG. 5. Probability density functions $f_{\Theta|L_{\text{target}}}$ for the lateral position variable Θ for multiple monaural levels, which are similar to the plots in Fig. 4. In the no-distractor condition, Θ is infinite and is not shown. Panels (A)–(D) therefore correspond to the fixed, roving-phase, roving-level, and double-rove conditions, respectively.

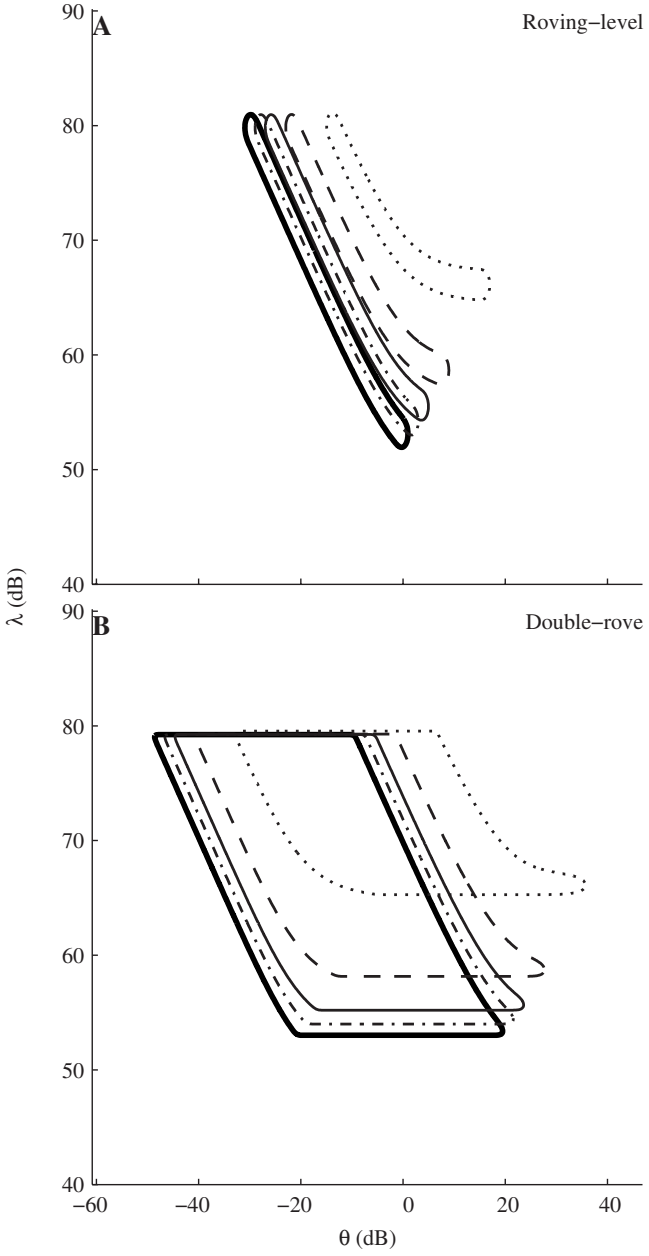


FIG. 6. Contours that enclose the region for which $f_{\Lambda,\Theta|L_{\text{target}}}$ is greater than 0.0001. Note that within these regions, $f_{\Lambda,\Theta|L_{\text{target}}}$ is not uniform. Panels (A) and (B) correspond to the roving-level and double-rove conditions, respectively. As in Figs. 4 and 5, five different values of ΔL are shown. The five different values of ΔL are 0 dB (thick solid), 2 dB (dot dash), 4 dB (thin solid), 8 dB (dashed), and 16 dB (dotted).

C. Model predictions

The Λ alone, Θ alone, and Λ and Θ together ideal-observer models can be used to make predictions of both thresholds and psychometric functions. The predicted thresholds for the three models are included along with the measured thresholds in Fig. 1. The predicted psychometric functions for the five different distractor conditions are shown in Fig. 7 along with the average empirical psychometric functions for the subjects. We consider the predicted thresholds of the ideal observer of Θ alone first, which is followed by the ideal observer of Λ alone and then the ideal observer of Λ and Θ together. Finally, we consider the predicted psychometric functions for all three models.

The ideal observer of Θ alone performs best in the fixed condition with a threshold of σ_Θ (assumed here to be 0.5 dB). The ideal observer of Θ alone never obtains threshold performance in the no-distractor condition since Θ is equal to minus infinity in this case independent of the target level. This is consistent with the position providing no information about the target level in monotic conditions. In the roving-phase, roving-level, and double-rove conditions, the thresholds are 16.9, 12.8, and 16.9 dB, respectively. The Θ alone model is not a good predictor of performance in both the no-distractor condition and the roving-phase condition. In all but the fixed condition, there is insufficient information in Θ to predict the measured thresholds. Although the perceived lateralization of the stimulus may be influencing the subjects' decisions, the Θ alone model is clearly an exceedingly poor predictor of the measured discrimination thresholds.

The ideal observer of Λ alone performs best in the no-distractor condition with a threshold equal to σ_Λ (assumed here to be 0.5 dB). In both the fixed and roving-phase conditions, the predicted thresholds are 0.9 dB; the predicted performance is slightly worse due to the added energy of the distractor. In both the roving-level and double-rove condi-

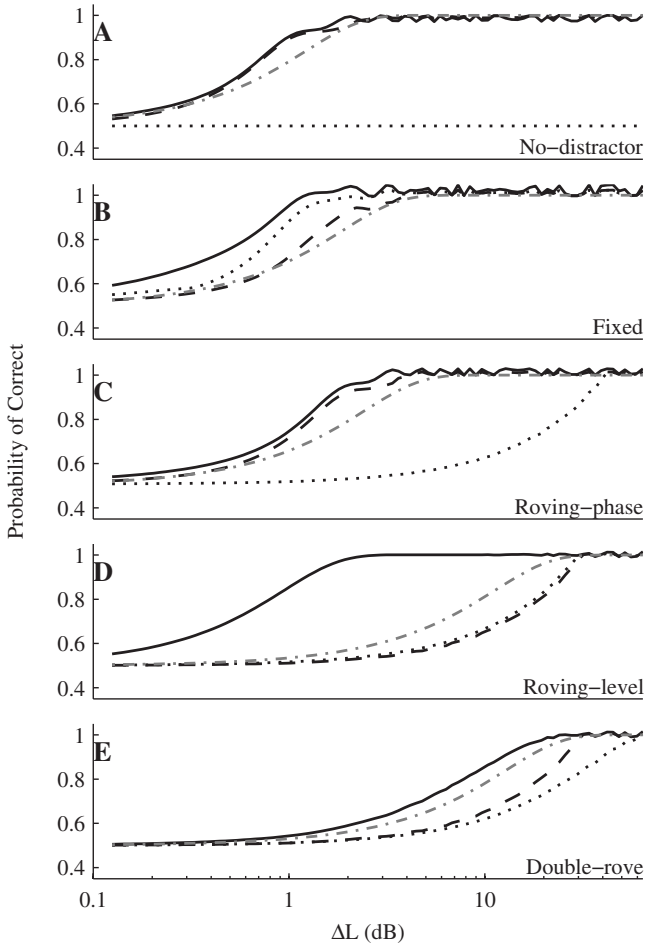


FIG. 7. Predicted psychometric functions for the five conditions for the ideal observer of Λ alone (dashed), Θ alone (dotted), and Λ and Θ together (solid). Also shown is the across-subject average fitted psychometric function (dot dash). Panels (A)–(E) correspond to the no-distractor, fixed, roving-phase, roving-level, and double-rove conditions, respectively.

tions, the predicted thresholds are 13.7 dB and are predominately determined by the range of the level rove. For all the conditions, the predicted thresholds are in general agreement with the measured thresholds. The empirical fact that roving the level (with or without roving the phase) causes such a large elevation in the measured thresholds suggests that the subjects heavily relied on the overall loudness for some of their decisions. Although the Λ alone model is a better predictor of performance than the Θ alone model, it is important to note that the predictions in the roving-level and double-rove conditions are significantly worse than the measured thresholds. This means that the subjects could not have been basing their decisions solely on Λ.

The ideal observer of Λ and Θ together performs best in the fixed condition with a threshold of 0.3 dB since the two internal noises N_Λ and N_Θ are statistically independent. In the no-distractor condition, the threshold is 0.5 dB. In the roving-phase, roving-level, and double-rove conditions, the thresholds were, respectively, 0.9, 0.5, and 4.8 dB. For the Λ and Θ together model (unlike the Θ alone and Λ alone models), the predicted thresholds are seen to be less than or equal to the average measured thresholds in all of the conditions. In other words, there is enough information in Λ and Θ taken together to perform as well or better than the subjects. However, as indicated by the predicted threshold being substantially lower than the measured thresholds in the fixed and roving-level conditions, the subjects are failing to make use of all the available information.

Figure 7 shows the predicted psychometric functions of the ideal observer of Λ alone (dashed), Θ alone (dotted), and Λ and Θ together (solid), along with the average empirical psychometric functions (dash dot) for the subjects. The differences among the psychometric functions predicted by the three models in the five conditions are indicative of changes in the information carried by Λ and Θ across the conditions. By comparing the model predictions and the empirical data over the whole range of ΔL (not only at the threshold), additional discrepancies between the predictions and empirical data become evident. Some of the theoretical psychometric functions differ from the empirical ones not only in the position on the abscissa (i.e., threshold) but also in shape.

V. DISCUSSION

It is clear from the results shown in Fig. 1 that the ability to discriminate the level of a monaural target can be severely degraded by the introduction of a contra-aural distractor. This contra-aural interference cannot be predicted by a model that includes monaural processing channels, which by definition are unaffected by stimulation of the other ear. In the models evaluated in this study, the measured performance is not predicted by the model based on lateral position; however, both models based on overall loudness and on loudness and position together show some promise in predicting the measured effects.

Although the Λ alone model reasonably predicts much of the data, there are fundamental problems. The main problem with this model occurs in the roving-level and double-rove conditions, where the subjects are seen to do modestly (but significantly) better than the model. Although the deviations between the model and data for these conditions are not large, they cannot be eliminated simply by decreasing the internal noise parameter. Even if all the internal noise were eliminated, the predicted performance in the double-rove condition would still be worse than that achieved by the subjects. Within the context of our models, this in turn implies that the subjects' decisions cannot be based solely on Λ but that Θ must also be considered.

The Λ and Θ together model predictions are reasonably consistent with the data in the no-distractor, roving-phase, and double-rove conditions (although the predictions are slightly lower than the average data in all three of these conditions). However, for the fixed condition, the prediction is substantially too low, and for the roving-level condition, it is monstrously too low. Thus, although it appears (as discussed above) that the subjects must be making some use of Θ as well as Λ, it is clear that their ability to use both together is clearly suboptimal. It seems unlikely that the less-than-optimum performance can be simply explained by assuming a degradation (i.e., increased internal noise) in performance caused by the need to simultaneously estimate two variables and the resulting problem of divided attention (e.g.,

Bonnel and Hafter, 1988). Rather, consistent with Fig. 6, it appears that to obtain a good performance in the roving-level condition requires the observer of both Λ and Θ to precisely combine the two observations. Therefore, in order to accurately predict the psychophysical results, the modeled observer must be modified such that it cannot precisely combine two observations.

The Λ and Θ together model assumes that the contra-aural interference arises because the subjects do not have access to the output of monaural processing channels. In the next paragraphs, several alternative hypotheses for the causes of poor performance are discussed and rejected. Specifically, we consider the effects of acoustic cross talk, subject confusion, and inadequate training.

The measured contra-aural interference could occur because the distractor corrupts the inputs to the monaural processing channels. Specifically, that acoustic cross talk results in the level at the target ear being influenced by the distractor. The effects of cross talk can be approximated by assuming that an attenuated and delayed version of the distractor is added (in units of pressure) to the target. Since the amplitude of the sum of two identical-frequency sinusoids depends on both the relative levels and phases, the level of the corrupted target is variable in the roving-phase, roving-level, and double-rove conditions. To limit the performance of the ideal observer, which is degraded only by the cross talk, to the empirically measured threshold in the double-rove condition, the amount of attenuation across the head would need to be less than 15 dB (substantially less than the typically assumed 40 dB).

It is also unlikely that the measured contra-aural interference is simply a result of the subjects being confused. The measured thresholds in the no-distractor condition agree with previous measures of monotic level discrimination thresholds (Viemeister, 1988). The thresholds in the fixed condition are similar to ILD discrimination thresholds, where the ILD is imposed by incrementing the level at one ear and decrementing the level at the other ear (Blauert, 1997). Additionally, if the subjects were confused, then introducing across-presentation variability to the distractor should degrade performance, but roving the distractor phase (i.e., perceived lateral position) did not substantially increase the thresholds of the subjects. Performance was only degraded by introducing across-presentation variability to the distractor level. However, introducing across-presentation variability to the overall level has little effect on spectral profile analysis (Green, 1988) or ILD discrimination (Bernstein, 2004). It seems that in the current task, performance is only degraded when the across-presentation distractor variability affects the information about the target level in both the overall loudness and the lateral position.

Finally, it also seems unlikely that subjects have been simply inadequately trained. The testing scheme that was employed in these experiments (the 4I-2AFC paradigm with trial-by-trial correct-answer feedback) is known to provide rapid learning and good performance in situations that may be confusing or difficult to learn (Trahiotis, 1992). Consistent with this established property of the testing scheme is that despite large differences in the relevant previous experience of the subjects (e.g., S1 was the first author of this article while S3 and S4 were naive), the measured discrimination performance did not appreciably vary across the subjects. Further, all our measured psychometric functions appear smooth and monotonically increasing with ΔL. Thus, the measured increases in the threshold due to the inclusion of the distractor are unlikely to be caused by artifacts associated with the adaptive procedure or by major perceptual changes as ΔL is varied. Rather, it seems likely that the contra-aural interference occurs because the subjects do not have access to the output of monaural processing channels.

VI. CONCLUDING REMARKS

The results of the experiments presented here strongly support the idea that when judging changes in the level of a target presented at one ear, listeners are unable to ignore a contra-aural distractor. Therefore, models in which it is assumed that the listener has access to monaural processing channels (either directly or through secondary perceptual attributes such as the time image or spatial width combined with the overall level) cannot be successfully applied to this experiment. Specifically, listeners' thresholds were increased by an order of magnitude (relative to the measured monaural performance) in some conditions for our 600 Hz stimuli. In general, subjects perceived a single compact image and reported that they used the loudness and the lateral position of this image for many of their judgments. A model based on the joint use of decision variables that correspond to the loudness and the lateral position of the primary image was consistent with many of the results, although performance in some conditions was notably poorer than that predicted when the internal noise was chosen to be consistent with the typical discrimination threshold for ITD and overall level. It is speculated that, although both the loudness and the lateral position can be used, there are difficulties in simultaneously using both for refined decisions. Furthermore, when a wider set of experiments is considered, it becomes clear that a model that only includes the loudness and the lateral position is not complete and that additional decision variables are needed to match the observed performance in other experiments. Overall, one can conclude that an adequate theory will likely involve degraded processing of the loudness and lateral position variables, the inclusion of further binaural decision variables, and the exclusion (at least in some conditions) of access to monaural channels.

ACKNOWLEDGMENTS

This research was supported by NIH/NIDCD Grant Nos. R01 DC00100, P30 DC004663, and F31 DC006769. The authors would also like to thank Dr. Frederick Gallun, Dr. Andrew Oxenham, and Dr. Bertrand Delgutte for reading a previous version of this manuscript. The comments by Dr. Armin Kohlrausch and two anonymous reviewers were also very helpful.

APPENDIX A

In this appendix, a general expression for the probability of a correct response in a 4I-2AFC task is calculated for

three observers in the $\{\Delta,\Theta\}$ space, where Λ and Θ are defined by Eqs. (2) and (3) in the text. Three models are considered: the ideal observer of Λ and Θ together, the ideal observer of Λ alone, and the ideal observer of Θ alone. The task requires the observer to discriminate the level of the target, L_{target}. On a given interval, the target is either unincremented, such that L_{target} is equal to L_0 (the reference level), or the target is incremented, such that L_{target} is equal to the sum of L_0 and ΔL. To calculate performance we note the following: (1) on each interval, there is a single observation of both Λ and Θ and (2) on a single trial, there are eight total observations (four of Λ and four of Θ). Due to the experimental paradigm, the observations in the first and last intervals carry no information for the ideal observers,[1] and therefore, only four observed values (two pairs) are relevant. The observation of Λ on the second interval is denoted λ_2; the observation of Λ on the third interval is denoted λ_3. Similarly, the observation of Θ on the second interval is denoted θ_2 and that on the third interval is denoted θ_3. The ideal observer of Λ and Θ together is considered first since the performance of the ideal observers of Λ alone and Θ alone follow from the ideal observer of Λ and Θ together.

The ideal observer of both Λ and Θ depends on two four-dimensional joint probability functions. The first is the probability densities of λ_2, λ_3, θ_2, and θ_3 given that an increment with size ΔL occurred on the second interval; the second is the probability density of λ_2, λ_3, θ_2, and θ_3 given that an increment with size ΔL occurred on the third interval. These four-dimensional joint probability functions can be written as the product of two two-dimensional joint probability functions by noting that when the interval in which the increment occurred is given, the observation of λ_2 is independent of λ_3 and the observation of θ_2 is independent of θ_3. Since there are two intervals in which the target level can be incremented, there are four relevant two-dimensional probability density functions.

The relevant two-dimensional joint probabilities are the probability of the observed values of Λ and θ on a particular interval given a target level. The log-likelihood ratio $\eta_{\Lambda,\Theta}$ is defined in terms of these probabilities as

$$\eta_{\Lambda,\Theta}(\lambda_2,\theta_2,\lambda_3,\theta_3,\Delta L) = 10\log_{10}\left(\frac{f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_2,\theta_2|L_0)f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_3,\theta_3|L_0+\Delta L)}{f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_2,\theta_2|L_0+\Delta L)f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_3,\theta_3|L_0)}\right).$$

The notation in this equation is designed to distinguish the identity of the functions from the values and variable in the arguments. Thus, the subscripts on $\eta_{\Lambda,\Theta}$ identify that this likelihood ratio applies to the case when both variables are available and it has five arguments.

The ideal observer is represented by a binary indicator function $\psi_{\Lambda,\Theta}$, which is calculated from the likelihood ratio. Specifically, when the *a priori* probabilities of each interval are equal and when the goal is to maximize the probability of a correct response, the optimum decision is determined by the sign of the log-likelihood ratio $\eta_{\Lambda,\Theta}$: when $\eta_{\Lambda,\Theta}$ is positive, the second interval is most likely to have the incremented target and that is the optimum decision. When $\eta_{\Lambda,\Theta}$ is negative, the third interval is more likely to have the incremented target and that is the optimum decision. This decision rule can be represented by the binary indicator $\psi_{\Lambda,\Theta}$, which is equal to unity when the optimum decision is the second interval and zero otherwise. Mathematically, the indicator function is

$$\psi_{\Lambda,\Theta}(\lambda_2,\theta_2,\lambda_3,\theta_3,\Delta L) = \begin{cases} 1 & \text{when } \eta_{\Lambda,\Theta}(\lambda_2,\theta_2,\lambda_3,\theta_3,\Delta L) \geq 0 \\ 0 & \text{when } \eta_{\Lambda,\Theta}(\lambda_2,\theta_2,\lambda_3,\theta_3,\Delta L) < 0 \end{cases}.$$

Then, the probability that the ideal observer of both Λ and Θ achieves the correct answer is a function of ΔL and can be written as

$$P_{\text{correct}}(\Delta L) = \frac{1}{2}\int\int\int\int [\psi_{\Lambda,\Theta}(\lambda_2,\theta_2,\lambda_3,\theta_3,\Delta L)f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_2,\theta_2|L_0)f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_3,\theta_3|L_0+\Delta L)]d\theta_2 d\lambda_2 d\theta_3 d\lambda_3$$
$$+\frac{1}{2}\int\int\int\int [(1-\psi_{\Lambda,\Theta}(\lambda_2,\theta_2,\lambda_3,\theta_3,\Delta L))f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_2,\theta_2|L_0+\Delta L)f_{\Lambda,\Theta|L_{\text{target}}}(\lambda_3,\theta_3|L_0)]d\theta_2 d\lambda_2 d\theta_3 d\lambda_3.$$

Thus, for the optimum decision rule, the probability of a correct response depends only on the joint probability density function of Λ and Θ given L_{target} since the indicator function is also defined in terms of this density function. This joint probability density function $f_{\Lambda,\Theta|L_{\text{target}}}$ is approximated with numerical methods and the details of this approximation are contained in Appendix B. The probability of a correct response for the ideal observers of Λ alone or of Θ alone is calculated in an analogous manner and the derivation is not presented.

APPENDIX B

In this appendix, analytical and numerical techniques are used to approximate the joint density function of Λ and Θ, as defined in Eqs. (2) and (3), for a target level L_{target}. As outlined in Appendix A, knowledge of the joint density allows the calculation of the probability of a correct response in our experiment. This appendix derives a relatively simple expression that can be evaluated using standard numerical functions and techniques. Before the details of the derivation of $f_{\Lambda,\Theta|L_{\text{target}}}$ are outlined, the model variables are related to

the experimental variables. Specifically, the values that are appropriate for the psychophysical experiment are substituted into Eqs. (2) and (3). All levels are in decibels. In the experiment, L_{left} is always the level of the target L_{target} and L_{right} is the level of the distractor. The level of the target is the sum of a reference level L_0 and an increment ΔL, such that $L_{\text{left}}=L_{\text{target}}=L_0+\Delta L$, where ΔL is zero when the target is not incremented. The distractor has a level that is equal to the reference level plus a random variable A; the level of the right ear can, therefore, be written as $L_{\text{right}}=L_0+A$. The interaural time difference T is the phase delay, which is defined as the negative of the distractor phase (also a random variable) divided by the radian frequency ω (i.e., $T=-\Phi/\omega$). The psychophysical experiment specifies that A has a uniform probability density function between $a_{\min}$ and $a_{\max}$. The experiment also specifies that Φ has a uniform probability density function between $\phi_{\min}$ and $\phi_{\max}$. Using this notation, when the distractor level is roved, $a_{\min}=0$ and $a_{\max}=30$, and when the distractor level is fixed, $a_{\min}=a_{\max}=0$. Similarly, when the distractor phase is roved, $\phi_{\min}=-\pi/2$ and $\phi_{\max}=\pi/2$, and when the distractor phase is fixed, $\phi_{\min}=\phi_{\max}=0$. Making these substitutions into Eqs. (2) and (3) results in

$$\Lambda = 10\log_{10}(10^{L_{\text{target}}/10}+10^{(L_0+A)/10})+N_\Lambda \tag{B1}$$

and

$$\Theta = L_{\text{target}}-(L_0+A)-\frac{k}{\omega}\Phi+N_\Theta. \tag{B2}$$

Our derivation of $f_{\Lambda,\Theta|L_{\text{target}}}$ begins by using the definition of conditional probability to expand the joint density function to

$$f_{\Lambda,\Theta|L_{\text{target}}}=f_{\Lambda|L_{\text{target}}}(\lambda|L_0+\Delta L)f_{\Theta|\Lambda,L_{\text{target}}}(\theta|\lambda,L_0+\Delta L).$$

Then, by using the fact that $f_{\Theta|\Lambda,L_{\text{target}}}$ can be obtained by integrating $f_{\Theta,A,\Phi|\Lambda,L_{\text{target}}}$ over all values of a and ϕ representing the values of the variables A and Φ, one obtains

$$f_{\Lambda,\Theta|L_{\text{target}}}=f_{\Lambda|L_{\text{target}}}(\lambda|L_0+\Delta L)\times\int\int[f_{\Theta|A,\Phi,\Lambda,L_{\text{target}}}(\theta|a,\phi,\lambda,L_0+\Delta L)\times f_{A,\Phi|\Lambda,L_{\text{target}}}(a,\phi|\lambda,L_0+\Delta L)]d\phi da.$$

Making a substitution of Θ based on Eq. (B2) and using the definition of conditional probability yields

$$f_{\Lambda,\Theta|L_{\text{target}}}=f_{\Lambda|L_{\text{target}}}(\lambda|L_0+\Delta L)\int\int f_{N_\Theta}\left(\frac{k}{\omega}\phi-\mu_\Theta(a)\right)\times f_{A,\Phi|\Lambda,L_{\text{target}}}(a,\phi|\lambda,L_0+\Delta L)d\phi da,$$

where $\mu_\Theta(a)$ is equal to $\Delta L-a-\theta$. Using the definition of conditional probability for $f_{A,\Phi|\Lambda,L_{\text{target}}}$ and then noting the independence of Φ and A, Λ, and L_{target} gives

$$f_{\Lambda,\Theta|L_{\text{target}}}=f_{\Lambda|L_{\text{target}}}(\lambda|L_0+\Delta L)\int\int f_{N_\Theta}\left(\frac{k}{\omega}\phi-\mu_\Theta(a)\right)\times f_{A|\Lambda,L_{\text{target}}}(a|\lambda,L_0+\Delta L)f_\Phi(\phi)d\phi da.$$

By using the definition of conditional probability on $f_{A|\Lambda,L_{\text{target}}}$, noting the statistical independence of $f_{A|L_{\text{target}}}$ and L_{target}, and simplifying, one obtains

$$f_{\Lambda,\Theta|L_{\text{target}}}=\int f_A(a)f_{\Lambda|A,L_{\text{target}}}(\lambda|a,L_0+\Delta L)\times\int f_\Phi(\phi)f_{N_\Theta}\left(\frac{k}{\omega}\phi-\mu_\Theta(a)\right)d\phi da.$$

By making use of the uniform probability density functions of the random variables A and Φ, $f_{\Lambda,\Theta|L_{\text{target}}}$ can be rewritten as

$$f_{\Lambda,\Theta|L_{\text{target}}}=c\int_{a_{\min}}^{a_{\max}}f_{\Lambda|A,L_{\text{target}}}(\lambda|a,L_0+\Delta L)\times\int_{\phi_{\min}}^{\phi_{\max}}f_{N_\Theta}\left(\frac{k}{\omega}\phi-\mu_\Theta(a)\right)d\phi da,$$

where c is equal to $1/(a_{\max}-a_{\min})(\phi_{\max}-\phi_{\min})$.

From Eq. (B1) it follows that

$$f_{\Lambda|A,L_{\text{target}}}=f_{N_\Lambda}(\lambda-10\log_{10}(10^{(L_0+\Delta L)/10}+10^{(L_0+a)/10})).$$

Making substitutions for the density functions of N_Λ and N_Θ yields

$$f_{\Lambda,\Theta|L_{\text{target}}}=\frac{c}{2\pi\sigma_\Theta\sigma_\Lambda}\int_{a_{\min}}^{a_{\max}}e^{-((\lambda-\mu_\Lambda(a))^2/2\sigma_\Lambda^2)}\times\int_{\phi_{\min}}^{\phi_{\max}}e^{-(\phi-\omega/k\mu_\Theta(a))^2/2(\omega/k\sigma_\Theta)^2}d\phi da, \tag{B3}$$

where $\mu_\Lambda(a)$ is the conditional expected value of Λ given that A is equal to a, which is equal to $10\log_{10}(10^{(L_0+\Delta L)/10}+10^{(L_0+a)/10})$.

Further analytical manipulations of $f_{\Lambda,\Theta|L_{\text{target}}}$ do not appear to reduce the complexity of the solution, but $f_{\Lambda,\Theta|L_{\text{target}}}$ as represented by Eq. (B3) above can be numerically approximated. The first step of the numerical implementation is to approximate the definite integral over A through a finite summation. Let us denote $a[n]$ as a sampled version of the continuous random variable A. Further, let $a[1]$ equal $a_{\min}$ and $a[N]$ equal $a_{\max}$. The probability density function $f_{\Lambda,\Theta|L_{\text{target}}}$ can then be numerically approximated as

$$f_{\Lambda,\Theta|L_{\text{target}}}\approx\frac{1}{N(\phi_{\max}-\phi_{\min})2\pi\sigma_\Theta\sigma_\Lambda}\times\sum_{n=1}^{N}e^{-((\lambda-\mu_\Lambda(a[n]))^2/2\sigma_\Lambda^2)}\times\int_{\phi_{\min}}^{\phi_{\max}}e^{-(\phi-(\omega/k)\mu_\Theta(a[n]))^2/2((\omega/k)\sigma_\Theta)^2}d\phi.$$

Note that the integral of the exponential function can be represented as the error function so that the whole expression

is easily numerically evaluated. One should note that in the limit when $a_{max}-a_{min}=0$ or $\phi_{max}-\phi_{min}=0$, one cannot simply evaluate this expression at zero but must rather evaluate the expression in the limit as the difference approaches zero.

[1]Although the first and last intervals convey no information for the ideal observer, these intervals may aid the nonideal subjects.

Bernstein, J. G., and Oxenham, A. J. (**2003**). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?," J. Acoust. Soc. Am. **113**, 3323–3334.

Bernstein, L. R. (**2004**). "Sensitivity to interaural intensive disparities: Listeners' use of potential cues," J. Acoust. Soc. Am. **115**, 3156–3160.

Bernstein, L. R., and Trahiotis, C. (**1982**). "Detection of interaural delay in high frequency noise," J. Acoust. Soc. Am. **71**, 147–152.

Bilsen, F. A. (**1977**). "Pitch of noise signals: Evidence for a 'Central spectrum,'" J. Acoust. Soc. Am. **61**, 150–161.

Bilsen, F. A., and Raatgever, J. (**2000**). "On the dichotic pitch of simultaneously presented interaurally delayed white noises: Implications for binaural theory," J. Acoust. Soc. Am. **108**, 272–284.

Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge).

Bonnel, A. M., and Hafter, E. R. (**1998**). "Divided attention between simultaneous auditory and visual signals," Percept. Psychophys. **60**, 179–190.

Brungart, D. S., and Simpson, B. D. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–2995.

Buus, S., and Florentine, M. (**1991**). "Psychometric functions for level discrimination," J. Acoust. Soc. Am. **90**, 1371–1380.

Colburn, H. S., and Durlach, N. I. (**1978**). "Models of binaural interaction," in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), vol. IV.

Durlach, N. I., and Colburn, H. S. (**1978**). "Binaural phenomena," in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), vol. IV.

Green, D. M. (**1988**). *Profile Analysis: Auditory Intensity Discrimination* (Oxford University Press, New York).

Hafter, E. R. (**1971**). "Quantitative evaluation of a lateralization model of masking-level differences," J. Acoust. Soc. Am. **50**, 1116–1122.

Hafter, E. R., and Jeffress, L. A. (**1968**). "Two-image lateralization of tones and clicks," J. Acoust. Soc. Am. **44**, 563–569.

Hafter, E. R., and Carrier, S. C. (**1970**). "Masking-level differences obtained with a pulsed tonal masker," J. Acoust. Soc. Am. **47**, 1041–1047.

Hartmann, W. M., and Constan, Z. A. (**2002**). "Interaural level differences and the level-meter model," J. Acoust. Soc. Am. **112**, 1037–1045.

Heller, L. M., and Trahiotis, C. (**1995**). "The discrimination of samples of noise in monotic, diotic, and dichotic conditions," J. Acoust. Soc. Am. **97**, 3775–3781.

Kidd, G. J., Mason, C. R., Arbogast, T. L., Brungart, D. S., and Simpson, B. D. (**2003**). "Informational masking caused by contralateral stimulation," J. Acoust. Soc. Am. **113**, 1594–1603.

Koehnke, J., and Besing, J. M. (**1992**). "Effects of roving level variation on monaural detection with a contralateral cue," J. Acoust. Soc. Am. **92**, 2625–2629.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Mills, J. H., Dubno, J. R., and He, N.-J. (**1996**). "Masking by ipsilateral and contralateral maskers," J. Acoust. Soc. Am. **100**, 3336–3344.

Rowland, R. C. J., and Tobias, J. V. (**1967**). "Interaural intensity difference limen," J. Speech Hear. Res. **10**, 745–756.

Ruotolo, B. R., Stern, R. M. J., and Colburn, H. S. (**1979**). "Discrimination of symmetric time-intensity traded binaural stimuli," J. Acoust. Soc. Am. **66**, 1733–1737.

Sivonen, V. P., and Ellermeier, W. (**2006**). "Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation," J. Acoust. Soc. Am. **119**, 2965–2980.

Stern, R. M., and Trahiotis, C. (**1997**). "Models of binaural perception," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey, and T. R. Anderson (Erlbaum, Mahwah).

Taylor, M. M., and Clarke, D. P. J. (**1971**). "Monaural detection with contralateral cue (MDCC). II. Interaural delay of cue and signal," J. Acoust. Soc. Am. **49**, 1243–1253.

Taylor, M. M., Clarke, D. P. J., and Smith, S. M. (**1971a**). "Monaural detection with contralateral cue (MDCC). III. Sinusoidal signals at a constant performance level," J. Acoust. Soc. Am. **49**, 1795–1804.

Taylor, M. M., Smith, S. M., and Clarke, D. P. (**1971b**). "Monaural detection with contralateral cue (MDCC). IV. Psychometric functions with sinusoidal signals," J. Acoust. Soc. Am. **50**, 1151–1161.

Trahiotis, C. (**1992**). "Developmental considerations in binaural hearing experiments," in *Developmental Psychoacoustics*, edited by L. A. Werner and E. W. Rubel (American Psychological Association, Washington, DC), vol. 1.

Viemeister, N. F. (**1988**). "Psychophysical aspects of auditory intensity coding," in *Auditory Function: Neurobiological Bases of Hearing*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (J. Wiley, New York).

Whilby, S., Florentine, M., Wagner, E., and Marozeau, J. (**2006**). "Monaural and binaural loudness of 5- and 200-ms tones in normal and impaired hearing," J. Acoust. Soc. Am. **119**, 3931–3939.

Yost, W. A. (**1972**). "Tone-on-tone masking for three binaural listening conditions," J. Acoust. Soc. Am. **52**, 1234–1237.

Yost, W. A., Penner, M. J., and Feth, L. L. (**1972**). "Signal detection as a function of contralateral sinusoid-to-noise ratio," J. Acoust. Soc. Am. **51**, 1966–1970.

Zurek, P. M. (**1979**). "Measurements of binaural echo suppression," J. Acoust. Soc. Am. **66**, 1750–1757.

Zwicker, E., and Zwicker, U. T. (**1991**). "Dependence of binaural loudness summation on interaural level differences, spectral distribution, and temporal distribution," J. Acoust. Soc. Am. **89**, 756–764.

Zwislocki, J. J. (**1972**). "A theory of central auditory masking and its partial validation," J. Acoust. Soc. Am. **52**, 644–659.

Influence of supraglottal structures on the glottal jet exiting a two-layer synthetic, self-oscillating vocal fold model

James S. Drechsel and Scott L. Thomson[a)]
Department of Mechanical Engineering, Brigham Young University, Provo, Utah 84602

(Received 9 August 2007; revised 16 January 2008; accepted 21 February 2008)

A synthetic two-layer, self-oscillating, life-size vocal fold model was used to study the influence of the vocal tract and false folds on the glottal jet. The model vibrated at frequencies, pressures, flow rates, and amplitudes consistent with human phonation, although some differences in behavior between the model and the human vocal folds are noted. High-speed images of model motion and flow visualization were acquired. Phase-locked ensemble-averaged glottal jet velocity measurements using particle image velocimetry (PIV) were acquired with and without an idealized vocal tract, with and without false folds. PIV data were obtained with varying degrees of lateral asymmetric model positioning. Glottal jet velocity magnitudes were consistent with those measured using excised larynges. A starting vortex was observed in all test cases. The false folds interfered with the starting vortex, and in some cases vortex shedding from the false folds was observed. In asymmetric cases without false folds, the glottal jet tended to skew toward the nearest wall; with the false folds, the opposite trend was observed. rms velocity calculations showed the jet shear layer and laminar core. The rms velocities were higher in the vocal tract cases compared to the open jet and false fold cases. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2897040]

PACS number(s): 43.70.Aj, 43.70.Bk, 43.70.Gr [AL] Pages: 4434–4445

I. INTRODUCTION

As the vocal folds cyclically open and close, an orifice-modulated jet is formed. This glottal jet has been the subject of numerous studies since it comprises an essential source of sound in speech. A better understanding of the glottal jet and its interactions with supraglottal boundaries (e.g., the vocal tract lumen and the false folds) will yield insight into the glottal sound source. Improved understanding of the physical characteristics of the glottal jet can also yield insight into general vocal fold behavior and can benefit future development of analytical and computational laryngeal models.

Excised larynges have been used to study the glottal jet, for example, using hot-wire anemometry (e.g., Alipour and Scherer, 1995; Alipour *et al.*, 1996) and particle image velocimetry [(PIV), Khosla *et al.* (2007)]. In the latter, PIV and high-speed imaging were used to evaluate, respectively, the flow structures immediately downstream of the glottal exit and the vocal fold motion for three different canine larynges. These experiments studied the glottal jet without a vocal tract for one subglottal pressure. Phase-locking was used to obtain two-dimensional average velocity vector fields over 30 phase positions; 10 averages at each phase were used. Various glottal jet vortical structures were described.

Experiments such as these that have used real vocal folds have the advantage of anatomical and physiological "realism." For example, present in excised models is the surrounding cartilaginous framework, the multiple tissue layers of the vocal folds, and the anistropic, nonhomogeneous nature of the tissue itself. However, *in vivo* human and canine studies offer only limited instrument access, and *in vivo* human studies are limited by potential health hazards. Excised larynges are much more accessible, but are typically only able to be vibrated for a few minutes and require additional conditioning (e.g., flow heating and humidification, in addition to periodic direct wetting of the excised larynx tissue). Both *in vivo* and excised models have limited potential for use in parametric studies involving geometry and/or material properties. Synthetic models overcome the above-noted challenges in that they have long usable lifetimes and are relatively easily parametrized.

Numerous studies using static and driven synthetic vocal fold models have been performed, e.g., Scherer *et al.* (1983, 2001), Pelorson *et al.* (1994), Hofmans *et al.* (2003), Shinwari *et al.* (2003), and Triep *et al.* (2005). These have been performed to study aspects of phonation such as intraglottal pressure distribution and glottal jet characteristics. Shadle *et al.* (1991) constructed a dynamic, life-size vocal fold model primarily for the purpose of flow visualization. The model consisted of one of a pair of thin shutters representing the vocal folds driven in simple harmonic motion by an electromechanical shaker; the glottal profile was rectangular. This configuration was described as representing one functioning and one paralyzed vocal fold. This setup also included a vocal tract and a set of hemisphere-shaped false vocal folds. It was observed that the false folds had a "straightening effect" on the glottal jet, even when the glottis was off-center. Barney *et al.* (1999) and Shadle *et al.* (1999) used a modification of this setup to investigate the influence of glottal velocities on the glottal acoustic source. A vocal tract was included, but false folds were not.

Mongeau *et al.* (1997) developed a dynamic, driven, life-size, open-jet vocal fold model with a convergent glottis profile. One goal was to validate the quasi-steady approximation assumed in static studies. Measurements using this

[a)]Electronic mail: thomson@byu.edu

 0001-4966/2008/123(6)/4434/12/$23.00

model generally agreed with data presented using the static models of Scherer (1981) and with excised canine experiments of Alipour and Scherer (1995). The quasi-steady assumption was shown to be valid for all but roughly 1/5th of the cycle (during glottal opening and closing). Zhang *et al.* (2002a, 2004) used this same model to further validate the quasi-steady assumption. A vocal tract was included, although the focus was primarily on acoustic measurements and no comparison was made between the open jet and vocal tract cases.

Erath and Plesniak (2006a, b, c) used a static model with a dynamic downstream boundary condition to study pulsatile jet flow through a divergent glottis. These studies were performed in a wind tunnel on a 7.5× scaled-up model and included a downstream length to simulate a vocal tract. An unsteadiness generator (rotating shutters downstream of the model test section) was used to create pulsatile flow. These studies used PIV to visualize and quantify the jet flow. They found that the jet tendency to attach to one side over the other was influenced by slight asymmetry in the glottis. At divergence angles of 10° and 20°, the jet exhibited bimodal behavior. However, at 40° divergence angle, the jet generally did not attach to either wall; this was a consequence of the jet being in the fully developed two-dimensional stall regime. It was found that asymmetric glottis conditions enhanced stability of the glottal jet.

Self-oscillating synthetic vocal fold models have also been studied. These types of models are important because they vibrate due to coupled fluid–solid–acoustic interactions, as the human vocal folds do. Titze *et al.* (1995) and Chan and Titze (1997) used a dynamic, synthetic vocal fold model that consisted of a stainless steel "body" covered by a silicone "epithelium" to study phonation threshold pressure. A hemilarynx configuration was used. Between the epithelium and the body was a cavity into which fluids of varying viscosity were injected; this was to simulate the superficial lamina propria. The model cover self-oscillated, but the body was static. This study did not use a vocal tract. These studies generally confirmed the previously developed model of phonation threshold pressure of Titze (1988) for large glottal widths, but showed that smaller widths produced results that deviated from the analytical model. Chan and Titze (2006) refined this model by using viscoelastic biomaterials in the cover layer and including a supraglottal tube. They found that the presence of a vocal tract consistently reduced phonation threshold pressure. The glottal jet itself was not studied.

Thomson *et al.* (2005) used a self-oscillating, synthetic vocal fold model made of a three-part silicone solution. The model was one-layer (homogeneous), isotropic, and was generally geometrically similar to the real vocal folds, but with a uniform cross section. A manufacturing process was described in which the model could be manufactured in separate layers of different material properties. The model was used in combination with a corresponding computational model of a similar one-layer model to quantify the aerodynamic energy transfer from the glottal jet to the vocal folds. It was demonstrated that the net energy transfer during glottal opening was greater than that during closing. Riede *et al.* (2008) applied this process to manufacture a two-layer vocal fold model. Zhang *et al.* (2006a) used the one-layer, isotropic model to study the effect of subglottal acoustics on synthetic vocal fold model vibration. It was shown experimentally and analytically that the one-layer model vibration was strongly coupled to acoustical resonances of the subglottal tract. In a follow-up study, Zhang *et al.* (2006b) investigated aerodynamically and acoustically driven modes of vibration. Restraining the vertical motion of the lateral regions of the one-layer model superior surface resulted in the model vibrating in response to aerodynamic coupling with the glottal jet rather than to acoustic coupling with the subglottal system. Neither Thomson *et al.* (2005) or Zhang *et al.* (2006a, b) used a vocal tract in their experiments.

Neubauer *et al.* (2007) used this one-layer model to examine flow structures immediately downstream of the glottal exit. PIV and high speed imaging was used to quantify velocities and to visualize near-field flow structures. Flow structures were extracted from the jet flow using principal component analysis. Measured results included oscillating jet angle, shear layer structure, flapping of the turbulent region, and coherent structures. This study observed vortex generation, vortex convection, and jet flapping. This study was essentially for an open jet case. The study was limited to a few quasi-phase-locked flow images (i.e., instantaneous, rather than ensemble-averaged, data). They observed asymmetric flow in the glottal jet, identified the Coanda effect as being present in vibrating model, observed evidence of feedback/feedforward coupling mechanisms between the glottal flow and the downstream near field, and suggested that "flapping" of the glottal jet in the turbulent region of the glottal jet was due to large-scale vortical motion.

Few glottal jet studies using self-oscillating models, real or synthetic, have compared flow structures with and without a vocal tract. This paper describes a series of experiments performed to investigate the influence of different supraglottal loading configurations on the jet exiting the vocal folds. A two-layer [body-cover, Hirano and Kakita (1985)], life-sized, self-oscillating vocal fold model was used. Supraglottal configurations included an open jet and eight cases of varying vocal tract asymmetry (five without false folds and three with false folds); all of these were studied at multiple subglottal pressures. Glottal exit velocity was measured quantitatively using PIV; model motion and the resulting glottal jet was observed qualitatively using high speed imaging. The model motion was also captured using high speed imaging.

II. METHODS

A. Synthetic model

The synthetic model used was a recently developed two-layer version of the Thomson *et al.* (2005) one-layer model (Riede *et al.*, 2008). The model was created using three-component addition-cure silicone (single-part silicone thinner and two-part Ecoflex 0030, Smooth-On, Inc., Easton, PA). The model consisted of two layers of differing modulus, as illustrated in Fig. 1. The model anterior-posterior length was approximately 1.7 cm. The cover layer was approximately 2 mm thick. The cover and body layers had Young's

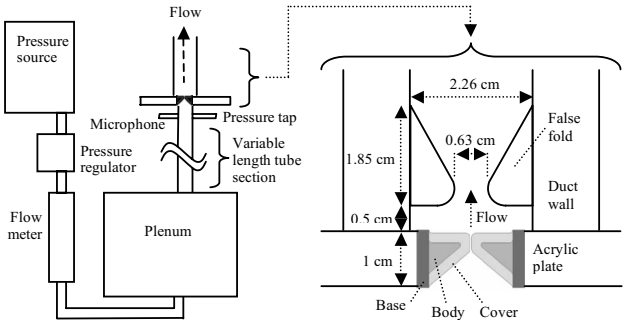


FIG. 1. Schematic and approximate dimensions of the synthetic, two-layer vocal folds and the false folds used in this study. The medial surfaces of the false folds were rounded as shown with a radius of approximately 0.33 cm.

moduli of approximately 4.1 and 22.5 kPa, respectively. In the human vocal folds, the Young's modulus is nonlinearly dependent on strain (due primarily to the presence of collagen fibers) and also depends on strain direction (due to tissue anisotropy). Relatively little data are available regarding the transverse Young's modulus of vocal fold tissue. In one study, Tran *et al.* (1993) measured the transverse Young's modulus of the human vocal fold and found the mean value to be 12.7 kPa (with no recurrent laryngeal nerve stimulation). Further details regarding the model fabrication process, including model geometry, can be found elsewhere (Riede *et al.*, 2008; Drechsel, 2007).

B. Experimental model setup

The model was placed in a test fixture and not moved over the measurement duration. Two vocal fold models were each attached to a separate rectangular acrylic plate [similar to that of Thomson *et al.* (2005)]. These acrylic plates supported the vocal fold models and provided a surface for mounting to the subglottal and supraglottal tracts. The gap between the acrylic plates was sealed using closed-cell foam. The two acrylic pieces were brought together, compressing the foam and bringing the medial surfaces together so that the midregions were just touching. In this model, a small gap existed at one end of the resting orifice (perhaps similar to a posterior commissure). Bolts through the acrylic pieces were tightened to control and maintain the initial glottal gap.

The acrylic plate assembly was mounted to a subglottal test section, illustrated in Fig. 1. A 60 cm cylindrical subglottal tube with a cross-sectional area of 5 cm^2 was used for these tests. Zhang *et al.* (2006a) studied the acoustic coupling of one-layer model vibration with subglottal acoustics. Other research (Drechsel, 2007) showed that the two-layer model exhibited similar coupling with subglottal acoustics as the one-layer models. The 60 cm subglottal tube length was selected on this basis to minimize the model onset pressure. A differential pressure transducer (Omega PX138-001D5V) was placed approximately 3 cm upstream of the acrylic plate. An expansion plenum (30.5 cm on each side) was placed at the upstream end of the subglottal tube. Shop air was used as a flow source, with a pressure regulator (Pneufine 26129-1C-19) and a flow meter (Matheson 605) located between the source and the plenum.

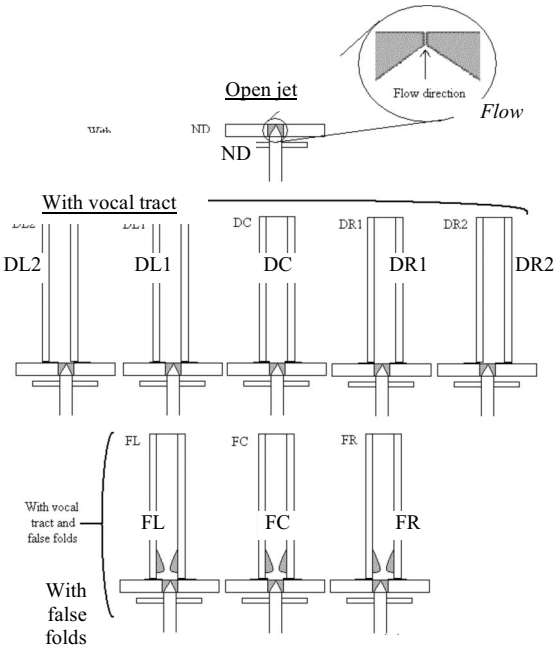


FIG. 2. Vocal tract configurations. ND denotes the no duct, or open jet, case. DL2 and DR2 denote the extreme left and right vocal tract offset cases (±4 mm) without the false folds, respectively, DL1 and DR1 denote the left and right offset cases (±1.5 mm), respectively, and DC denotes the symmetric vocal tract case, all without the false folds. FL, FC, and FR denote cases with false folds and the vocal tract positioned to the left (−1.5 mm), center, and right (+1.5 mm), respectively, of the vocal folds.

An idealized vocal tract was manufactured from 1.25-cm-thick aluminum on three sides, with the fourth wall made of 0.32-cm-thick glass for optical access. It was 30.5 cm long and had a square cross-sectional area of 5 cm^2. When used, the vocal tract was mounted to the acrylic plate assembly using a plastic-laminated grid for lateral positioning.

A model of the false vocal folds (shown in Fig. 1) was manufactured out of clear rigid acrylic according to the mean values for an adult male reported in Agarwal *et al.* (2003) (see Drechsel, 2007 for further details). The lateral gap between the false folds was 0.63 cm; the vertical gap between the top of the (nonvibrating) vocal folds and the bottom of the false folds was 0.5 cm. The false folds were polished to improve transmission of the laser sheet into the space between the vocal folds and false folds.

The supraglottal test configurations (see Fig. 2) included an open jet (no vocal tract), a vocal tract without false folds, and a vocal tract with false folds. A 32 cm × 40 cm × 57 cm shroud was used with the open jet configuration to increase seed density in the regions around the jet and thus enable velocity vector calculation; this was deemed to be sufficiently large to exert negligible influence on the glottal jet.

For the vocal tract without false folds configuration, five different cases were studied: one with the duct centered over the vocal folds (symmetric case), and four in which the vocal tract was laterally offset from the vocal fold medial plane (±1.5 and ±4 mm). For the vocal tract with false folds, three cases were studied: one with the duct centered over the vocal folds (symmetric case), and two in which the vocal tract was laterally offset from the vocal fold medial plane

(± 1.5 mm). The nomenclature used to denote these various cases is given in Fig. 2. These cases were tested for three mean subglottal pressures: 1.25, 1.5, and 1.9 kPa, although only the 1.25 and 1.9 kPa results are presented here.

As mentioned earlier, one wall of the vocal tract was made of glass. This presented a potential problem when seed particles were introduced to the airflow. If either the seed density or the driving pressure were sufficiently high, oil would accumulate on the glass, inhibiting imaging. DEHS oil (di-ethyl-hexyl sebacate, CAS#122-62-3) was used as the seed particle because it accumulated less on the glass than often-used olive oil. Measurements of the particle diameter were not made directly, but manufacturer data specify that the generator produces particles ranging from 0.2 to 1 μm. The Stokes number (St) was calculated to estimate the particles' ability to follow the flow; the Stokes number, $\mathrm{St}=\rho d^2 U/(18\mu L)$, was calculated, where ρ is particle density (912 kg/m^3), d is particle diameter, U is fluid velocity (60 m/s used here for a high-end estimate), μ is air viscosity (1.8×10^{-5} Pa s), and L is a flow length scale (the streamwise intraglottal length of 3 mm was used here). Using these parameters, Stokes number estimates ranged from 0.0023 to 0.056 for 0.2 to 1 μm sized particles, respectively, thus satisfying the $\mathrm{St}\ll 1$ criteria.

Generally, as the pressure increased, the seed density could be reduced to minimize oil accumulation. However, oil accumulation persisted at the 1.9 kPa pressure, resulting in fewer images being able to be acquired (see Sec. II E).

C. High-speed imaging setup

High-speed images of the vocal folds and of the glottal jet were acquired using a Photron FASTCAM-APX RS high-speed camera system. The resolution varied depending on the view of interest. Three views were imaged, with resolutions as follows: 640×448 pixels (for top view of the vocal fold motion), 896×624 pixels (flow visualization, side view), and 768×784 (flow visualization, front view), with corresponding frame rates of 10 000, 5000, and 5000 frames/s, respectively. High-intensity white LEDs (Visual Instrumentation Corporation, single-LED array model 200800, controller model 200900) were pointed directly at the vocal folds for imaging the motion, and directed in a forward-scattering mode for jet illumination in the flow visualization studies. LEDs were used to minimize heating of the vocal folds. For flow visualization the flow was seeded using a LaVision Aerosol Generator and DEHS oil.

D. Particle imaging velocimetry experimental setup

For the present research, two-dimensional PIV was used, employing the double-image/cross-correlation method, in which the camera captured two successive images separated by time dt. This technique was used by Khosla *et al.* (2007) and Neubauer *et al.* (2007). PIV assumes that the seed particles have negligible inertia with respect to the fluid momentum; this and the neutrally buoyant seed particle criteria were satisfactorily verified for these experiments (Drechsel, 2007). Compared to other velocity measurement techniques, PIV is relatively nonintrusive, is easily calibrated, and yields a significant quantity of flow information. The main drawback is that acquisition rates of only about 5 Hz are typical, requiring phase-locking and ensemble averaging for higher frequency events.

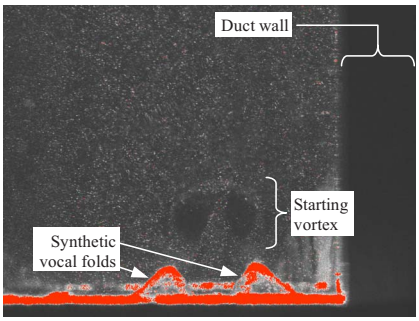


FIG. 3. (Color online) Example PIV image obtained using the current setup.

A LaVision PIV system was used, including a LaVision Imager Intense CCD camera, a LaVision Aerosol Generator, and a New Wave Research Solo II Nd:YAG laser. The camera resolution was 1376×1040 pixels. The laser sheet was approximately 0.5–1.0 mm thick. Again, DEHS oil was used for flow seeding. The LaVision software package, DAVIS, was used for image acquisition and processing.

The laser sheet was in the model frontal plane, approximately centered anterior-posteriorly. The camera was positioned so that the top of the vocal fold model was just above the bottom of the field of view, with the glottal jet approximately centered left-to-right. Care was taken to make all surfaces in the field of view as dark as possible to minimize laser reflection into the camera. The vocal fold fixture was often checked to ensure that the model was correctly centered and level in the camera frame. The camera field of view was adjusted to extend to an area as wide as the inside of the vocal tract and just higher than the false folds, or about 22 mm (horizontal) $\times$ 16 mm (vertical). Calibration was accomplished by placing a ruler in the camera view. A typical calibration constant was approximately 54 pixels/mm. Figure 3 shows a typical image obtained using this setup using PIV (shown with an offset vocal tract configuration).

E. Data analysis

1. PIV analysis

The model frequency was fairly consistent; however, even small fluctuations in frequency would inhibit successful ensemble averaging based on a constant frequency data acquisition. Therefore, a custom phase-locking circuit trigger the PIV image acquisition using the subglottal pressure sensor as the triggered input (see Drechsel, 2007 for details).

The PIV software allowed for adjustable time delay following a trigger input. Using this feature, 30 phases of vocal fold oscillation were interrogated, with either 100 or 50 image pairs (depending on pressure as discussed in the following) collected at each phase. The dt was set to 3 μs. The image pairs were analyzed using a nonweighted, double-pass, decreasing window size (32×32 to 16×16 pixels), 50% overlap, cross-correlation algorithm. No postprocessing was performed to correct for potentially spurious vectors in

the resulting vector fields. The vector fields were then ensemble-averaged and the rms velocity fields were calculated as discussed in the following.

The nine vocal tract configuration cases repeated at three different pressures required approximately 4.5 h of nearly continuous synthetic model vibration. Considering the additional time involved in experiment preparation, the model was actually run for many more hours. This is much longer than excised larynx models allow and demonstrates one aspect of synthetic model usefulness. Separate studies (Drechsel, 2007) have confirmed that the model vibratory behavior was fairly consistent over such long duration.

2. rms velocity calculation

The rms of the velocity is a spatially resolved measure of the fluctuation of the instantaneous velocity fields, and is calculated according to

$$|\mathrm{rms}| = \sqrt{\frac{1}{n-1}\left(\sum_i (u_i - \bar{u})^2 + (v_i - \bar{v})^2\right)} \quad (1)$$

where n is the number of averaged image pairs at a given phase, u_i and v_i are the x and y components of the instantaneous velocity of the ith image pair, and $\bar{u}$ and $\bar{v}$ are the local average x- and y-velocity components. Calculation of the $|\mathrm{rms}|$ velocity field is useful in identifying the jet core and shear layer regions and in quantifying the local jet velocity fluctuations.

3. Required number of images

Ensemble averaging was used to reduce noise and to calculate average and rms velocity quantities. The number of image pairs necessary for convergence of these quantities was studied using two points in the flow field. The first point was in the jet core, a short distance above the glottal exit. The second point was located at the same vertical distance as the first point, but positioned laterally in the jet shear layer. Figure 4 shows the results of this study.

According to Fig. 4, the average and rms velocities appear to have stabilized somewhat between 100 and 200 averages for both point locations, although some fluctuations persist. For example, the jet core average velocity fluctuations are within about 0.3 m/s between 100 and 200 averages, or less than 2% of the jet core velocity. The shear layer average velocity fluctuations are within about 0.2 m/s between 100 and 200 averages, or about 20% of the shear layer velocity. As the number of averages increases, so also do the time and computer storage requirements. Further, in cases with a vocal tract, more averages allowed for more time for oil to accumulate on the glass. As a compromise, 100 averages were used for the 1.25 kPa pressure study, and 50 averages were used for the 1.9 kPa pressure study. It is noted that a more rigorous study of the number of data points required to calculate the true mean and rms velocities would entail performing this similar calculation not only at two points, but at each point in the flow field, and would also require more averages. This would potentially require data sets numbering in the high hundreds or thousands. Thus the present analysis serves as only an estimate of the mean and rms quantities.

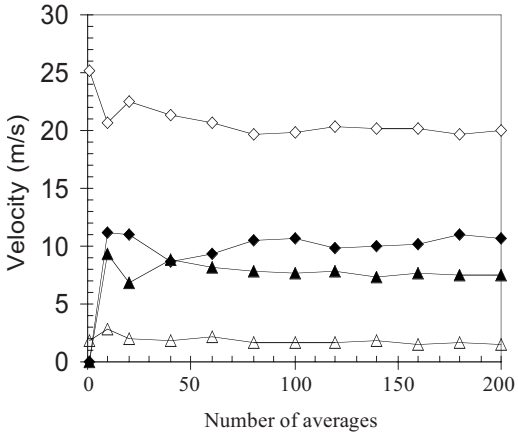


FIG. 4. Convergence of measured velocities versus numbers of images used in ensemble averaging (◇=jet core average velocity, ◆=jet core |rms| velocity, △=shear layer average velocity, ▲=shear layer |rms| velocity).

4. Jet centerline

Erath and Plesniak (2006a) described a method for determining the jet deflection angle at a given point downstream of the glottis. Their approach was extended here to identify the entire jet centerline over a complete cycle of oscillation at each phase. This jet centerline finding algorithm was implemented using the average velocity data and a MATLAB script as follows.

First, spurious vectors along the right and left edges of each average velocity data set were set to a value of zero (e.g., in the duct walls). The following steps were then performed at each vertical position above the glottal exit:

(1) A minimum velocity threshold was determined based on the jet growth, and this threshold was used to determine when jet began and decayed (7 m/s was used here).
(2) Smoothing was accomplished in the horizontal direction by taking the average of each velocity vector with its six neighboring vectors (three on each side).
(3) The maximum jet velocity was identified using the smoothed velocity profiles.
(4) The jet core was defined as all lateral vectors that had a value of greater than 80% of the maximum jet velocity.
(5) The jet center was defined as the geometric center of all vectors within one standard deviation of the jet core mean value.

The jet centerline was constructed from the individual jet centers over the entire range of vertical positions above the glottis, and this "rough" centerline was then smoothed in the vertical direction using an average of the individual centerline point and its four (two on each side) neighboring centerline points. Figure 5 shows the examples at three phases of the centerline using this algorithm, where the located centerline can be seen to satisfactorily follow the jet center.

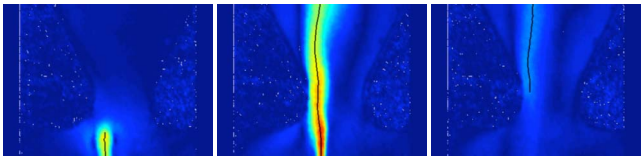


FIG. 5. (Color online) Selected phases from jet centerline finding algorithm applied to FR case with $p=1.25$ kPa. Color denotes velocity magnitude and the black line is the calculated centerline position. Shown are examples of starting jet (left, phase ≈0.07 of one cycle), full jet (middle, phase ≈0.33), and decaying jet (right, phase ≈0.40). In the extreme sides of the images, potentially spurious vectors have been replaced with zero vectors.

III. RESULTS

The average model vibration frequency during all measurements was measured to be 132 Hz over the 1.25–1.9 kPa subglottal pressure range. Time-averaged flow rates were 103, 149.5, and 232.3 ml/s for subglottal pressures of 1.25, 1.5, and 1.9 kPa, respectively.

A. High-speed images

Figure 6 shows a top view of one cycle of model vibration for the 1.25 and 1.9 kPa cases; Fig. 7 shows two views of the glottal jet using flow visualization for the same pressures. A small steady flow can be seen in the 1.25 kPa flow visualization images (see top rows of Fig. 7) where the orifice did not completely close during collision. Surface tackiness can be seen to have interfered somewhat with the model opening. For the higher pressure, the glottal orifice first divided into two separate anterior and posterior areas during glottal opening (see Fig. 6). Relative to the 1.9 kPa case, the inferior-superior motion of the vocal folds at 1.25 kPa was small (see Fig. 7, side views). Observation of the flow through several sequential cycles did not yield any evidence of significant variation in the overall jet pattern (including jet direction).

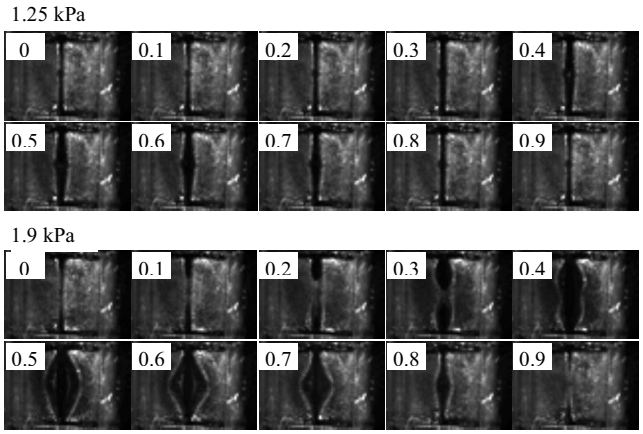


FIG. 6. Top view of high-speed images of synthetic vocal fold model at $p=1.25$ kPa (left) and $p=1.9$ kPa (right). The approximate phase (relative to one cycle) is shown.

A fairly coherent starting vortex is apparent at the higher pressure. The starting vortex and the lateral motion both occurred in the early part of the cycle. The adhesion facilitated

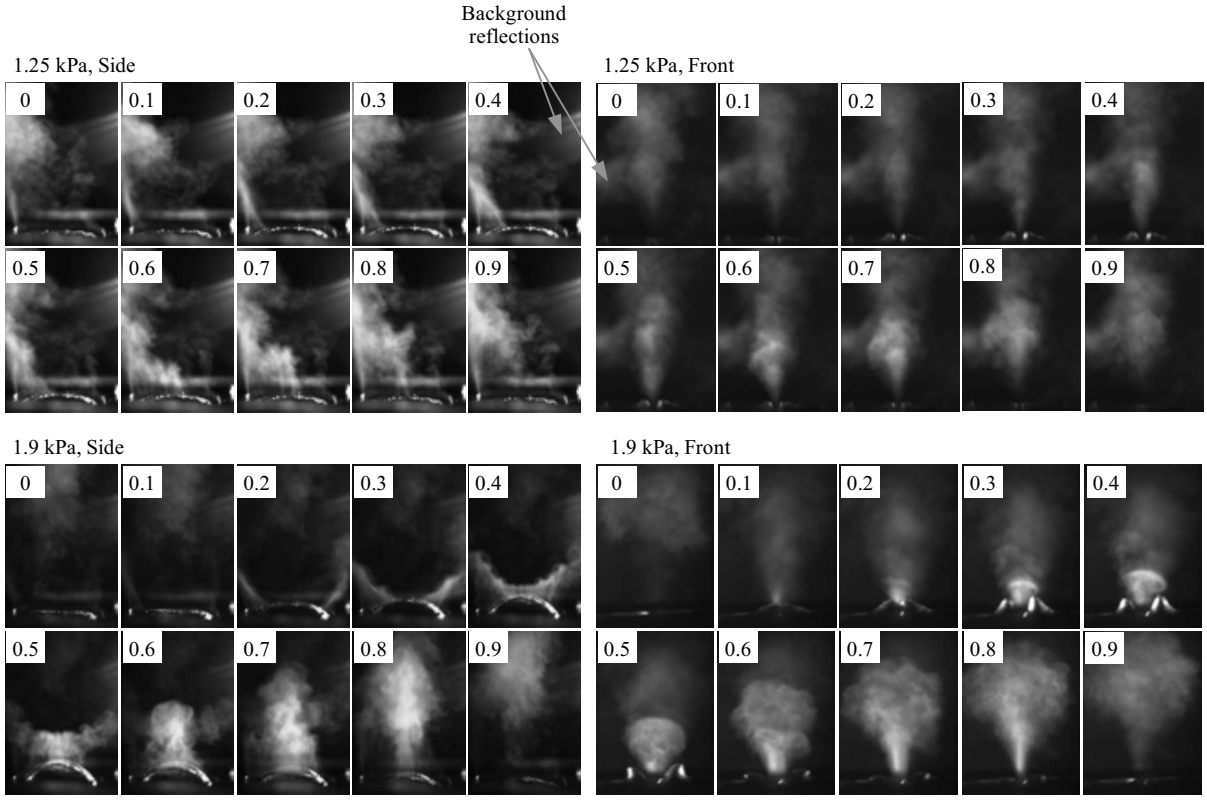


FIG. 7. Side view (left columns) and front view (right columns) of high-speed flow visualization at $p=1.25$ kPa (top) and $p=1.9$ kPa (bottom). The phase proceeds from left to right and from top to bottom for each image. Note that the side and front views were not obtained simultaneously, so that the phases shown (relative to one cycle) are only approximate.

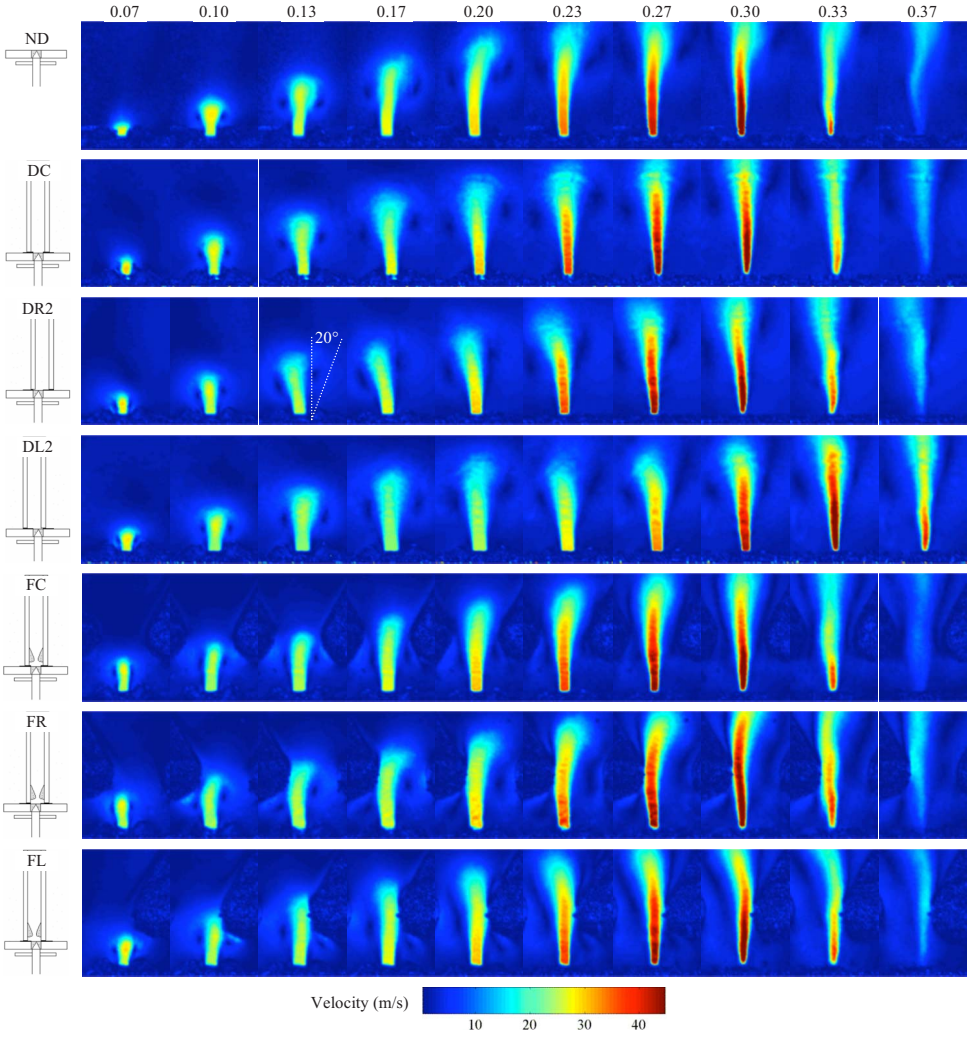


FIG. 8. (Color online) Ensemble-averaged velocity over the glottal cycle at pressure $p=1.25$ kPa. The icons outside of each image set show the respective test geometry show the respective test geometry and are defined in Fig. 3. The top row of text denotes approximate phase (referenced to fraction of one period).

rapid lateral acceleration at the vocal fold midplane, which seemed to contribute to the impulsive development of the vortex. The jets were all roughly (although not perfectly) symmetric in the medial-lateral sense. Also of interest is the double-jet formation during glottal opening, followed by subsequent merging into a single jet, at the higher pressure (Fig. 7, side view).

B. Average velocity

Maximum average velocity magnitudes measured in this study using PIV were in the range of 50–61.5 m/s. These compare well with magnitudes measured using an excised larynx reported by Alipour and Scherer (2006). Reynolds numbers based on the glottal jet velocity and estimates of the maximum orifice area were calculated to be in the range of 727–1621. These Reynolds numbers suggest the flow exiting the orifice should be laminar; rms calculations (to follow) support this observation for most (but not all) configurations, pressures, and phases.

Figure 8 shows ensemble-averaged velocity plots for all of the cases at ten phases for the 1.25 kPa case. For all pressures and all cases, a starting vortex was observed. The vortex appeared to last for more than half of the jet lifetime in most of the test cases. The vortex tended to grow laterally as the jet developed. For both pressures, the duct influenced the direction of the jet in the vocal tract cases, causing the jet to tend toward the closest wall. This is discussed further in Sec. III C.

Comparing the false fold cases with the open jet case for $p=1.25$ and $p=1.9$ kPa (not shown), the false folds clearly interfered with the starting vortex. This effect was repeated for both asymmetric false fold cases, with the vortex nearest to the vocal tract center persisting longer than that nearest to the false folds. A small vortex was also sometimes seen in the space between the true and false folds (the laryngeal ventricle, although it is noted that the laryngeal ventricle likely extends further laterally in human phonation). This jet interaction with the false folds may constitute a sound source through fluid–solid interaction. Further, the "grazing" of the jet over the laryngeal ventricle could act as a kind of acoustic resonator (Zhang *et al.*, 2002b). It is emphasized that interference of the vortex with the glottal jet was observed for both symmetric and asymmetric false fold cases. This presents evidence that for average adult human male false fold size and position, the false folds may prematurely disrupt the coherent glottal vortex, although further experiments using real vocal folds and vocal tract should be performed to confirm this observation.

In the asymmetric false fold cases, the jet tended to deflect away from the nearest false fold as the jet neared the

bottom surface of a false fold. This deflection could have been due to either the jet impingement or to the vortex interaction, or a combination of the two. The FR case at 1.9 kPa showed a second vortex shed from the downstream side of the near false fold as the jet passes the false fold. This was not seen in the 1.25 kPa cases, and it was not clear whether it occurred in the 1.9 kPa FL case (not shown). However, it is possible that this occurrence was related to the higher pressure. Also, the jet width is wider for the false fold cases than for the vocal tract or open jet cases.

The PIV experiments of Erath and Plesniak (2006b) showed jet bimodal (lateral "flapping") behavior. Examination of the instantaneous PIV velocity data in this study did not show any of this behavior. Erath and Plesniak (2006b) stated that the flapping behavior was never observed for divergent glottal profiles of 40° or greater. It is possible that the profiles in these experiments reached such angles, although the glottal profiles were not measured in this study. Other differences in experimental setup and models may also account for this difference (their model consisted of rotating shutters downstream of a static vocal fold model with a divergent profile).

C. Jet centerline

Plots of jet centerline versus phase for different duct configurations and for both pressures are shown in Fig. 9; the z coordinate in the vertical axes labels denotes streamwise location. These plots provide more convenient identification of the jet direction than the velocity plots.

The open jet at $p=1.25$ kPa tended to skew toward the right. With the vocal tract, the jet initially skewed slightly toward the nearer wall. Eventual slight straightening of the glottal jet is seen. The same trends were found for $p=1.9$ kPa, only the more pronounced influence of the duct wall resulted in greater jet deflection. The closer the jet was to the wall, the more it tended to skew toward the wall in its initial development. For the DL2 case, the 1.25 kPa jet skewed only slightly toward the wall only in the initial development; the 1.9 kPa DL2 case clearly tended toward the wall throughout the cycle. The reason for this difference in behavior at different pressures is not clear, although it is likely attributable to a combination of possible asymmetric motion and higher flow rate.

The false fold 1.25 and 1.9 kPa cases showed a different tendency. When the false folds were offset with respect to the vocal fold model, the jet initially tended to skew away from the near wall and from the near false fold toward the duct centerline, then straighten, and then be further skewed by the near false fold boundary. Again, this is more pronounced at the higher pressure. This behavior is attributed to the additional fluid resistance caused by the presence of the near false fold. It is noted that the centerlines for the symmetric false folds cases and the open jet cases are very similar.

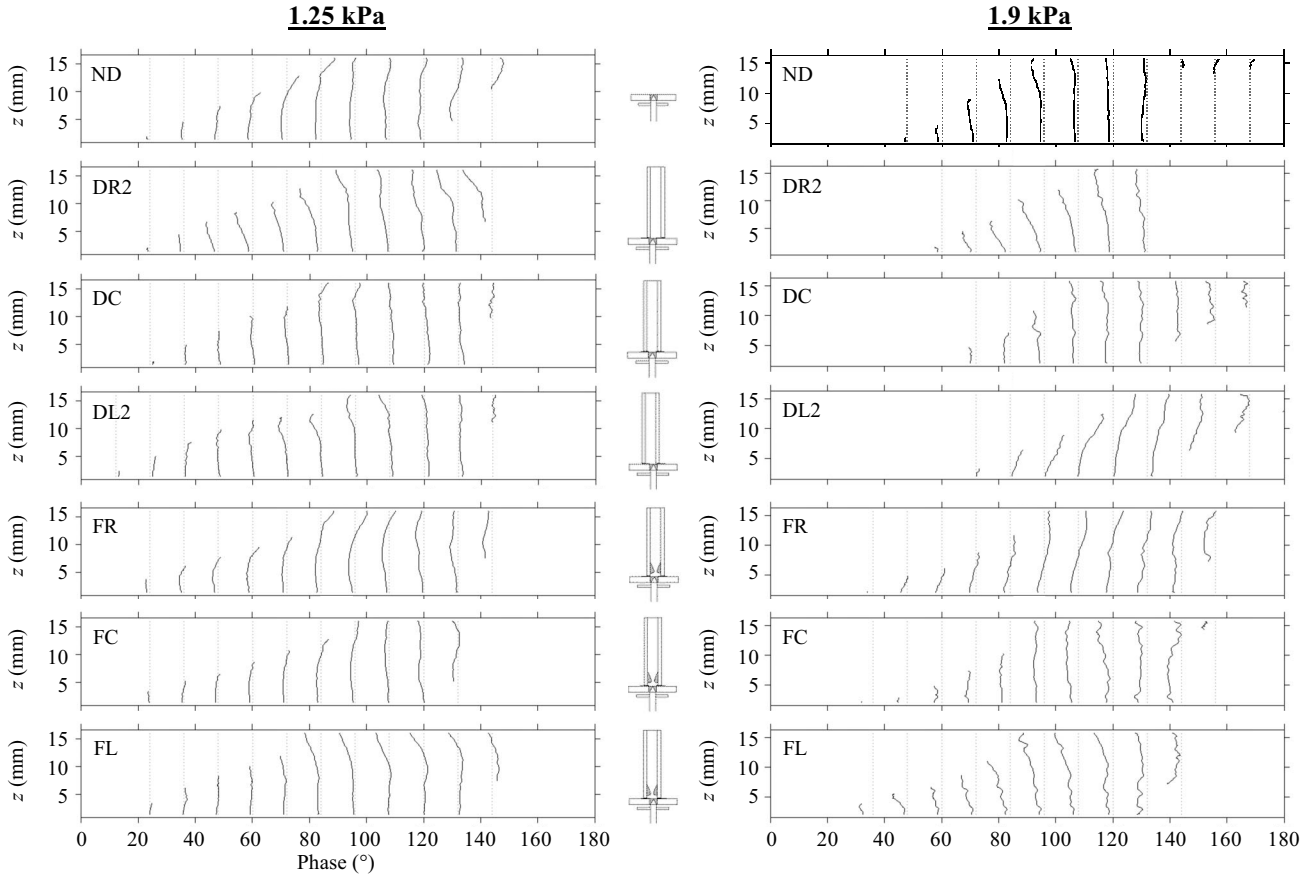


FIG. 9. Jet centerline plots versus phase for duct cases at $p=1.25$ kPa (left) and $p=1.9$ kPa (right). The respective configuration labels and icons are shown. The z coordinate in the vertical axes denotes streamwise location.

D. rms velocity

|rms| results are shown in Figs. 10 and 11 for select cases. As would be expected, |rms| was highest in the jet shear layers and lowest in the jet core. The high |rms| values outside of the jet in the no-duct cases were attributed to low seed particle density outside of the open jet. Similarly, high |rms| values in the laryngeal ventrical may have been due to reduced illumination.

The results show consistently higher |rms| values in the leading vortex with the vocal tract than without, suggesting that the leading vortex may emerge more consistently (in terms of velocity magnitude and perhaps bearing) in the open case than in the confined case. In the lower pressure case, a laminar core that persisted a short distance downstream of the glottis is evident over most of the cycle. This laminar core disappeared during glottal closing; this is consistent with observations of Mongeau *et al.* (1997) and Zhang *et al.* (2002a) that the quasisteady assumption begins to break down during glottal closing due to jet turbulence. They also indicated that the quasisteady assumption is not as good during opening, and the results here indicate that fluctuations or unsteadiness in the starting vortex may contribute to this. Asymmetry in the vocal tract position did not seem to noticeably influence the |rms| values.

The |rms| plots assist in visualizing the interaction of the glottal jet with the false folds. Comparison of the FC and FR cases (in particular at the higher pressure) suggests that the vortex interaction with the false folds may have reduced the overall |rms| values in the starting vortex, suggesting a potentially stabilizing influence of the false folds. Other features of the jet, such as laminar core characteristics, do not seem to have been significantly altered.

For the $p=1.9$ kPa cases, the laminar core was nearly nonexistent when the jet was nearest the duct walls. Less asymmetric cases (DR1, DL1, not shown here) did show a distinct laminar core. The symmetric case showed a smaller laminar core than the DR1 and DL1 cases. With the presence of the false folds, the laminar core was evident over some parts of the cycle, although the jet appeared to be turbulent during closing.

IV. DISCUSSION

A. Comparison with previous models

The present model's behavior is here compared with previous models, specifically with the models of Neubauer *et al.* (2007) and Khosla *et al.* (2007). Several aspects of this model's behavior are considered, some of which are in agreement with these models and some of which are not.

1. Synthetic model vibration

The model was similar to the human vocal folds in terms of frequency of oscillation, flow rate, and amplitude of vibration. However, some differences regarding mucosal wave and adhesion are noted.

An important aspect of vocal fold motion is the mucosal wave, i.e., the wave-like transition of a convergent profile to a divergent profile over the vibration period. One-layer models have been shown to demonstrate a degree of this change in profile (Thomson *et al.*, 2004). Analysis of the present two-layer model's medial surface dynamics is not available, but some information can be gathered from the images shown here. Evaluation of Figs. 6 and 7 shows that during the closed phase of oscillation, the medial surface of the present model was at least straight (because of collision), and possibly slightly convergent during opening. The model was divergent during part of the phase, which is especially evident in Fig. 6 for the 1.9 kPa case, where in phases 0.5–0.7 the glottal entrance can clearly be seen to be smaller than the glottal exit. This divergent motion is also somewhat evident in Fig. 7 (1.9 kPa case, front view) and Fig. 8 (1.25 kPa, DR2 case); in Fig. 8 the divergent angle is estimated to be approximately 20°. Thus the convergent-divergent behavior was evident. However, the "wavelike" motion typical of vocal fold vibration did not seem to be significantly manifest in this model. More data should be acquired to quantify the medial surface dynamics of these models. Further, it is recommended that efforts to develop synthetic models that more closely match the vibration patterns seen in *in vivo* and excised human vocal folds should be pursued.

Another limitation of the present model is the adhesion of the surfaces and the resulting change in vibration. This adhesion seemed to be influenced by the amount of DEHS oil wetting the surface (potentially explaining the differences in jet duration seen in Figs. 9 and 11). While a liquid bridge can sometimes be seen between the human vocal folds (e.g., Hsiao *et al.*, 2002; Hsiung, 2004), this bridge typically does not cause the vocal folds to adhere to the extent seen in the 1.9 kPa case. This adhesion contributed to the development of the vortex, to the double-jet formation during glottal opening and subsequent merging, and to a shorter open quotient time. Pressed phonation is characterized by short quotient times, and thus may potentially yield similarly stronger initial vortices. The double-jet formation is not expected to be representative of typical phonation, although it may exist when abundant mucus results in a liquid bridge remaining intact throughout the phonation cycle. This double-jet formation was also evident in the flow visualization images of Neubauer *et al.* (2007).

Also worth mentioning is the degree of asymmetry in the present model. Qualitatively the motion was generally symmetric, but small asymmetries were present as evidenced by the slightly asymmetric starting vortex (see Fig. 7, 1.9 kPa case, front view, phases 0.3 and 0.4). However, this asymmetry did not seem to be nearly as pronounced, from either structural or aerodynamic perspectives, as in the model used by Neubauer *et al.* (2007). In the present model, the asymmetry seemed to be more pronounced at higher flow rates than at lower flow rates. Erath and Plesniak (2006c) showed that a glottal jet tended to attach to one side of the glottis when the flow acceleration was zero, and that a symmetric jet was formed at times of maximum acceleration. Thus considering the impulsive nature of the present jet due to adhesion, the strong vortex suggests a strong acceleration, and this may in part contribute to the lack of asymmetric fluid motion.

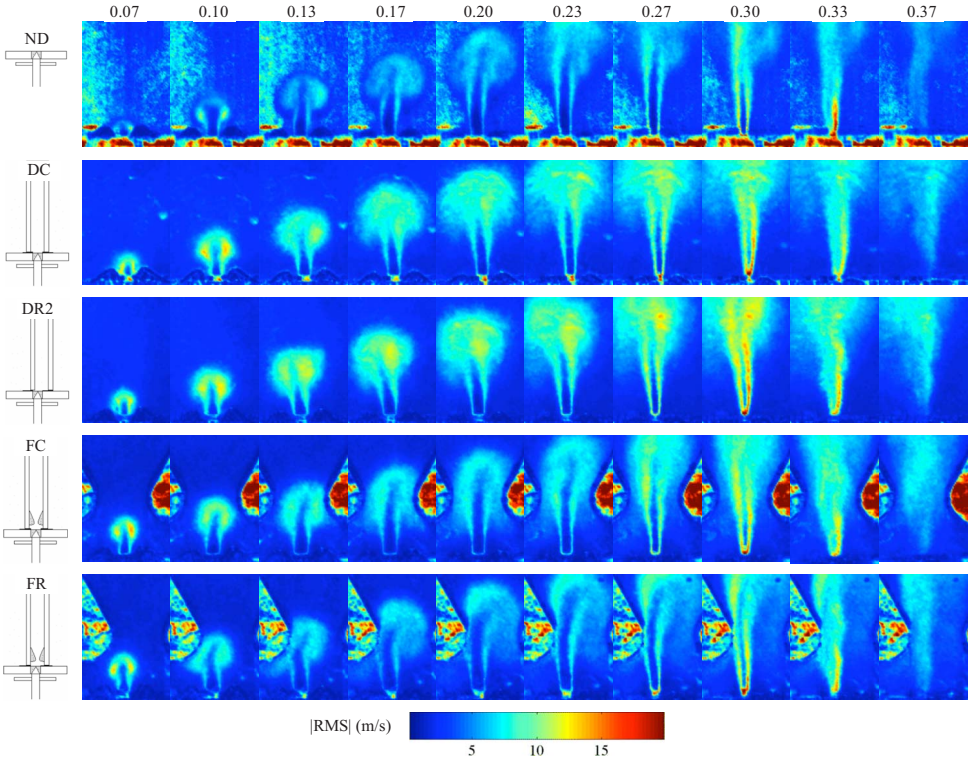


FIG. 10. (Color online) Jet centerline plots versus phase for duct cases at $p=1.25$ kPa (left) and $p=1.9$ kPa (right). The respective configuration labels and icons are shown. The z coordinate in the vertical axes denotes streamwise location.

2. Flow patterns

a. Axis switching. Axis switching occurs for various shapes of noncircular orifices. It was observed by Khosla *et al.* (2007) and was seen in the present model. This can be seen in the flow visualization images by comparing phases 0.6–0.8 of the front and side views of the 1.9 kPa case (Fig. 7). The major axis of the jet near the vocal fold exit is oriented in the anterior-posterior direction, whereas further downstream it is oriented in the medial-lateral direction.

b. Starting vortex. The starting vortex seen in the present model was also observed by Khosla *et al.* (2007) and Neubauer *et al.* (2007), although it perhaps lasted longer in the present case. The model used in this paper appeared to have a shorter open quotient time than the Khosla *et al.* (2007) and Neubauer *et al.* (2007) models. It is therefore likely that the momentum build-up during the closed phase led to a stronger starting vortex, which may account for the longer duration.

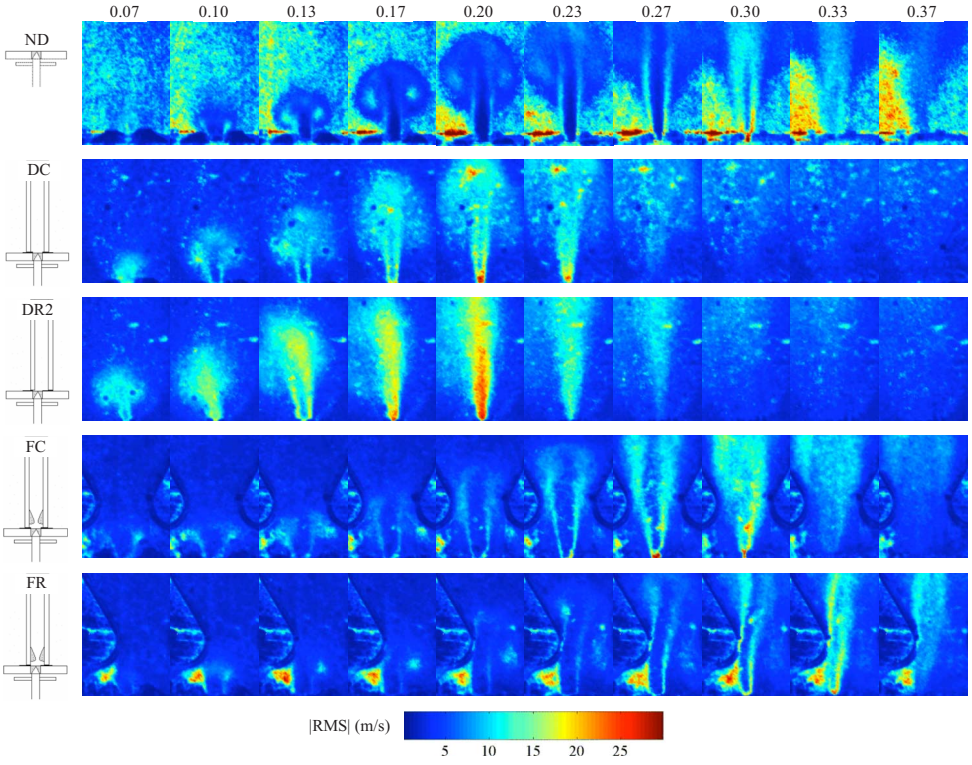


FIG. 11. (Color online) |rms| over the glottal cycle at pressure $p=1.9$ kPa. The icons outside of each image set show the respective test geometry. The top row of text denotes approximate phase (referenced to fraction of one period).

c. Kelvin–Helmholtz vortex structures. Using laser-illuminated flow visualization, Neubauer *et al.* (2007) showed the presence of Kelvin–Helmholtz instability in the supraglottal jet, as evidenced by distinct coherent vortices forming along the shear layer of the jet. Khosla *et al.* (2007) also observed these structures. The flow visualization in the present work (Fig. 7) used high-speed video imaging with dc illumination. A longer shutter speed of 1/5000 s was required for adequate light exposure using the dc illumination. This resulted in the particles traveling a relatively long distance during image exposure. Consequently, "smearing" of the flow field occurred, inhibiting the visibility of coherent structures potentially located in the shear layer (such as Kelvin–Helmholtz vortices). Thus the present data are not conclusive regarding the presence of such flow structures in this model.

d. Intraglottal flow separation. Intraglottal flow separation, a feature seen in both experimental and numerical results, and which is an important component of vocal fold self-oscillation (Titze, 1988), was evident in these models, e.g., phases 0.6–0.7 of Fig. 7 (1.9 kPa, front view).

e. Glottal jet skewing. In Figs. 7 and 8, the jet can be seen to be slightly skewed during opening. In Fig. 8, phase 0.37 no duct case, the jet is beginning to experience "flapping" motion in the downstream regions similar to that discussed by Neubauer *et al.* (2007). Khosla *et al.* (2007) and Neubauer *et al.* (2007) observed skewing of the glottal jet during closing, consistent with the expected behavior of a divergent orifice. It is possible that slight skewing in the immediate near field just prior to closing occurred in the present model, but it does not seem to be significant; further investigation into this behavior could be beneficial.

B. Influence of supraglottal loading

Symmetric positioning of the vocal tract, with and without false folds, created jet centerline patterns that were generally similar to the open jet case. When the vocal tract was asymmetrically positioned with respect to the vocal fold model, the resulting jet centerline initially skewed toward the nearest wall in the vocal tract cases, but away from the nearest wall (and toward the vocal tract centerline) in the false fold cases. This can be explained as follows. In the vocal tract cases, as the gap between the jet and the nearest wall decreases, the fluid available to be entrained into the jet flow decreases. With less fluid to be entrained, the fluid that is entrained experiences greater velocity, which leads to a decrease in pressure. This pressure imbalance between the far and near walls forces the jet to initially tend toward the nearer wall. In the false fold cases, the flow between the jet and the nearest vocal tract wall is theorized to be substantially slower than in the vocal tract cases. Due to the obstructing influence of the inferior false fold surface, there is no outlet for the flow toward the nearest wall. Also, the vocal folds impose additional surface area for boundary layer and friction effects to occur. These effects combine to cause a decrease in velocity in this region, with a corresponding increase in pressure predicted. This would tend to force the jet away from the wall toward the vocal tract midplane. Further experimentation could determine at what value of vertical gap between the inferior false fold surface and the glottal exit the jet transitions from skewing toward the nearest wall to skewing away from the nearer wall.

V. CONCLUSIONS

The glottal jet exiting a two-layer synthetic, self-oscillating vocal fold model was investigated using high-speed imaging and PIV at multiple pressures. The behavior of the open jet case, commonly reported in the literature, was compared with the case which includes the vocal tract and the false folds.

Adhesion between the medial surfaces of the vocal fold model was evident, particularly at the highest pressure. This adhesion created a double orifice over a portion of the cycle (also noted by Neubauer *et al.*, 2007) and appeared to delay the jet formation, which in turn contributed to an impulsive starting vortex.

Measurements of glottal jet velocity magnitudes generally agreed with earlier measurements using excised larynges. The average velocity fields from different vocal tract configurations were compared. A starting vortex was observed in all cases. In the false fold cases, the presence of the false folds in the vocal tract obstructed the downstream convection of the starting vortex. In asymmetric false fold cases, the side of the vortex nearest a false fold was prevented from developing with the opposing vortex. In the symmetric case, the false folds equally obstructed both sides of the starting vortex. In one case, a new vortex structure was seen to be shed from the surface of one of the false folds.

The jet core centerline was calculated. When the vocal tract was asymmetrically positioned with respect to the vocal fold model, the resulting jet centerline generally skewed toward the nearest wall in the vocal tract (no false fold) cases, but away from the nearest wall (and toward the vocal tract centerline) in the false fold cases. This effect was more pronounced at the higher pressures. Symmetric positioning of the vocal tract, with and without false folds, created jet centerline patterns that were generally similar to the open jet cases.

Finally, the rms velocity magnitudes, calculated from ensemble-averaged PIV measurements, clearly showed jet core (with relatively low |rms| values) and shear layer (with relatively high |rms| values) regions close to the glottal exit. Further downstream, the |rms| values increased and the jet core disappeared (as would be expected). The jets in the vocal tract cases contained higher |rms| values than in the open jet cases. When the false folds were included in the vocal tract, the |rms| values decreased to more closely match the open jet case. This suggests a potentially destabilizing influence of the vocal tract and a stabilizing influence of the false folds on the fluctuating glottal jet velocities in the vocal tract.

Several areas of future work are suggested, including comparison of the medial surface dynamics of one- and two-layer models, development of synthetic models with improved geometry (e.g., geometry derived from MRI data), investigation of the supraglottal jet with more anatomically correct supraglottal vocal tract geometry, and a more in-depth study of the cycle-to-cycle variations in the supraglottal jet issuing from symmetric and asymmetric models.

ACKNOWLEDGMENTS

This work was supported in part by NIH/NIDCD Grant No. R01 05788. The authors gratefully acknowledge Jake Munger's assistance in preparing the experimental setup and in testing material properties.

Agarwal, M., Scherer, R., and Hollien, H. (**2003**). "The false vocal folds: Shape and size in frontal view during phonation based on laminagraphic tracings," J. Voice **17**, 97–113.

Alipour, F., and Scherer, R. (**1995**). "Pulsatile airflow during phonation: An excised larynx model," J. Acoust. Soc. Am. **97**, 1241–1248.

Alipour, F., and Scherer, R. (**2006**). "Characterizing glottal jet turbulence," J. Acoust. Soc. Am. **119**, 1063–1073.

Alipour, F., Scherer, R., and Knowles, J. (**1996**). "Velocity distributions in glottal models," J. Voice **10**, 50–58.

Barney, A., Shadle, C., and Davies, P. (**1999**). "Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory," J. Acoust. Soc. Am. **105**, 444–455.

Chan, R., and Titze, I. (**1997**). "Further studies of phonation threshold pressure in a physical model of the vocal fold mucosa," J. Acoust. Soc. Am. **101**, 3722–3727.

Chan, R., and Titze, I. (**2006**). "Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics," J. Acoust. Soc. Am. **119**, 2351–2362.

Drechsel, J. (**2007**). "Characterization of synthetic, self-oscillating vocal fold models," Master's thesis, Brigham Young University, Provo, UT.

Erath, B., and Plesniak, M. (**2006a**). "An investigation of bimodal jet trajectory in flow through scaled models of the human vocal tract," Exp. Fluids **40**, 683–696.

Erath, B., and Plesniak, M. (**2006b**). "The occurrence of the Coanda effect in pulsatile flow through static models of the human vocal folds," J. Acoust. Soc. Am. **120**, 1000–1011.

Erath, B., and Plesniak, M. (**2006c**). "An investigation of jet trajectory in flow through scaled vocal fold models with asymmetric glottal passages," Exp. Fluids **41**, 735–748.

Hirano, M., and Kakita, Y. (**1985**). "Cover-body theory of vocal fold vibration," in *Speech Science: Recent Advances*, edited by R. G. Daniloff (College-Hill Press, San Diego), pp. 1–46.

Hofmans, G. C. J., Groot, G., Ranucci, M., Graziani, G., and Hirschberg, A. (**2003**). "Unsteady flow through in-vitro models of the glottis," J. Acoust. Soc. Am. **113**, 1658–1675.

Hsiao, T. Y., Liu, C. M., and Lin, K. N. (**2002**). "Videostrobolaryngoscopy of mucus layer during vocal fold vibration in patients with laryngeal tension-fatigue syndrome," Ann. Otol. Rhinol. Laryngol. **111**, 537–541.

Hsiung, M. W. (**2004**). "Videolaryngostroboscopic observation of mucus layer during vocal cord vibration in patients with vocal nodules before and after surgery," Acta Oto-Laryngol. **124**, 186–191.

Khosla, S., Muruguppan, S., Gutmark, E., and Scherer, R. (**2007**). "Vortical flow field during phonation in an excised canine larynx model," Ann. Otol. Rhinol. Laryngol. **116**, 217–228.

Mongeau, L., Franchek, N., Coker, C., and Kubli, R. (**1997**). "Characteristics of a pulsating jet through a small modulated orifice, with application to voice production," J. Acoust. Soc. Am. **102**, 1121–1133.

Neubauer, J., Zhang, Z., Miraghaie, R., and Berry, D. (**2007**). "Coherent structures of the near field flow in a self-oscillating physical model of the vocal folds," J. Acoust. Soc. Am. **121**, 1102–1118.

Pelorson, X., Hirschberg, A., van Hassel, R. R., and Wijnands, A. P. J. (**1994**). "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," J. Acoust. Soc. Am. **96**, 3416–3431.

Riede, T., Tokuda, I. T., Munger, J. B., and Thomson, S. L. (**2008**). "Mammalian laryngeal air sacs add variability to the vocal tract impedance physical and computational modeling," J. Acoust. Soc. Am. (in press).

Scherer, R., Shinwari, D., De Witt, K., Zhang, C., Kucinschi, B., and Afjeh, A. (**2001**). "Intraglottal pressure profiles for a symmetric and oblique glottis with a divergence angle of 10 degrees," J. Acoust. Soc. Am. **109**, 1616–1630.

Scherer, R., Titze, I., and Curtis, J. (**1983**). "Pressure-flow relationships in two models of the larynx having rectangular glottal shapes," J. Acoust. Soc. Am. **73**, 668–676.

Scherer, R. C. (**1981**). "Laryngeal fluid mechanics: Steady flow considerations using static models," Ph.D. thesis, University of Iowa, Iowa City, IA.

Shadle, C., Barney, A., and Thomas, D. (**1991**). "An investigation into the acoustics and aerodynamics of the larynx," in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, edited by J. Gauffin and B. Hammarberg (Singular, San Diego), pp. 78–80.

Shadle, C., Barney, A., and Davies, P. (**1999**). "Fluid flow in a dynamic mechanical model of the vocal folds and tract. II. Implications for speech production studies," J. Acoust. Soc. Am. **105**, 456–466.

Shinwari, D., Scherer, R., DeWitt, K., and Afjeh, A. (**2003**). "Flow visualization and pressure distributions in a model of the glottis with a symmetric and oblique divergent angle of 10 degrees," J. Acoust. Soc. Am. **113**, 487–497.

Thomson, S. L., Mongeau, L., and Frankel, S. (**2005**). "Aerodynamic transfer of energy to the vocal folds," J. Acoust. Soc. Am. **118**, 1689–1700.

Thomson, S. L., Mongeau, L., Frankel, S. H., Neubauer, J., and Berry, D. A. (**2004**). "Self-oscillating laryngeal models for vocal fold research," *Proceedings of the Eighth International Conference on Flow-Induced Vibrations*, Ecole Polytechnique, Paris, France, 5–9 July 2004, Vol. **2**, pp. 137–142.

Tran, Q. T., Berke, G. S., Gerratt, B. R., and Kreiman, J. (**1993**) "Measurement of Young's modulus in the in vivo human vocal folds," Ann. Otol. Rhinol. Laryngol. **102**, 584–591.

Titze, I. R. (**1988**).. "The physics of small-amplitude oscillation of the vocal folds," J. Acoust. Soc. Am. **83**, 1536–1552.

Titze, I., Schmidt, S., and Titze, M. (**1995**). "Phonation threshold pressure in a physical model of the vocal fold mucosa," J. Acoust. Soc. Am. **97**, 3080–3084.

Triep, M., Brücker, C., and Schröder, W. (**2005**). "High-speed PIV measurements of the flow downstream of a dynamic mechanical model of the human vocal folds," Exp. Fluids **39**, 232–245.

Zhang, Z., Mongeau, L., and Frankel, S. (**2002a**). "Experimental verification of the quasi-steady approximation for aerodynamic sound generation by pulsating jets in tubes," J. Acoust. Soc. Am. **112**, 1652–1663.

Zhang, Z., Mongeau, L., Frankel, S., Thomson, S. L., and Park, J. (**2004**). "Sound generation by steady flow through glottis-shaped orifices," J. Acoust. Soc. Am. **116**, 1720–1728.

Zhang, Z., Neubauer, J., and Berry, D. (**2006a**). "The influence of subglottal acoustics on laboratory models of phonation," J. Acoust. Soc. Am. **120**, 1558–1569.

Zhang, Z., Neubauer, J., and Berry, D. (**2006b**). "Aerodynamically and acoustically driven modes of vibration in a physical model of the vocal folds," J. Acoust. Soc. Am. **120**, 2841–2849.

Zhang, C., Zhao, W., Frankel, S. H., and Mongeau, L. (**2002b**). "Computational aeroacoustics of phonation. II. Effects of flow parameters and ventricular folds," J. Acoust. Soc. Am. **112**, 2147–2154.

Duration differences in the articulation and acoustics of Swiss German word-initial geminate and singleton stops[a)]

Astrid Kraehenmann[b)]
Universität Konstanz, FB Sprachwissenschaft, D186, 78457 Konstanz, Germany

Aditi Lahiri
Centre for Linguistics and Philology, University of Oxford, Walton Street, Oxford OX1 2HG, United Kingdom

(Received 24 September 2007; revised 28 March 2008; accepted 7 April 2008)

Stops in Swiss German contrast only in quantity in all word positions; aspiration and voicing play no role. As in most languages with consonant quantity contrast, geminate stops are produced with significantly longer closure duration (CD) than singletons in an intersonorant context. This holds word medially as well as phrase medially, e.g., [on**i** **tt**o:sə] "without roar" versus [on**i** **t**o:sə] "without can." Since the stops are voiceless, no CD cue distinguishes geminates from singletons phrase initially. Nevertheless, do speakers utilize articulatory means to maintain the contrast? By using electropalatography, the articulatory and acoustic properties of word-initial alveolar stops were investigated in phrase-initial and phrase-medial contexts. The results are threefold. First, as expected, CD and contact duration of the articulators mirror each other within a phrase: Geminates are longer than singletons. Second, phrase initially, the contact data unequivocally establish a quantity distinction. This means that—even without acoustic CD cues for perception—geminates are articulated with substantially longer oral closure than singletons. Third, stops are longer in phrase-initial than phrase-medial position, indicating articulatory strengthening. Nevertheless, the difference between geminates and singletons phrase initially is proportionately less than in phrase-medial position. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2916699]

PACS number(s): 43.70.Aj, 43.70.Kv [AL] Pages: 4446–4455

I. INTRODUCTION

In this investigation, we focus on the articulatory properties of word-initial voiceless geminate and singleton stops in Swiss German, contrasting them in two phrase-medial contexts and one phrase-initial context. We raise three related questions. First, although no acoustic information regarding closure duration (CD) is available for listeners phrase initially, do speakers still make an articulatory distinction between geminates and singletons in natural speech? Second, if the consonantal quantity contrast is indeed maintained, is there any additive effect distinguishing the contrasting sounds phrase initially as compared to phrase medially? Third, phrase medially, are geminates and singletons articulated differently in different contexts, namely, after vowel-final words versus after obstruent-final words?

A number of earlier acoustic studies revealed that stops in Swiss German contrast in the duration of their closure phase (CD), i.e., not in closure voicing or the duration of the closure release or voice onset time (VOT) or any combination of the two. This has variously been called a phonological distinction in terms of fortis versus lenis (e.g., Dieth and Brunner, 1943; Fulop, 1994; Willi, 1996), voicing (e.g., Ham, 1998), and quantity (e.g., Kraehenmann, 2001, 2003), the latter of which we will adopt. It roughly corresponds to the standard German contrast variously analyzed as a fortis-lenis (cf. Kohler, 1984), a [spread glottis] (cf. Iverson and Salmons, 1995; Jessen and Ringen, 2002), or a voicing contrast (cf. Wiese, 1996), which, however, is primarily realized in phonetic terms as difference in VOT, not CD. The Swiss German quantity contrast occurs in all word positions (Table I).

On purely acoustic grounds, Kraehenmann (2001, 2003) showed that word-initial geminate CDs were about twice the length of singletons when in an intersonorant context. However, when a preceding word ended in an obstruent consonant, the contrast was neutralized by geminates having become shorter and singletons longer. Based on these results, we expect in the present study, where we combine articulatory and acoustic facts, that, within a phrase, the articulatory electropalatography (EPG) measures go hand in hand with the acoustic CD measures. That is, in phrase-medial position after a vowel, geminate durations should be longer than those of singletons, while geminate and singleton durations should be indistinguishable after an obstruent. Our expectations concerning neutralization and maintenance of the long-short contrast in this phrase-medial position rest on the syllabification of these consonants. We briefly sketch our assumptions in Fig. 1.

According to standard assumption, geminates are part of two syllables. Word medially, they close one syllable and build the onset of the following syllable (d.i.), while single-

[a)]Portions of this work were presented in "Non-neutralizing quantity in word-initial consonants: articulatory evidence," Proceedings of the 16th International Conference of the Phonetic Sciences, Saarbrücken, Germany, August 2007.

[b)]Electronic mail: astrid.kraehenmann@uni-konstanz.de

 0001-4966/2008/123(6)/4446/10/$23.00

TABLE I. Swiss German quantity contrast in three places of articulation and three word positions.

	Labial		Alveolar		Velar	
Initial	/**pp**a:ʀ/	"couple"	/**tt**ipp/	"tip"	/**kk**a:ʀ/	"tour bus"
	/**p**a:ʀ/	"bar"	/**t**ipp/	"dip"	/**k**a:ʀ/	"cooked"
Medial	/su**pp**əʀ/	"great"	/ma**tt**ə/	"mat"	/ma**kk**ə/	"tic"
	/su**p**əʀ/	"clean"	/ma**t**ə/	"maggot"	/ma**k**ə/	"stomach"
Final	/al**pp**/	"alp"	/vɛl**tt**/	"world"	/ʃnɛ**kk**/	"snail"
	/xal**p**/	"calf"	/fɛl**t**/	"field"	/vɛ:**k**/	"way"

tons only build the onset (d.ii.). Similarly, word initially—provided there is room—geminates close the final syllable of a preceding word (c.i.) and, thus, are part not only of two syllables but also of two words. The phonological quantity contrast is then realized both word medially and word initially in a vocalic context by significantly longer phonetic duration of geminates as compared to singletons. If, however, the final syllable of a preceding word is already closed by an obstruent consonant (b.i.), the first part of geminates cannot be syllabified, remains unassociated, and subsequently deletes. As a result, the phonological distinction between geminates and singletons disappears, as does the phonetic length difference, leading to contrast neutralization.

In phrase-/utterance-initial position, the syllabification of geminates is a moot point. It would be possible that neutralization similar to the consonantal context occurs since there is no preceding syllable available. However, following Kraehenmann (2001), we assume that the first part of initial geminates is prosodically associated (a.i.), although at the word rather than the syllable level. This representation would allow both for phonological and articulatory/acoustic maintenance of the contrast in this position. We hypothesize that

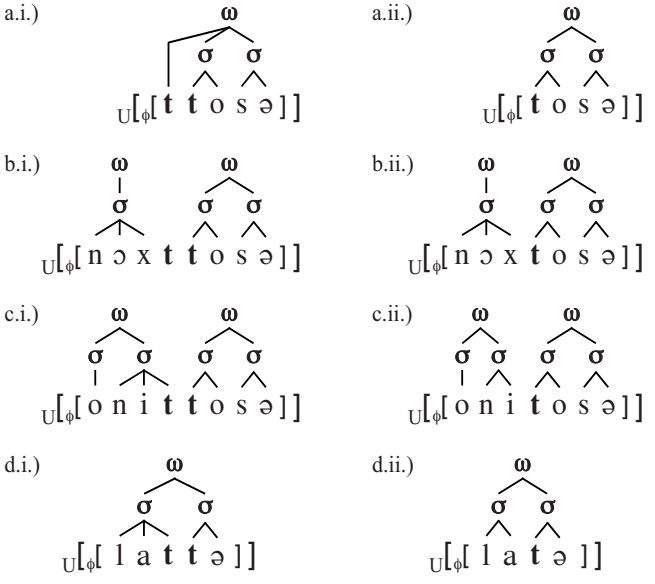


FIG. 1. Sample syllabifications and phrasings of word-medial (d) and word-initial geminates and singletons phrase/utterance initially (a) and phrase medially (b) and (c). (U=utterance phrase; ϕ=phonological phrase; ω=phonological word; σ=syllable): (a.i.) /**tt**o:sə/ "roar," (a.ii.) /**t**o:sə/ "can," (b.i.) /nɔ**x** **tt**o:sə/ "after roar," (b.ii.) /nɔ**x** **t**o:sə/ "after can," (c.i.) /on**i** **tt**o:sə/ "without roar," (c.ii.) /on**i** **t**o:sə/ "without can," (d.i.) /la**tt**ə/ "crossbar," and (d.ii.) /la**t**ə/ "shop."

geminates have a longer linguopalatal contact than singletons—in spite of the fact that in this position, CD information is missing.

Continuing on issues involving phrase-/utterance-initial contexts, in more recent literature, investigations have heavily focused on what influence the edges of prosodic domains—such as the syllable, the phonological word, the phonological phrase, etc.—have on the articulation of speech sounds. Consistent durational effects, which are of most relevance to our study, are primarily reported at the beginning of prosodic domains and seem to increase in force as the height of the domain in the prosodic hierarchy increases (e.g., Fujimura, 1990; Byrd *et al.*, 2005). For example, Fougeron and Keating (1997) found initial strengthening, i.e., longer and more extreme lingual articulation of initial consonants in English CV syllables with a cumulative effect. Likewise, a boundary effect on domain-initial segments (alveolar stops and fricatives) in Dutch was found by Cho and McQueen (2005). Also, Jun (1993), Cho and Keating (2001), and Keating *et al.* (2003) report that VOT measures in Korean stops are longest phrase initially, somewhat shorter word initially within a phrase, and shortest word medially within a phrase. While all these studies investigated the phonetic effects of different prosodic contexts for individual word-initial segments, the present study will take this one step further by examining not only absolute differences but also differences in the way a phonological contrast is realized.

In literature focusing specifically on word-initial voiceless geminates and singletons, Abramson (1986, 1987, 1991, 1999) established in a series of studies on Pattani Malay that CD was the primary acoustic cue. In its absence, i.e., phrase initially, listeners relied on two combined secondary cues, namely, rms amplitude of the first syllable and fundamental frequency ($F0$) of the vowel following the word-initial consonant, to successfully recover the phonemic difference. Similarly, by following up on work by Tserdanelis and Arvaniti (2001) on the word-medial quantity contrast in Cypriot Greek, Muller (2003) also found consistent CD and VOT differences in word-initial stops. She found geminates to have both longer CDs and longer VOTs phrase medially and that phrase initially the longer VOTs were sufficient secondary cues for native listeners to perceive the phonological difference. Ridouane's (2007) comprehensive study on the consonant system of Tashlhiyt Berber revealed two sets of results for word-initial stops. First, in terms of acoustics, release duration (i.e., positive VOT) did not significantly dif-

fer for geminates and singletons, and there was only a tendency for geminates to show greater rms amplitude during the release phase. The CD information could not be determined because the word-initial consonants only occurred phrase initially in the acoustic study. Second, in terms of articulation, an independent EPG study found that word-initial geminates were systematically articulated longer than their singleton counterparts, both phrase initially and medially. Moreover, phrase-initial stops were longer than phrase-medial ones, which Ridouane (2007) interpreted as prosodic lengthening.

While the studies on domain-initial prosodic strengthening only considered languages without a consonantal quantity contrast and had their primary focus on articulation, the studies on languages with a quantity contrast primarily focused on the acoustics. There are only few studies that combine the two (e.g., Lehiste *et al.*, 1973; Farnetani, 1990; Dunn, 1993; Löfqvist, 2005, 2007; Payne, 2006), and only Ridouane (2007) who discusses the articulation results with reference to prosodic strengthening issues. However, unfortunately, he had separate data sets for his articulatory and acoustic analyses. He investigated, on the one hand, the articulation of word-initial geminate and singleton stops in phrase-initial and phrase-medial position, and on the other hand, the acoustics of word-initial, word-medial, and word-final consonants. In our study, we intend to establish both the articulation and the acoustics of word-initial geminates and singletons, as they appear at different domain edges, i.e., phrase/utterance initially and phrase medially. Based on the existing literature mentioned, we expect to find phrase-initial articulatory strengthening as compared to the phrase-medial context. That is, we anticipate to ascertain longer articulations for geminates and singletons alike, possibly both in terms of the stop closure and the release phase.

In sum, by comparing Swiss German word-initial voiceless geminate versus singleton stops in differing phrasal positions, we expect (a) contrast maintenance phrase medially after a vowel and phrase/utterance initially, (b) contrast neutralization phrase medially after an obstruent, and (c) articulatory strengthening phrase initially as compared to phrase medially.

II. METHOD

A pilot EPG study reported by Kraehenmann and Jaeger (2003) had shown (a) that in phrase-initial position Swiss German geminates were distinguished from singletons and (b) that phrase medially there was neutralization in an obstruent context as opposed to contrast maintenance in a sonorant context. However, the pilot study only reported on data from a single speaker, and the carrier sentences used proved not to be ideal because they could potentially have had a phrase break at the crucial point of interest ([i ha elf_ksaitt] "I said eleven_;" [i ha t͡svai_ksaitt] "I said two_"). Our study avoided this by embedding the target words with initial geminates and singletons inside a prepositional phrase where the connection to the preceding preposition was prosodically fairly tight. In normal speech, a prepositional phrase always constitutes a single phonological phrase (ϕ).

TABLE II. Prosodic and segmental contexts.

Phrase-initial environment		
∅/**tt**o:sə/	"roar"	Isolation context
∅/**t**o:sə/	"can"	
Phrase-medial environment		
/nɔ**x**/ /**tt**o:sə/	"after roar"	C-context
/nɔ**x**/ /**t**o:sə/	"after can"	
/on**i**/ /**tt**o:sə/	"without roar"	V-context
/on**i**/ /**t**o:sə/	"without can"	

A. Recording material and procedure

Although Swiss German distinguishes geminate and singleton stops at three places of articulation (see Table I), we restrict our investigation to the alveolar stops /tt t/ in order to get the clearest results possible, particularly for the articulatory investigation. We chose 61 minimal and near-minimal pairs of words, of which 52 were proper names and 70 were common nouns (see Appendix A for a list of the individual names and nouns). We included proper names because they have been known to display special linguistic characteristics. Specifically, they are associated with processing difficulties (e.g., Brédart, 1993; Izaute, *et al.*, 2002) and language impairments (Evrard, 2002) and also show phonological peculiarities. For example, in Greek, proper names derived from common nouns show recessive word stress (Kurylowicz, 1966): *karpós* "fruit" versus *Kárpos* person name. Since there is evidence for the differential access and representation properties of proper names, it is feasible that speakers either overemphasize the quantity differences in proper names, or choose to disregard the contrast—either way distinguishing proper names from other common nouns.

We had two different prosodic environments: (a) phrase initial (henceforth *isolation context*) and (b) phrase medial. The phrase-medial condition had two segmental contexts: In the first, the preceding word ended in an obstruent (henceforth *C-context*), while in the second the preceding word ended in a vowel (henceforth *V-context*). Thus, each noun and name occurred in three distinct contexts: isolation, consonantal, and vocalic (Table II).

In order to be able to compare whatever effect we would get within the word-initial contrast with the typologically more common word-medial contrast, we recorded 42 additional nouns. These nouns contained the alveolar stops /tt t/ in the word-medial position between vowels (e.g., /k͡xø:**tt**ɐ/ "mutt,"/k͡xø:**t**ɐ/ "lure," see Appendix B for full list). The full set of items presented to the speakers consisted of two-thirds target words and one-third fillers. All target and filler items consisted of two syllables and carried main stress on the first.

Our subjects were given custom-fitted EPG palates a few weeks before the day of the recording to give them a chance to get accustomed to talking as uninhibitedly and naturally as possible with the palates in place. They also had at least 15 min warm-up time before the recording began.

We recorded four female Swiss German speakers, ranging in age between 27 and 42. The subjects read the test

TABLE III. Number of tokens used in each category.

	Names	Nouns		
	Initial	Initial	Medial	
Singletons	537	692	140	
Geminates	462	693	146	
	999	1385	286	2670

items in the three different contexts as they appeared in a random order on a computer screen. After a short break, the sets were read a second time. The EPG and audio signals were directly recorded onto the computer.

For all subsequent duration measures, we had a total of 2670 tokens—122 word-initial stops, 4 speakers, 2 repetitions, 3 contexts; 42 word-medial stops, 4 speakers, 2 repetitions, 1 context. The distribution is as listed in Table III below.

Since we needed both the EPG and the acoustic data combined, we discarded tokens in which, due to various factors, either one or the other was not usable. For example, for the sets in the isolation context, the subjects were asked to start with their mouth slightly open and the tongue not touching the palate to ensure that the first contact of the articulators corresponded to the beginning of the word. In quite a number of cases, this instruction was not followed, and thus the EPG could not be considered. Tokens were also discarded if there were hesitations, pauses, and/or noise interferences at the crucial points of interest. The distribution of tokens by context is given in Table IV.

B. Electropalatography

The articulatory goal of the study was to ascertain how articulation of word-initial long and short voiceless stops changes as a function of their preceding context. The measure we used was the duration of contact during the constriction of the consonants between the two main articulators, namely, the tip of the tongue and the anterior portion of the hard palate. To obtain these contact duration measures, we used the EPG system WINEPG (Articulate Instruments Ltd., Edinburgh, UK). In this system, 62 electrodes, embedded in a thin custom-fitted acrylic palate, are scanned for tongue-to-palate information at a sampling interval of 10 ms. Simultaneously, the audio signal was recorded at a sampling rate of 48 kHz via a Sennheiser MKH20P48 microphone. Data analysis of the articulatory duration measure was done with the ARTICULATE ASSISTANT software (Version 1.12), while analysis of the acoustic duration measures was done with the MULTI SPEECH software (Kay Elemetrics, Version 2.2).

C. Measurement

We annotated the EPG and speech files of our test items such that the articulatory and acoustic duration measures could be extracted and statistically analyzed.

1. Articulatory parameter

The articulatory annotations were done with the ARTICULATE ASSISTANT software. The annotation marked the interval between the first and last EPG frames in which 100% of the electrodes in the first row or at least 80% in the two front rows combined indicated contact of the articulators. We call this measure the *duration of maximum contact* (DMC) (cf. Kraehenmann and Jaeger, 2003). We give an illustration of this in Fig. 2 below.

2. Acoustic parameters

From the audio signal, two measures were annotated by using the MULTI SPEECH software. The first was CD. With the help of spectrograms and wave forms, tag 1 was set in the C-context data at the offset of the random noise, particularly in the higher frequencies, of the preceding velar fricative [x]. In the data for the V-context and the tokens with medial contrast, the offset of the vowel was taken to be the point at which there was sudden drop in amplitude along with the disappearance of higher harmonics in the wave form. In the isolation context, tag 1 could not be set because there was no preceding sound. Tag 2 designated the point of closure release in all data. The second acoustic measure, VOT,[1] designated the interval between the stop release (tag 2) and the beginning of the regular wave form pattern (i.e., voicing) of the following vowel (tag 3). Although CD is the most relevant cue also cross-linguistically, we include VOT as a measure because it has been shown to play a role in some languages with a quantity contrast, for example, in Cypriot Greek by Tserdanelis and Arvaniti (2001) and Muller (2003) or in Turkish by Lahiri and Hankamer (1988). Furthermore, there could be phrase-initial strengthening of this measure similar to the Korean VOT measures by Jun (1993), Cho and Keating (2001), and Keating *et al.* (2003).

D. Statistical analysis

An ANOVA was separately performed for the words with initial and medial contrast (using the statistical software suite JMP; SAS Institute, 2003; MAC version 5.0.1.2) with the following independent factors: *speaker* (as random factor),

TABLE IV. Distribution of used tokens by context.

	Names initial		Nouns initial		Nouns medial		
	Sing	Gem	Sing	Gem	Sing	Gem	
C-	202	160	242	236			
V-	186	171	231	232	140	146	
Iso	149	131	219	225			
	537	462	692	693	140	146	2670

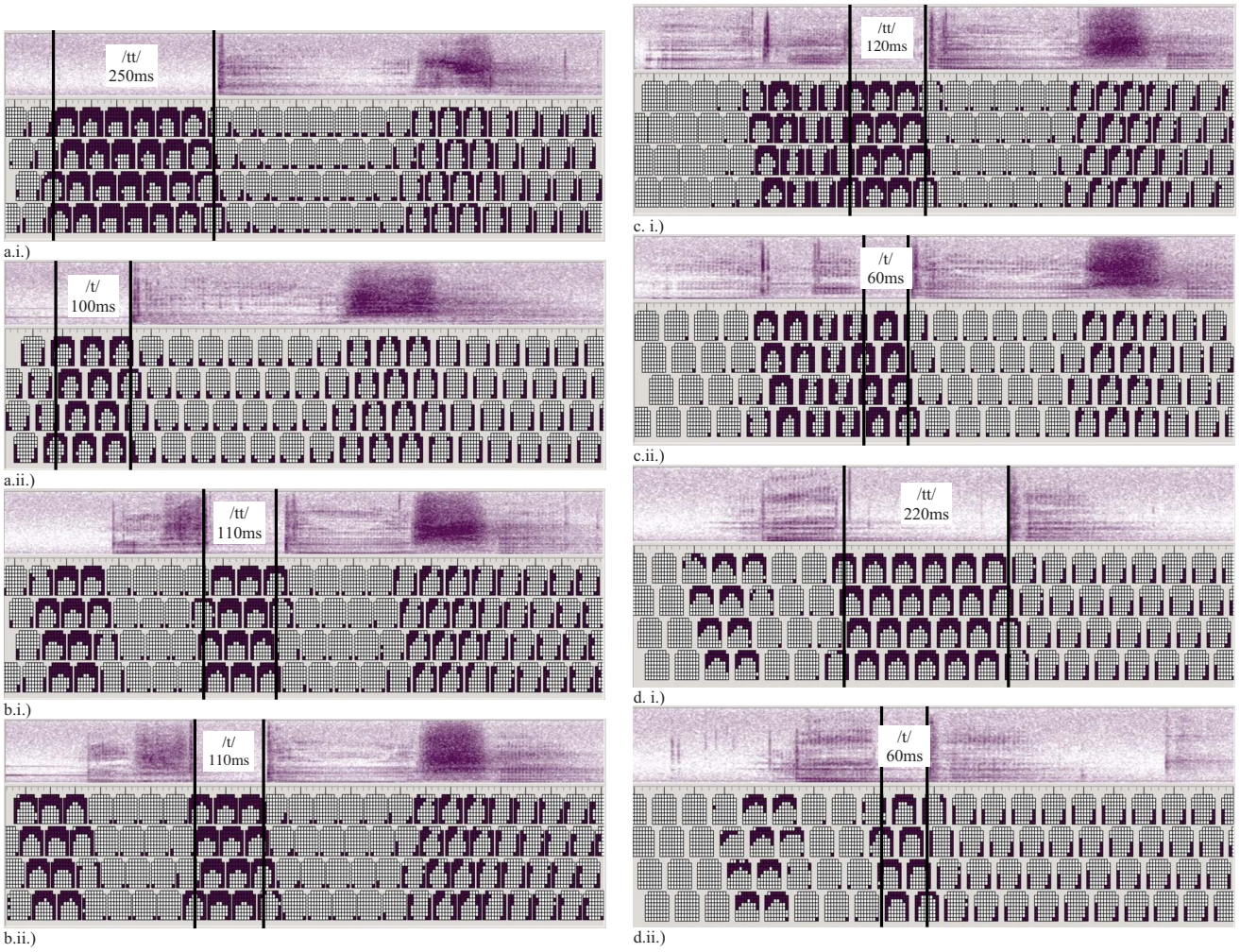


FIG. 2. (Color online) EPG illustration of quantity contrast in (a) isolation, (b) consonantal, (c) vocalic context, and (d) in word-medial context. The duration can be calculated as follows: (number of frames within lines—1) multiplied by 10 ms: (a.i.) /**tt**o:sə/ "roar," speaker 3, (a.ii.) /**t**o:sə/ "can," speaker 3, (b.i.) /nɔ**x** **tt**o:sə/ "after roar," speaker 2, (b.ii.) /nɔ**x** **t**o:sə/ "after can," speaker 2, (c. i.) /on**i** **tt**o:sə/ "without roar," speaker 2, (c.ii.) /on**i** **t**o:sə/ "without can," speaker 2, (d.i.) /la**tt**ə/ "crossbar," speaker 2, and (d.ii.) /la**t**ə/ "shop," speaker 2.

quantity (singleton, geminate), *condition* (isolation, consonantal, vocalic), and *noun type* (common noun or proper name) in a standard least squares design by using the restricted maximum likelihood (REML) estimation. The dependent variables were DMC, CD, and VOT. Significance was computed at the 5% level, and asterisks in the graphs and after the probability values indicate significant value differences.

III. RESULTS

A. Articulatory parameter: DMC

For the DMC measure there was no main effect for noun type [$F(1,2376)=2.47$, $p=0.7344$]. We found main effects for quantity, $F(1,2376)=1155.01$, $p=0.0002$*, and for condition, $F(2,2376)=1597.17$, $p=0.0009$*. Articulator contact was on average 55 ms longer for geminates (167 ms) than for singletons (112 ms). It was longest in the isolation context (207 ms), shorter in the C-context (109 ms), and shortest in the V-context (102 ms).

A *post hoc* test revealed that geminate contact was significantly longer in all three contexts (Fig. 3; Table V).

However, an additional *post hoc* test showed that the difference in DMC between geminates and singletons was significantly smaller in the C-context as compared to both the V- and the isolation context ($p<0.0001$*).

In comparison with the DMC values for word-medial geminates and singletons, the word-initial contrast in the V-context spanned a smaller range: The ratio of geminates to singletons was roughly 2:1 in the latter, as opposed to 3:1 for

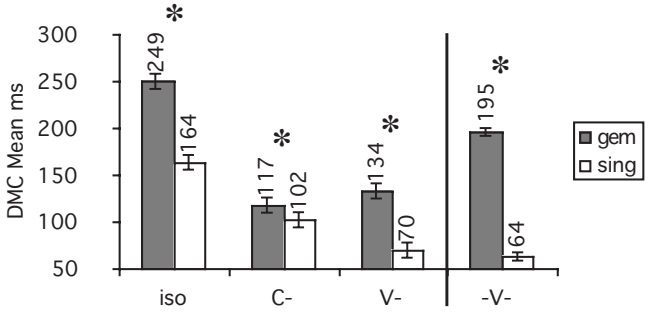


FIG. 3. DMC least squares means (ms) for *quantity* within *context*. Error bars: ±1 standard deviation(s).

TABLE V. DMC least squares means (ms), standard error (ms), difference (ms), and probability values for geminates and singletons in the three word-initial contexts (iso, *C*-, *V*-) and one word-medial context (-*V*-).

			LSM	Std error	*diff*	*p*
Initial	iso	Gem	249	16.6	85	<0.0001*
		Sing	164	16.6		
	C-	Gem	117	16.6	15	<0.0001*
		Sing	102	16.6		
	V-	Gem	134	16.6	64	<0.0001*
		Sing	70	16.6		
Medial	-*V*-	Gem	195	8.6	131	<0.0001*
		Sing	64	8.7		

the former. However, the difference displayed the same high level of statistical significance for both value sets.

B. Acoustic parameters: CD and VOT

As with DMC, there was a main effect for the CD measure for quantity, $F(1,1653)=1278.35$, $p=0.0022$*. CD was on average 42 ms longer for geminates (132 ms) than for singletons (90 ms). There was no main effect for condition, $F(1,1653)=78.27$, $p=0.0776$, nor for noun type, $F(1,1653)=2.45$, $p=0.1181$.

The factors quantity and condition significantly interacted. Geminates had longer CDs than singletons both in the C- and the V-context (Fig. 4; Table VI).

Here, too, the difference in CD between geminates and singletons was significantly smaller in the C-context as compared to the V-context ($p<0.0001$*).

The comparison to the CD values of medial geminates and singletons was virtually the same as in the articulatory data.

For VOT, there was no main effect for quantity, $F(1,2376)=34.57$, $p=0.4398$, nor for noun type, $F(1,2376)=4.73$, $p=0.3917$. However, there was an effect for condition, $F(2,2376)=4323.15$, $p=0.0209$* (Fig. 5, Table VII). Similar to the two other length measures, we found the shortest average duration in the C-context. The VOT values were significantly smaller in the C-context as compared to both the V-context ($p=0.0133$*) and the isolation context ($p=0.0147$*). There was no significant difference between the vocalic and the isolation context ($p=0.9782$).

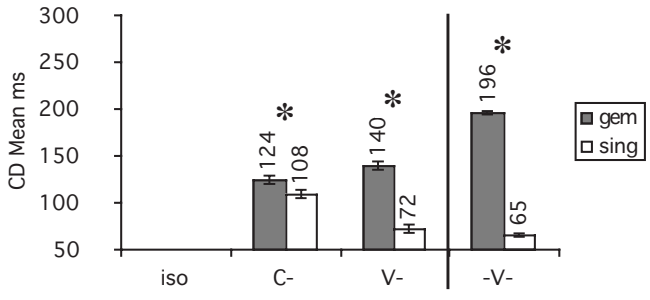


FIG. 4. CD least squares means (ms) for *quantity* within *context*. Error bars: ±1 standard deviation(s).

IV. DISCUSSION

We investigated Swiss German alveolar geminate and singleton stops in the word-initial position with respect to their articulatory and acoustic duration properties. Our main interest was directed toward finding out how the quantity distinction is manifested in articulator contact, closure duration, and release duration and how they vary as a function of different phrasal contexts.

The one variable which played the most minor role in our study was noun type: The speakers treated the word-initial geminates and singletons of proper names just like those in common nouns in all three conditions tested. Geminates, therefore, seem to be just as different and distinguished from singletons in proper names as they are in common nouns.

Regarding the other variables in our data, the three hypotheses we entertained were the following: (a) phrase-initial geminates would maintain longer contact than singletons although the acoustic correlate of closure duration was unavailable, (b) both phrase-initial geminates and singletons are produced longer than phrase-medial ones, and (c) the contrast between phrase-medial geminates and singletons in the consonantal context is neutralized while it is maintained in the vocalic context. We discuss each in turn.

(a) Our first hypothesis is confirmed, i.e., the contrast between word-initial geminates and singletons is maintained in phrase-/utterance-initial position in articulation. The average DMC measure is 249 ms for geminates and is 164 ms for singletons. This result may seem surprising if viewed with acoustics only in mind, because the quantity contrast is between voiceless stops and thus there is no acoustic cue of closure duration in this position.

(b) Second, there is a marked increase in the duration of linguopalatal contact for both geminates and singletons in phrase initial as compared to the phrase-medial contexts (see Fig. 3 and Table V). This finding confirms the pilot results by Kraehenmann and Jaeger (2003) and is similar to Tashlhiyt Berber, as discussed by Ridouane (2007). In *absolute* measures, geminates as well as singletons are articulated roughly 100 ms longer than in the phrase-medial vocalic context. These results replicate previous findings of articulatory strengthening at the beginning of a higher prosodic domain for a range of other languages (e.g., Fougeron and Keating, 1997; Keating *et*

TABLE VI. CD least squares means (ms), standard error (ms), difference (ms), and probability values for geminates and singletons in two of the three word-initial contexts (iso, *C*-, *V*-) and one word-medial context (-*V*-).

			LSM	Std error	*Diff*	*p*
Initial	*C*-	Gem	124	9.1	16	0.0002*
		Sing	108	9.1		
	V-	Gem	140	9.1	68	<0.0001*
		Sing	72	9.1		
Medial	-*V*-	Gem	196	3.7	131	<0.0001*
		Sing	65	3.7		

al., 2003; Byrd *et al.*, 2005; Cho and McQueen, 2005; Ridouane, 2007).

However, other than the absolute differences, we are also interested in the realization of the phonological contrast. In *proportional* terms, the highly significant difference between geminate and singleton articulations has become considerably smaller in the phrase-initial isolation context as opposed to the phrase-medial vocalic context: 1.5:1 versus 2:1. Thus, for the articulatory measure, the difference between geminates and singletons decreases rather than increases in phrase-initial position. This means that, although the articulation is heightened within both categories, the contrast between the categories is not.

(c) Turning now to the comparison between the two phrase-medial conditions, articulation and the acoustic measure CD parallel each other (cf. Figs. 3 and 4) as was established in earlier studies (Dieth and Brunner, 1943; Kraehenmann and Jaeger, 2003). As expected, geminates and singletons are clearly distinguished in the V-context.

Contrary to our expectations, however, the contrast also seems to be realized in the C-context. Kraehenmann (2001) and Kraehenmann and Jaeger (2003) did not find any articulatory or acoustic length differences for geminates and singletons in this condition. In other words, the contrast was neutralized in their data. While the difference of about 16 ms between geminates and singletons is statistically significant in this study, it is highly questionable—and subject to further study—whether it is also linguistically significant, i.e., whether it is sufficient for the phonological contrast to be recoverable in perception. At any rate, as mentioned above, the difference in articulatory and acoustic length is significantly smaller in the C-context than in the V-context. In terms of proportion, there is a 2:1 ratio in the V-context as compared to a 1.1:1 ratio in the C-context. The shortening of the geminate was expected based on the assumption that a syllable position is lost [cf. Fig. 1(b.i.)]. What the syllabification account cannot explain is the fact that singletons also lengthen in the C-context, which means that they strengthen although the prosodic structures are identical in both contexts [cf. Fig. 1(b.ii.) versus Fig. 1(c.ii.)]: The singletons begin the same word-initial syllable. This lengthening suggests that speakers are attempting to approach similar duration values which are within the ambiguous range of geminates and singletons, namely, around the 90–110 ms mark in this data set. Kraehenmann (2003) reports in her investigation that the CD values of geminates and singletons in the C-context (77.6 ms versus 70.5 ms) are comparable to the CD values of stops in word-medial consonant clusters, such as /nix**tə**/ "niece" (78.1 ms). Our data unfortunately do not contain word-medial clusters and thus we cannot verify whether our values are comparable.

Note that the DMC measures for word-medial stops are three times as long for the geminates as for the singletons, confirming results of earlier acoustic work (Kraehenmann, 2001, 2003). The word-medial contrast, however, was only used as a control to establish that the DMC measures would be parallel to the acoustic measures established in earlier research. We did not vary the words with medial contrast in different phrasal contexts and, hence, there is no phrasal articulatory strengthening issue. Within a word, one could vary the prosodic environment in terms of the number of syllables, such as [kʀavá**tt**ə] "tie" versus [má**tt**ine] "matinee," but that would be another study.

A final comment concerns the acoustic measure VOT (Fig. 5). Our results are as anticipated, considering that there

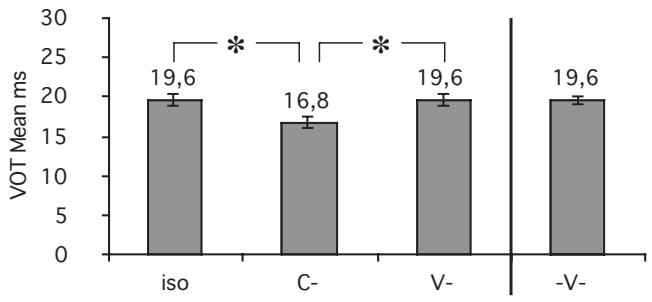


FIG. 5. VOT least squares means (ms) for *context*. Error bars: ±1 standard deviation(s).

TABLE VII. VOT least squares means (ms) and standard error (ms) across the three word-initial contexts (iso, *C*-, *V*-) and in one word-medial context (-*V*-).

		LSM	Std error
Initial	iso	19.6	1.5
	C-	16.8	1.5
	V-	19.6	1.5
Medial	-*V*-	19.6	0.7

was no main effect for quantity. This means that there was no statistical difference in the duration of the closure release for geminates and singletons, which is consistent with findings in earlier work (cf. Kraehenmann, 2001; 2003; Kraehenmann and Jaeger, 2003; Staeheli, 2005) and was also found for the voiceless stops of Tashlhiyt Berber by Ridouane (2007). Therefore, VOT can be ruled out as attributing to the phonological quantity contrast. What is surprising is the way VOT measures differ across—rather than within—contexts. If VOT patterned like the measures by Jun (1993), Cho and Keating (2001), and Keating *et al.* (2003), we would expect longer values in the phrase-initial context as opposed to the phrase-medial ones. While they are indeed significantly longer than in the C-context ($p=0.0147$*), they are indistinguishable from the ones in the V-context ($p=0.9782$). Thus, what we found is not an instance of language-specific enhancement of a phonetic feature (cf. Cho and McQueen, 2005), since it does not make a phonological contrast (i.e., quantity) more pronounced. Rather, it seems that it marks a certain phonetic context, the context in which stops are shortest in their primary correlate, namely, the duration of the articulatory and acoustic closure: After an obstruent-final word, word-initial stops have the shortest closure as well as the shortest VOT. With a difference of barely 3 ms, it appears very unlikely that it is more than a mechanical effect of the shorter closure gesture. The fact that VOT does not lengthen phrase initially [where we have the longest closures (cf. Fig. 3)] in comparison to phrase medially in the vocalic context leads us to conclude that release duration of any sort is phonologically as well as phonetically absolutely inert in this language.

V. CONCLUSIONS

Our study showed that a contrast in word-initial voiceless geminate and singleton stops is clearly maintained phrase initially where the main acoustic cue, closure duration, is missing. There is no articulatory neutralization of the word-initial quantity contrast. Whether this articulatory difference can be exploited in perception is subject to further investigations. In the phrase-medial position, the quantity contrast of the word-initial stops is considerably reduced in absolute terms, both regarding the acoustic CD and articulatory contact. Moreover, the difference is sensitive in the segmental context. When the preceding word ends with a vowel (V-context), the duration measures are much longer for geminates than for singletons (approximately 70 ms, 2:1 for DMC and CD). In comparison when the preceding word ends in an obstruent consonant, the differences—although significant—are marginal (approximately 15 ms.; 1.1:1 for CD and DMC). One could, therefore, claim that the word-initial contrast is enhanced phrase initially and that Swiss German shows the same domain-initial articulatory strengthening as found in other languages (e.g., Fougeron and Keating, 1997; Keating *et al.*, 2003; Byrd *et al.*, 2005; Cho and McQueen, 2005; Ridouane, 2007). Nevertheless, in proportional terms, the results are ambiguous. Although both DMC and CD are much greater for both geminates and singletons in phrase-/utterance-initial position (e.g., DMC geminates 249 ms, singletons 164 ms), the difference between them is 1.5:1, as compared to either the phrase-medial vocalic context, where the difference is 2:1, or the word-medial context, where the difference is 3:1. Thus, the quantity contrast itself is not enhanced phrase initially, although an overall strengthening effect at a prosodic boundary is undoubtedly there.

ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation: SFB 471, Leibniz Prize awarded to Aditi Lahiri) and the Ministry of Science and Culture, Baden-Württemberg. First, we would like to thank our four subjects for their generous cooperation. Thanks also go to Madeleine Staeheli for her extensive help in the laboratory, to Achim Kleinmann for his technical assistance, and to Henning Reetz and Willi Nagl for their support with the statistics. Furthermore, we would like to acknowledge Dr. Med. Dent. Thomas Hörmeyer and his laboratory team in Konstanz for their free service on the dental models for the acrylic palates. Thanks also to Arthur Abramson, Frans Plank, Rachid Ridouane, Allison Wetterlin, two anonymous JASA reviewers, the JASA Associate Editor Anders Löfqvist, and the participants of the ICPhS 2007 in Saarbrücken for their helpful comments and discussions on earlier versions of this work.

APPENDIX A: NAMES AND NOUNS WITH WORD-INITIAL CONTRAST

Names /tt/		Names /t/	
Tahar	Timon	Dagi	Dimo
Taina	Tinette	Dagmar	Dina
Taleb	Tito	Dagon	Dino
Tamar	Titus	Daina	Dogan
Tamra	Tobi	Dani	Domi
Tanja	Tomi	Daphne	Donald
Tara	Toni	Dara	Donat
Tatian	Tonya	David	Doris
Tela	Toris	Delfons	Dorit
Telka	Tory	Delma	Dunja
Telmo	Tunja	Denja	Durgun
Tero	Tünde	Derik	Dylan
Tessa	Türkkan	Deta	Dürke
Nouns /tt/		Nouns /t/	
Taler	"old coin"	Dame	"lady"
Tate	"deeds"	Datum	"date"
Tackel	"dachshund"	Daune	"down"
Taucher	"diver"	Delle	"dent"
Teller	"plate"	Delta	"delta"
Tecki	"cover"	Denker	"thinker"
Teflon	"Teflon"	Deppe	"dorks"
Tempel	"temple"	Dessin	"pattern"
Tesseer	"dessert"	Detail	"detail"
Ticki	"thickness"	Dichter	"poet"
Tiger	"tiger"	Dichti	"density"
Tili	"ceiling"	Dichtig	"seal"

Tinte	"ink"	Dinar	"denar"
Tipex	"Tipex"	Diner	"dinner"
Tischler	"carpenter"	Dischtle	"thistle"
Tischli	"table (DIM)"	Disel	"diesel"
Toggel	"pawn"	Diwan	"divan"
Toner	"toner"	Dogge	"mastiff"
Tonner	"thunder"	Doole	"jackdaw"
Totzet	"dozen"	Dooping	"doping"
Toose	"roar"	Doppel	"duplicate"
Tuume	"thumb"	Dose	"can"
Tuure	"tours"	Dosis	"dosage"
Tunell	"tunnel"	Dossier	"file"
Tuusig	"thousand"	Double	"double"
Tuusis	"Thusis"	Duden	"dictionary"
Tääler	"valleys"	Dumping	"dumping"
Töönig	"tinge"	Duuma	"Duma"
Töörli	"gate (DIM)"	Dääne	"Danes"
Tüle	"pie"	Dööner	"kebab"
Tümpel	"pool"	Döösli	"can (DIM)"
Tüüfi	"depth"	Dübel	"peg"
Tüürig	"inflation"	Dünger	"fertilizer"
Tüüschig	"deception"	Düse	"nozzle"
		Düüter	"interpreter"
		Düütig	"interpretation"

APPENDIX B: NOUNS WITH WORD-MEDIAL CONTRAST

Medial /tt/		Medial /t/	
Butter	"butter"	Adel	"nobility"
Chette	"chain"	Badi	"bath"
Chittel	"frock"	Bode	"floor"
Chlette	"barnacle"	Flider	"lilac"
Chutte	"cowl"	Jodel	"yodel"
Flotte	"fleet"	Liide	"affliction"
Foti	"photograph"	Luuder	"hussy"
Hütte	"hut"	Moode	"fashion"
Jute	"jute"	Pudel	"poodle"
Leiter	"leader"	Sooda	"soda"
Motte	"moth"	Model	"model"
Latte	"bar"	Lade	"store"
Kööter	"mutt"	Kööder	"lure"
Vatter	"father"	Fade	"thread"
Matte	"mat"	Made	"maggot"
Fuetter	"feed"	Fueder	"cart load"
Lette	"mud"	Leder	"leather"
Patte	"flap"	Paddel	"paddle"
Ruete	"rod"	Rueder	"oar"
Schatte	"shadow"	Schade	"damage"
Wetter	"weather"	Wedel	"frond"

[1]The combined use of *after closure time* (ACT), *superimposed aspiration* (SA), and CD has been established as a more accurate means of quantifying the difference between voiced and voiceless consonants (Mikuteit and Reetz, 2007; cf. also Clements and Khatiwada, 2007), lessening the confusion between CD definitions overlapping with lead VOT or negative VOT, positive VOT overlapping with aspiration, and so on. However, we chose to continue by using the term VOT for ACT since we are not dealing with a voicing contrast. The consonants are all voiceless and neither prevoicing nor aspiration play any role in the quantity distinction. Consequently, VOT essentially means *positive VOT*, i.e., consists of the duration of the burst release, and is interchangeable with ACT.

Abramson, A. S. (**1986**). "The perception of word-initial consonant length: Pattani Malay," J. Int. Phonetic Assoc. **16**, 8–16.

Abramson, A. S. (**1987**). "Word-initial consonant length in Pattani Malay," *Proceedings of the 11th International Congress of the Phonetic Sciences*, Tallinn, 68–70.

Abramson, A. S. (**1991**). "Amplitude as cue to word-initial consonant length: Pattani Malay," *Proceedings of the 12th International Congress of the Phonetic Sciences*, Aix-en-Provence, 98–101.

Abramson, A. S. (**1999**). "Fundamental frequency as cue to word-initial consonant length: Pattani Malay," *Proceedings of the 14th International Congress of the Phonetic Sciences*, San Francisco, 591–594.

Brédart, S. (**1993**). "Retrieval failures in face naming," Memory **1**, 351–366.

Byrd, D., Lee, S., Riggs, D., and Adams, J. (**2005**). "Interacting effects of syllable and phrase position on consonant articulation," J. Acoust. Soc. Am. **118**, 3860–3873.

Cho, T., and Keating, P. A. (**2001**). "Articulatory and acoustic studies on domain-initial strengthening in Korean," J. Phonetics **29**, 155–190.

Cho, T., and McQueen, J. M. (**2005**). "Prosodic influences on consonant production in Dutch: effects of prosodic boundaries, phrasal accent and lexical stress," J. Phonetics **33**, 121–157.

Clements, G. N., and Khatiwada, R. (**2007**). "Phonetic realization of contrastively aspirated affricates in Nepali," *Proceedings of the 16th International Congress of the Phonetic Sciences*, Saarbrücken, 629–632.

Dieth, E., and Brunner, R. (**1943**). "Die Konsonanten und Geminaten des Schweizerdeutschen experimentell untersucht (The consonants and geminates of Swiss German examined experimentally)," Romanica Helvetica **20**, 757–762.

Dunn, M. H. (**1993**). *The Phonetics and Phonology of Geminate Consonants: A Production Study* (UMI Dissertation Services, Ann Arbor, MI).

Evrard, M. (**2002**). "Ageing and lexical access to common and proper names in picture naming," Brain Lang **81**, 174–179.

Farnetani, E. (**1990**). "V-C-V lingual coarticulation and its spatiotemporal domain," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 93–130.

Fougeron, C., and Keating, P. A. (**1997**). "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740.

Fujimura, O. (**1990**). "Methods and goals of speech production research," Lang Speech **33**, 195–258.

Fulop, S. A. (**1994**). "Acoustic correlates of the fortis/lenis contrast in Swiss German plosives," Calgary Working Papers in Linguistics **16**, 55–63.

Ham, W. (**1998**). *Phonetic and Phonological Aspects of Geminate Timing* (CLC, Ithaca, NY).

Iverson, G. K., and Salmons, J. C. (**1995**). "Aspiration and laryngeal representation in Germanic," Phonology **12**, 369–396.

Izaute, M., Chambres, P., and Larochelle, S. (**2002**). "Feeling-of-knowing for proper names," Can. J. Exp. Psychol. **56**, 263–272.

Jessen, M., and Ringen, C. (**2002**). "Laryngeal features in German," Phonology **19**, 186–218.

Jun, S.-A. (**1993**). "The phonetics and phonology of Korean prosody," Ph.D. thesis, Ohio State University.

Keating, P. A., Cho, T., Fougeron, C., and Hsu, Ch.-Sh. (**2003**). "Domain initial articulatory strengthening in four languages," in *Papers in Labora-*

tory Phonology VI: Phonetic Interpretation, edited by J. Local, R. A. Ogden, and R. Temple (Cambridge University Press, Cambridge), pp. 145–163.
Kohler, K. J. (**1984**). "Phonetic explanation in phonology: The feature Fortis/Lenis," Phonetica **41**, 150–174.
Kraehenmann, A. (**2001**). "Swiss German Stops: Geminates all over the word," Phonology **18**, 109–145.
Kraehenmann, A. (**2003**). *Quantity and Prosodic Asymmetries in Alemannic: Synchronic and Diachronic Perspectives* (Mouton de Gruyter, Berlin).
Kraehenmann, A., and Jaeger, M. (**2003**). "Initial geminate stops: Articulatory evidence for phonological representation," *Proceedings of the 15th International Congress of the Phonetics Sciences*, Barcelona, 2725–2728.
Kurylowicz, J. (**1966**). "La position linguistique du nom propre (The linguistic status of proper nouns)," in *Readings in Linguistics II*, edited by E. P. Hamp, F. W. Householder, and R. Austerlitz (University of Chicago Press, Chicago), pp. 362–370.
Lahiri, A., and Hankamer, J. (**1988**). "The timing of geminate consonants," J. Phonetics **16**, 327–338.
Lehiste, I., Morton, K., and Tatham, M. A. A. (**1973**). "An instrumental study of consonant gemination," J. Phonetics **1**, 131–148.
Löfqvist, A. (**2005**). "Lip kinematics in long and short stop and fricative consonants," J. Acoust. Soc. Am. **117**, 858–878.
Löfqvist, A. (**2007**). "Tongue movement kinematics in long and short Japanese consonants," J. Acoust. Soc. Am. **122**, 512–518.
Mikuteit, S., and Reetz, H. (**2007**). "Caught in the ACT," Lang Speech **50**, 247–277.
Muller, J. S. (**2003**). "The production and perception of word initial geminates in Cypriot Greek," *Proceedings of the 15th International Congress of the Phonetic Sciences*, Barcelona, 1867–1870.
Payne, E. (**2006**). "Non-durational indices in Italian geminate consonants," J. Int. Phonetic Assoc. **36**, 83–95.
Ridouane, R. (**2007**). "Gemination in Tashlhyt Berber: An acoustic and articulatory study," J. Int. Phonetic Assoc. **37**, 119–142.
Staeheli, M. (**2005**). "Affrikate und Obstruenten-Geminaten im Thurgauer Dialekt (Affricates and obstruent geminates in the dialect of Thurgovian)," M.S. thesis, University of Konstanz.
Tserdanelis, G., and Arvaniti, A. (**2001**). "The acoustic characteristics of geminate consonants in Cypriot Greek," *Proceedings of the Fourth International Conference on Greek Linguistics*, 29–36.
Wiese, R. (**1996**). *The Phonology of German* (Claredon, Oxford).
Willi, U. (**1996**). *Die segmentale Dauer als phonetischer Parameter von 'fortis' und 'lenis' bei Plosiven im Zürichdeutschen (Segment duration as a phonetic parameter of 'fortis' and 'lenis' stops in Zurich German)* (Franz Steiner, Stuttgart).

Phrase boundary effects on the temporal kinematics of sequential tongue tip consonants[a)]

Dani Byrd, Sungbok Lee,[b)] and Rebeka Campos-Astorkiza[c)]
Department of Linguistics, University of Southern California, 3601 Watt Way, GFS 301, Los Angeles, California 90089-1693

(Received 12 July 2007; revised 28 March 2008; accepted 31 March 2008)

This study evaluates the effects of phrase boundaries on the intra- and intergestural kinematic characteristics of blended gestures, i.e., overlapping gestures produced with a single articulator. The sequences examined are the juncture geminate [d(#)d], the sequence [d(#)z], and, for comparison, the singleton tongue tip gesture in [d(#)b]. This allows the investigation of the process of gestural aggregation [Munhall, K. G., and Löfqvist, A. (1992). "Gestural aggregation in speech: laryngeal gestures," J. Phonetics **20**, 93–110] and the manner in which it is affected by prosodic structure. Juncture geminates are predicted to be affected by prosodic boundaries in the same way as other gestures; that is, they should display prosodic lengthening and lesser overlap across a boundary. Articulatory prosodic lengthening is also investigated using a signal alignment method of the functional data analysis framework [Ramsay, J. O., and Silverman, B. W. (2005). Functional Data Analysis, 2nd ed. (Springer-Verlag, New York)]. This provides the ability to examine a time warping function that characterizes relative timing difference (i.e., lagging or advancing) of a test signal with respect to a given reference, thus offering a way of illuminating local nonlinear deformations at work in prosodic lengthening. These findings are discussed in light of the π-gesture framework of Byrd and Saltzman [(2003) "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," J. Phonetics **31**, 149–180]. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2912444]

PACS number(s): 43.70.Bk [AL] Pages: 4456–4465

I. INTRODUCTION

A. Articulatory overlap and juncture geminates

Coproduction in speech occurs when two gestures temporally overlap. When these gestures share all or some of their articulators, they have been described as *blended* in that the gestural parameters are (basically) averaged (Saltzman and Munhall, 1989). When two abutting consonants are identical, they are called juncture geminates. The temporal behavior of juncture geminates is important to study as they are canonical *sequences* but are produced as *single-articulator* constriction movements. When consonants abutting at word edges use different articulator sets, we can observe their coproduction rather clearly; we see two different constrictions being made in a temporally overlapped fashion (Browman and Goldstein, 1992). However, when the two consonants share the same articulator, as in the case of juncture geminates, their individual characteristics become more difficult to observe. This type of coproduction has been called *gestural aggregation* (Munhall and Löfqvist, 1992). The fact that two individual gestures are being coarticulated may or may not be obvious from the articulatory movements. For example, Munhall and Löfqvist (1992) examining the aggregation of two overlapping laryngeal gestures across a word boundary observe a single smooth movement at fast speaking rates.

For cases when only one continuous movement is observed, two analyses seem possible. First, overlap with blending of two separate gestures could simply result in a movement with only a single displacement extremum, i.e., a single smooth trajectory of movement (Saltzman and Munhall, 1989). However, summation of the two underlying gestures could also result in such a trajectory (Munhall and Löfqvist, 1992). Previous studies suggest that juncture geminates are the result of extreme overlap and blending rather than a summation process. The degree of overlap between coproduced consonants varies with the speech rate so that overlap increases as the rate gets faster (Byrd and Tan, 1996; Hardcastle, 1985). Munhall and Löfqvist (1992) looked at laryngeal gestures in juncture geminates across different speech rates. In fact, at slower rates, Munhall and Löfqvist observed two distinct laryngeal movements; at faster rates, a single smooth trajectory was present. These results suggest that the single movement in fast speech was the consequence of great overlap between the two gestures. Comparable results were reported by Löfqvist and Yoshioka (1981).

Juncture geminates can also be differentiated from single consonants in terms of their durational characteristics and their degree of articulator displacement. Kelso and Tuller (1987) note that the results of gestural summation would be a larger gesture with increased amplitude and steeper onset and offset slopes. Byrd (1995) used electropalatography to

[a)] A preliminary and abbreviated portion of this paper appeared in "Gestural deaggregation via prosodic structure," Proceedings of the Seventh International Seminar on Speech Production, Ubatuba, Brazil, 2006.
[b)] Electronic mail: dbyrd@usc.edu
[c)] Present address: The Ohio State University Department of Spanish and Portuguese, Hispanic Linguistics program.

 0001-4966/2008/123(6)/4456/10/$23.00

investigate lingual juncture geminates in English. Linguapalatal contact patterns indicate that these juncture geminates were produced with a single raising and lowering of the tongue. Byrd (1995) found the coproduced articulation for juncture geminates to be longer than the movement for a single gesture. Byrd's results also indicated no consistent increase in maximum contact for the geminated consonants relative to single consonants syllable onsets. Also, constriction formation and release contact slopes of the juncture geminates, i.e., the temporal pattern of increase or decrease in linguapalatal contact as indicated by the number of electrodes contacted on the palate, appeared comparable to those of singleton onset and coda slopes, respectively. These data suggest that a blending process that averages the spatial target values for two overlapping consonantal gestures is at work (Byrd, 1995). Munhall and Löfqvist (1992) also found no consistent tendency for the combined single movement to be larger than an individual (noncoproduced) movement, although a simulated summation of the gestures predicts such a difference. Vaxelaire (1995) examined x-ray film data on constrictions for lingual juncture geminates in French and found slightly different results in that the extent of contact between the tongue and palate is greater for the French juncture geminates than for the similar single consonant. Geminates had longer articulatory durations than singletons, as predicted from previous studies.

B. Prosodic effects on articulatory gestures

Prosodic structure affects the spatial and temporal characteristics of individual gestures, as well as the relative coordination among different gestures (Byrd and Saltzman, 2003; Byrd *et al.*, 2000). With respect to the influence on the intragestural characteristics, acoustic and articulatory studies have shown that gestures become longer near prosodic boundaries (Oller, 1973; Klatt, 1976; Wightman *et al.*, 1992 in the acoustic domain; Edwards *et al.*, 1991; Beckman and Edwards, 1992 in the articulatory domain). Also, gestures in the vicinity of prosodic boundaries become more extreme and larger (Fougeron and Keating, 1997; Byrd and Saltzman, 1998; Cho, 2005; Cho and Jun, 2000; Cho and Keating, 2001; Tabain, 2003; Keating *et al.*, 2004).

Articulatory studies have examined overlap patterns under the influence of prosodic structure and found that temporal overlap is less among gestures separated by or adjacent to a boundary (McClean, 1973; Byrd *et al.*, 2000; Byrd and Saltzman, 1998; Byrd and Choi, in press). Byrd *et al.* (2000) analyzed the timing patterns across a phrasal boundary of the tongue tip and upper lip gestures for two nasal consonants [m, n] in Tamil. The time between the gestural onsets was weakly affected by the presence of a boundary. However, the time between extrema, i.e., maximum closings, was significantly longer in the boundary condition. As for the relative timing, the word-final gesture reached its extremum position significantly earlier into the word-initial gesture when a phrase intervened. This indicates that the first gesture was less overlapped with the second gesture under this condition. These results, those by Byrd and Choi (in press), and simulations by Byrd and Saltzman (2003) show that gestures are pulled apart across a phrasal boundary, resulting in less temporal coproduction between gestures.

C. The π-gesture: A dynamical system approach to prosodic structure

Critically, these findings have been modeled as a boundary-adjacent slowing of the time course of gestural activation within Byrd and Saltzman's π-gesture model of phrasal structure in speech production (Byrd and Saltzman, 2003). Byrd and Saltzman suggest that phrase boundaries can be understood as a transitory warping of the local speech rate. In this model, prosodic boundaries are represented as cognitive control structures called π-gestures (prosodic gestures) that, when active, locally slow the clock that controls the overall temporal pacing of gestural activation. This local slowing is hypothesized to apply to all gestures coproduced with the π-gesture activation interval. Furthermore, the strength or amount of slowing (a consequence of the strength of activation of the π-gesture) is modeled as increasing as the phrase edge approaches and waning as it recedes in time. Both simulation and empirical work (Byrd and Saltzman, 2003; Byrd *et al.*, 2000) have shown that under the clock-slowing influence of a prosodic gesture, constriction gestures become both longer and less overlapped with one another. Under the π-gesture account of boundary effects on articulation, juncture geminates, just like the sequences of nonidentical consonants in the above studies, should show lengthening and lesser overlap when they occur across a phrase boundary. If the change in overlap is sufficient, the juncture geminate constrictions may even display signs of deaggregation or "pulling apart."

D. Goals of the present study

This study evaluates the effects of phrasal boundaries on the intra- and intergestural characteristics of blended gestures produced with the same articulator. This allows us to investigate the process of gestural aggregation and the manner in which it is affected by prosodic structure. Juncture geminates are expected to be affected by prosodic boundaries in the same way as other gestures. This means that they should display lengthening and a more extreme articulation across a boundary. This effect would be reflected in their constriction formation duration and plateau duration, which are predicted to become longer under the boundary condition. The intergestural changes in overlap when the juncture geminates occur across a phrase boundary are expected to be manifested in longer plateau duration as the two abutted gestures slide apart, possibly yielding longer total sequence duration. In fact, the intervening phrase boundary condition may even be associated with a two-peaked gestural trajectory.

In addition to the durational study of the effect of phrase boundary, the lengthening effect is also investigated using a signal alignment method provided in the functional data analysis (FDA) framework (Ramsay and Silverman, 2005). In general, the purpose of signal alignment is to minimize phase or timing differences among signals that are generated by a shared underlying process (e.g., repetitions of a same

TABLE I. Stimuli used in experiment.

Intervening boundary?	Sentences
	Juncture geminates [dd] (gestural aggregation):
Yes	JIMMY loved Do**dd**. **D**eb bet DAD did not know that.
No	JIMMY loved Do**dd**-**D**eb. Bet TAD did not know that.
	Same-articulator, non-geminate [dz]:
Yes	KIMMY loved Do**dd**. **Z**eb bet DAD did not know that.
No	KIMMY loved Do**dd**-**Z**eb. Bet TAD did not know that.
	Different-articulators, singleton [d]:
Yes	TIMMY loved Do**dd**. **B**ub bet TAD did not know that.
No	TIMMY loved Do**dd**-**B**ud. Bet DAD did not know that.

utterance). It provides an optimal way to estimate signal average and variability along the time axis, and the FDA signal alignment technique has been used in a number of studies for that purpose (Ramsay *et al.*, 1996; Lucero *et al.*, 1997; Lucero and Koenig, 2000; Koenig and Lucero, 2002; Lucero and Löfqvist, 2005). In addition to providing such signal processing advantages, an interesting extension of the signal alignment is the ability to examine a time warping function that characterizes relative timing difference (i.e., lagging or advancing) of a test signal with respect to a given reference. Lee *et al.* (2006) have shown that computation and examination of such time warping functions provide a comprehensive way to investigate the lengthening effect of a phrase boundary. The traditional, i.e., piecewise, method of comparing the durations of intervals delineated by kinematic landmarks such as velocity extrema or zero crossings does not give such a view of timing differences along the entire, continuous kinematic trajectories.

II. METHOD

A. Stimuli and subjects

The experiment stimuli were constructed to test the effects of an utterance-level phrase boundary between two consonants C_1 and C_2, produced using the same articulator. The stimulus sentences are given in Table I. Each stimulus was formed by two sentences related in semantic content. The subject of the first sentence and the object of the second sentence are in focus in order to curtail the possibility of accents at the boundary. While the supralaryngeal articulatory correlates of the boundary tone at the edge of the intonational phrase are part of the timing phenomena being investigated, it is preferable not to have additional pitch accents on the target words. Contrast in prosodic structure was generated via the use of nominal compounds, specifically proper name compounds, versus sequences of proper names spanning a phrase boundary. Subjects were instructed to model compound name productions on names like "Sue-Ann" and were given sufficient practice with the compound names before the experiment.

The stimuli were controlled for prosodic boundary and consonant sequence. The target consonants were either part of the same phrase (i.e., a compound name) or separated by an intonational phrase boundary. As for the sequences of interest, these were the juncture geminate [dd] for the same articulator and manner condition, [dz] for the same articulator but different manner condition, and [db] for the different articulator condition. The sentences were blocked by boundary condition, and ten consecutive repetitions of each sentence were recorded, yielding a total of 60 tokens for each subject. Three native speakers of American English with no known hearing or speech disorder participated in the study. Speakers will be referred to as N, K, and J.

B. Data collection

The electromagnetic midsagittal articulometer (EMMA) system (Perkell *et al.*, 1992) was used to track the horizontal (x) and vertical (y) movements of transducers adhered to the tongue tip and lips. Transducers were placed on the nose, upper and lower teeth (maxilla and jaw, respectively), upper and lower lips, and the tongue tip. (Also, irrelevantly for this dataset, for speakers N and J, three receivers were placed on the tongue body, and for speaker K two were). Of these points, only the tongue tip and lips trajectory will be relevant for the present study. The transducer trajectory data were sampled at a 625 Hz rate and the acoustic data at 20 kHz. The data were corrected for head movement using reference transducers on the nose and maxilla and were rotated to the occlusal plane. After voltage-to-distance conversion (with a filter cutoff of 17 Hz), correction for head movement (using the nose and maxillary reference transducers), and rotation to the occlusal plane, the position signals were subject to 25 Hz smoothing. The EMMA trajectory data for speakers N and J had quantization noise, and the corresponding velocity signals were subject to an additional smoothing routine of a lowpass filter at 25 Hz.

C. Durational analysis

In this paper, the focus is on the kinematic behavior of the tongue tip (TT) gesture produced in each of the target sequences, i.e., [dd], [dz], [db], and the lip movement in the [db] sequence. In the first part of Sec. III, we examine the articulation of these sequences from an Articulatory Phonology perspective (e.g., Browman and Goldstein, 1992)—specifically, we identify kinematic landmarks that are related to events in constriction formation and release that are thought to be important from a gestural control perspective, such as kinematic points related to gestural onset and target achievement. In the second part of Sec. III, we adopt a different approach to examining kinematic trajectories that is not limited to identification of kinematic landmarks and the piecewise durations between them—this is FDA, discussed in Sec. II D below.

For the lip movement in [db], a derived signal was created corresponding to lip aperture (LA). This signal was calculated as the Euclidean distance between the lower and the upper lip. Using the trajectory analysis software MVIEW (under development by Mark Tiede), five points were defined in the TT and LA trajectories for the tongue tip and bilabial closing gestures, respectively: gestural onset, plateau onset, maximum constriction, plateau offset, and gestural offset. Time and spatial values for these landmarks were derived

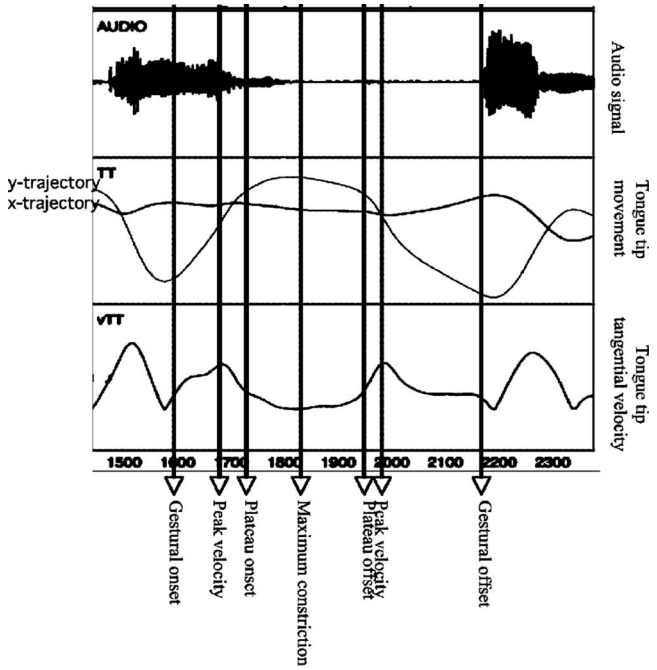


FIG. 1. Illustration of time point markings from a TT gesture "Jimmy loved Dodd. Deb bet Dad didn't know that."

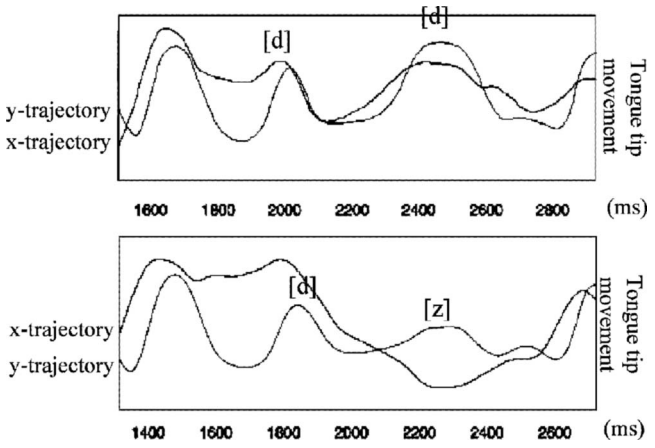


FIG. 2. Example of TT trajectory corresponding to "Dodd. Deb" (top) and "Dodd. Zeb" (bottom) for speaker J, illustrating this speaker's two-constriction production pattern for sequences [d#d] and [d#z] across a boundary.

from the TT tangential velocity and LA (one-dimensional) velocity trajectories. The gestural onset and offset correspond to the beginning of the constriction formation and the end of the release. They were calculated as threshold-crossing points in the tangential velocity trajectory for the TT and the velocity trajectory for LA, where the threshold was defined as a percentage of the range between the maximum and the minimum local velocity. The threshold was set to 20%. The plateau onset and offset correspond to the beginning and end of the constriction plateau. These were also calculated as threshold-crossing points, with a value of 30%, of the local velocity range.[1] The maximum constriction time point was of course not thresholded. Additionally, peak tangential velocity time points during the closing and opening movements were recorded. These were not thresholded and correspond to time points with velocity maxima. Figure 1 illustrates the seven measured time points from a TT gesture. Several tokens showed multiple peak velocities either in the constriction formation or the release movement. In those cases, the fastest peak was chosen. 17 tokens were excluded from the analysis because they were unusual, unusable, or missing.

It is worth noting that for [dd] and [dz], a single plateaued movement of the tongue tip was produced. However, speaker J, unlike the other two speakers, produced the sequences [d#z] and [d#d] in the boundary condition (only) with two distinct tongue tip gestures. Figure 2 illustrates these two-gesture productions of [d#z] and [d#d] across a phrasal boundary. For these tokens, the landmarks from both gestures were combined for the analysis. Specifically, the gestural onset, first peak velocity, and plateau onset were taken from the first tongue tip gesture, and the plateau offset, second peak velocity, and gestural offset were selected from the second gesture. The maximum constriction point was taken from the gesture with the highest plateau. For [d#z], this point was taken from the first gesture, and for [d#d], it was located in the second gesture.

Based on the TT and LA movement landmarks, three articulatory intervals were derived from the measured time points and used as dependent variables.

(1) Constriction formation duration for TT gesture: time from gestural onset to plateau onset.
(2) Plateau duration:
   (a) For [dd] and [dz] sequences: time from TT plateau onset to TT plateau offset.
   (b) For [db] sequences: time from TT plateau onset to LA plateau offset.

(3) Peak-velocity-to-peak-velocity duration:
   (a) For [dd] and [dz] sequences: time from TT peak velocity during closing movement to TT peak velocity during opening movement.
   (b) For [db] sequences: time from TT peak velocity during closing movement to LA peak velocity during opening movement.

The articulatory landmarks and derivable variables are represented in Fig. 3. Constriction formation duration is taken as an indicator of the closing gesture duration, and plateau and peakvel-to-peakvel duration serve as an index of gestural overlap. Sequence type ([dd], [dz], [db]) is not predicted to have an effect on constriction formation duration but may affect plateau and peakvel-to-peakvel duration. The effect of the presence or absence of a phrase boundary intervening between the two consonants is hypothesized to be significant for both duration variables but may differ in degree for the plateau and peakvel-to-peakvel duration depending on the sequence type.[2]

Individual two-factor, repeated measure ANOVAs for each subject were conducted in order to evaluate the effects of *boundary condition* and *sequence type* on constriction formation duration, plateau duration, and peakvel-to-peakvel duration, with boundary condition treated as the repeated measure across sequence types. PLSD *post hoc* tests were

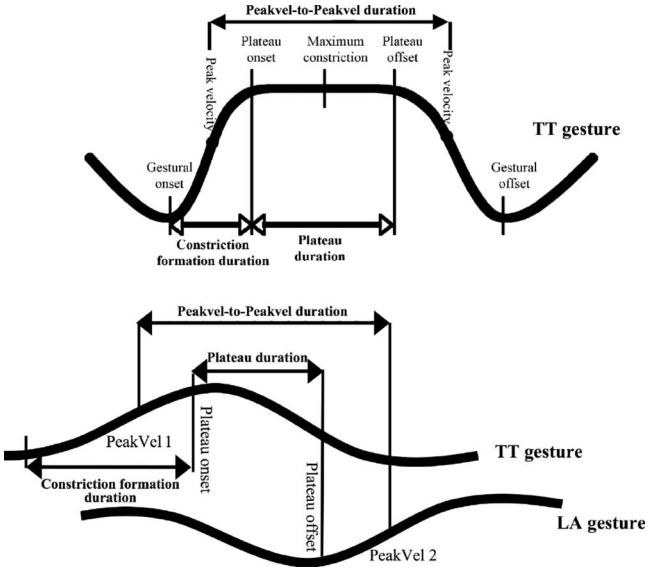


FIG. 3. (Top) Articulatory landmarks and derived dependent variables for the [dd] and [dz] sequences and (bottom) for the [db] sequence.

also carried out for pairwise comparisons among the different sequence types. The significance level for all the statistical tests was set at $p<0.01$, rather than a less conservative $p<0.05$ level, in order to lessen the effect of a larger degree of freedom on significance level that occurs by regarding each token as an experimental unit (cf., Max and Onghena, 1999), arising due to power requirements given the typically small number of subjects in articulatory kinematic experiments.

D. Timing pattern difference analysis based on the FDA signal alignment method

This study further investigates the effects of boundary and consonant sequence type by applying the FDA signal alignment, or time registration, method to the entire TT velocity pattern over time (Ramsay and Silverman, 2005; Lee *et al.*, 2006), rather than utilizing only specific kinematic time points. Here, the focus is on the velocity trajectories of the tongue tip gesture produced in each of the target sequences of [dd] and [dz] by the two subjects K and N. Velocity patterns are chosen because these have traditionally been used for the analysis of skilled movements as underlying kinematic signatures of important control events (see Nelson, 1983). Importantly, a velocity pattern has well-defined landmarks (i.e., extrema and zero crossings) that facilitate the FDA landmark time registration (Ramsay and Silverman, 2005; Lee *et al.*, 2006). We briefly describe next the conceptual outline of the FDA time registration method and how it is applied in this study.

The purpose of FDA time registration is to find a smooth time warping function $h(t)$ that minimizes the difference between test and reference signals, as shown in Eq. (1), where $h(t)$ is the time warping function to be determined, λ is a smoothing or regularization parameter, $w(t)$ is a smoothness control function for $h(t)$, and T is the end point of the time path.

$$D(x,y,\lambda,w)=\int_0^T [x(h(t))-y(t)]^2 dt+\lambda\int_0^T w(t)^2 dt. \quad (1)$$

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 constrained by Eq. (2).

$$\frac{d^2h(t)}{dt^2}=w(t)\frac{dh(t)}{dt}. \quad (2)$$

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 $h(t)$. For instance, when $w(t)$ is positive, the rate of internal time change of the test signal $h(t)$ is elongated when compared to the physical time [i.e., $h(t)>t$], and thus the test signal runs "late." That is, the same landmark in the test signal occurs later in clock time than that landmark in the reference signal. For further mathematical details, the reader can refer to Ramsay and Silverman (2005).

Because our interest is in timing, the landmark time registration option has been chosen for this study in order to take advantage of the clear landmark locations (extrema and zero crossings) observed in the velocity patterns. The landmark time registration accepts predetermined signal landmark time points as break points and performs time alignment between two successive landmark points by linear shifting and scaling. In this study, 12 B-splines of order 6 and a λ value of 10^{-12} are used to represent $w(t)$. All computations are performed with the MATLAB implementations of the FDA smoothing and time registration algorithms that are publicly available (Ramsay, 2007).

Each interval to which the time registration is applied extends from the maximum constriction of the TT gesture in the first /d/ in [dodd] to the LA minimum for the bilabial stop /b/ in the next word. Then, the control and test velocity signals are processed for each subject using the FDA time registration procedure as follows. First, a linear time normalization is applied to each individual velocity signal by resampling after smoothing so that each signal has 200 equally sampled data points.

A reference signal for each subject is then determined from the phrase-boundary utterance signals as follows. Initially, an average of these signals is computed and used as an initial reference signal for time alignment. Then, after time alignment, an average of the time-aligned test signals is computed again and used as a final reference pattern. Next, each no-phrase-boundary control signal is subjected to the landmark time registration with respect to the boundary reference signal, and each time warping function is computed against this reference.[3] After registration, a time deformation function $F(t)$ is computed as follows:

$$F(t)=h_{\text{no-bound}}(t)-h_{\text{bound}}(t). \quad (3)$$

A negative value of F represents a situation in which events of the no-boundary condition occur earlier relative to the timing of the like boundary condition event, i.e., earlier relative to the internal clock time of a no-boundary test signal with respect to the boundary reference. It is also noted that because a linear time normalization is done before the time

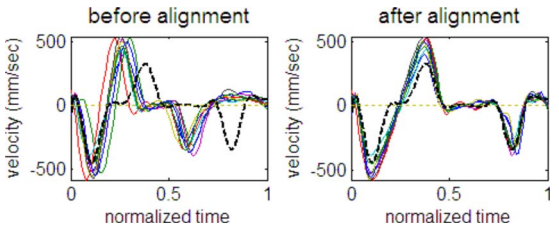


FIG. 4. (Color online) (left) Control velocity trajectories (i.e., no-boundary condition, [dodd deb]) before alignment to the averaged test (i.e., boundary condition, [dodd#deb]) velocity trajectory (dashed line) as reference. (right) After alignment. It can be observed that the velocity trajectories of the control no-boundary signals are well aligned with and expanded relative to the boundary reference.

alignment, the resulting time warping function reflects nonlinear, local timing modulations in tongue tip movement due to the presence of the phrasal boundaries.

Example plots of signals before and after time registration and corresponding time warping functions are shown in Fig. 4 for [Dodd#deb] repetitions by subject K. The magnitude of the negative time deformation functions that will be presented in Sec. III can be interpreted as the amount of clock time of no-boundary signals that needs to be expanded or slowed to match the clock time of the boundary pattern. Equivalently, if absolute value is taken, each time deformation function represents the relative amount of clock slowing due to the presence of the phrase boundary.

One should note that because end points for this analysis are anchored or "pinned" at the edges of the interval of interest and the two end points correspond to the time points of the same articulatory events common in both control (without boundary) and test (with boundary) signals, the time deformation function represents the nonlinear gestural execution timing difference interior to the two end points. Therefore, the area under the time deformation curve is the measure of the amount of nonlinear timing deformation between no boundary and boundary utterances.

III. RESULTS

A. Constriction formation interval

A main effect of boundary on the constriction formation duration is obtained for all the subjects: speaker J [$F(1,24)=64.1$, $p<0.0001$], speaker K [$F(1,22)=35.2$, $p<0.0001$], and speaker N [$F(1,21)=180.6$, $p<0.0001$]. The constriction formation interval is longer in the presence of a phrase boundary. Figure 5 shows the constriction formation duration for the different boundary and sequence conditions split by speaker. All means and standard deviations for all variables are given in the Appendix. The independent variable of sequence type has a significant effect on constriction formation only for speaker N [$F(2,21)=18.0$, $p<0.0001$]. According to a PLSD *post hoc* test, speaker N's constriction formation for [dz] is different from [dd] and [db] ($p<0.0001$), so that [dz] presents a longer duration; [dd] and [db] are not significantly different from each other. Speakers J and K do not show a significant effect of sequence type. The interaction between boundary and sequence is not statistically significant for any speaker, but for speaker J, there was a trend toward an interaction ($p=0.046$), as is seen in Fig. 5.

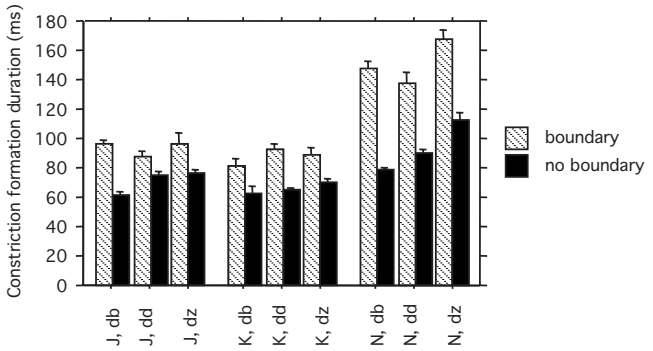


FIG. 5. Boundary and sequence effects on the constriction formation duration (means and standard deviation error bars).

B. Plateau duration

All speakers show a main effect of boundary on the plateau duration: speaker J [$F(1,24)=395.6$, $p<0.0001$], speaker K [$F(1,22)=61.2$, $p<0.0001$], and speaker N [$F(1,21)=87.1$, $p<0.0001$]. The plateau is longer in the boundary condition. The type of sequence shows a main significant effect only for speaker K [$F(2,22)=5.1$, $p=0.001$]. These data are shown in Fig. 6. The results from a PLSD *post hoc* test show that for speaker K, the plateau duration for [dd] is longer than for [db] and [dz] at a significance level of $p<0.013$. [db] and [dz] are not significantly different from each other. As for the apparent interaction between sequence and boundary, it failed to reach significance. Figure 6 shows each speaker's plateau duration for the different boundary and sequence conditions.

Speaker J shows a large mean and standard deviation for the [d#z] and [d#d] boundary condition. As noted above, J produced this set of stimuli with two distinct tongue tip gestures, unlike the two other speakers. This fact accounts for the exceptionally long plateaus for these particular sequences. Also, note that all speakers display a high standard deviation for the three sequences in the boundary condition. This indicates that the plateau duration across a boundary shows more variability than when it falls within the same phrase. The scattergram in Fig. 7 shows the distribution of plateau durations for all the speakers pooled. Byrd *et al.* (2000) also found greater variability in intergestural timing

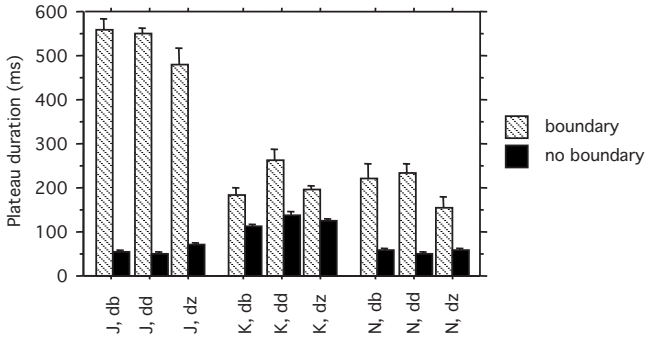


FIG. 6. Boundary and sequence effects on the plateau duration (means and standard deviation error bars).

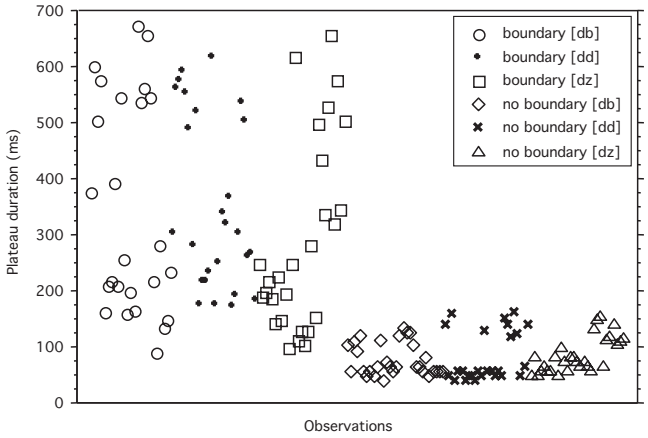


FIG. 7. Scattergram showing the distribution of plateau durations (all speakers pooled).

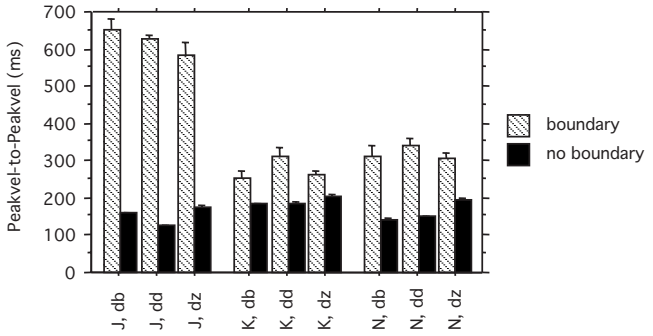


FIG. 8. Boundary and sequence effects on peakvel-to-peakvel duration (means and standard deviation error bars).

under the boundary condition. They interpret this as an indication that intergestural timing is less constrained when the consonants are in separate phrasal domains.

C. Peak velocity to peak velocity

All speakers show a main effect of boundary on the peakvel-to-peakvel interval duration: speaker J [$F(1,24)=470.7$, $p<0.0001$], speaker K [$F(1,22)=48.3$, $p<0.0001$], and speaker N [$F(1,21)=158.9$, $p<0.0001$]. The peakvel-to-peakvel duration is longer in the boundary condition. The type of sequence does not have a significant effect for any of the speakers. The interaction of sequence and boundary was significant for the three speakers: speaker J [$F(2,24)=114.2$, $p=0.001$], speaker K [$F(2,22)=5.99$, $p<0.008$], and speaker N [$F(2,21)=6.54$, $p<0.006$]. This magnitude interaction is such that the peakvel-to-peakvel interval for [dd] is more lengthened by the presence of a boundary than that for [dz] and [db]. Figure 8 shows each speaker's peakvel-to-peakvel duration for the different boundary and sequence conditions. The Appendix reports the means and standard deviations.

D. Timing difference patterns examined by the FDA time registration

In Fig. 9, time deformation functions are shown for [dodd#deb] and for [dodd#zeb], respectively, for speaker K. In Fig. 10, they are shown for speaker N.

One can clearly observe the detailed patterns of slowing of the control utterances relative to the utterances having a phrase boundary, as indicated by the negative time deformation functions. Although there are differences in the amount of time deformation among repetitions due to the inherent noise associated with the control system, the patterns are fairly similar across repetitions and across speakers. Generally, the temporal modifications of the TT velocity trajectories due to the presence of a phrase boundary increase in magnitude over time. This can be seen by the skew of the deformation functions, indicating greater slowing later or closer to the phrase edge (see also Lee *et al.* 2006). (Recall that the final end point is fixed and the function after the maximal negative inflection is not informative.) This finding of progressive nonlinear lengthening supports the predictions of the π-gesture model of boundary-adjacent slowing (Byrd and Saltzman, 2003). Also, note that in some but not all instances, there are two pulses of slowing, suggesting that the activation function of the π-gesture may not be simply smoothly/monotonically rising. This also demonstrates that FDA can provide a view of underlying structure in speech articulation that may not be available from the articulatory kinematic trajectories alone, which generally showed a single smooth constriction interval in the juncture geminates.

IV. DISCUSSION

The data presented above confirm the predictions of the overlap account of gestural aggregation. According to this account, aggregated gestures should display patterns for constriction formation durations similar to single productions but show longer plateau duration than noncoproduced gestures. Our results are in accordance with these predictions. We find no difference for [d]s in [dd] and [dz] as compared to the singleton [d] in [db] in constriction duration. Speaker N did show an exceptional effect of sequence but this was limited to [dz] being longer. However, even for this speaker, the juncture geminate and the singleton [d] have similar constriction formation durations. Thus, in line with previous studies, we find that gestural aggregation is not a summation process but rather the result of temporal overlap with blending. Furthermore, in Byrd *et al.* (2006), we examined the plateau duration for just the *singleton* [d] in the [db] sequence and found that aggregated [dd] gestures show longer plateau duration than that of the noncoproduced singleton [d] gesture, also in accordance with the predictions of the overlap account of gestural aggregation. Note, however, that the present experiment cannot definitely speak to whether the combined activation would increase the magnitude of the lingual gesture differences in the single [d] productions and the abutted [d] realizations since the hard palate limits the observed movement amplitude. A summation process predicts that gestural amplitude should be greater in coproduced contexts. However, based on our findings and Munhall and Löfqvist's (1992), one might expect to find similar movement amplitude for comparable singleton and abutted gestures.

The results reported here shed light on effects of phrase structure on gestural aggregation. All subjects show blended

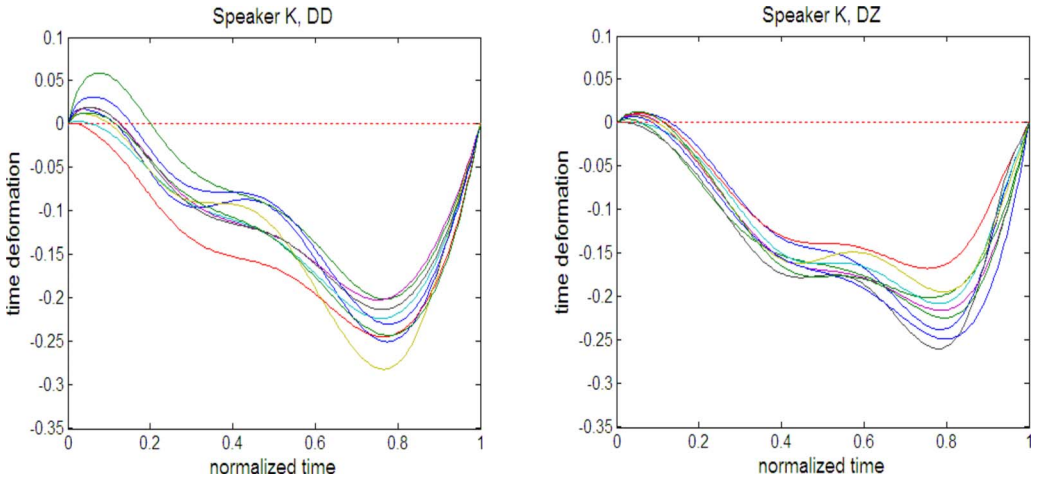


FIG. 9. (Color online) Time deformation functions of (left) [dodd#deb] and (right) [dodd#zeb] for subject K. Note that each line represents a single production of the signal.

or aggregated tongue tip constriction gestures in our [dd] and [dz] phrase medial data. When these sequences span a phrase boundary, we see them "pulled apart" as indicated by our measures of overlap. For some, they remain aggregated in that they have only a single-peaked smooth trajectory; for one subject, however, we see deaggregation sufficient to result in two peaks in the trajectory. The π-gesture model provides a mechanism, we suggest, that may be able to account for these two qualitatively different patterns as arising from a single underlying control mechanism. Byrd and Saltzman (2003) have shown that longer durations and lesser overlap result from the local slowing at phrase boundaries engendered by a π-gesture. Here, our data suggest that phrase boundary effects on juncture geminates are comparable to those on singleton gestures, as predicted within this framework since a π-gesture slows gestural activations of all gestures with which it is concurrently active. The experiment presented here demonstrates that aggregated or blended gestures spanning two phrasal domains result in longer constriction formations, longer plateaus, and longer peakvel-to-peakvel intervals than in cases when the juncture geminates are phrase medial. These longer constriction formations, plateaus, and peakvel-to-peakvel intervals in juncture geminates can be straightforwardly understood as the result of intragestural lengthening and lesser intergestural overlap driven by prosodic structure.

Moreover, the effect of boundary on deaggregation can sometimes result in two-peak productions for juncture geminates, as found for speaker J. This suggests that prosodically driven modulation of gestural overlap (Byrd and Saltzman, 2003) can result not only in quantitative but also qualitative differences in output articulatory kinematics. The presence of a phrase boundary can even lead to total deaggregation of two gestures involved in the production of a juncture geminate. Further data and modeling will be necessary to explore these relations between π-gesture and deaggregation, but it is possible that the interspeaker differences we have observed could be understood as the result of different degrees of activation strength of the π-gesture depending on the particular boundary strength implementation of each speaker. This would, in turn, result in different degrees of overlap, with very minimal overlap yielding in the deaggregation pattern.

Finally, the FDA allows us to examine the time course of the phrasal slowing hypothesized to be driven by a π-gesture. These functions indicate that consistently across subjects the amount of slowing waxes as the boundary ap-

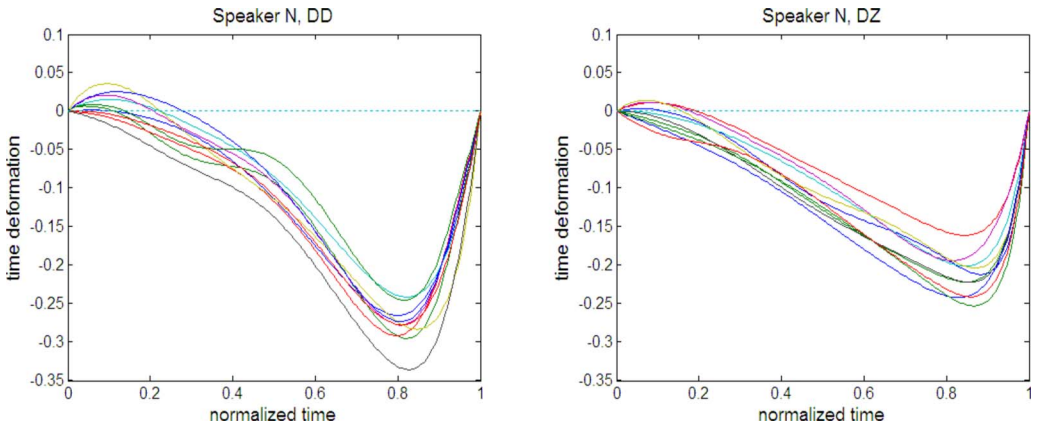


FIG. 10. (Color online) Time deformation functions of (left) [dodd#deb] and (right) [dodd#zeb] for subject N. Note that each line represents a single production of the signal.

proaches, as predicted by Byrd and Saltzman (2003). It also presents a novel way to examine local nonlinear perturbations to the temporal structuring of speech that can offer insight beyond the traditional piecewise, durational-comparison approaches. In particular, it is intriguing to see that the temporal deformations may not be smoothly distributed across the entirety of the constrictions preceding and following the boundary. If this preliminary result is supported in further study, it will serve to illuminate the optimal control characterization for prosodic clock-slowing mechanisms such as the π-gesture.

To conclude, this contribution of this experiment is threefold. First, we provide further evidence supporting the overlap account of gestural aggregation. Second, we analyze prosodic effects on juncture geminates, showing that phrase boundaries effect juncture geminates in comparable ways to their effect on singleton gestures and that gestural deaggregation can occur across a phrase boundary due to a sufficient prosodically driven decrease in overlap. Lastly, we present FDA time deformation functions that are consistent with predictions made by the π-gesture model of phrasal lengthening and which provide insight beyond the traditional piecewise comparison of kinematic durations.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NIH grant No. DC03172 and of Haskins Laboratories, where these data were collected. They also thank Gary Weismer and an anonymous reviewer.

APPENDIX: MEANS AND STANDARD DEVIATIONS (MS)

[b=boundary; nb=no boundary]

Constriction formation duration

		[db]		[dd]		[dz]	
		b	nb	b	nb	b	nb
Speaker J	Mean	96	61.6	87.1	75.4	96.8	76
	Std. dev.	9.9	8.5	14.1	4.3	20.1	9.4
Speaker N	Mean	148	78.4	138	89.6	167	112.8
	Std. dev.	12.1	6.3	21.3	10.5	20.7	13.3
Speaker K	Mean	80.5	63	92.2	64.7	89.3	69.8
	Std. dev.	10.3	15.9	12.6	6.4	14.7	8.6

Plateau duration

		[db]		[dd]		[dz]	
		b	nb	b	nb	b	nb
Speaker J	Mean	556.8	56	549.3	51.4	480.8	71.2
	Std. dev.	82.7	5.3	41.8	4.3	119.7	12.2
Speaker N	Mean	221	60	235	49.6	156	60
	Std. dev.	88.2	11.5	54.4	8.3	69.3	12.1

Plateau duration

		[db]		[dd]		[dz]	
		b	nb	b	nb	b	nb
Speaker K	Mean	184	113.8	264.2	139.3	193.8	124.9
	Std. dev.	46	12.5	67.7	15.3	34.5	17.7

Peakvel-to-peakvel duration

		[db]		[dd]		[dz]	
		b	nb	b	nb	b	nb
Speaker J	Mean	652.8	159.2	624.9	124.6	581.6	175.2
	Std. dev.	80.4	10.3	41.8	6.3	107.5	11
Speaker N	Mean	309	140.8	341	148.8	307	194.4
	Std. dev.	91.7	10.8	53.1	8.6	39.4	16.9
Speaker K	Mean	252.6	183.2	311.1	185.8	260.7	202
	Std. dev.	46.1	8.4	69	16.4	30	16.5

[1]These threshold values allowed for clear and consistent results within and between subjects. Different values for onsets and plateaus were used simply because the kinematic characteristics at these points in the trajectories differed and because it was important that small velocity fluctuations during the plateaus be captured within the plateau area.

[2]It is possible that plateau duration might behave differently for blended single-articulator gestures than for a sequence of gestures using different articulators, and for this reason, we examine separately it in addition to the fairly common measure of overlap in sequences of peakvel-to-peakvel duration.

[3]It should be noted that we have chosen to align the control *no-boundary* intervals to the test *phrase boundary* reference intervals. This particular directional choice was made to reflect that our interest is in the *lengthening* effect due to the presence of a phrase boundary, not shortening due to the absence of the boundary. This also yields a more robust performance of FDA alignment algorithm due to the expansion or interpolation nature of the alignment direction. (Note that this choice differs from that made by Lee *et al.* 2006 but does not represent any fundamental difference in method.)

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.

Browman, C., and Goldstein, L. (**1992**). "Articulatory Phonology: An Overview," Phonetica **49**, 155–180.

Byrd, D. (**1995**). "Articulatory characteristics of single and blended lingual gestures," in Proceedings of the XIIIth International Congress of Phonetic Sciences, edited by K. Elenius, and P. Branderud, pp. 438–441.

Byrd, D., Campos-Astorkiza, R., and Shepherd, M. (**2006**). "Gestural deaggregation via prosodic structure," Proceedings of the Seventh International Seminar on Speech Production, Ubatuba, Brazil.

Byrd, D., and Choi, S. (in press). "At the juncture of prosody, phonology, and phonetics—The interaction of phrasal and syllable structure in shaping the timing of consonant gestures," Papers in Laboratory Phonology 10, Mouton.

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, Cambridge, pp. 70–87.

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.
Byrd, D., and Tan, C. C. (**1996**). "Saying consonant clusters quickly," J. Phonetics **24**, 263–282.
Cho, T. (**2005**). "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, New York).
Cho, T., and Jun, S. A. (**2000**). "Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean," Chicago Linguistics Society **36**, 31–44.
Cho, T., and Keating, P. (**2001**). "Articulatory and acoustic studies on domain-initial strengthening in Korean," J. Phonetics **29**, 155–190.
Edwards, J., Beckman, M. E., and Fletcher, J. (**1991**). "The articulatory kinematics of final lengthening," J. Acoust. Soc. Am. **89**, 369–382.
Fougeron, C., and Keating, P. (**1997**). "Articulatory strengthening at edges of prosodic domains," J. Acoust. Soc. Am. **101**, 3728–3740.
Hardcastle, W. J. (**1985**). "Some phonetic and syntactic constraints on lingual coarticulation during /kl/sequences," Speech Commun. **4**, 247–263.
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.
Kelso, J. A. S., and Tuller, B. (**1987**). "Intrinsic timing in speech production: Theory, methodology, and preliminary observations," in *Sensory and Motor Processes in Language*, edited by Keller and Gopnik (Erlbaum, Hillsdale, NJ), pp. 203–222.
Klatt, D. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.
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.
Lee, S., Byrd, D., and Krivokapić, J. (**2006**). "Functional data analysis of prosodic effects on articulatory timing," J. Acoust. Soc. Am. **119**, 1666–1671.
Löfqvist, A., and Yoshioka, H. (**1981**). "Laryngeal activity in Icelandic obstruent production," Nordic Journal of Linguistics **4**, 1–18.
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.
Max, L., and Onghena, P. (**1999**). "Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research," J. Speech Lang. Hear. Res. **42**, 261–270.
McClean, M. (**1973**). "Forward coarticulation of velar movement at marked junctural boundaries," J. Speech Lang. Hear. Res. **16**, 286–296.
Munhall, K. G., and Löfqvist, A. (**1992**). "Gestural aggregation in speech: laryngeal gestures," J. Phonetics **20**, 93–110.
Nelson, W. L. (**1983**). "Physical principles for economies of skilled movements," Biol. Cybern. **46**, 135–147.
Oller, K. D. (**1973**) "The effect of position in utterance on speech segment duration in English," J. Acoust. Soc. Am. **54**, 1235–1247.
Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M. (**1992**). "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096.
Ramsay, J. O. (**2007**). ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/ last accessed 5/6/2008.
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. (**2005**). *Functional Data Analysis*, 2nd ed. (Springer-Verlag, New York).
Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecological Psychol. **1**, 333–382.
Tabain, M. (**2003**). "Effects of prosodic boundary on /aC/ sequences: articulatory results." J. Acoust. Soc. Am. **113**, 2834–2849.
Vaxelaire, B. (**1995**). "Single vs. double (abutted) consonants across speech rate: X-ray and acoustic data from French," in Proceedings of the XIIIth International Congress of Phonetic Sciences, edited by K. Elenius, and P. Branderud, pp. 384–387.
Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (**1992**). "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**, 1707–1717.

A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/

Xinhui Zhou[a] and Carol Y. Espy-Wilson[b]
Speech Communication Laboratory, Institute of Systems Research, and Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742

Suzanne Boyce[c]
Department of Communication Sciences and Disorders, University of Cincinnati, Mail Location 0394, Cincinnati, Ohio 45267

Mark Tiede[d]
Haskins Laboratories, 300 George Street Suite 900, New Haven, Connecticut 06511

Christy Holland[e]
Department of Biomedical Engineering, Medical Science Building 6167, Mail Location 0586, University of Cincinnati, Cincinnati, Ohio 45267

Ann Choe[f]
Department of Radiology, University Hospital G087C, Mail Location 0761, University of Cincinnati Medical School, Cincinnati, Ohio 45267

(Received 16 April 2007; revised 26 February 2008; accepted 29 February 2008)

Speakers of rhotic dialects of North American English show a range of different tongue configurations for /r/. These variants produce acoustic profiles that are indistinguishable for the first three formants [Delattre, P., and Freeman, D. C., (1968). "A dialect study of American English r's by x-ray motion picture," Linguistics **44**, 28–69; Westbury, J. R. *et al.* (1998), "Differences among speakers in lingual articulation for American English /r/," Speech Commun. **26**, 203–206]. It is puzzling why this should be so, given the very different vocal tract configurations involved. In this paper, two subjects whose productions of "retroflex" /r/ and "bunched" /r/ show similar patterns of F1–F3 but very different spacing between F4 and F5 are contrasted. Using finite element analysis and area functions based on magnetic resonance images of the vocal tract for sustained productions, the results of computer vocal tract models are compared to actual speech recordings. In particular, formant-cavity affiliations are explored using formant sensitivity functions and vocal tract simple-tube models. The difference in F4/F5 patterns between the subjects is confirmed for several additional subjects with retroflex and bunched vocal tract configurations. The results suggest that the F4/F5 differences between the variants can be largely explained by differences in whether the long cavity behind the palatal constriction acts as a half- or a quarter-wavelength resonator.
 [DOI: 10.1121/1.2902168]

PACS number(s): 43.70.Bk, 43.70.Fq, 43.70.Aj [BHS] Pages: 4466–4481

I. INTRODUCTION

It is well known that different speakers may use very different tongue configurations for producing the rhotic /r/ sound of American English (Delattre and Freeman, 1968; Hagiwara, 1995; Alwan *et al.*, 1997; Westbury *et al.*, 1998; Espy-Wilson *et al.*, 2000; Tiede *et al.*, 2004). While the picture of variability in tongue shape is complex, it is generally agreed that two shapes, in particular, exhibit the greatest degree of contrast: "retroflex" /r/ (produced with a raised tongue tip and a lowered tongue dorsum) and "bunched" /r/ (produced with a lowered tongue tip and a raised tongue dorsum). Figure 1 shows examples of these shapes drawn from our own studies of two different speakers producing their natural sustained /r/ (as in "pour"). Similar examples of this contrast may be found from Delattre and Freeman (1968) and Shriberg and Kent (1982). These examples are typical in showing three supraglottal constrictions along the vocal tract: a narrowing in the pharynx, a constriction along the palatal vault, and a constriction at the lips. However, the locations of constrictions and the degrees and lengths of constriction significantly differ, especially along the palate. At first glance, the degree of difference between the two configuration types for /r/ appears to be similar to that between, say, /s/ and /ʃ/ or /i/ and the unrounded central vowel /ɨ/. Thus, it might be expected that the two types of /r/ would show clear acoustic and perceptual differences. However, the question of an acoustic correlation between formant frequencies and tongue shape was investigated by Delattre and Freeman (1968) and, more recently, by Westbury *et al.* (1998).

[a]Electronic mail: zxinhui@glue.umd.edu
[b]Electronic mail: espy@glue.umd.edu
[c]Electronic mail: boycese@uc.edu
[d]Electronic mail: tiede@haskins.yale.edu
[e]Electronic mail: Christy.Holland@uc.edu
[f]Electronic mail: Ann.Choe@uc.edu

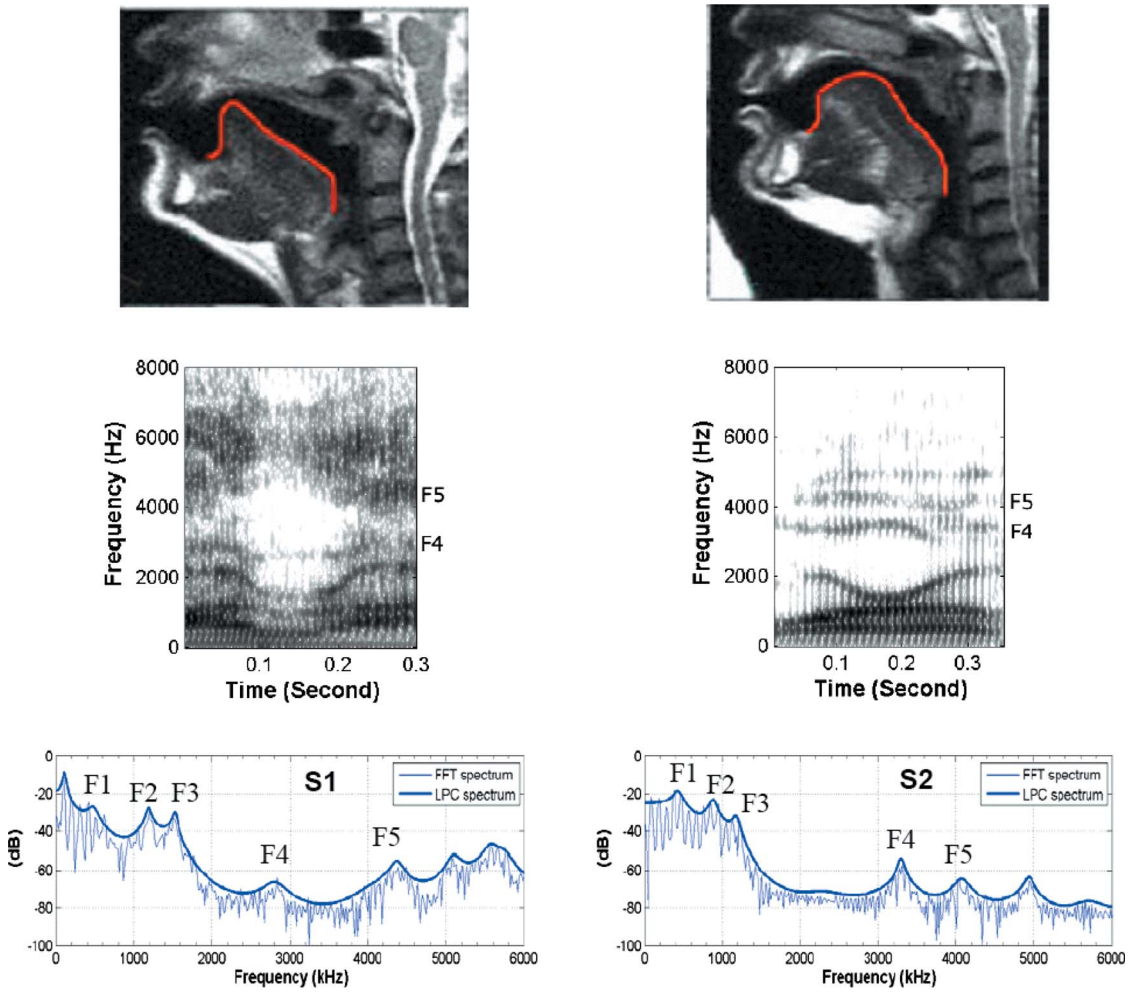


FIG. 1. (Color online) Top panel: Midsagittal MR images of two tongue configurations for American English /r/. Middle panel: Spectrograms for nonsense word "warav." Lower panel: Spectra of sustained /r/ utterance. The left side is for S1 and the right side is for S2.

Interestingly, no consistent pattern was found. In a recent perceptual study, Twist *et al.* (2007) found that listeners also appear to be insensitive to the difference between retroflex and bunched /r/.

American English /r/ is characterized by a lowered third formant frequency (F3) sitting in the region between 60% and 80% of average vowel F3 (Hagiwara, 1995) and often approaching F2 (see Lehiste, 1964; Dalston, 1975; Espy-Wilson, 1987). This low F3 is the most salient aspect of the acoustic profile of /r/ (Lisker, 1957; O'Connor *et al.*, 1957). F1 and F2 typically cluster in the central range of a particular speaker's vowel space, consistent with the common symbol of hooked schwa (or schwar) for /r/ when it acts as a syllabic nucleus.

As noted above, previous attempts have failed to find a correlation between formant frequency values and tongue shapes for /r/. However, these previous studies focused on the first three formants, F1–F3. In recent years, Espy-Wilson *et al.* have suggested that the higher formants may contain cues to tongue configuration and vocal tract dimensions (Espy-Wilson and Boyce, 1999; Espy-Wilson, 2004). Typically, researchers have not looked at higher formants such as F4 and F5 because their lower amplitude in the spectrum can make them difficult to identify and measure. In addition, the process of speech perception appears to largely depend on the pattern of the first three formants. However, higher formants are particularly responsive to smaller cavities in the vocal tract (e.g., piriform sinuses, sublingual spaces, the laryngeal cavity), and thus may give more detailed information regarding the vocal tract shape. Such knowledge may contribute to human speech perception and speaker identification to some extent. In addition, detailed knowledge of the vocal tract shape from acoustics is desirable for automatic speech and speaker recognition purposes.

In this paper, we investigate a case of two subjects with similar vocal tract anatomy who produce very different bunched and retroflex tongue shapes for /r/. These are the subjects shown in Fig. 1. As the middle panel of the figure shows, the subjects' acoustic profiles resemble those discussed by Delattre and Freeman (1968) and Westbury *et al.* (1998) in that their F1–F3 values are similar. However, the two subjects also show very different patterns for F4 and F5. In particular, the distance between F4 and F5 for the retroflex /r/ is double that of the bunched /r/. The lower panel of Fig. 1 shows examples of the same F4/F5 pattern drawn from running speech, this time from production of the nonsense word /warav/. In this paper, we investigate the question of whether different patterns of the higher formants are a con-

sistent feature of bunched versus retroflex tongue shape. If so, this difference in acoustic signatures may be useful for a number of purposes that involve the mapping between articulation and acoustics, i.e., speaker recognition, articulatory training, speech synthesis, etc. Alternatively, the different patterns of F4 and F5 may derive from structures independent of tongue shape, for instance, additional cavities in the vocal tract such as the laryngeal vestibule (Kitamura *et al.*, 2006; Takemoto *et al.*, 2006a) or the piriform sinuses (Dang and Honda, 1997). The key piece of evidence is whether such structures differ in such a way as to explain the F4/F5 patterns across /r/ types.

In this paper, we approach the task of understanding this difference in formant pattern in the following way. First, magnetic resonance imaging (MRI) is used to acquire a detailed three-dimensional (3D) geometric reconstruction of the vocal tract. Second, we used the finite element method (FEM) to simulate the acoustic response of the 3D vocal tract and to study wave propagation properties at different frequencies. Third, we derive area function models from the FEM analysis of our 3D geometry. The resulting simulated acoustic response is verified against the 3D acoustic response. The area function models are then used to isolate the effects of formant-cavity affiliations. The results of the simulation are compared to actual formant values from the subjects.

II. MATERIALS AND METHODOLOGIES

A. Subjects

The data discussed in this paper were obtained as part of a larger study on the variety of tongue shapes in productions of American English /r/ and /l/. For the purposes of this paper, we concentrate on /r/ data from two native speakers of American English, referred to here as S1 and S2.[1] As Fig. 1 shows, S1 produces a retroflex /r/ and S2 produces a bunched /r/. Both subjects are male. S1 was 48 years old and S2 was 51 at the time the data were collected. S1 had lived in California, Minnesota, and Connecticut and S2 had lived in Texas, Massachusetts, and Southwestern Ohio. Both spoke a rhotic dialect of American English.[2] The subjects were similar in palate length, palate volume, overall stature, and vocal tract length (see Table I).[3] We also compare the data from S1 and S2 to that from other subjects with similar retroflex or bunched tongue shapes for /r/ collected in the larger study. These subjects are referred to as S3–S6. The articulatory data collected for all subjects include MRI scans of the vocal tract for sustained natural /r/, dental cast measurements, computed tomography (CT) scans of the dental casts, and acoustic recordings made at various points in time.

B. Image acquisitions

MR imaging was performed on a 1.5 T G.E. Echospeed MR scanner with a standard phased array neurovascular coil at the University Hospital of the University of Cincinnati, OH. Subjects were positioned in supine posture, with their heads supported by foam padding to minimize movement. The subjects were instructed to remain motionless to the extent possible during and between scans. For hearing protection and comfort, subjects wore earplugs during the entire session. In addition, the subjects' ears were covered by padded earphones.

TABLE I. Dimension sizes of S1 and S2 in overall height, and volume, length, depth, and width of the palate. The measurements of the palate are based on the dental casts of the subjects. The width of the palate is the distance between edges of the gum between the second premolar and the first molar on both sides of the upper jaw. The length of the palate is the distance of the edges of the gum between the upper middle two incisors and the cross section of the posterior edge of the back teeth. The depth of the palate is the distance from the floor of the mouth to the cross section with the lateral plane. The volume of the palate is the space surrounded by the margin between the teeth and gums, the posterior edge of the back teeth, and the lateral plane. We used several techniques to calculate the volume, all of which gave the same answer within a certain range, and the average volume as a matter of displacement in water is reported here. That measure was done three times.

	S1	S2
Height of subject	188 cm	188 cm
Length of palate	35.8 mm	33.6 mm
Depth of palate	16.1 mm	13.2 mm
Width of palate	25.5 mm	25.0 mm
Av. volume of palate	29.1 mm^3	29.1 mm^3
Maxillary teeth volume	3.4 mm^3	3.3 mm^3

Localization scans were performed in multiple planes to determine the optimal obliquities for orthogonal imaging. A midsagittal plane was identified from the brain morphology. Axial and coronal planes were then oriented to this midsagittal plane. During each subsequent scan, the subject was instructed to produce sustained /r/ as in "pour" for a defined period of time (between 5 and 25 s depending on the sequence). T2 weighted 5 mm single shot fast spin echo images were obtained in the midline sagittal plane with two parasagittal slices. T1 weighted fast multiplanar spoiled gradient echo images (repetition time (TR) of 100–120 ms, echo delay time (TE) of 4.2 ms, 75° flip angle) were obtained in the coronal and axial planes with a 5 mm slice thickness. There was no gap between adjacent slices. The scanning regions for the coronal and axial planes include the region from the surface of the vocal folds to the velopharyngeal port and the region from the rear wall of the velopharynx to the outside edge of the lips. Depending on the dimensions of the subjects' vocal tract, the data set comprised 24–33 images in the axial and coronal planes. For all images, the field of view was 240×240 mm^2 with an imaging matrix of 256×256 to yield an in-plane resolution of 0.938 mm per pixel.

The MR imaging technique we used does not distinguish between bony structures such as teeth and air due to the low levels of imageable hydrogen. Thus, to avoid overestimation of oral tract air space, CT scans of each subject's dental cast were acquired on a GE Lightspeed Ultra multidetector scanner with a slice thickness of 1.25 mm, subsequently superimposed on the volumes derived from MRI as described below. Images were resampled to 1.25 mm at 0.625 mm intervals to optimize 3D modeling. The field of view was 120 mm with an imaging matrix of 512×512 to yield an in-plane image resolution of 0.234 mm per pixel.

C. Acoustic signal recording

During the MRI sessions, the subject's phonation in the supine position was recorded using a custom-designed microphone system (Resonance Technology Inc.) and continuously monitored by a trained phonetician to ensure that the production of /r/ remained consistent over the course of the experiment. Subjects were instructed to begin phonation before the onset of scanning and to continue to phonate for a period after scanning was complete. A full audio record of the session was preserved using a portable DAT tape recorder (SONY TD-800). Due to the noise emitted by the scanner during the scans, the only portions of the subject's productions of /r/ that can be reliably analyzed occur in 500 ms after phonation began, before the scanner noise commenced, and in 500 ms after the scanner noise ceased while the subject continued to phonate. The recordings are still quite noisy, but it was possible to measure F1–F3 with reasonable accuracy during most scans.

Subjects were also recorded acoustically in separate sessions in a sound-treated room by using a Sennheiser headset microphone and a portable DAT tape recorder (SONY TD-800). Subjects recorded a set of utterances encompassing sustained productions of /r/ plus a number of real and nonsense words containing /r/. As in the MR condition, subjects were instructed to produce /r/ as in "pour." In addition, they recorded sustained /r/ as in "right," "reed," and "role." For the sustained productions, subjects were recorded in both upright and supine postures. The nonsense words were "warav," "wadrav," "wavrav," and "wagrav," repeated with stress either on the first syllable or the second syllable. The real words included /r/ in word-initial, word-final, and intervocalic positions. For the real and nonsense words, subjects were recorded in the upright posture. Acoustic data recorded in the sound-proofed room are referred to as sound booth acoustic data. Recording conditions were such that, in addition to F1–F3, F4, and F5 could be reliably measured.

D. Image processing and 3D vocal tract reconstruction

We used the software package MIMICS (Materialise, 2007) to obtain a 3D reconstruction of the vocal tract. This software has been widely employed in the medical imaging field for processing MRI and CT images, for rapid prototyping, and for 3D reconstruction in surgery.

Our reconstruction proceeded in four steps. Step (1) involved segmentation between the tissue of the vocal tract and the air space inside the vocal tract for each MR image slice in the coronal and axial sets. Because the cross section of the oral cavity is best represented by the coronal slice images, and the cross section of the pharyngeal and laryngeal cavities are best represented by the axial slices, we used the following procedure to weight them. First, the segmented axial slices were transformed into a 3D model. Then, the coronal slices were overlapped with the axial-derived model. As in the study by Takemoto *et al.* (2006b), we extended the cross-sectional area of the last lip slice with a closed boundary halfway to the last slice in which the upper and lower lips are still visible. The coronal slice segmentation in the pharyngeal and laryngeal cavities was then corrected by reference to the axial slice 3D model.

Step (2) involved compensation for the volume of the teeth using the CT scans, which were made in the coronal plane. The CT images were segmented to provide a 3D reconstruction of the mandible and the maxillae with the teeth. (This process was considerably easier than for the MR slices described above, given the straightforward nature of the air/tissue boundary in that imaging modality.) The 3D reconstruction of the dental cast was then overlapped with the MRI coronal slices. The reconstruction of the maxilla cast was positioned on the MR images by following the curvature of the palate. The reconstruction of the mandible cast was positioned with reference to the boundary provided by the lips. In step (3), the final segmentation was translated into a surface model in stereolithography (STL) format (Lee, 1999). Finally, the 3D geometry surface was smoothed using the MAGICS software package (Materialise, 2007). The validity of the reconstructed 3D vocal tract geometry was evaluated by comparing midsagittal slices created from the reconstructed 3D geometry to the original midsagittal MR images. We also used this method to check for the possibility that subjects had changed their vocal tract configuration for /r/ across scans. The data sets of all the subjects in this study show very good consistency, and overall boundary continuity between the tissue and the airway was successfully achieved.

As noted above, the difference in the F4/F5 formant pattern between S1 and S2 must be derived from a difference in vocal tract dimensions, either in small structures such as the piriform sinuses and laryngeal vestibule (Dang and Honda, 1997; Kitamura *et al.*, 2006; Takemoto *et al.*, 2006a) or in tongue shape differences. The laryngeal vestibule cavities were included in the 3D model, but given the resolution of the MR data, the representation is relatively crude. The dimensions of the piriform sinuses were measured and found to be similar to the range in length of 16–20 mm and in volume of 2–3 cm^3 reported by Dang and Honda (1997).[4] Because no significant differences were found between the subjects for either structure, we conclude that the tongue shape differences between S1's retroflex and S2's bunched /r/ are likely the major factor determining their differences in the F4/F5 pattern. Possibly, these cavities at the glottal end of the vocal tract are less influential for /r/ than for vowels due to the greater number, length, and narrowness of constrictions involved.

E. 3D finite element analysis

The FEM analysis was used in this study to obtain the acoustic response of the 3D vocal tract and to obtain the wave propagation at different frequencies. The pressure isosurfaces at low frequency were used to extract area functions. The governing equation for this harmonic analysis is the Helmholtz equation,

$$\nabla \cdot \left(\frac{1}{\rho} \nabla p \right) + \frac{\omega^2 p}{\rho c^2} = 0, \quad (1)$$

where p is the acoustic pressure, ρ (1.14 kg/m^3) is the density of air at body temperature, c (350 m/s) is the speed of sound, and ω is the angular frequency ($\omega = 2\pi f$, where f is the vibration frequency in hertz and the highest frequency in our harmonic analysis is 8000 Hz). The boundary conditions for the 3D finite element analysis are as follows: for the glottis, a normal velocity profile as sinusoidal signal at various frequencies; for the wall, rigid; for the lips, the radiation impedance Z of an ideal piston in an infinitely flat baffle (Morse and Ingard, 1968),

$$Z = \rho c(1 - J_1(2ka)/(ka) + jK_1(2ka)/(2ka)), \quad (2)$$

where $k = 2\pi f/c$, $a = \sqrt{A_1/\pi}$ (A_1 is the area of the lips opening), J_1 is the Bessel function of order 1, and K_1 is the Struve function of order 1. The volume velocity at the lips is measured by velocity integration over the cross section at the lips, and the acoustic response of the vocal tract is defined as the volume velocity at the lips divided by the volume velocity at the glottis. Note that for the purpose at hand, the ideal piston model has been shown to be computationally equivalent to a 3D radiation model at the lips (Matsuzaki *et al.*, 1996).

The finite element (FEM) analysis was performed using the COMSOL MULTIPHYSICS package (Comsol, 2007). The mesh for FEM was created using tetrahedral elements as in the STL format.

F. Area function extraction

Area functions are generated by treating the vocal tract as a series of uniform tubes with varying areas and lengths. The extraction of area functions from imaging data is typically an empirical process. Baer *et al.* (1991), Narayanan *et al.* (1997), and Ong and Stone (1998) based their area function extractions on a semipolar grid (Heinz and Stevens, 1964). In contrast, Chiba and Kajiyama (1941), Story *et al.* (1996), and Takemoto *et al.* (2006b) extracted area functions by computing a centerline in air space and then evaluating the cross-sectional areas within planes chosen to be perpendicular to the centerline extending from the glottis to the mouth.

In general, because our area functions were derived from the 3D FEM, it might be expected that the area function simulation and the simulated acoustic response from the 3D model should be the same. However, it should be noted that area function extraction, by transforming the bent 3D geometry of the vocal tract into a straight tube with varying cross-sectional areas (Chiba and Kajiyama, 1941; Fant, 1970), necessarily involves considerable simplification. An additional and related problem is that it assumes planar wave propagation, and thus tends to neglect cross-mode wave propagation and potential antiresonances or zeros. Thus, we expect some small differences between the simulation results using area function analysis and planar wave propagation from simulation results directly obtained from the corresponding 3D geometry (Sondhi, 1986).

In this study, we used the low-frequency wave propagation properties resulting from the 3D finite element analysis to guide the area function extraction from the reconstructed 3D geometry. This approach is quite similar to the centerline approach. The logic of this procedure was as follows. As noted above, area-function-based vocal tract models assume planar wave propagation. Finite element analysis at low frequencies such as 400 Hz (around F1 for /r/) produces pressure isosurfaces that indicate approximate planar acoustic wave propagation. Thus, a tube model derived from area functions whose cutting plane follows these pressure isosurfaces should constitute a reasonable one-dimensional model for the 3D vocal tract. In this study, as the curvature of the vocal tract changes, the cutting orientation in our method was adjusted to be approximately parallel to the pressure isosurface at 400 Hz. This procedure was performed by recording the coordinates of the isosurfaces. Those coordinates are then used to determine the cutting planes. The distance between two sampling planes was set to be the distance between their centroids. The vocal tract length was estimated as the cumulative sum of the distance between the centroids. The cutting plane gap was about 3 mm. Since this method was based on the 3D reconstructed geometry instead of sets of MR images, pixel counting and other manipulations such as reslicing of images were not needed. The area calculation was based on the geometric coordinates of the reconstructed vocal tract.

As noted above, the reduction of a vocal tract 3D model to area functions requires considerable simplification. To assess the degree to which our area function extraction preserved essential aspects of the vocal tract response, we compared the simulation output from the 3D FEM to the acoustic response of Vocal Tract Acoustic Response (VTAR), a frequency-domain computational vocal tract model (Zhou *et al.*, 2004) which takes area functions of the vocal tract as input parameters and includes terms to account for energy losses due to the yielding wall property of the vocal tract, the viscosity and the heat conduction of the air, and the radiation from the lips. The vocal tract response from the 3D model and from VTAR were, in turn, evaluated by comparison with formant measurements from real speech produced by the subjects, as described below.

G. Formant measurement of /r/ acoustic data

Formants from both sound booth and MR acoustic recordings were measured by an automatic procedure that computed 24th order LPC (Linear Prediction Coding) spectrum over a 50 ms window from a stable section of the sustained production. The 50 ms window for the MR acoustic data was taken from the least noisy segment of the approximately 500 ms production preceding the onset of MR scanning noise. Only F1–F3 were measured in the MR acoustic recording because the noise in the high-frequency region masked the higher formants very effectively. Both sets of measurements are shown in Tables III and IV. To maximize the comparability of the MR and sound booth acoustic measures, the latter were measured from productions recorded when the subjects were in supine posture. The formant val-

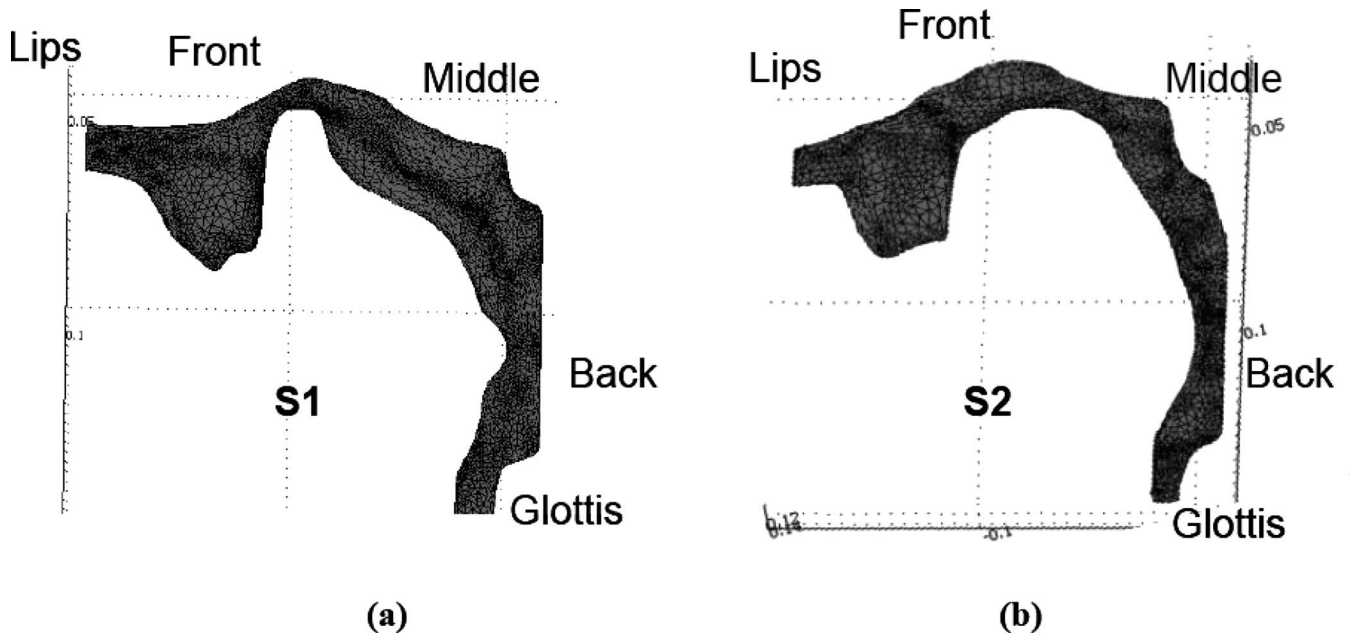


FIG. 2. FEM mesh of the reconstructed 3D vocal tract. (a) The retroflex tongue shape. (b) The bunched tongue shape.

ues of the sustained /r/ in MRI sessions are the average of the measurements from all the scans including midsagittal, axial, and coronal scans.

The difference in F4/F5 pattern between the subjects previously alluded to is clearly shown in Tables III and IV for the sound booth recording of sustained /r/. As Fig. 1 shows, the pattern in question was even more strongly evident in the more dynamic real word condition. We concluded from these data that the patterns shown in sustained /r/ are representative of patterns shown in running speech.

H. Reconstructed 3D vocal tract geometries

The reconstructed 3D vocal tract shapes for the retroflex /r/ of S1 and the bunched /r/ of S2 are shown in Fig. 2. The two shapes are significantly different in several dimensions that are likely to cause differences in cavity affiliations. First, S1's retroflex /r/ has a shorter and more forward palatal constriction, leading to a slightly smaller front cavity. At the same time, the lowered tongue dorsum of the retroflex /r/ leads to a particularly large volume of the midcavity between the palatal and pharyngeal constrictions. Further, the transition between the front and midcavities is sharper for the retroflex /r/. This difference makes it more likely that the front and midcavities are decoupled for the retroflex /r/ of S1 than for the bunched /r/ of S2. Unlike the speakers analyzed by Alwan *et al.* (1997) and Espy-Wilson *et al.* (2000), neither S1 nor S2 shows a sublingual space whose geometry is clearly a side branch to the front cavity. However, the two subjects' overall vocal tract dimensions from the 3D model are very similar. These dimensions are shown in Table II.

I. FEM-based acoustic analysis

In previous work, FEM analysis has been used to study the acoustics of the vocal tract for open vocal tract sounds, i.e., vowels (Thomas, 1986; Miki *et al.*, 1996; Matsuzaki *et al.*, 2000; Motoki, 2002). Zhang *et al.* (2005) applied this approach to a two-dimensional vocal tract for a schematized geometry based on a single subject producing /r/. In this study, we extend the work of Zhang *et al.* (2005) by computing the pressure isosurfaces at various frequencies to 3D vocal tract shapes based on S1's retroflex and S2's bunched /r/. As Fig. 3 shows, the retroflex and bunched /r/ shapes have similar wave propagation. For both, as expected, the wave propagation is almost planar up to about 1000 Hz. Between 1500 and 3500 Hz, a second wave propagates almost vertically to the bottom of the front cavity. Above 4500 Hz, the isosurface becomes more complex and part of the acoustic wave propagates to the two sides of the front cavity. The results show that the wave propagation property should be kept in mind when assuming planar wave propagation along the vocal tract, particularly for antiresonances. Note that for both subjects, F4 and F5 occur in the transition region below 4500 Hz. This will be discussed later. The cutting orientations for the area functions based on the pressure isosurfaces are shown in the upper panel of Fig. 4 as grid lines. The area functions themselves are shown in the lower panel of Fig. 4.

III. RESULTS

The purpose of this study was to determine if the F4/F5 difference in pattern between bunched and retroflex /r/ occurs as a result of tongue shape differences. The approach involves comparing the results of calculations to acoustic

TABLE II. Measurements on the reconstructed 3D vocal tract in surface model (STL file format).

	S1	S2
X dimension	51 mm	46 mm
Y dimension	106 mm	107 mm
Z dimension	106 mm	100 mm
Volume	62 909 mm^3	48 337 mm^3
Surface area	14 394 mm^2	12 243 mm^2

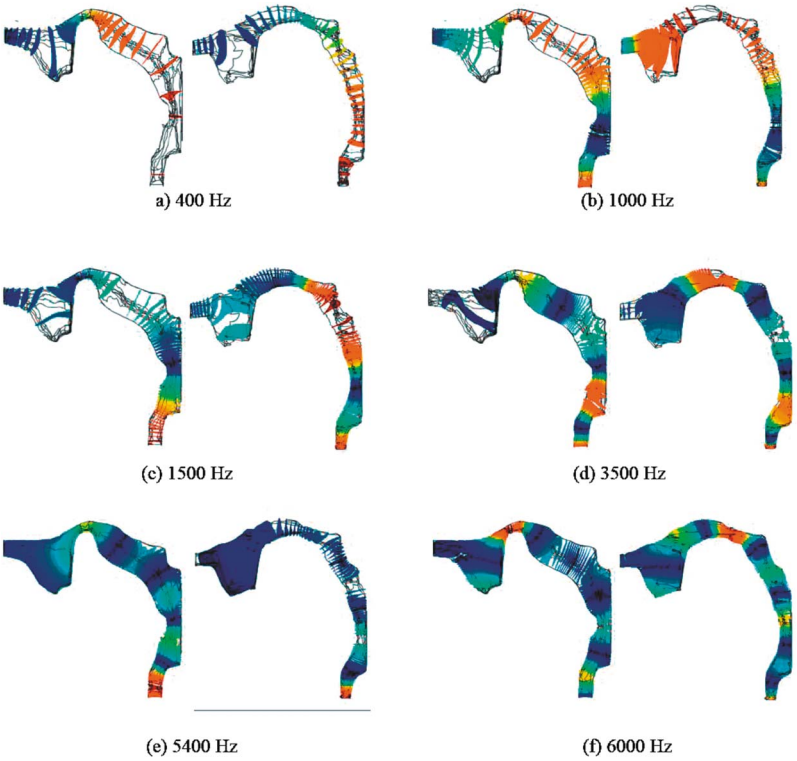


FIG. 3. (Color online) Pressure isosurface plots of wave propagation inside the vocal tracts of the retroflex /r/ (S1 on the right side) and the bunched /r/ (S2 on the right side) at different frequencies. (Pressure isosurfaces are coded by color: the red color stands for high amplitude and the blue color stands for low amplitude.) (a) 400 Hz, (b) 1000 Hz, (c) 1500 Hz, (d) 3500 Hz, (e) 5400 Hz, and (f) 6000 Hz.

spectra from actual productions by the subjects during (a) MR and (b) sound booth acoustic sessions, respectively. The calculated results include (c) generating an acoustic response from the FEM analysis based on the 3D model, (d) generating an acoustic response from the VTAR computational model using FEM-derived area functions, (e) generating sensitivity functions for better understanding of formant-cavity affiliations and manipulating the VTAR computation model to isolate the effects of particular cavities and constrictions, and (f) generating simple-tube models to understand the types of resonators that produce the formants. The FEM analysis makes no assumptions regarding planar wave propagation, whereas the area functions are derived from cutting planes determined by the FEM at low frequency. The isolation of cavity/constriction influences is done by using VTAR to synthesize changes in the dimensions of a particular cavity/constriction while holding the rest of the vocal tract constant. In effect, we compare the acoustic responses from the 3D FEM and the area functions with the subjects' actual production.

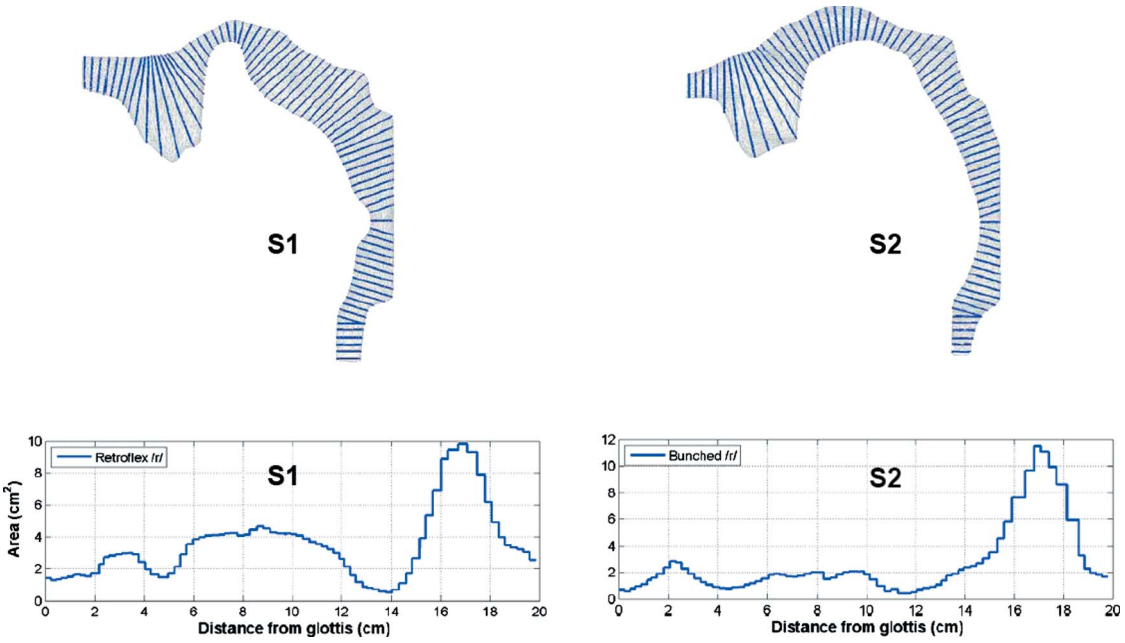


FIG. 4. (Color online) Top panel: Grid lines for area function extraction inside the vocal tract. Lower panel: Area function based on the grid lines. (In each panel, the left side is for S1 and the right side is for S2.)

TABLE III. Formants measured from S1's retroflex /r/ compared with calculated values from the 3D FEM, tube model with area function model, and simple-tube model, respectively (Unit: Hz). The percentage difference between the FEM formant values and the actual subject formant values from MR (Δ1) and sound acoustic (Δ2) sessions are also given. Note that due to background noise, only F1–F3 could be consistently measured from the MRI acoustic data.

	Retroflex /r/ (S1)							
	MRI acoustic data	Second both supine position	Sound booth upright position	3D FEM			Area function tube model	Simple tube model
				Formant	Δ1 (%)	Δ2 (%)		
F1	522	391	438	380	27.2	2.81	383	418
F2	1075	1234	1188	1160	7.91	6.0	1209	1262
F3	1534	1547	1563	1580	3.0	2.13	1609	1660
F4		2797	2828	2940		5.11	3002	2936
F5		4328	4234	4280		1.11	4366	4233
F5-F4		1531	1406	1340			1364	1297

MR versus sound booth acoustic data. Because the FEM analysis and area functions are both based on MR data, the F4/F5 patterns would ideally have been extracted from the simultaneously recorded acoustic signal ("MR acoustic data"). As noted previously, however, F4 and F5 are masked in the MRI condition by the noise of the scanner. Hence, acoustic data recorded in a sound booth (from the supine posture) were used for comparisons with the calculated acoustic response results. Comparison between the MR and sound booth acoustic data for the first three formants show that the subjects' productions are, for the most part, highly similar, as shown in Tables III and IV. There are notable deviations in the F1 and F2 produced by S1 and in the F3 produced by S2. While these differences probably indicate a slight difference in articulatory configuration for sustained /r/, this same alternation between formant values can also be seen in their running speech for both real and nonsense words.[5] In all cases, the characteristic F4/F5 pattern is maintained.

The difference in F4/F5 patterns between the retroflex configuration of S1 and the bunched configuration of S2 is also observed when subjects produce /r/ in the upright posture. This is shown for running speech in Fig. 1. In addition, the formant values from sound booth acoustic sustained productions recorded in the upright posture are reported in Tables III and IV, for comparison to the values recorded in supine posture.

Comparison of actual formants to acoustic response from FEM and area function. In Fig. 5, spectra from the subjects' actual productions are shown along with acoustic responses from the models for S1 and S2. As shown in Figs. 5(a) and 5(c) (in addition to Tables III and IV), the FEM provides formant values for F1–F3 similar to those measured from actual productions in MRI sessions by each speaker. The percentage differences (between modeled and measured acoustics) are also given in Tables III and IV. As Fig. 5(b) and Tables III and IV also show, the spacing between F4 and F5 in the sound booth data for actual speaker production is much larger for the retroflex /r/ than for the bunched /r/ (a difference of 1531 Hz versus 796 Hz for the supine position, and 1469 Hz versus 651 Hz for the upright position). Notably, the FEM also replicates this pattern of different spacing between F4 and F5. A similar difference in spacing is also predicted by the VTAR computer model using the extracted area functions (see Tables III and IV). Thus, these results support our methods for deriving a 3D model. They also

TABLE IV. Formants measured from S2's bunched /r/ compared with calculated values from the 3D FEM, area function model, and simple-tube model, respectively (Unit: Hz). The percentage difference between the FEM formant values and the actual subject formant values from MR (Δ1) and sound acoustic (Δ2) sessions are also given. Note that due to background noise, only F1–F3 could be consistently measured from the MRI acoustic data.

	Bunched /r/ (S2)							
	MRI acoustic data	Second both supine position	Sound booth upright position	3D FEM			Area function tube model	Simple tube model
				Formant	Δ1 (%)	Δ2 (%)		
F1	445	453	391	480	7.87	5.96	457	472
F2	1008	906	891	1040	3.17	14.79	998	1047
F3	1469	1203	1219	1660	13.0	37.99	1626	1680
F4		3313	3281	3260		1.60	3330	3190
F5		4109	4016	4000		2.65	3912	3841
F5-F4		796	735	740			582	651

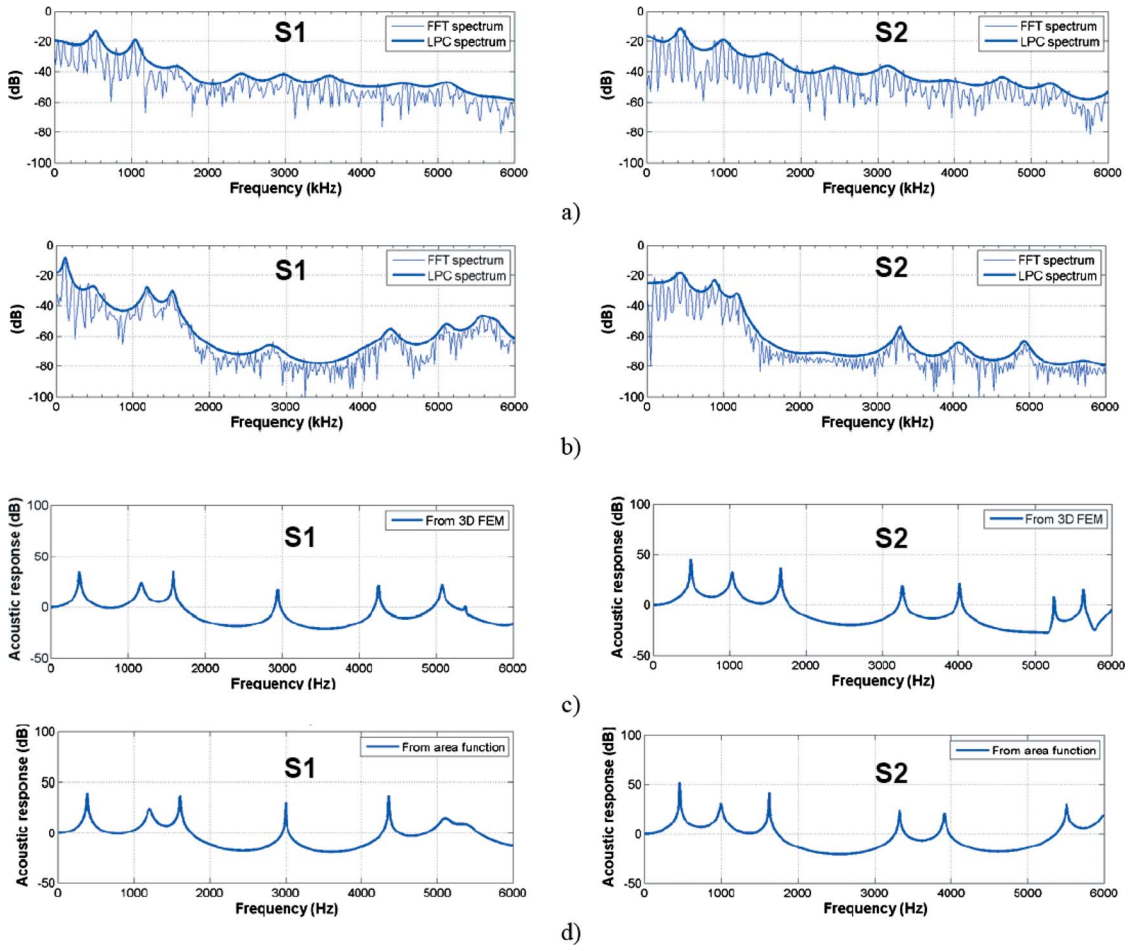


FIG. 5. (Color online) For S1 and S2: (a) Spectrum of sustained /r/ utterance in MRI session, (b) spectrum of sustained /r/ utterance in the sound booth acoustic data, (c) the acoustic response based on 3D FEM, and (d) the acoustic response based on the area function.

suggest that the source of the differences in the F4/F5 pattern between the bunched and retroflex /r/ follows from their respective differences in overall tongue shape.

A. FEM-derived area functions

Spectra generated from 3D FEM and area function sources are shown in Figs. 5(c) and 5(d). Formant values generated are shown in Tables III and IV. Both comparisons show that the results from the two methods match within 5% of each other. Note, however, that although the FEM produces zeros above 5000 Hz, they are not produced by the area function vocal tract model because it does not contain side branches and is based on only plane wave propagation.

B. Sensitivity functions and simple-tube modeling based on FEM-derived area functions

To gain insight into formant-cavity affiliations, the area function models were used to obtain sensitivity functions for F1–F5. Additionally, the area function models were simplified to arrive at models consisting of 3–8 sections (as opposed to about 70 sections) in order to gain insight into the types of resonators from which the formants originate and the effects of area perturbations of these resonators. These will be referred to as simple-tube models.

1. Sensitivity functions for F1–F5

The sensitivity functions of the formants are calculated as the difference between the kinetic energy and potential energy at the formant frequency as a function of distance starting from the glottis, divided by the total energy of kinetic and potential energies in the system (Fant and Pauli, 1974; Story, 2006). The relative change of the formant that corresponds to the change in the area function can be described as

$$\frac{\Delta F_n}{F_n} = \sum_{i=1}^{N} S_n(i)\frac{\Delta A_i}{A_i}, \tag{3}$$

where F_n is the nth formant, ΔF_n is the change of the nth formant, S_n is the sensitivity of the nth formant, A_i is the area of the ith section, and ΔA_i is the area change of the ith section. Section I is the first section starting from the glottis, and N is the last section number at the lips.

The calculated sensitivity functions are shown in Fig. 6 (the left panel is for S1 and the right panel is for S2). At a point where a curve for a given formant passes through zero, a perturbation in the cross-sectional area will cause no shift in the formant frequency. Otherwise, the curve shows how the formant will change if the area is increased at that point. If S_n is positive at a certain point, increasing the area at that point will increase the value of the nth formant. If S_n is

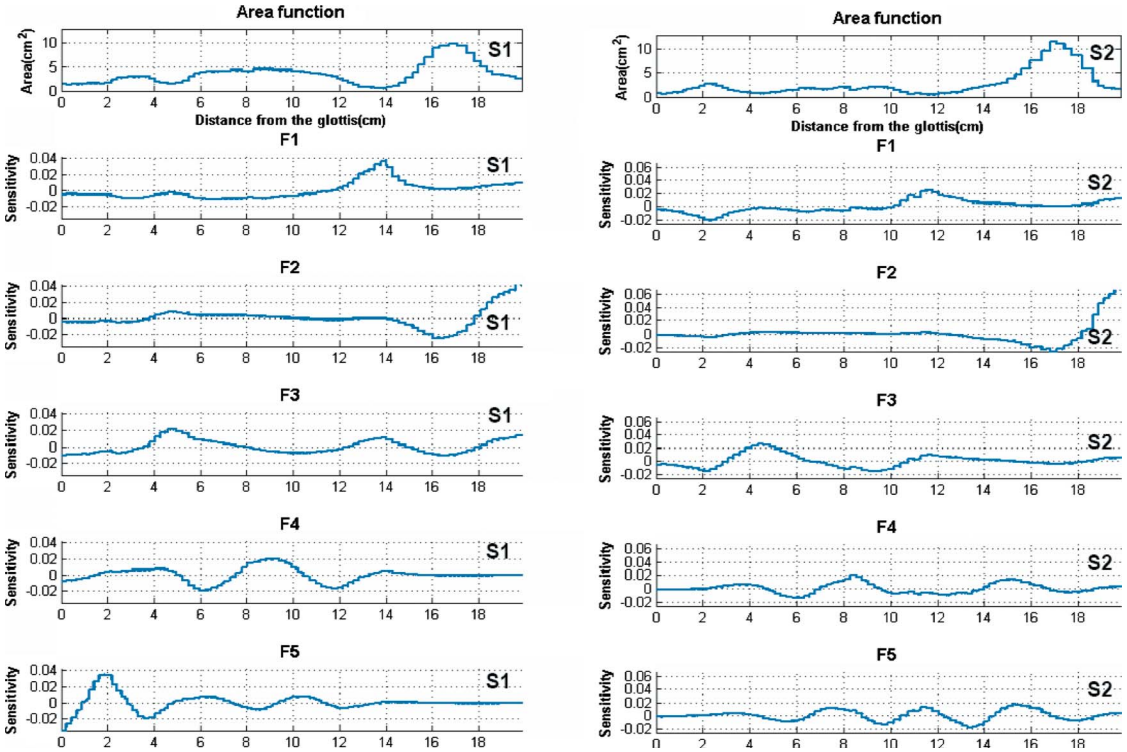


FIG. 6. (Color online) Acoustic sensitivity functions of F1–F5 for the retroflex /r/ of S1 and S2.

negative at a certain point, increasing the area at that point will decrease the value of the *n*th formant. The number of such zero crossings on a curve is equal to $2N-1$ (1, 3, 5, 7, and 9 for F1–F5, respectively) (Mrayati *et al.*, 1988), where N is the formant number for that curve.

As shown in Fig. 6, the sensitivity functions for F1–F3 have some similarities in their patterns for both the retroflex /r/ and the bunched /r/. In both cases, F2 is mainly affected by the front cavity where the lip constriction with small area plus the large posterior volume between the lip constriction and the palatal constriction act as a Helmholtz resonator. The frequency of a Helmholtz resonator is given by

$$F_H = \frac{c}{2\pi}\sqrt{\frac{A_1}{l_1 A_2 l_2}},$$

where A_1 and l_1 are the area and length of the lip constriction and A_2 and l_2 are the area and length of the large volume behind the lip constriction. From this equation, F_H will increase if the area of the lip constriction increases or if the area of the large volume behind the lip constriction decreases. The sensitivity functions for F2 show this behavior since it is significantly positive during the portion of the tube that corresponds to the lip constriction and, conversely, significantly negative during the portion of the tube that corresponds to the large volume.

This conclusion is supported by the spectra in Figs. 7 and 8. Figures 7 and 8 compare the spectra from the full vocal tract model with the spectra from the shortened vocal tract that includes only the front cavity as highlighted (acoustic responses were calculated with radiation at the lips) and the spectra from the shortened vocal tract that includes only the back cavity as highlighted (pressure on the front side is assumed to be zero). As can be seen, the first resonance of the front cavity is F2 from the full vocal tract for both subjects.

Based on the area function data of S1, Fig. 9 shows how the F2/F3 cavity affiliations switch when the front cavity volume is changed by varying its length. When the front cavity volume exceeds about 17 cm^3, there is a switch in formant-cavity affiliation between F2 and F3. The front cavity resonance is so low that it becomes F2 and the resonance of the cavity posterior to the palatal constriction becomes F3. It seems that the front cavity resonance may be F2 or F3 depending on the size of the volume of the Helmholtz resonator. This conclusion is supported by the findings from two different subjects showing bunched configurations discussed by Espy-Wilson *et al.* (2000). In that study, F3 was clearly derived from the Helmholtz front cavity resonance. However, the subjects in that study had much smaller front cavity volumes (of 5 and 8 cm^3) relative to those of the current subjects S1 and S2 (of 24 and 27 cm^3), respectively.

Due to coupling between cavities along the vocal tract, F1 and F3 of both retroflex and bunched /r/ can be affected by area perturbation along much of the vocal tract. However, there are differences. The F1 sensitivity function for S1's retroflex /r/ shows a prominent peak in the region of the palatal constriction (between 12.6 and 14.6 cm), whereas the F1 sensitivity function for S2's bunched /r/ shows a prominent peak and large positive value in the region of the palatal constriction (between 10.7 and 12.3 cm) and also a prominent peak dip in the region posterior to the pharyngeal constriction (between 1.6 and 2.8 cm). This difference in the F1 sensitivity functions of the retroflex and bunched /r/ is due to the differences in the area functions posterior to the front

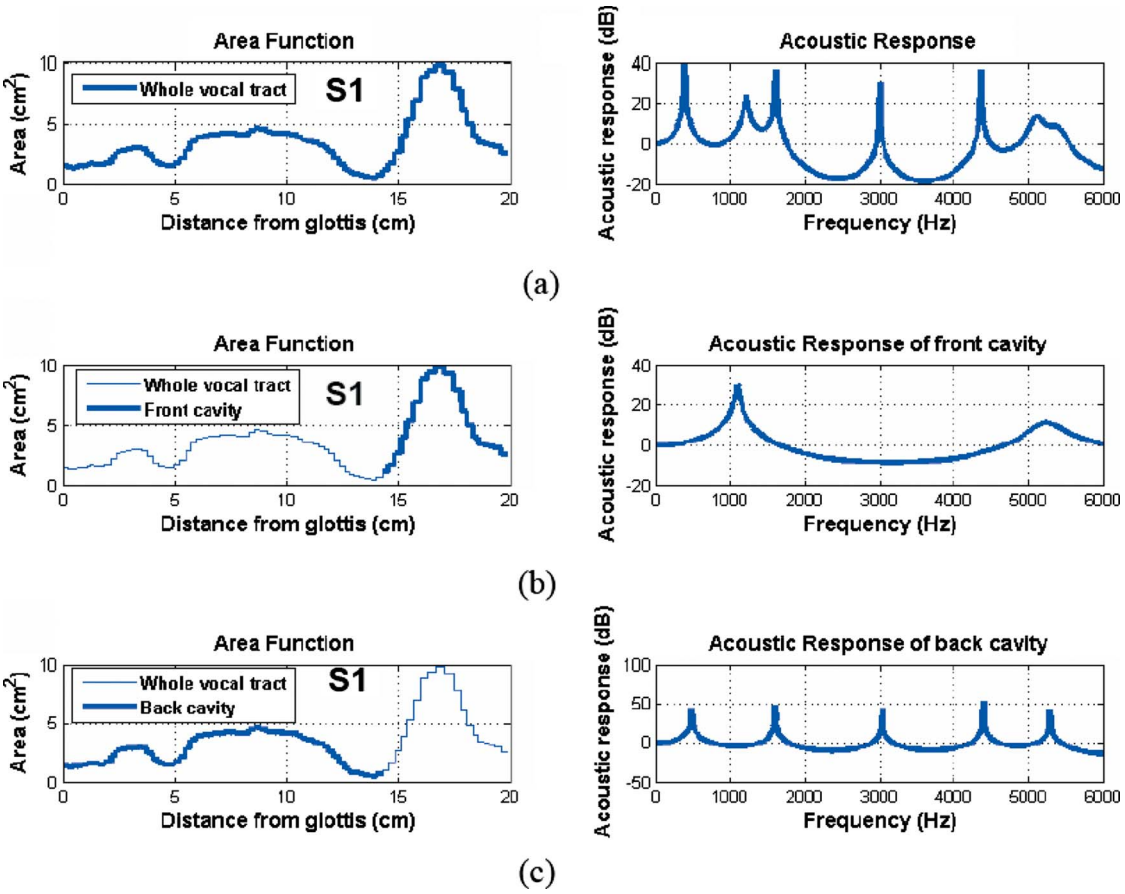


FIG. 7. (Color online) Acoustic response of S1's retroflex /r/ area function with front and back cavities separately modeled. (The left side is the area function and the right side is the corresponding acoustic response). (a) Area function of the whole vocal tract and its corresponding acoustic response. (b) Area function of the front cavity and its corresponding acoustic response. (c) Area function of the back cavity and its corresponding acoustic response.

cavity. In the retroflex /r/, the areas of the palatal constriction are much smaller than the areas of the back cavity posterior to the palatal constriction. This shape is more like a Helmholtz resonator for F1. In the bunched /r/, the overall shape of the area function posterior to the front cavity is similar to that of the retroflex /r/. However, the areas are more uniform so that F1 is the first resonance of a uniform tube (see discussion of simple-tube modeling below).

As the sensitivity functions indicate, F3 can be decreased by narrowing at each of the three constriction locations along the vocal tract. Note, however, that in both of these cases, F3 is most sensitive to the perturbation of the pharyngeal constriction. It is relatively much less sensitive to the palatal constriction and even less to the lip constriction. This result confirms the finding of Delattre and Freeman (1968) that the percept of /r/ depends strongly on the existence of a constriction in the pharynx.

Sensitivity functions for F4 and F5 have very different patterns for the retroflex /r/ and the bunched /r/. In the retroflex /r/, F4 and F5 are only minimally affected by the area perturbation of the front cavity, starting at the location about 14.8 cm from the glottis, which means that they are resonances of the cavities posterior to the palatal constriction. This conclusion is supported by the spectra in Fig. 7 which shows that the first four resonances of that part of the vocal tract behind the palatal constriction are close to F1–F5. In the bunched /r/, F4 and F5 are not sensitive to the area perturbation of the cavity posterior to the pharyngeal constriction and they are affected to some extent by the front cavity. Again, this sensitivity to the front cavity is probably due to a higher degree of coupling between the back and front cavities for the bunched /r/ relative to the retroflex /r/. Given the more gradual transition between the back and front parts of the vocal tract for the bunched /r/, Fig. 8 shows two possible divisions. In one case, the front cavity is assumed to start at 11.8 cm from the glottis. In the other case, it starts 2.9 cm further forward, at 14.7 cm from the glottis. In both cases, the first resonance (a Helmholtz resonance formed by the lip constriction and the large volume behind it) of the front cavity is around 1000 Hz, the frequency of F2 in the spectrum derived from the full vocal tract. However, this choice of a division point has a significant effect on the location of the second resonance (a half-wavelength resonance of the large volume between the lip constriction and the palatal constriction) from the front cavity. If the front cavity starts at 11.8 cm, the second resonance is around 3300 Hz, the region of F4 from the full vocal tract spectrum. If the front cavity starts around 14.7 cm, the second resonance of the front cavity is around 5500 Hz, which corresponds to the region around F6 in the spectrum derived from the full vocal tract.

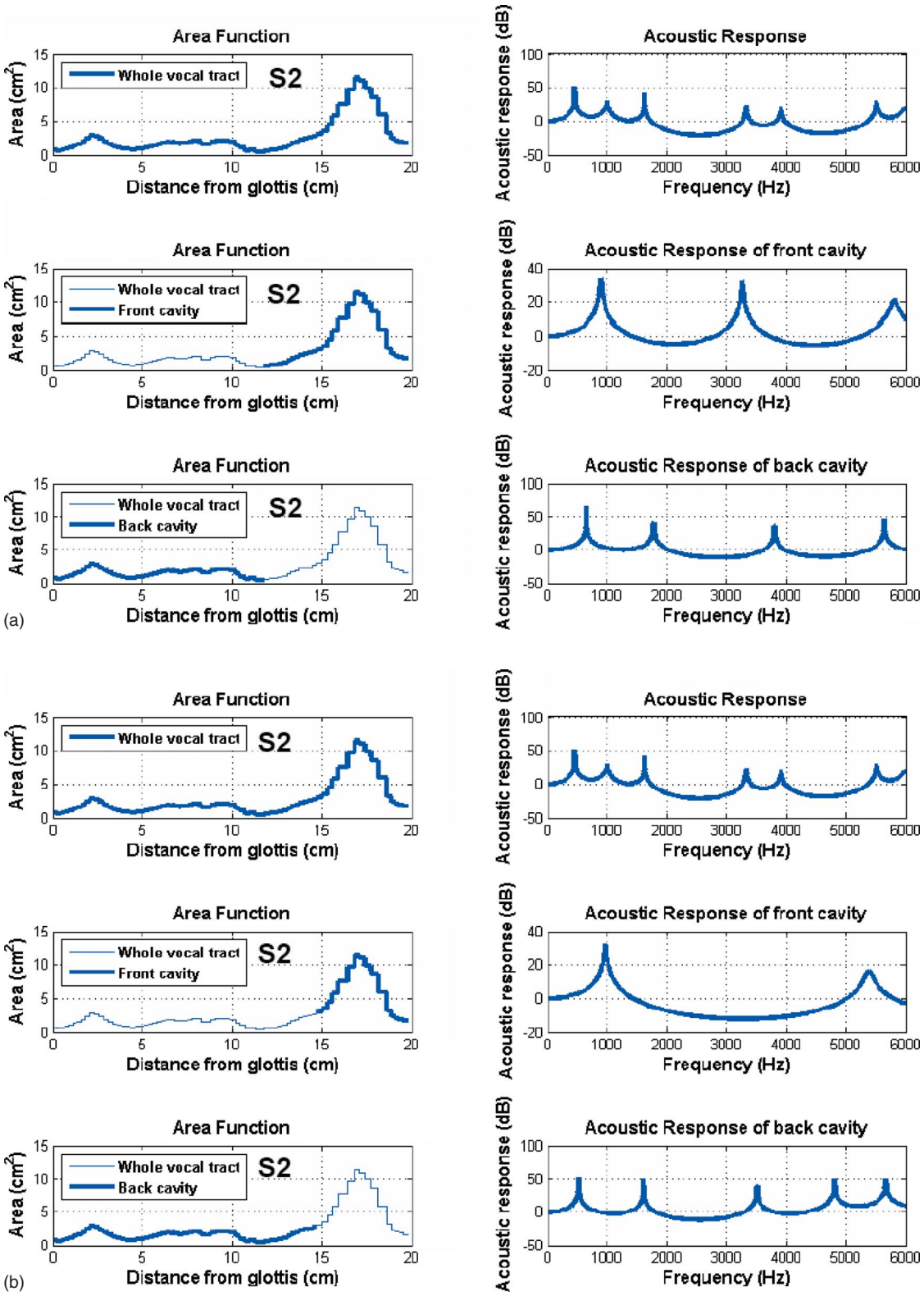


FIG. 8. (Color online) Acoustic response of S2's bunched /r/ area function with front and back cavities separately modeled. (The left side is the area function and the right side is the corresponding acoustic response). (a) The dividing point between the front cavity and the back cavity at about 12 cm. (b) The dividing point between the front cavity and the back cavity at about 15 cm.

2. Simple-tube models based on FEM-derived area functions

Figure 10 shows simple-tube models for the retroflex and bunched /r/ along with the original area functions and the corresponding acoustic responses. In the first case of the retroflex /r/, as shown in Fig. 10(a), the simple model consists of four tubes: a lip constriction, a large volume behind the lip constriction, a palatal constriction, and a long tube posterior to the palatal constriction [see Fig. 10(a)]. Henceforth, the area forward of the palatal constriction will be

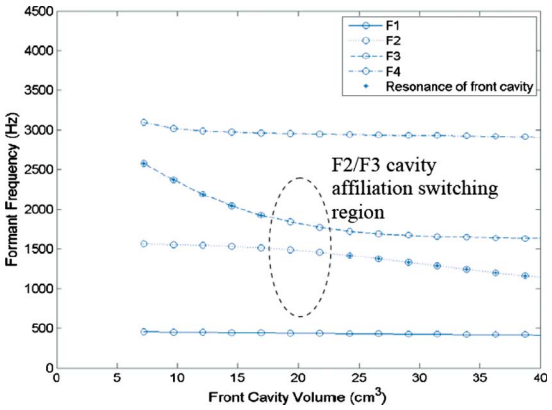


FIG. 9. (Color online) F2/F3 cavity affiliation switching with the change of the front cavity volume by varying its length (based on the area function data of S1).

referred to as the front cavity, while the area from the palatal constriction backward to the glottis will be referred to as the long back cavity. As we saw from the sensitivity functions, F2 comes from the front cavity, acting like a Helmholtz resonator at low frequencies. F1 comes from the long back cavity plus the palatal constriction, which together act as a Helmholtz resonator at low frequencies. F3–F5 are half-wavelength resonances of the long back cavity. The fact that the three formants are fairly evenly spaced [see Figs. 10(a) and 10(b)] is thus explained. Refinement of the simple tube, by allowing additional discrete sections as in Fig. 10(b), indicates that if we include the pharyngeal narrowing in our model, F3 is further lowered in frequency. In addition, if we include the narrowing in the laryngeal region above the glottis, F4 and F5 rise in frequency. The net results from these perturbations can be seen in Fig. 10(b). These formant-cavity affiliations agree well with our understanding from the sensitivity functions. Further, Tables III and IV show that there is close agreement between the formant frequencies measured from the actual acoustic data and those predicted both by the FEM-derived area functions and the simple-tube model.

In the case of the bunched /r/, the long back cavity has a wide constriction in the pharynx and is more uniform overall, so that we model it initially as a quarter-wavelength tube [see Fig. 10(c)]. If we then account for the pharyngeal narrowing, F3 is lowered and F5 is raised. If we include the palatal constriction itself, F4 is raised and F5 is lowered. Finally, including the laryngeal narrowing in the model raises F4 and (to a lesser extent) F5. The net results of these manipulations are shown in Fig. 10(d). Again, Tables III and IV show that there is close agreement between the formant frequencies predicted by both the FEM-derived area functions and the simple-tube model and measured from the actual acoustic data.

C. Formants in acoustic data of sustained /r/ and nonsense word "warav"

At this point, it appears plausible that the F4/F5 pattern shown by S1 and S2 is a function of their retroflex and bunched tongue shapes. As a partial confirmation of this hypothesis, we investigated acoustic data from sustained /r/ data for two subjects (S3 and S4) who have retroflex /r/ tongue shapes similar to S1 and two subjects (S5 and S6) who have bunched /r/ tongue shapes similar to S2. The averaged spectra (from a 300 ms segment of sound booth acoustic recordings) of the sustained /r/ sounds produced by the six subjects in the upright position are shown in Fig. 11. As can be seen, the retroflex /r/ has a larger difference in F4 and F5 than the bunched /r/. The differences between F4 and F5 for S3 and S4 are about 1900 and 2000 Hz, respectively, while the differences between F4 and F5 for S5 and S6 are about 500 and 600 Hz, respectively. These results are consistent with the results obtained from S1 and S2 in that the spacing between F4 and F5 is larger for the retroflex /r/ than for the bunched /r/.

In addition, the formant trajectories of the nonsense word "warav" for all the six subjects are shown in Fig. 12 (note that the spectrograms of Fig. 1 are repeated here for comparison). The differences between F4 and F5 of /r/ at the lowest point of F3 for S1, S3, and S4 are about 2100, 1500, and 1600 Hz, respectively, while the differences between F4 and F5 of /r/ at the lowest point of F3 for S2, S5, and S6 are about 700, 900, and 600 Hz, respectively. These results indicate that, for these subjects, the difference between F4 and F5 for the retroflex /r/ in dynamic speech is relatively larger than that in the bunched /r/ and provides additional support for the simulation result from the 3D FEM and computer vocal tract models based on the area functions.

IV. DISCUSSION

In this paper, we investigate the relationship between acoustic patterns in F4 and F5 and articulatory differences in tongue shape between subjects. The primary data come from S1 and S2, who produce sharply different bunched and retroflex variants of /r/ associated with different patterns of F4 and F5. S1 and S2 are particularly comparable because they resemble each other in terms of vocal tract length and oral tract dimensions. The results suggest that bunched and retroflex tongue shapes differ in the frequency spacing between F4 and F5. Further, the F4/F5 patterns produced by S1 and S2 can be derived from a very simple aspect of the difference between the two vocal tract shapes. For both S1's retroflex /r/ and S2's bunched /r/, F4 and F5 (along with F3) come from the long back cavity. However, for S1, these formants are half-wavelength resonances, while for S2, these formants are quarter-wavelength resonances of the cavity. Additionally, the finding of an F4/F5 difference in pattern is replicated in the acoustic data from an additional set of four subjects, two with bunched and two with retroflex tongue shapes for /r/. These results suggest that acoustic cues based on F4-F5 spacing may be robust and reliable indicators of tongue shape, at least for the classic (tongue tip down) bunched and (tongue dorsum down) retroflex shapes discussed here.

It appears that this spacing between F4 and F5 is due to the difference in long back cavity dimension/shape. In the case of the retroflex /r/, there is one long back cavity posterior to the palatal constriction. Our simple-tube modeling and the sensitivity functions show that F4 and F5 are half-

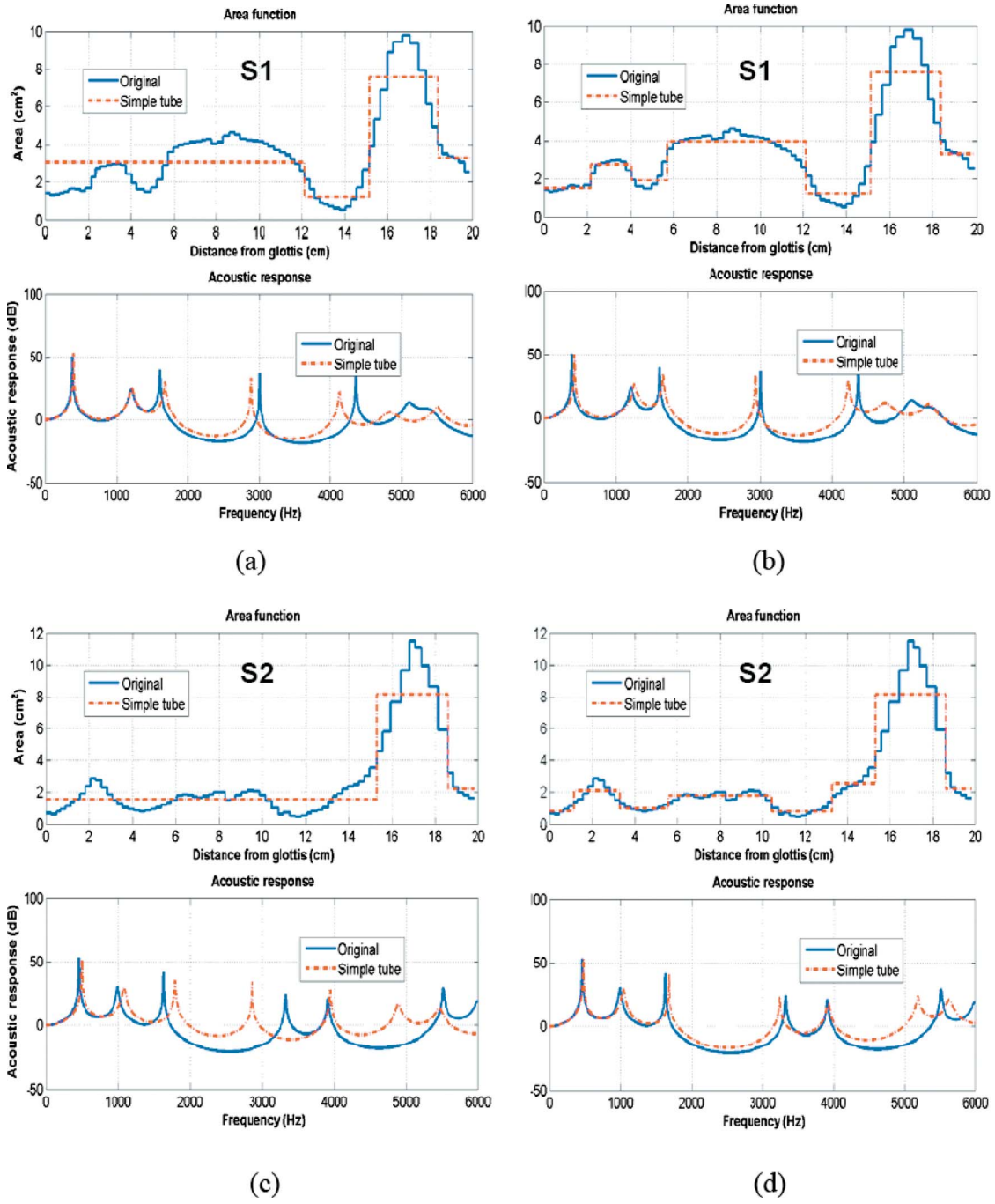


FIG. 10. (Color online) Simple-tube models overlaid on FEM-derived area functions (top panel) and corresponding acoustic responses (bottom panel). (a) Four element simple-tube model of the retroflex /r/ of S1. (b) Seven element simple-tube model of the retroflex /r/ of S1. (c) Three element simple-tube model of the bunched /r/ of S2. (d) Eight element simple-tube model of the bunched /r/ of S2.

wavelength resonances of the back cavity. In fact, F4 and F5 are the second and third resonances of the back cavity (F3 is the first resonance of this cavity). For S1, this half-wavelength cavity is about 12 cm long which gives a spacing between the resonances of about 1460 Hz. The narrowing in the laryngeal regions shifts F4 and F5 upward by different amounts so that the spacing changes to about 1300 Hz. This spacing agrees well with the 1469–1531 Hz measured from S1's sustained /r/. For the bunched /r/, the back cavity can be modeled as a quarter-wavelength tube. Our simple-tube modeling shows that F4 and F5 are the third and fourth resonances of this cavity. The sensitivity functions, on the other hand, show that F4 and F5 are influenced by the front cavity. This is probably due to the higher degree of coupling between the front and back cavities for the bunched /r/ of S2. The length of the back cavity for S2 is about 15 cm. Thus, the spacing between F4 and F5 for the bunched /r/ should be about 1150 Hz. However, the narrowing in the laryngeal, pharyngeal, and palatal regions decreases this difference to about 650 Hz, as seen in Fig. 10(d). This formant difference agrees well with the value of 651–796 Hz measured from S2's sustained /r/. As a point of interest, the spacing between F4 and F5 in the spectrograms of Fig. 12 is generally greater across all of the consonants and vowels for the speakers who

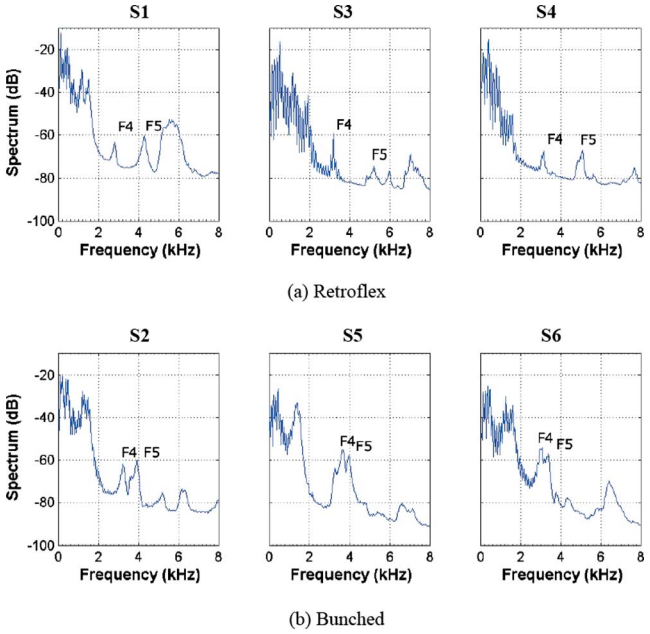


FIG. 11. (Color online) Spectra of sustained /r/ utterances from six speakers (three retroflex /r/ and three bunched /r/). (a) Retroflex /r/ (left: S1; middle: S3; right: S4.) (b) Bunched /r/ (left: S2; middle: S5; right: S6).

produce the retroflex tongue shape for /r/ than it is in the spectrograms for the speakers who produce the bunched tongue shape for /r/. However, the difference does appear to be considerably enhanced during the /r/ sounds with the lowering of F4 and the slight rising of F5 during the retroflex /r/, and the rising of F4 for S2 during the bunched /r/.

The relationship of tongue shapes for /r/ to specific acoustic properties as found in this study may be useful for the development of speech technologies such as speaker and speech recognition. For example, knowledge-based approaches to speech recognition heavily rely on acoustic information to infer articulatory behavior (Hasegawa-Johnson *et al.*, 2005; Kinga *et al.*, 2006; Juneja and Espy-Wilson, 2008). In addition, speakers appear to use tongue shapes in very consistent ways (Guenther *et al.*, 1999). Thus, the use of a particular tongue shape for /r/ may produce acoustic characteristics that are indicative of a speaker's identity, even if these characteristics are not relevant to the phonetic content.

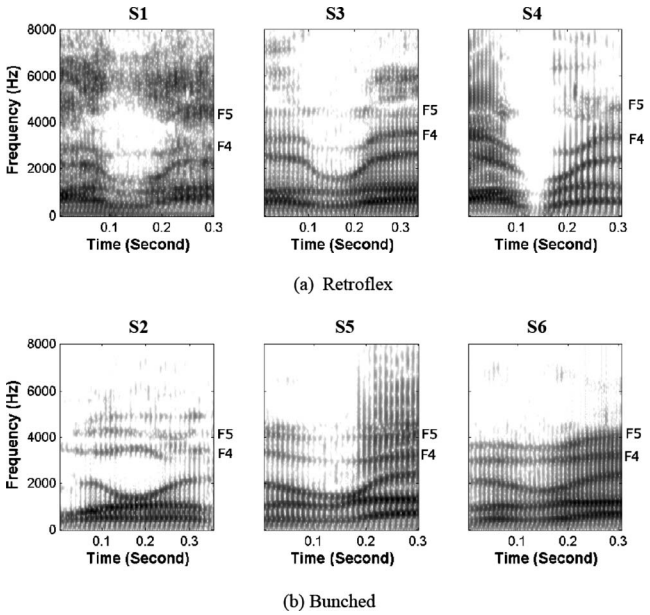


FIG. 12. Spectrograms for nonsense word "warav" from six speakers (three retroflex /r/ and three bunched /r/; only portions of spectrograms are shown in the figure with /r/ in the middle). (a) Retroflex /r/ (left: S1; middle: S3; right: S4). (b) Bunched /r/ (left: S2; middle: S5; right: S6).

ACKNOWLEDGMENT

This work was supported by NIH Grant No. 1-R01-DC05250-01.

[1]In the larger study (see Tiede *et al.*, 2004), subjects S1, S2, S3, S4, S5, and S6 are coded as subjects 22, 5, 1, 20, 17, and 19, respectively.

[2]Linguists distinguish between rhotic dialects, in which /r/ is fully pronounced in all word conditions, and nonrhotic dialects, in which some postvocalic /r/s are replaced by a schwalike vowel. Nonrhotic dialects are typically found throughout the southern states and in coastal New England.

[3]Ideally, productions of both a retroflex and bunched /r/ from a single speaker would be compared. Some speakers do indeed change their productions between true retroflex and bunched shapes in different phonetic contexts (Guenther *et al.*, 1999). However, this behavior appears to be a reaction to coarticulatory pressures in dynamic speaking conditions and is not easily elicited or trained in a sustained context. We in fact trained S2 to produce /r/ with his tongue tip up, and we collected a full set of MRI data for this production, in addition to the set with his natural /r/ production. However, even with training, S2 was not able to produce /r/ without a raised tongue dorsum as well as a raised tongue tip; thus, we were not able to compare 3D models of both a bunched configuration and a true retroflex tongue shape. While S1 was able to produce bunched /r/ in context, he was not able to sustain it consistently. At the same time, all of our speakers produced the same tongue shape consistently when asked to produce their natural sustained /r/. Thus, we contrast sustained bunched and retroflex /r/ as produced by subjects whose age and vocal tract dimensions are as similar as possible.

[4]Measured piriform dimensions: for S1, 18 mm in length and 2.3 cm^3 in volume; for S2, 12 mm in length and 2 cm^3 in volume.

[5]For subject S2, we collected a separate session of sound booth acoustic data in which his tongue shape for /r/ was monitored via ultrasound (Aloka SD-1000, 3.5 MHz probe held under the jaw). In all cases (upright running speech, supine and upright sustained /r/), S2 used a bunched tongue configuration with the tongue tip down when producing his natural /r/.

Alwan, A., Narayanan, S., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," J. Acoust. Soc. Am. **101**, 1078–1089.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," J. Acoust. Soc. Am. **90**, 799–828.

Chiba, T., and Kajiyama, M. (**1941**). *The Vowel: Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo).

Comsol (**2007**). COMSOL MULTIPHYSICS (http://www.comsol.com, accessed 12/20/2007).

Dalston, R. M. (**1975**). "Acoustic characteristics of English /w,r,l/ spoken correctly by young children and adults," J. Acoust. Soc. Am. **57**, 462–469.

Dang, J. W., and Honda, K. (**1997**). "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101**, 456–465.

Delattre, P., and Freeman, D. C. (**1968**). "A dialect study of American English r's by x-ray motion picture," Linguistics **44**, 28–69.

Espy-Wilson, C. Y. (**1987**). Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Espy-Wilson, C. Y. (**2004**). "Articulatory strategies, speech acoustics and variability," Proceedings of Sound to Sense: Fifty+ Years of Discoveries in Speech Communication, pp. B62–B76.

Espy-Wilson, C. Y., and Boyce, S. E. (**1999**). "The relevance of F4 in distinguishing between different articulatory configurations of American English /r/," J. Acoust. Soc. Am. **105**, 1400.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., and Alwan,

A. (**2000**). "Acoustic modeling of American English /r/," J. Acoust. Soc. Am. **108**, 343–356.
Fant, G. (**1970**). *Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations* (Mouton, The Hague).
Fant, G., and Pauli, S. (**1974**). "Spatial characteristics of vocal tract resonance modes," Proceedings of the Speech Communication Seminar, pp. 121–132.
Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., and Perkell, J. S. (**1999**). "Articulatory tradeoffs reduce acoustic variability during American English /r/ production," J. Acoust. Soc. Am. **105**, 2854–2865.
Hagiwara, R. (**1995**). "Acoustic realizations of American /r/ as produced by women and men," UCLA Working Papers in Phonetics, Vol. 90, pp. 1–187.
Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchhoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., and Tianyu, W. (**2005**). "Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 213–216.
Heinz, J. M., and Stevens, K. N. (**1964**). "On the derivation of area functions and acoustic spectra from cinéradiographic films of speech," J. Acoust. Soc. Am. **36**, 1037–1038.
Juneja, A., and Espy-Wilson, C. Y. (**2008**). "Probabilistic landmark detection for automatic speech recognition using acoustic-phonetic information," J. Acoust. Soc. Am. **123**, 1154–1168.
Kinga, S., Frankel, J., Livesc, K., McDermott, E., Richmon, K., and Wester, M. (**2006**). "Speech production knowledge in automatic speech recognition," J. Acoust. Soc. Am. **121**, 723–742.
Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., and Honda, K. (**2006**). "Cyclicity of laryngeal cavity resonance due to vocal fold vibration," J. Acoust. Soc. Am. **120**, 2239–2249.
Lee, K. (**1999**). *Principles of CAD/CAM/CAE Systems* (Addison-Wesley, Reading, MA).
Lehiste, I. (**1964**). *Acoustical Characteristics of Selected English Consonants* (Indiana University, Bloomington).
Lisker, L. (**1957**). "Minimal cues for separating /w,r,l,y/ in intervocalic position," Word **13**, 256–267.
Materialise (**2007**). Trial versions of Mimics and Magics (http://www.materialise.com, accessed 12/20/2007).
Matsuzaki, H., Miki, N., and Ogawa, Y. (**2000**). "3D finite element analysis of Japanese vowels in elliptic sound tube model," Electron. Commun. Eng. **83**, 43–51.
Matsuzaki, H., Miki, N., Ogawa, Y., Matsuzaki, H., Miki, N., and Ogawa, Y. (**1996**). "FEM analysis of sound wave propagation in the vocal tract with 3D radiational model," J. Acoust. Soc. Jpn. (E) **17**, 163-166.
Miki, N., Mastuzaki, H., Aoyama, K., and Ogawa, Y. (**1996**). "Transfer function of 3-D vocal tract model with higher mode," Proceedings of the Fourth Speech Production Seminar (Autrans), pp. 211–214.
Morse, P. M., and Ingard, K. U. (**1968**). *Theoretical Acoustics* McGraw-Hill, New York.
Motoki, K. (**2002**). "Three-dimensional acoustic field in vocal-tract," Acoust. Sci. & Tech. **23**, 207–212.
Mrayati, M., Carré, R., and Guérin, B. (**1988**). "Distinctive regions and modes—a new theory of speech production," Speech Commun. **7**, 257–286.
Narayanan, S. S., Alwan, A. A., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," J. Acoust. Soc. Am. **101**, 1064–1077.
O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1957**). "Acoustic cues for the perception of initial /w,j, r,l/ in English," Word **13**, 24–43.
Ong, D., and Stone, M. (**1998**). "Three dimensional vocal tract shapes in [r] and [l]: A study of MRI, ultrasound, electropalatography, and acoustics," Phonoscope **1**, 1–13.
Shriberg, L. D., and Kent, R. D. (**1982**). *Clinical Phonetics* (Macmillan, New York).
Sondhi, M. M. (**1986**). "Resonances of a bent vocal tract," J. Acoust. Soc. Am. **79**, 1113–1116.
Story, B. H. (**2006**). "Technique for 'tuning' vocal tract area functions based on acoustic sensitivity functions (L)," J. Acoust. Soc. Am. **119**, 715–718.
Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**, 537–554.
Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., and Honda, K. (**2006a**). "Acoustic roles of the laryngeal cavity in vocal tract resonance," J. Acoust. Soc. Am. **120**, 2228–2238.
Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (**2006b**). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," J. Acoust. Soc. Am. **119**, 1037–1049.
Thomas, T. J. (**1986**). "A finite element model of fluid flow in the vocal tract," Comput. Speech Lang. **1**, 131–151.
Tiede, M., Boyce, S. E., Holland, C., and Chou, A. (**2004**). "A new taxonomy of American English /r/ using MRI and ultrasound," J. Acoust. Soc. Am. **115**, 2633–2634.
Twist, A., Baker, A., Mielke, J., and Archangeli, D. (**2007**). "Are 'covert' /r/ allophones really indistinguishable?," Selected Papers from NWAV 35, University of Pennsylvania working Papers in Linguistics **13**(2).
Westbury, J. R., Hashi, M., and Lindstrom, M. J. (**1998**). "Differences among speakers in lingual articulation for American English /r/," Speech Commun. **26**, 203–226.
Zhang, Z., Espy-Wilson, C., Boyce, S., and Tiede, M. (**2005**). "Modeling of the front cavity and sublingual space in American English rhotic sounds," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 893–896.
Zhou, X. H., Zhang, Z. Y., and Espy-Wilson, C. Y. (**2004**). "VTAR: A MATLAB-based computer program for vocal tract acoustic modeling," J. Acoust. Soc. Am. **115**, 2543.

Selective acoustic cues for French voiceless stop consonants

Anne Bonneau[a] and Yves Laprie[b]
LORIA/CNRS, Speech Team, Vandœuvre 54506, France

(Received 9 March 2007; revised 7 April 2008; accepted 7 April 2008)

The objective of this study is to define selective cues that identify only certain realizations of a feature, more precisely the place of articulation of French unvoiced stops, but have every realization identified with a very high level of confidence. The method is based on the delimitation of "distinctive regions" for well chosen acoustic criteria, which contains some exemplars of a feature and (almost) no other exemplar of any other feature in competition. Selective cues, which correspond to distinctive regions, must not be combined with less reliable acoustic cues and their evaluation should be done on reliable elementary acoustic detector outputs. A set of selective cues has been defined for the identification of the place of /p,t,k/, and then tested on a corpus of sentences. The cues were estimated from formant transitions and the transient segment (an automatic segmentation of the transient part of the burst has been designed). About 38% of the feature realizations have been identified by selective cues on the basis of their very distinctive patterns. The error rate, which constitutes the crucial test of our approach, was 0.7%. This opens the way to interesting applications for the improvement of oral comprehension, lexical access, or automatic speech recognition. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2916693]

PACS number(s): 43.70.Fq [DOS] Pages: 4482–4497

I. INTRODUCTION

It is well accepted that acoustic cues do not identify every realization of a phonetic feature with the same level of confidence. On the one hand, previous studies have always shown (sometimes large) overlaps between the acoustic region coding different phonetic categories (Öhman, 1966). On the other hand, it appears that certain realizations of a feature are identifiable in a highly reliable way, both from a perceptual and acoustical point of view, as shown by Davis and Kuhl (1992) for the perception of the voicing feature of velar stops. These studies concerned speech produced in well controlled phonetic contexts, but it is also evident, from the know how of spectrogram readers, that some realizations of sounds embedded in sentences and produced by an unknown speaker can be identified with reliability (Lamel, 1988). The object of this study is to define acoustic cues that identify certain realizations of a phonetic feature, more precisely one of the three places of articulation of French unvoiced stops, with a very high level of confidence and in an automated way. To that purpose, we develop a specific methodology that can be applied to other classes of sounds. It may be worth noting that the aim is to identify a feature from a given realization and not to identify the quality of the realization itself. These cues, based on phonetic knowledge, should be available for sounds extracted from sentences produced by various speakers. From now on, they will be called selective cues.

Beside its phonetic interest, the highly reliable identification of certain exemplars of a feature has potential applications in domains such as automatic speech recognition (ASR) or the improvement of oral comprehension (auditory deficiencies, language learning, noise, etc.). In ASR, this identification would generate "confidence islands" and would allow the system to prune lexical searches by favoring the words containing features identified by selective cues (Coste-Marquis, 1994). The interest of phonetic features for lexical access has been stressed by Stevens (2002). The improvement of oral comprehension can be obtained through the enhancement or the slowing down of certain well defined parts of speech signals, e.g., transitions, or certain phonetic classes, e.g., stops or fricatives (Hazan and Simpson, 1998). It would probably be interesting to apply these speech modifications to sounds with highly reliable cues.

Both aspects of our study, i.e., the search for a highly reliable automatic identification, on the one hand, and the selection of certain exemplars of a feature, on the other hand, have common points with recent studies. We present some of them below, with the aim of highlighting both these common points and the divergences with respect to our study.

Recent studies in phonetic decoding have focused on the exploitation of very reliable information. Landmarks proposed by Liu (1996) define regions in an utterance when the acoustic correlates of distinctive features are the most salient. Since the landmarks only give temporal information, it is necessary to attach acoustic cues to them in order to identify speech sounds. Consonantal acoustic cues attached to landmarks were presented by Stevens (2000). The detection of landmarks was fairly reliable since the error rate was approximately 14%. However, global error rate should take into account both the detection error rate of landmarks and the identification error rate of acoustic cues. Unlike the selective cues, the cues attached to landmarks do not involve any selective behavior. Our aim is to identify only certain realizations of a feature, but have every realization detected by a selective cue identified with a very high level of confidence.

[a]Electronic mail: anne.bonneau@loria.fr. URL: http://parole.loria.fr
[b]Electronic mail: yves.laprie@loria.fr

 0001-4966/2008/123(6)/4482/16/$23.00

Studies by Davis and Kuhl (1992) about the perception of the voicing feature, or by Iverson and Kuhl (1995) about the perception of synthetic two formant vowels, tend to show that certain exemplars of a feature are considered as the "best exemplars" of this feature. According to the "perceptual magnet theory" (Kuhl, 1991), this intraphonemic distinction plays an important role in perception (language acquisition and foreign language learning), helping the learners to construct phonetic categories (see Lotto *et al.*, 1998, for a criticism of this theory). Our search for selective cues share common points with the search for best exemplars since we operate distinctions between different realizations of a same sound. But we are only looking for exemplars that exhibit very distinctive acoustic patterns and do not want to establish any correspondence between the acoustical realizations detected by selective cues and the best exemplars of a feature.

Acoustic decoding systems, "global" or knowledge-based systems cannot identify certain realizations of a feature with a very high level of confidence. This is quite normal for global systems, such as Hidden Markov Model (HMM), where decisions are made by statistical models. Knowledge-based systems, no longer developed today, could have identified some well realized feature exemplars with a high level of confidence. However, acoustic decoding, as it was carried out by these systems, was based on acoustic cues, which provided both partial and uncertain informations. The lack of reliability came from acoustic cues themselves—the values of one acoustic cue for different features overlap—and, at a lower level, from acoustic detectors, such as segmentation, voicing detection, or formant tracking. Hence, whenever, for a given realization, an acoustic cue indicated the presence of a feature with a very high level of confidence, the quality of this information was diminished by the uncertainty linked to acoustic detectors as well as by the interference of other cues.

We believe that it is possible to identify certain realizations of an acoustic feature in a highly reliable manner, provided we adopt the following methodology. First, we should research whether there exists a region in a given acoustic space, containing certain exemplars of this feature and no other exemplar of any other feature in competition. From now on, such a region will be called a "distinctive region." Selective cues should be triggered only on realizations belonging to a distinctive region. Secondly, in an automatic system, all the procedures leading to the estimation of selective cues must be reliable, so only acoustic cues that can be estimated from reliable detector outputs should be eligible as selective cues. Finally, selective cues must not be combined with less reliable cues. In return, it is possible to define more than one selective cue for a given feature.

During selective cue tuning, we must fulfill two conflicting requirements: making almost no error and correctly identifying a substantial number of realizations. The higher the number of realizations the distinctive region might contain, the more efficient the corresponding selective cue is. In this exploratory study, we did not know the approximate level we could hope for. Since selective cues are activated only on realizations belonging to a distinctive region, the identification rate expected is clearly lower from those obtained in other systems. One important result (a way of judging the approach) is that the error rate should be very low.

Theoretically, if the distinctive regions are well determined, the corresponding selective cues should make no error. Practically, this is quite impossible: firstly, because the acoustic detectors cannot be entirely reliable, secondly, because a mispronunciation, not taken into consideration during the corpus labeling, or a very large effect of the phonetic context, not easily predictable, can lead to a false alarm of the acoustic cues, and, finally, because the distinctive region determination should be made from a corpus which is at the same time very large and representative, which is hardly feasible. Practically, we consider that an error rate below 1% is acceptable. A selective cue for a class of sounds is thus intended to identify certain exemplars of this class with as few errors as possible.

We have chosen to define selective cues for the places of articulation of voiceless stop consonants. French unvoiced stops are either labial (/p/), dental (/t/), or velar (/k/). Note that /t/ is considered as alveolar in English, and that /k/ could be postpalatal before front vowels. This choice was motivated by the feature itself, coded by numerous acoustic cues, which are complex (such as the burst spectrum), difficult to detect (especially formants at stop-vowel boundary), and very sensitive to the vocalic context. Therefore, if we can find efficient selective cues for this class of sounds, it may be possible to do so for other classes of sounds. As said before, all the procedures leading to place identification must be automatic and reliable. That is the reason why we have elaborated acoustic detectors devoted to the segmentation and the analysis of the transient and used previous work for formant extraction (Laprie and Berger, 1996). To evaluate the relevance of selective cues, we have chosen to define them from a small training corpus, with well-controlled contexts and test them on a larger one with new speakers and natural sentences.

In order to exploit all the information provided by acoustic signals to the optimum, we have also defined acoustic cues devoted to the elimination of a feature with quasicertainty. These cues are called "selective negative cues" or "selective exclusion cues" (so as to distinguish them from "selective positive cues," which are devoted to the identification of a feature). Negative cues could be exploited in automatic speech recognition to prune lexical searches.

In the following sections, we will present the protocol and acoustic detectors (second section), then the selective cues defined from the burst and the transitions (Secs. III and IV), the results (Sec. V), and the discussion and the concluding remarks (Sec. VI).

II. EXPERIMENTAL PROTOCOL

A. Corpus

Three corpora made up of unvoiced stops followed by oral vowels, and appearing in syllable-initial or intervocalic position, have been elaborated for the training and the test stages. They were all recorded by a group of male French speakers, aged from 20 to 45 years old, and different for

TABLE I. Composition of the training and test corpora. W. and S. stand for words and sentences.

	Training corpus			Test corpus		
Corpus	C1 (W.)	C2.a (S.)	Tot.	C2.b (S.)	C3 (S.)	Tot.
Speakers	4	4	8	4	10	14
Stops	190	219	409	472	1813	2285

each corpus. They were segmented and labeled by hand. The number of items per corpus is given in Table I.

The first corpus (C1), extracted from the French database BDSons, contained monosyllabic CV or CVC words made up of an initial unvoiced stop followed by an oral vowel and, for CVC words, any final consonant. Final stops were not investigated. Four speakers uttered each word once in an anechoic chamber. Since the same stop-V sequence appeared in different CV[C] words (e.g., /kɔ/ appeared in "code," "cor," "col,"etc.), different repetitions of the same CV sequence by a single speaker were present in the corpus. The second corpus (C2), especially designed for this study, contained 16 sentences uniquely made up of stops in different vocalic contexts. Most stops were in an intervocalic position. The sentences were built so that each stop was followed by each vowel at least one time. Four speakers uttered each sentence three times, in an anechoic chamber. The third corpus (C3), made up of V-unvoiced stop-V or unvoiced stop-V sequences, was extracted from a set of sentences recorded three times by each of the ten speakers. The corpus was recorded in an office, and some small marks of echo and of noise were present in the signal. This corpus was exploited to test selective cues in natural sentences and informal recording conditions. As can be seen, there is a gradual level of difficulty with regard to phonetic identification from the first corpus to the third one due to phonetic contexts (syllables or sentences) and recording quality. The first corpus (C1) and the first repetition of the second one (C2a) have been reserved for the training stage. The two remaining repetitions of the second corpus (C2b) and the third one have been kept for the test.

All the corpora were recorded at 16 kHz, and the frequency range considered was 0–8000 Hz. Further details about the first corpus can be found in (Carré *et al.*, 1984). The second and third corpus were recorded by means of a Sony mini-DAT and an Electret microphone (Sony ECM-44B).

For this exploratory study, we analyzed stops followed by the vowels /u, o, ɔ, y, ø, œ, a, ɛ/. To take into account the important influence of vowels on stop spectra (Suomi, 1985), we split the vocalic context into three classes according to the vowel labialization and degree of frontness: front rounded vowels /y, ø/, front open and central vowels /a, ɛ, œ/ (henceforth referred to as central vowels, for short), and back vowels /u, o, ɔ/. Theoretically, the vowel /æ/ should be put in the front-rounded class. Since this vowel is often realized as a rather neutral vowel, with a weak degree of lip protrusion, we have chosen to put it in the central class. Of course, the ideal classification should be determined by trying to detect the degree of rounding of each realization (with F3 frequency value, essentially), which was not within the scope of this study. Table II gives the number of stops as a function of the vocalic contexts taken into account for stop identification. There were very few data for the front-rounded context. This was partially due to the fact that we put the vowel /œ/ in central contexts, as well as to the chance of sentence formation.

TABLE II. Number of data as a function of the vocalic context and the consonant place.

	Training			Test		
Vowel class	P	T	K	P	T	K
Back	21	39	61	242	279	318
Central	46	110	57	545	443	308
Front rounded	23	35	17	44	68	38
Total	90	184	135	831	790	664

B. Acoustic detectors for stop place identification

We present here the detectors devoted to the identification of stop place by the set of selective cues: (1) the detection and segmentation of the transient from the burst, (2) the extraction of pertinent acoustic parameters from the transient spectrum, and (3) the set of acoustic parameters extracted from the formant transitions.

1. Segmentation of the transient

The segmentation of the transient from the burst fricative noise was calculated from a wide-band spectrogram. A Hamming window of 4 ms was applied onto the preemphasized signal, which was zero padded before the application of a 256 points Discrete Fourier Transform (DFT). The zero padding merely introduced spectral smoothing but did not alter the definition of spectrum. This smoothing improved our ability to resolve the spectrum (i.e., there are more points outlining the same curve) which was interesting afterwards to segment the burst in homogeneous segments. We chose a 1 ms time shift in order to offer better time resolution.

Burst noise of stops can be decomposed in three segments (Fant, 1973), the transient, the frication noise, and the aspiration (generally absent in French). Formant transitions are often visible in frication noise and aspiration. In CV sequences, the burst begins at the release of the articulation (at the end of the closure) and ends with the first voiced period. The frication noise corresponds to the transition from the consonant articulation to the following sound, generally a vowel. Consequently, the spectrum of the burst rapidly changes and, in most of the cases, is more discriminant in its initial part, i.e., the transient. Correlatively, the closer to the vowel the spectrum is calculated, the more it is dominated by the vowel formants. This is particularly visible in the spectra of voiced stops calculated by Krull (1990) at two different instants (from 0 to 10 ms and from 10 to 20 ms after stop release). It thus seems judicious to evaluate the place of articulation of stops just after the release of the consonant articulation. Since the segmentation of the transient is not easy

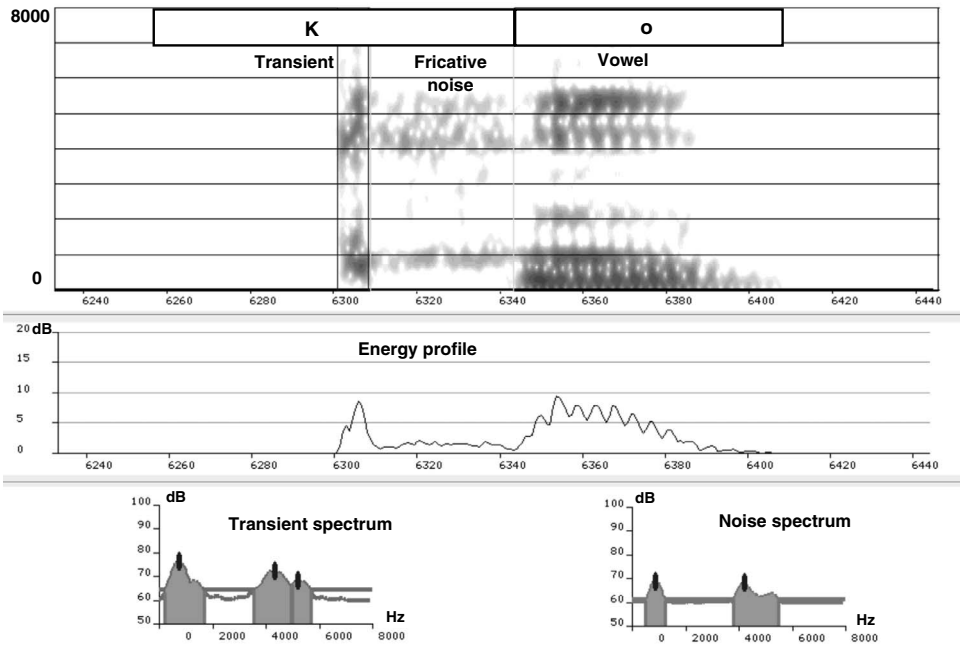


FIG. 1. Analysis of the transient part of the burst (syllable /ko/). Top: segmentation of the burst into two parts, the transient and the fricative noise, three vertical bars delimit these events, the time is indicated in ms. Middle: energy above the silence threshold calculated with a 4 ms window. Bottom: transient spectrum (bottom left) and fricative noise spectrum (bottom right); the horizontal line onto each spectrum represents the average spectrum energy level; only peaks above this line are represented and analyzed.

to perform, only a fixed length segment, at the beginning of the burst, is generally taken into account [10–15 ms for Zue (1976)]. The objective of the segmentation algorithm we designed is to locate the transient. Practically, this consists of decomposing the burst into two homogeneous parts. The technique calls on the computation of wide-band spectra.

$X(e^{j\omega_k})$ denotes the Fourier transform of the windowed signal and $S(e^{j\omega_k})=\max(20\log_{10}|X(e^{j\omega_k})|-dB_{\text{silence}},0)$, the energy above the silence threshold used in the spectrogram calculation and $\bar{S}(e^{j\omega_k})$ the average value of $S(e^{j\omega_k})$ over the speech segment determined by instants t_0 and t_1

The homogeneity criterion is the spectral variance computed over each segment of the burst,

$$\text{VarSpec}(t_0,t_1)=\frac{1}{t_1-t_0}\sum_{t=t_0}^{t=t_1}\sum_{k=K_0}^{k=K_1}[S_t(e^{j\omega_k})-\bar{S}(e^{j\omega_k})]^2,$$

where K_0 and K_1 delimit the frequency domain used for segmentation. The smaller it is, the higher the homogeneity. The best segmentation corresponds to the time instant where the sum of the spectral variances of the two segments is minimal. The criterion to be minimized is thus $\text{VarSpec}(t_i,\text{limit})+\text{VarSpec}(\text{limit},t_f)$; where t_i represents the instant where energy becomes higher than the threshold set for speech detection and t_f is the beginning of the first voiced period.

This segmentation criterion gives good results when the transient is very intense as well as when transient and noise spectra are different (see Fig. 1 for an example of burst segmentation).

Segmentation errors stem from either the acoustic structure of the transient or the context in which the segmentation is used. We shall now present methods intended to increase the segmentation robustness *vis-à-vis* errors in the determination of the burst or voicing onset. Indeed, the segmentation algorithm relies on the idea that a burst can be decomposed into two parts. It may therefore make errors when this hypothesis is not verified because either the voicing onset is detected too late or a spurious noise comes before the burst.

The first difficulty is related to the inaccuracy of the determination of the voicing onset. Taking into account one or several voicing periods in addition to the burst may indeed modify the boundary position because it leads to a false burst with three parts with different spectral and energy characteristics: the transient, the frication noise, and the first voicing periods. As the extreme regions (the transient and the first voicing periods) concentrate most of the acoustic energy, the segmentation generally gives a boundary somewhere during the frication noise. We thus reduced the influence of energy at the end of burst to eliminate this risk of errors. Practically, we replaced the deviation term of the variance calculation by $[S_t(e^{j\omega_k})-\bar{S}(e^{j\omega_k})]^2\times\cos^{\alpha}[(t-t_i)/t_f-t_i]\pi/2$, where t_i and t_f are burst extremities. This means that $t_i\leqslant t_0\leqslant t_1\leqslant t_f$. The exponent α is used to adjust the influence of the last spectra and thus depends on the precision of the voicing determination.

As the reduction of the last spectra influence increases the contribution of energy at the burst onset, the transient boundary may be detected too early. Practically, the end of the transient always corresponds to a sensitive lowering of energy, and the decrease of the last spectra energy thus does not modify the boundary position.

The second difficulty is linked to the presence of spurious noises during the closure, which trigger false alarms. These noises are detected because the energy threshold for detecting burst onsets is weak, so that bursts without a strong transient can be located. Schematically, these false alarms correspond to the presence of a small noise followed by a silence and then the true burst. Several noises may come before the burst, but this does not change the procedure described below. The spurious noise is always considered as the transient and the silence that comes just after may be

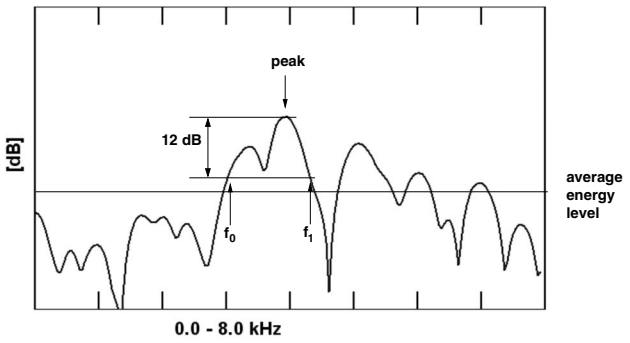


FIG. 2. Determination of the dominant peaks in the spectrum.

merged with either the transient or the frication noise. These two cases are distinguished by the correction procedure.

The spurious noise and the silence form the transient. The average spectrum of the transient is calculated after segmentation. If the silence has been merged with the noise the average spectrum is weak or even smaller than the spectrogram silence threshold, i.e., dB_{silence}. This enables this kind of error to be easily detected. The segmentation procedure is thus reiterated until the transient spectrum becomes at least 3 dB higher than the spectrogram silence threshold.

The spurious noise has been considered as the transient. This means that the silence and the true transient have been merged together and that they wrongly represent the frication noise. The frication noise is thus segmented. If there is a significantly long silence (more than 20 ms), this means that the frication noise actually consists of a silence followed by the true burst which is then segmented as expected.

This double correction concerning both the transient and the frication noise eliminates most of the errors due to spurious noises but does not eliminate multiple transients which correspond to peaks in energy significantly more intense and followed by a short silence (less than 10 ms).

The interest of the transient for the identification of stop place is discussed in Sec. III B 1.

2. Extraction of acoustic parameters from the transient

The average spectrum was calculated from all the short term spectra (4 ms) estimated for the transient. Only the spectral energy above an arbitrary reference threshold, chosen to roughly correspond to the minimum audible field, was taken into account. Relevant peaks, which must be above the average spectrum energy level, were then searched in the resulting spectrum. They were detected according to their spectral energy, the most intense peak being located as a priority. The extension of each peak corresponds to the spectral region it dominates (see Fig. 2). It was determined by scanning the spectrum downwards from the peak maximum until the energy is more than 12 dB below this maximum or lower than the average spectrum energy. This dominance determination allowed small peaks to be merged with the most prominent ones (see Fig. 1 for the estimation of burst peaks).

For each peak determined in this way, a set of parameters was evaluated: the maximum (maximal value of the peak energy), the frequency, (evaluated at the peak maximum), and the "width." The width of a peak was estimated by calculating the second order momentum with respect to the frequency of the peak,

$$\text{width} = \frac{\sqrt{\int_{f_0}^{f_1}[S(f)-\bar{S}]\times(f-f_{\max})^2 df}}{\sqrt{\int_{f_0}^{f_1}[S(f)-\bar{S}]df}},$$

$\bar{S}$ is the average level of the spectrum, $S(f)$ the energy level at frequencies f, f_0, and f_1 are the lower and higher boundaries of the peak, and $f_{\max}$ the frequency of the maximum. The higher the compactness of the peak energy, the smaller the width corresponding to the denominator. This calculation of the width penalized peaks that spread over a large spectral region or included other small peaks. Of course, the most prominent peak of the spectrum is particularly important in the determination of stop places. In addition to these parameters, provided by the transient, we estimated the duration of the entire burst, from the beginning of the transient to the first period of the following vowel.

3. Extraction of acoustic parameters from formant transitions

We evaluated F2 and F3 formant frequencies at both sides of the stop consonant: the vowel onset for CV transitions and the vowel offset for VC transitions. The automatic evaluation of formant frequencies at stop boundaries is very difficult, but these parameters constitute important information for stop place (Delattre *et al.*, 1955). We also estimated the formant slopes between the vowel boundaries—the onset or the offset—and the vowel centers; they were evaluated in Hz/ms. Since we wanted very reliable detector outputs, we decided to consider only F2 transitions to define selective acoustic cues, F3 being more difficult to detect, and less crucial for stop place identification.

C. Evaluation methodology

1. Detectors

The acoustic detectors, more precisely the transient segmentation and the formant tracking, have been elaborated and tested on "clean" speech, i.e., speech recorded in quiet conditions, and have not yet been adapted to speech recorded in an office. Consequently, for the third corpus, recorded in an office, we verified the transient segmentation by hand. Transition cues were defined and tested on the first and second corpora (C2b for the test). For these corpora, the procedures linked to the identification of stop place by selective cues were entirely automatic.

2. Evaluation of the results

The cues were separately evaluated for each segment (burst, left and right transitions) and simultaneously for all the segments. Here is how the set of cues was evaluated.

Given two consonant places Ci and Cj, given that totCi represents the number of realizations of Ci, NbCiCj the number of realizations of Ci identified by at least one positive cue devoted to the identification of Cj, and NbCiExCj the num-

ber of realizations of Ci allowing the exclusion of Cj (i.e., the number of realizations of Ci from which at least one cue excluding Cj is triggered),

R1i . rate of correct identification of Ci

= NbCiCi/totCi,

R2ji . rate of correct exclusion of Cj from Ci

= NbCiExCj/totCi.

Since there are three places of articulation for French unvoiced stops, we estimated two correct exclusion rates (R2) for a given place of articulation. For instance, for the consonant /k/, we evaluated R2tk and R2pk, corresponding to the correct exclusion of /t/ and of /p/ from the realizations of /k/. We also estimated the actual information provided by each exclusion cue. This rate was calculated as the previous one (R2), but we did not take into account the activation of the cue when it happened on a realization already identified by a positive cue. Indeed, in this case, the information provided by the negative cue was redundant. The exclusion rate evaluated in this way gave an idea of the "additional" information, which was supplied by the exclusion cue in the absence of direct identification through positive cues.

Given Nb'CiExCj is the number of realizations of Ci, not identified by a positive cue, and allowing the exclusion of Cj,

R3ji . rate of correct exclusion of Cj from Ci in the absence of identification = Nb′CiExCj/totCi.

Then we evaluated the "overall information" provided by the set of cues from a given place (Ci): that is the relative number of realizations that were correctly identified and/or allowed the correct exclusion of one or two other places (say, Cj and Ck).

R4i . overall information rate from Ci

= R1i + R3ji + R3ki.

The error rate committed by the positive cues for a given place (Ci) corresponded to the relative number of realizations of Ci identified as Cj or Ck. The error rate for the exclusion cues corresponded to the relative number of realizations of Ci incorrectly excluded.

R5i . rate of error in the identification of Ci

= (nbCiCj + nbCiCk)/totCi,

R6i . rate of error in the exclusion of Ci

= NbCiExCi/totCi.

Particular cases. If two cues excluding two different consonant places were triggered on the same realization, the right consonant place was indirectly identified and, for the estimation of the results, the cues were considered as a positive cue. In fact, we will see (Sec. V) that this case was rare and did not occur when the segments were separately tested.

A conflict between the cues happened either when two positive cues identifying a different place or one positive cue and one exclusion cue for the same place were triggered on a same realization. When a conflict was observed, no decision was taken, and the activations of the cues on the realization were not taken into account.

III. CUES PROVIDED BY THE BURST

A. Literature review

The acoustic cues for stop places of articulation are provided by the burst spectrum, the transitions, the duration, and amplitude of the burst (Edwards, 1981). The main cues derived from the burst spectrum are the frequency of the most prominent peak, probably the most efficient cue to stop places (Ali *et al.*, 2001), its prominence, and the spectral compactness. The slopes and onsets (or offsets) of the second and third formant transitions at both sides of the consonant also provided important information; they are described in Sec. IV.

The duration of the burst (measured from the release to the first period of the following vowel) is a well-known cue for stop place. Among unvoiced stops, velars (/k/) tend to have the longest bursts and labials (/p/) the shortest [see Wajskop (1979) for a study of the duration of French stop bursts]. Nevertheless, there are very important overlaps between the values of burst durations for different places of articulation, even for a single phonetic context, so it is not possible to define distinctive regions from this cue. Figure 2 in Edwards (1981) illustrated these important overlap areas for English stops well.

Two sets of spectral acoustic cues, related to the compact/diffuse and acute/grave features, characterize stop places ["preliminaries to speech analysis" Jackobson *et al.* (1952)]: velar stops (/k,g/) are considered as compact, dentoalveolars (/t,d/) as diffuse and acute, and labials (/p,b/) as diffuse and grave. Cues related to spectral acuteness are often based on the frequency of the most prominent peak (henceforth referred to as "main peak," for short). Edwards (1981) observed that this peak was generally below 2500 Hz for labial stops and above this frequency for dentoalveolar stops. The main peak frequency of velars is close to the vowel F2 before front rounded or back vowels and even higher than the vowel F3 before high front unrounded vowels (Bonneau *et al.*, 1996; Fischer-Jörgensen, 1954). Blumstein and Stevens (1979) defined three invariant static cues modeling the overall shape of the spectrum: the "compact," the "diffuse-rising," and the "diffuse-falling" or "diffuse-flat" templates, representing, respectively, velar, dentoalveolar, and labial spectra. The first 26 ms following stop release were taken into account to calculate the spectrum.

Numerous works have called into question the invariance of static cues for coding stop places. First, it should be noted that many studies have been based on (American) English stops. Yet, the consonant /t/ is realized at two slightly different places according to the language—dental in French and alveolar in English. Lahiri *et al.* (1984) showed that many dental spectra, in opposition to those of alveolars,

could not be efficiently distinguished from labial spectra by a static cue based on spectral acuteness and did not match the diffuse-rising template.

Besides, the influence of the subsequent vowel on stop spectral cues should probably not be overlooked. In particular, the influence of the following vowel labialization is particularly striking. Fischer-Jörgensen (1954) showed that the anticipation of labialization not only substantially lowered the frequency of the main peak of Danish velar and alveolar stops but also increased the concentration of energy around the maximum of these consonants.

Thus, there is no simple relationship between stop spectral forms and stop places of articulation, especially for languages such as French, which is characterized by strong vocalic labialization, and the presence of dental consonants, often confused with labials. To handle the first problem—the effect of labialization—it seemed necessary to take into consideration the vocalic context. Considering the second problem, we had to determine whether there nevertheless existed a distinctive region for the dental /t/. Taking into account the vocalic context would probably decrease consonant confusions in this case too.

B. Definition of selective cues from the burst transient

This definition was carried out in two steps. First, we proposed acoustic cues as correlates to the acute/grave and compact/diffuse features and retained those that contributed to the emergence of distinctive regions from the data of the training corpus (Sec. III B 1). Then, we determined these regions, defined the corresponding selective cues, and discussed their phonetic relevance (Sec. III B 2). Let us recall that a distinctive region should contain certain exemplars of a class and (almost) no exemplar of any other class in competition for stop place identification. To put it in another way, the aim was to find a region where a substantial number of exemplars of a class could be distinguished from all the other exemplars of the other classes in competition without any confusion.

1. Acoustic cues

Transient segmentation. In order to test the interest of using the transient, segmented by our method, instead of the whole burst, we evaluated the frequency and prominence of the main peak of the spectrum averaged over the whole burst. We exploited the data of the training corpus. Results showed that the main peaks of most /t/ and /k/ were less prominent which led to a decrease in their discriminating power as a cue to stop place. Figure 1 illustrates the value in exploiting the transient information, the main peak of /k/ being clearly more salient in this segment than in the frication noise.

Frequency and prominence of the spectral maximum. The frequency of the main peak is one of the most important acoustic cues for stop place identification, and we chose it as the correlate to the acute/grave feature. We observed large overlaps between the areas covered by the values of this criterion for different places of articulation, so the peak frequency alone did not constitute highly reliable information, and must be associated with a criterion evaluating the peak prominence (this will clearly appear in Sec. III B 2). To estimate the prominence of the main peak, we calculated the difference between its maximal energy value and that of the second most prominent peak.

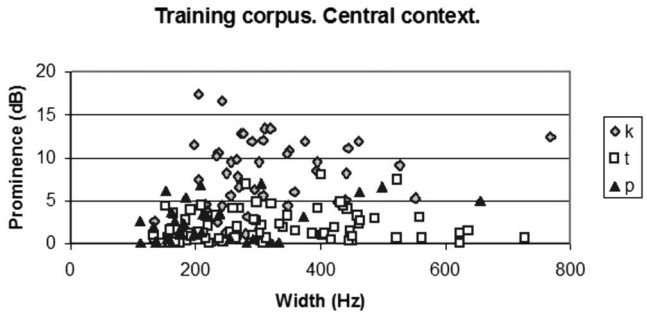


FIG. 3. Distribution of the values of the width of the most prominent peak of the burst spectrum as a function of its prominence (central context). See Sec. III B 1 for the prominence evaluation.

Spectral compactness. A sound is considered as compact if its spectrum is characterized by a large concentration of energy in a frequency region both relatively narrow and central. In fact, centrality is an abstract notion since it corresponded to different frequency regions for vowels and consonants in the preliminaries, and it covered a large region for Blumstein and Stevens' compact template (from 1200 to 3500 Hz, i.e., approximately the frequency region of the velar peak for English). Let us ignore the notion of centrality and focus on spectral compactness alone. A stop can be considered as compact if the most prominent peak of the spectrum conveys an important concentration of energy and if its width is not very large (to fit the "relatively narrow frequency region" criterion).

In order to characterize the energy of the main peak, we accepted the criterion evaluating the main peak emergence with respect to that of the second most prominent peak. When the value of this criterion was very high, it distinguished the most salient velars from other consonants whose main peak was in the typical velar frequency region. This criterion was adopted by Halle *et al.* (1957) to distinguish velars from labials in English.

Concerning the peak width (see Sec. II for its estimation), the examination of the data for each vocalic context showed that there were large overlaps between the values of the criterion for consonants from various places and no typical values for velars (Fig. 3 shows the distribution of the criterion values in central context). Consequently, we did not retain this criterion as a parameter for selective cue definition.

We thus used two criteria: the emergence and the frequency of the main peak.

2. Selective positive cues

We determined the distinctive regions revealed by the distribution of the cue values in the "frequency-emergence plan." Figure 4 presents the data after the application of a set of "safeguard" filters, described below, which eliminated po-

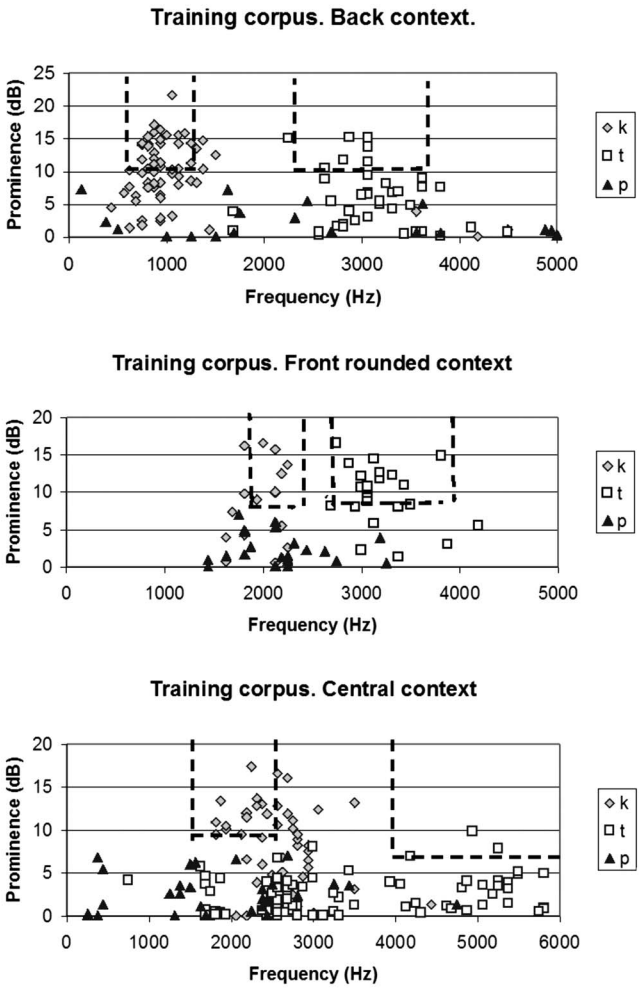


FIG. 4. Representation of stop consonants as a function of the frequency (Hz) and prominence (dB) of the most prominent peak of the burst spectrum. See Sec. III B 1 for the prominence evaluation. Data are separated according to the class of the following vowel (back, front-rounded or central contexts).

tential errors due to incorrect behavior of detectors. For the definition of selective cues, the regions retained were somewhat reduced with respect to the regions discernible on the figures in order to make the cues more robust when applied to other corpora (with different speech rates, contexts, etc.). However, the work presented here was more intended to test the feasibility of the method than to determine the best separations between classes. This test was realized with the corpora C2b and C3.

It appeared, from the data of all the vocalic contexts that no distinctive regions could be found for the place of /p/, at least from the set of parameters taken into account in this work. This was to be expected since labial spectra are not intense and do not have a very prominent main peak, probably because there is no cavity of resonance for labial sounds. Zue (1976) and Edwards (1981) also discarded labials from spectral analysis. So the definition of selective cues from burst was restricted to the places of /t/ and /k/. The selective cues for rounded contexts were introduced and then for the central context.

Rounded contexts (back and front rounded contexts). The acoustic configurations exhibited by /k/ and /t/ in these contexts (Fig. 4, top and center) were in agreement with the preceding description of stop spectra. Both classes were characterized by the presence of a prominent peak, situated in front of F2 for /k/ and above approximately 2200–2500 Hz for /t/. What was important for this study was that, beyond a certain level of emergence, these characteristic peaks turned out to be highly reliable cues for the places of these consonants. As it will be shown in the following paragraphs, /k/ and /t/ could be essentially distinguished from each other by the main peak frequency, whereas prominence was crucial for the distinction between /k/ and /p/. Both criteria were necessary to distinguish /t/ from /p/ without confusion.

Velars (/k/). Data clearly showed that velars (/k/) characterized by a very prominent peak in front of the vowel F2—nearly 700–1500 Hz for the back context and 1900–2400 Hz for the front rounded one—could be identified in a highly reliable way. As could be seen from Fig. 4, the frequency of the main peak was sufficient to distinguish velars which appeared in the expected region from all the dentals (/t/), no matter what the prominence of the peak is. On the contrary, only velars exhibiting a very prominent peak in the expected frequency region could be separated from all the labials.

Selective cue for the place of /k/: Rounded contexts. The main peak of the spectrum is situated in the F2 frequency region of the following vowel and clearly dominates the spectrum.

Practically, we could calculate the vowel F2 frequency at the boundary between the consonant and the vowel. Nevertheless, because of possible errors in F2 estimation and problems in determining the accepted degree of divergence between the consonant peak frequency and the vowel F2 frequency, we set *a priori* frequency limits for the F2 region.

To delimit the distinctive velar region, we adopted the following frequency limits and levels of emergence: 700–1500 Hz and 10 dB for the back context; 1900–2400 Hz and 8 dB for the front rounded one. These parameters led to firing rates of 66% and 53% on the training corpus, respectively for the back and the front rounded contexts, with no error.

Dentals (/t/). The main peak of most /t/ was situated in high frequencies (approximately 2600–4000 Hz). Although labial spectra were generally falling or flat, some /p/ had their main peak (not a very prominent one), in this high frequency area. In the back vocalic context, this was also the case for a few velars, whose main peak was at about 3500–4000 Hz. This peak was never very prominent since a low frequency peak was always present in the spectrum and relatively intense. Therefore, dentals with a prominent peak in high frequencies could be distinguished from all the other consonants, and then be identified in a highly reliable way.

Selective cue for the place of /t/: Rounded contexts. The main peak of the spectrum is situated in high frequencies and clearly dominates the spectrum.

To define the distinctive dental regions, we adopted the following frequency limits and levels of emergence: 2500–4000 Hz and 10 dB for the back vocalic context; 2700–4000 Hz and 8 dB for the front rounded vocalic con-

text. These parameters led to firing rates of 28% (back context) and 63% (front-rounded context) on the training corpus, with no error.

Central context. Contrary to classical descriptions, peaks in high frequencies did not often dominate the spectra of /t/, which appeared to be relatively flat. Flat spectra have been observed by Lahiri *et al.* (1984) for languages with dental consonants, such as French. However, the authors did not specify that they were more frequent in central contexts than in rounded ones, as was shown in our data. The main peak frequency of /t/, highly variable, was sometimes situated in front of F2 (1600–1800 Hz), more often in front of F3 (2600–3000 Hz) of central vowels, as well as in higher frequencies (approximately above 4000 Hz). These frequency values overlapped those of the main peak of /k/ situated between the F2 and the F3 of the following vowel. As a consequence, the frequency of the main peak was not as efficient a criterion as it was in rounded contexts, and the performance of selective cues was expected to be lower in central context.

As can be seen from Fig. 4 (bottom), velars with a very prominent peak in their characteristic frequency region could be distinguished from all the other stops. Only dentals with a prominent main peak situated in very high frequencies—approximately above 4000 Hz, which avoided both the main and the second peak of velars—could be separated with no error from the other stops. This represented relatively few dentals.

Selective cue for the place of /k/: Central context. The main peak of the spectrum is situated between the F2 and F3 of the following vowel and clearly dominates the spectrum.

To define the distinctive velar region, we adopted the following frequency limits and levels of emergence: 1500–2800 Hz and 9 dB. These parameters led to a firing rate of 33% on the training corpus, with no error.

Selective cue for the place of /t/: Central context. The main peak of the spectrum is above 4000 Hz, and relatively prominent.

To define the distinctive dental region, we adopted the following level of emergence: 6 dB. These parameters led to a firing rate of only 5% on the training corpus, with no error.

3. Exclusion cues

These cues were intended to eliminate a class of consonants from identification, with a very high level of confidence. Negative cues searched for spectral configurations that were never observed for a given class but were frequently observed for at least one of the others. Note that the regions of exclusion of a given feature should not be entirely confounded with a distinctive region of another one; otherwise negative cues would be redundant with positive ones and useless. The negative cues proposed below were based on the location of the main peak of the burst spectrum. In rounded contexts, as previously noted, the main peaks of /k/ and /t/ were situated in well characteristic frequency regions with no dispersion excepted for not very prominent peaks. This meant that /k/ and /t/ could be excluded from identification if the main peak of the consonant under analysis was situated outside its own typical region and was relatively prominent. Since velar consonants were always characterized by the presence of a maximum in their typical frequency region, we could also exclude them when no spectral peak was observed in the frequencies of the main velar peak. We did not define any exclusion cue for /k/ and /t/ in central context due to their poor distinctiveness on the frequency axis.

Exclusion cue for the place of /t/: Back and front rounded contexts. The main peak of the spectrum is situated in front of the vowel F2 and its prominence is relatively high.

Practically, the regions considered for F2 were 700–1200 Hz for the back context and 1900–2300 Hz for the front one (the region of the velar main peak), and the level of prominence was set to 5 dB. We chose to evaluate the performance of the cues through the rate R3 (Sec. II C 2) which gave an idea of the additional information provided by exclusion cues in the absence of direct identification. The consonant /t/ was correctly excluded in about 10% of the cases on the realizations of the consonant /k/ (back context) and in 8% of the cases on /k/ and /p/ (front rounded context).

Exclusion cue for the place of /k/: Back and front-rounded contexts. Velar consonants are excluded from identification if at least one of the two following spectral configurations is observed: (1) the main peak of the spectrum is situated in a frequency region which is above that of the low velar peak and below that of the high velar peak (at around 3500 Hz); and (2) there is no spectral peak in the frequencies of the main velar peak.

Practically, the frequency regions adopted for the first criteria were 2000–3200 Hz (back context) and 2800–3500 Hz (front-rounded context). The prominence level was set to 5 dB. Considering the additional information (R3), /k/ was correctly excluded in about 25% of the cases on /t/ and /p/ (back context) and 10% of the cases on /t/ (front rounded context).

4. Safeguard filters

In a few cases, the briefness of the burst, a segmentation problem, or an incorrect behavior of the acoustic detectors due to an aberrant spectral form could generate errors (selective cue false alarms). Let us mention the most frequent problems and the solutions we have found in each case. The safeguard filters, defined below, were available for all the contexts and were applied before the search for selective cues.

A first source of problems came from short segments. When the burst was very short, the explosion (transient) was close to the following vowel, and its spectrum was often dominated by the following vowel formants. If the peak corresponding to F2 or F3 was very prominent, a selective cue for /k/ could be triggered from a labial or a dental spectrum. Sometimes, a spurious noise (short, in most cases) preceded the burst explosion and was confused with the latter. The spectral form of this kind of noise was not foreseeable and could lead to a false alarm. To eliminate these two sources of potential errors, we have fixed a minimal duration for dental and velar bursts, below which selective cues could not be activated.

Another source of errors came from the automatic analysis of the burst and, in particular, the peak detection. When there were no sufficiently deep valleys between two consecutive spectral peaks, the analysis (see Sec. II) detected a long stretching peak. If the remaining part of the spectrum was not very intense, the prominence criterion took a high value, which could trigger a false alarm of a selective cue. We have therefore specified a maximal value for the peak width, beyond which a selective cue could not be activated.

Both filters (minimal burst duration and maximal peak width) eliminated potential errors without greatly reducing the firing rate of selective cues since consonants appearing in a distinctive region of a class generally had a substantial duration and a restricted width.

5. Concluding remarks for burst cues

When all the realizations of consonants /k/ and /t/, extracted from the training corpus, were considered together, the selective cues led to an identification of about 37% of the data (54% in rounded contexts and 16% in central context), with no error. The cues defined in rounded contexts were more efficient than in the central one because of the clear separation of the main peak of /t/ and /k/ on the frequency axis in the former contexts. This differentiation was probably due to the anticipation of labialization, which increases the concentration of energy in the spectrum (Fischer-Jörgensen, 1954). In order to verify their validity, the cues were tested on new data (Sec. V).

IV. CUES PROVIDED BY THE TRANSITIONS

A. Literature review

It was far from evident whether or not formant transitions, which vary a great deal depending on vocalic context, speaker, and speech rate could allow the place of certain stops to be identified in a highly reliable manner. Indeed, Kewley-Port (1982), in a study concerning stop-V isolated syllables uttered by one speaker, showed that F2 and F3 transitions combined together provided context-dependent cues for stop places, whereas F2 transitions alone failed to do so in two contexts. Moreover, she projected stops into planes determined by F2 and F3 onset frequencies—one plane per vocalic context—and remarked that, despite the very limited sources of variations present in the corpus, the boundaries of the acoustic regions representing different stop categories were very close together. She made the hypothesis that, at a perceptual level, "the distinctions between stop categories are likely to be lost." It seems also clear that, at an acoustical level, these distinctions should be lost or at least weakened, if more sources of variations have been taken into account. In the study concerning V-stop-V sequences produced by one speaker, Öhman (1966) showed that VC and CV second formant transitions were very deeply influenced by the vowel situated on the other side of the stop (the "transconsonantal" vowel). Indeed, F2 transitions could not indicate stop places in a reliable manner, even when the adjacent vowel was known. In some cases, the F2 transition slopes of sequences made up of a same consonant and a same vowel and differing only in the transconsonantal vowel were completely opposed. Although Lindblom (1996), reconsidering Öhman's data, found distinctive context-dependent regions for each stop category in a three-dimensional space determined by F2 and F3 onset frequencies and F2 at the vowel center, these regions were once more very close together. As observed from Kewley-Port's study, it is thus probable that the presence of more sources of variations in the data, due to speaker, speech rate and, for sentences, the neighboring sounds, among others, would generate large overlaps between the regions representing different places.

Before observing the impact of variation on the existence of distinctive regions for stop places, let us describe how transitions indicate stop places in minimal contexts. Since only very distinctive patterns are of interest for selective cue definition, we give below a rough description of transitions for each stop place in isolated CV syllables and discuss the effect of the transconsonantal vowel in VCV contexts when the direction of CV transitions might be deeply modified by this effect. As said in the protocol section, only F2 transitions were taken into account for selective cue definition.

Labial transitions. Labialization lowering F2 formant frequency, CV transitions are generally rising, but can be flat or slightly falling before back rounded vowels (Calliope, 1989; Kewley-Port, 1982). Nevertheless, in VCV contexts, when the first vowel is more anterior than the second one, and when the vowels are coarticulated, we can observe strongly falling slopes.

Dental transitions. Dentals are the only stop consonants to have a relatively constant F2 locus, no matter what the following vowel is (Fant, 1973). Let us recall that the locus is the virtual starting or ending point of transitions (Delattre *et al.*, 1955). Formant transition slopes are thus determined by the locus frequency (at about 1600 Hz) and the second formant frequency of the vowel: they are strongly falling before back vowels and rising before high front vowels.

Velar transitions. Well marked CV slopes appear before central vowels. They fall from the consonant to the following vowel and have a very high F2 onset frequency, at about 2200–2300 Hz in the data of Kewley-Port (1982), well above the dental locus. In VCV contexts, when the first vowel is back, the slope can be rising instead of being strongly falling. In back context, velar transitions are generally flat or slightly rising (Kewley-Port, 1982; Halle *et al.*, 1957).

This short description underlines the possible strong influence of the transconsonantal vowel. This influence was one of the most important obstacles in the definition of selective cues, when we did not use F3, the importance of which has been stated by Lindblom (1996). We have paid particular attention to the emergence and the handling of this coarticulation phenomenon. Since, in our study, transitions were extracted from sentences and produced by many speakers, we expected large overlaps between the regions representing different phonetic categories. The object of the following study was to determine whether, in spite of (probably) large overlapping areas, it was possible to find distinctive regions for each class of unvoiced stop conso-

nants. For this exploratory study, we limited the vocalic contexts to the back and central ones, which provided the most distinctive transition patterns.

B. Selective cues definition

The parameters chosen to define selective cues were the formant slopes, calculated at both sides of the consonant, as well as the vowel initial (onset) F2 frequency for CV transition and terminal (offset) F2 frequency for VC transition (see Sec. II for the estimation of all these parameters). We defined a set of selective cues for each vocalic context, taking into account both the data of the training corpus and, to ensure that the cues were phonetically founded, results reported in the phonetic literature. Let us recall that this corpus included a set of isolated monosyllabic words and a set of continuous sentences, both pronounced by four speakers, all different for each set.

The cues for VC transitions were deduced from CV cues. Although the correspondence between CV and VC cues is not straightforward (Ohde and Sharf, 1977), we supposed that, for a same vowel and a same consonant, VC transition slopes were the opposite of CV slopes and that offset VC formant frequencies were approximately the same as onset CV formant frequencies. We could not make such an assumption for velar VC transitions, since the velar articulation is very sensitive to the vocalic context and could be influenced very early by the following vowel. We thus defined left transition cues for /t/ and /p/ but not for /k/.

1. Selective positive cues

Back vocalic context. According to the above description, /p/ and /k/ followed by back vowels exhibit flat or slightly rising slopes (sometimes even slightly falling for /p/). The falling slopes of /t/, when they are sufficiently marked, as well as their high frequency onsets, linked to the dental locus, should distinguish this consonant from the others. Halle *et al.* (1957) noted that "before back vowels, the cue for /t/ is a markedly falling transition, while the absence of such an F2 transition is the cue for /p/ and /k/, which do not significantly differ from each other." Nevertheless, in the case of intervocalic coarticulation between a front or central vowel and a back one during the bilabial closure, we can observe strongly falling slopes for /p/, similar to those of /t/. The impact of this phenomenon on the definition of a distinctive dental region had to be carefully considered.

Data showed the existence of a distinctive region for /t/, which was made up of exemplars exhibiting high onsets and strongly falling slopes (Fig. 5, top). Apart from cases of very high onsets, taking into consideration both cues was necessary to optimize the region since we observed some high frequency onsets for /k/ (above 1200 Hz), coming from /kɔ/ contexts, and one clearly falling slope for /p/. These data came from the sequence /øpu/ and was probably due to intervocalic coarticulation. It is worth noting that such coarticulation, at least when it was revealed by a strong acoustic effect, was not very frequent since we observed only one

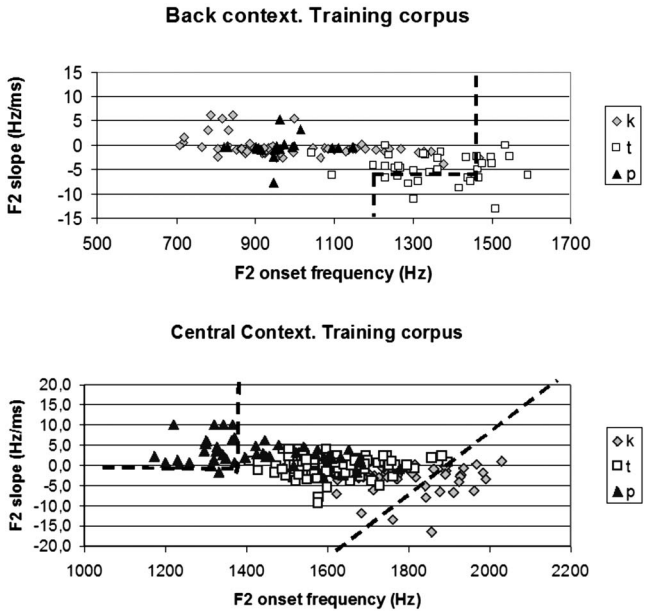


FIG. 5. Representation of stop consonants as a function of the following vowel F2 onset frequency (Hz) and formant slope (Hz/ms). Data are separated according to the class of the following vowel (back and central contexts).

hypothetic case in spite of five possible cases in the training corpus. Moreover, it did not impede the definition of a selective cue for /t/.

Selective cue for the place of /t/: Back context. Strongly falling slope associated with a high onset frequency or the presence of a very high onset.

If we selected consonants with onsets higher than 1200 Hz and slopes lower than −6 Hz/ms, as well as consonants with very high onsets (above 1450 Hz), the cue fired up in 46% of the cases, with no error.

Central context. According to the previous description, /k/ are characterized by falling slopes starting from high onsets and /p/ by rising slopes. Due to the existence of a dental locus at about 1600–1800 Hz, the transition slopes of /t/ are either slightly rising, when the vowel F2 is higher than the dental locus, or slightly falling when the vowel F2 is lower. The lower onsets should be found for (rising) labial transitions preceding vowels with a low F2. As a consequence, the onsets probably constitute an interesting cue to identify /p/ and /k/. Nevertheless, the effect of intervocalic coarticulation could change the direction of velar and labial slopes, as well as their respective onsets, and perturb the definition of distinctive regions for these consonants.

Data showed the existence of distinctive regions for /p/ and /k/. Indeed, in agreement with the above description, the lower onsets and most strongly rising slopes were exhibited by /p/, whereas the higher onsets and the most strongly falling slopes were exhibited by /k/ (Fig. 5, bottom). Nevertheless, there was a large overlap between the dental and the velar region and only velars characterized either by very high frequency onsets, or high onsets associated with a clearly falling slope could be distinguished from all the other consonants. Once more, the effect of intervocalic coarticulation appeared rather weak and did not impede the definition of

selective cues for /k/ and /p/. First, we did not see any case of intervocalic coarticulation radically changing the direction of the velar transition, i.e., generating a clearly falling slope. Then, we observed only four very slightly falling labial slopes in /ipœ/ sequences, possibly due to intervocalic coarticulation, in spite of numerous possible cases.

Selective cue for the place of /p/: Central context. The onset of F2 is low and the slope is rising.

If we selected onsets lower than 1350 Hz and positive slopes (higher than 0), the cue fired up in 37% of the cases.

Selective cue for the place of /k/: Central context. A very high frequency onset or a strongly falling slope coming from a high onset.

If we defined the distinctive velar region by the area at the right of the dotted line represented in Fig. 5, the cue fired up in 28% of the cases with no error. The line passes by the points (2000 Hz, 5 Hz/ms) and (1600 Hz, −20 Hz/ms).

2. Exclusion cues

Data presented in Fig. 5 clearly showed that some patterns never occurred for a given class of consonants, whereas they were observed for other classes, which allowed us to define exclusion cues. As we did for the positive cues, we defined the cues for CV contexts then derived VC cues from this set. As can be seen from Fig. 5, transition values for /t/ in back context exhibited little dispersion and, more importantly, no spreading in typical labial and velar regions. More precisely, /t/ had neither low onsets (under approximately 1050 Hz) nor positive slopes. Similarly, in central context, velar transitions had neither low onsets (under approximately 1600 Hz) nor positive slope. Finally, we could also exclude /p/ when we observed high onset frequencies and strongly falling slopes (back and central contexts). As we did for all the cues, safety margins were applied to make them more resistant to other sources of variation than the ones present in the training corpus. For all the cues, except the dental one, we have kept only the formant onset criterion.

Exclusion cue for the place of /t/: Back context. Low onsets (under 950 Hz) or rising slopes (above 5 Hz/ms).

Exclusion cue for the place of /k/: Central context. Low onsets (under 1500 Hz).

Exclusion cue for the place of /p/: Back and central contexts. High onsets (above 1800 Hz in central context, above 1300 Hz in back context).

The correct exclusion rates (R3, "additional information rate") reached about 50% of the cases for the exclusion of /t/ from /p/ and from /k/ in back context, as well as the exclusion of /k/ from /p/ in central context. The consonant /p/ was correctly excluded from velars (central context) and dentals (back context) in about 20% of the cases.

3. Concluding remarks for transition cues

It clearly appeared from the data of the training corpus that the three most typical F2 formant transitions allowed the definition of selective cues for the place of identification of unvoiced stops. The identification scores obtained with positive selective cues varied from 28% to 46% according to the context and the place, with no error (see Sec. IV B 1). The correct firing rate of the three positive selective cues defined from these transitions varied from 28% to 46% according to the context and the place, with no error. The weak effect of the transconsonantal vowel, which contrasted with Öhman's results, facilitated the definition of selective cues from transitions. The composition of our corpus, made up of voiceless stops, which were probably less subject to coarticulation with vowels than the voiced ones (Fant, 1973), could explain the difference between our results and those of Öhman, based on voiced stops. The validity of the cues for right transitions, as well as their relevance for left ones, was tested on new data (Sec. V).

TABLE III. Correct identification, correct exclusion, error, and overall information rates for burst cues in back, central front-rounded (frontR.) contexts, and across all contexts. See Sec. II C 2 for the method of evaluation. The place excluded from identification as well as the place wrongly identified are indicated in superscript. As an example, in the first table, "14^{t}" indicates that the dental place has been excluded from identification in 14% of cases for the realization of the consonant /k/. The last table displays the global results as a function of the corpus; the data for the identification of /t/ and /k/ only are given between parentheses (no cue was defined for the identification of /p/).

	Positive		Negative		
Burst	Id.	Err.	Exc.	Err.	Ov
P-back	⋯	0	34^{k}	0	34
T-back	38	0	25^{k}	0	63
K-back	42	0	14^{t}	0.9^{k}	56
P-central	⋯	0.2^{t}	⋯	⋯	⋯
T-central	16	0.4^{k}	⋯	⋯	⋯
K-central	19	0.4^{t}	⋯	⋯	⋯
P-frontR.	⋯	0	5^{t},5^{t}	0	10
T-frontR.	50	0	14^{k}	0	64
K-frontR.	35	0	13^{t}	0	48
C2b	18 (27)	0	9	0.4	27
C3	21 (33)	0.17	10.5	0.05	31.5
Global	20 (31.5)	0.13	10	0.13	30

V. RESULTS

In this section, we will first comment on the results for the set of selective cues provided by each speech segment (the burst, the right, and left transitions), and then all the cues together. The evaluation methodology is explained in Sec. II.

A. Cues provided by the burst

These cues were tested on the second and third corpora (see Tables I and II). Table III shows the identification, exclusion, and error rates obtained by positive and negative cues, for each unvoiced stop place, and each vocalic context. The realizations of both corpora have been merged, except for the global results.

As mentioned in the protocol section, the segmentation of the transient was entirely automatic for the second corpus, whereas the segmentation proposed by the detector was verified for the third corpus, which was recorded in an office.

TABLE IV. Number of data for testing the transition cues on the right and left hand sides of the consonant (corpus C2.b). B and C indicate the vocalic context (back or central).

	Right B	Left B	Tot B	Right C	Left C	Tot C	Tot
P	41	33	74	51	36	87	161
T	79	35	114	139	77	216	330
K	54	47	101	44	16	60	161
Tot	174	115	289	234	129	363	652

The automatic segmentation was corrected only when noise or small amounts of echo were visible in the signal; about 10% of the segmentation marks were modified. We did not correct the errors due to the presence of a parasite noise inside the stop closure if the speech signal was clean elsewhere. This led to three false alarms, which were taken into account in the error rates.

We have three main observations to make about the results. The first one concerns the low error rates, which amounted to 0.26% when we added the scores for positive (0.13%) and negative cues (0.13%). These rates are in agreement with our requirements. Concerning positive cues, designed for unvoiced stop place identification, there was no error in rounded contexts and very few errors in the central one (from 0.2% to 0.4%).

The second observation concerns correct identification rates. Across all the data, the identification rate was 20%. The scores obtained for /t/ and /k/ only (no cue was defined for /p/) were 33% and 27%, respectively, for the third corpus and the second corpus. Results obtained for the third corpus were particularly interesting since this corpus was new with regard to the corpora retained for the training phase (different speakers, different kinds of sentences and recording conditions). These results tend to prove that the cues are relevant, at least for isolated sentences.

The third remark concerns the effect of the context. Results confirmed the observations made from the training corpus, in particular, the higher performance of burst positive cues in rounded contexts—with scores varying from 35% up to 50%—than in central context (scores under 20%). In rounded contexts, the overall information rate (R4 in Sec. II C 2) for /t/ and /k/ varied from 48% to 64% depending on the context and the consonant.

B. Cues provided by the transitions

Left and right transition cues were tested on the second corpus (C2b), each kind of transition in its own vocalic context (Table IV).

The detection of formant onsets and offsets as well as the estimation of slopes at the boundaries between vowels and consonants were entirely automatic. Sometimes, the detector failed to find the formants and the cues could not be evaluated. This was essentially the case in the back context because the second formant of back vowels /u,o/ is not very intense and thus difficult to detect. This happened for 18% of the consonants in this context; most failed cases were observed on short vowels with very rapid transitions. All the data, whether the formant was detected or not, were taken into account in the results (Table V).

All (four) error rates concerning either positive or negative cues for left and right transitions, across the place and the context, were below 1%, which is in agreement with our requirements. The global identification rate, including contexts for which no cue was defined, were 23% for right tran-

TABLE V. Correct identification, correct exclusion, error, and overall information rates for left and right transition cues in back and central (cent.) contexts. For the meaning of superscripts, see Table IV..

	Positive		Negative		
Right Tr	Id.	Err.	Exc.	Err.	Ov
P-back	⋯	0	32^{t}	0	32
T-back	63	0	16^{p}	1.3	79
K-back	⋯	0	6^{p}, 37^{t}	0	43
P-cent	33	0	50^{k}	0	83
T-cent	⋯	0	8^{p}, 8^{k}	0	16
K-cent	61	0	32^{p}	0	93
Global	23	0	27	0.3	50
Left Tr	Positive		Negative		Ov
	Id.	Err.	Exc.	Err.	
P-back	⋯	0	18^{t}	0	18
T-back	63	0	6^{p}	0	69
K-back	⋯	2^{t}	77^{t}, 2^{p}	0	81
P-cent	36	0	⋯	3^{p}	36
T-cent	⋯	0	36^{p}	0	36
K-cent	⋯	6^{p}	31^{p}	0	37
Global	15	0.8	31	0.4	46

TABLE VI. Rates obtained with the whole set of cues—positive and negative—provided by the burst and the transitions. From top to bottom: identification rates, overall information rates—number of realizations for which at least one cue is triggered, leading to a correct identification of their place and/or a correct exclusion of another place—error rates, and global results.

Id.	Back	Centr.	Front
P	20	41	13
T	77	17	55
K	26	68	25
Ov.	Back	Centr.	Front
P	68	84	53
T	91	40	73
K	93	96	50
Err.	Back	Centr.	Front
P	0	1.8	0
T	0	0	0
K	3.7	0	0
Global	Id.	Err.	Ov.
/P,T,K/	38	0.7	70

sitions and 15% for left transitions, with overall information rates of 50% (right transitions) and 46% (left transitions). The correct identification rates achieved by the different cues varied from 33% to 63% according to the context and the place. We believe this is an encouraging result, given formant detection problems and the removal of F3 for the cue definition. These results tended to confirm the weak effect of the transconsonantal vowel, observed from the data of the training corpus.

There was no training for left transition cues, which were derived from the right ones. The identification rates achieved by left cues were very close to those obtained with the right ones for the same context and the same consonant. Nevertheless, all but one of the errors came from left transitions, so the inclusion of special training for left transitions would probably enhance the results.

C. Combination of all the cues

We report here the results obtained when all the cues—left and right transition cues, as well as burst cues—were simultaneously tested on each V-stop-V sequence, on the corpus C2b. Table VI gives the rates obtained for the three consonant places as a function of the right vocalic context. As an example, the result for the sequence /atu/ is presented in the back vocalic context, although the cues triggered on the left transitions were specific to the central context. Since we did not define any transition cues for front-rounded vowels, results for the front-rounded context were only based on burst cues and, when the preceding vowel was back or central, left transition cues.

The simultaneous firing of all the cues led to possible conflicts between them as well as redundancy. We will first consider these cases before commenting on the identification and the error rates. There was often more than one positive and one negative cue defined in a given context to identify or exclude a given place. We observed few cases of conflict (about 1.5% of the consonant realizations received conflicting decisions), which, interestingly, removed all but one of the errors made by transition cues. Let us recall that, in the cases of conflict, the judgment made by the cues was not taken into account.

The information provided by the system was sometimes redundant, which is proof of its richness. In order to give an idea of this redundancy, let us draw attention to the number of positive cues per realization. There was more than one positive cue per realization in 15% of the cases. This score reached 42% in the back context for the consonant /t/. In this context, there were three positive cues per realization (including positive cues due to double exclusions) in 10% of the cases. Otherwise, the activation of three positive cues simultaneously happened in 3% of the cases. Finally, the number of double exclusions leading to an identification reached 8% of the cases; most of them being triggered at the same time as another positive cue.

The overall information rates (rate R4 in Sec. II C 2) gave an idea of the (nonredundant) information provided by the set of cues. At least one cue, positive or negative, was triggered in 70% of all the realizations. The global identification and error rates were 38% (due to positive cues essentially) and 0.7% (including positive and exclusion cues); they are discussed in the next section. Identification scores greatly varied with the vocalic context and the place of articulation. Among the best results (more than 50%), we found consonants that were essentially identified by burst cues (/t/ in front rounded context) or transition cues (/k/ in central context). The frequent activation of both kinds of cues led to an identification rate of 77%—the highest score—for /t/ in back context. Left transition cues increased the results in most contexts and allowed the identification of consonants for which no cue was defined at the right hand side (such as labials followed by rounded vowels). These results tended to show that the search for reliable information should take into account the various segments of stop consonants and determine their respective efficiency as a function of the vocalic context. This is in agreement with the study by Dorman and Raphael (1980) concerning the distribution of acoustic cues for stop places in VCV syllables.

VI. DISCUSSION AND CONCLUDING REMARKS

This study has shown that it is possible to find distinctive (nonoverlapping) acoustic regions for each French unvoiced stop place and thus identify certain realizations of this feature in a highly reliable way. We defined a set of acoustic cues corresponding to these regions, called selective cues, which were automatically detected for sentences produced in quiet conditions, through detectors designed for stop place identification. The efficiency of selective cues varied with the speech segment, the vocalic context, and the place under consideration. Thus, a fine exploitation of all the information provided by speech sounds appeared necessary to maximize selective identification rate.

We used automatic acoustic detectors instead of extracting acoustic cues by hand for two reasons. First, we did not want to introduce any bias linked to our "phonetic judgment"

by correcting or editing formant trajectories or burst characteristics by hand. Using automatic acoustic detectors thus guaranteed the impartiality of the data processed to identify stop places. Second, using automatic detectors enables the estimation of a lower boundary of the performance that can be achieved through selective cues. The validation of the selective cues is thus all the more relevant since future improvements of automatic formant tracking and burst analysis algorithms will probably give rise to better identification results. This could allow us to exploit more cues, such as F3 transitions. We will continue our research in this direction.

When all the cues, provided by the burst transients and the transitions, were simultaneously tested, the overall identification and error rates were 38% and 0.7%, and the overall information rate, which took into consideration the firing rate of positive and negative cues, is 70%.

The identification rate obtained in this study (38%) is not easy to appreciate since, to the best of our knowledge, there is no other study concerning selective identification. Phonetic data found in the literature show the existence of very large overlapping areas between acoustic regions coding different speech sound categories [see, for example, the study by Öhman (1966) concerning stops produced by one speaker in VCV sequences]. For sounds extracted from sentences and produced by different speakers, one should expect important overlapping areas, which would limit the firing rate of selective cues. Note that this rate is also limited by the automatic nature of the detections.

As discussed in the Introduction, both the identification and the error rates obtained in this study were expected to be lower than those obtained by other acoustic decoding systems. Indeed, our aim is to identify only exemplars with highly reliable cues. Therefore, the identification achieved here is of a very specific nature and is voluntarily restricted to some exemplars of a class (those appearing in distinctive regions). Classical stochastic systems used in ASR try to identify every exemplar of a class. By doing so, they take a limited risk on every exemplar, including those that are not the object of our study. Most recent work about phonetic decoding reaches a phonetic accuracy between 75% and 76% (Deng *et al.*, 2005; Glass, 2003). It should be noted that these results are often obtained by adding a phone bigram model which slightly improves the intrinsic performances of the stochastic models. More specific studies have been devoted to the identification of presegmented stop consonants. That of Ali *et al.* (2001), dedicated to the identification of American stop consonants, has been evaluated on a subset of the Timit database. The hard-decision algorithm they designed within the framework of a knowledge-based approach exploits burst characteristics and formant information derived from an auditory-based front end. Unlike our approach, this algorithm does not rely on selective information and thus yields an average accuracy of 86% for the place of /p,t,k/. However, it is the level of error that constitutes the crucial test of our system, since it guarantees selective cues reliability. The error rate obtained by acoustic decoding systems, i.e., approximately 14%, although satisfying with respect to their objective (ASR), is clearly not acceptable with respect to our own objective. We believe that the error rate obtained by the set of selective cues proposed in this study (0.7%) clearly speaks for its reliability.

The interest of highly reliable cues could also be evaluated in relation with their potential exploitation. In the domain of automatic speech recognition, the most time consuming step corresponds to the exploration of all the possible lexical solutions for a sentence. Selective cues will favor the early emergence of the best lexical and sentence solutions. Guiding the exploration of lexical solutions by exploiting the selective cues spotted in the input sentence can thus give rise to a very substantial reduction of the time spent.

The enhancement of the acoustic structure of speech sounds has been investigated in the design of hearing aids and second language learning as well (Hazan and Simpson, 1998). One of the crucial points is to prevent the system from introducing perceptive biases in the transformed signal, which means that only speech sounds presenting salient acoustic cues should be enhanced. From this point of view, selective cues present a double interest. Their very reliable detection guarantees that very few acoustic artifacts will occur in the transformed signal. On the other hand, the enhancement of speech sounds presenting selective acoustic cues, i.e., more salient that other sounds, probably represents a fruitful direction of research.

The applications linked to selective cues would benefit from the extension of our work, devoted to the places of articulation of unvoiced stops, to other features. We have proposed a protocol that can be applied to any phonetic feature. The definition of selective cues could probably be extended to the places of voiced stops, with adaptation. We are convinced that they could also be efficient for the places of articulation of fricative sounds, often clearly indicated by the characteristics of fricative spectra.

Ali, A., der Spiegel, J. V., and Mueller, P. (**2001**). "Acoustic-phonetic features for the automatic classification of stop consonants," IEEE Trans. Acoust., Speech, Signal Process. **9**, 833–841.

Blumstein, S. E., and Stevens, K. N. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.

Bonneau, A., Djezzar, L., and Laprie, Y. (**1996**). "Perception of the place of articulation of french stop bursts," J. Acoust. Soc. Am. **100**, 555–564.

Calliope (**1989**). *La Parole et Son Traitement Automatique (Speech and Its Automatic Processing)* (Masson, Paris).

Carré, R., Descout, R., Eskénazi, M., Mariani, J., and Rossi, M. (**1984**). "The french language database: defining, planning and recording a large database," Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing 1984, Vol. 9, pp. 324–327, San Diego.

Coste-Marquis, S. (**1994**). "Interaction between most reliable acoustic cues and lexical analysis," Proceedings of International Conference on Spoken Language Processing, Yokohama, Japan, Vol. **4**, pp. 2187–2190.

Davis, K., and Kuhl, P. K. (**1992**). "Best exemplars of english velars stops: a first report," Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Canada, Vol. **1**, pp. 495–498.

Delattre, P., Liberman, A., and Cooper, F. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Deng, L., Yu, D., and Acero, A. (**2005**). "A generative modeling framework for structured hidden speech dynamics," Neural Information Processing System (NIPS) Workshop on Advances in Structured Learning for Text and Speech Processing, Whistler, BC, Canada.

Dorman, M., and Raphael, L. (**1980**). "Distribution of acoustic cues for stop consonant place of articulation in vcv syllables," J. Acoust. Soc. Am. **67**, 1333–1335.

Edwards, T. J. (**1981**). "Multiple features analysis of intervocalic English

plosives," J. Acoust. Soc. Am. **69**, 535–547.
Fant, G. (**1973**). "Stops in CV syllables," in *Speech Sounds and Features* (The MIT Press, Cambridge).
Fischer-Jörgensen, E. (**1954**). "Acoustic analysis of stop consonants," Miscellenea Phonetica **2**, 42–59.
Glass, J. (**2003**). "A probabilistic framework for segment-based speech recognition," Comput. Speech Lang. **17**, 137–152.
Halle, M., Hughes, G., and Radley, J.-P. (**1957**). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107–116.
Hazan, V., and Simpson, A. (**1998**). "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Commun. **24**, 211–226.
Iverson, P., and Kuhl, P. (**1995**). "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling," J. Acoust. Soc. Am. **97**, 553–562.
Jackobson, R., Fant, G., and Halle, M. (**1952**). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates* (MIT Press, Cambridge, MA).
Kewley-Port, D. (**1982**). "Measurements of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. **72**, 379–389.
Krull, D. (**1990**). "Relating acoustic properties to perceptual responses: a study of Swedish voiced stops," J. Acoust. Soc. Am. **88**, 2557–2570.
Kuhl, P. K. (**1991**). "Human adults and infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not," Percept. Psychophys. **50**, 93–107.
Lahiri, A., Gewirth, L., and Blumstein, S. (**1984**). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391–404.
Lamel, L. (**1988**). "Formalizing Knowledge used in Spectrogram Reading: Acoustic and Perceptual Evidence from Stops," Ph.D. thesis, Massachussets Institute of Technology.
Laprie, Y., and Berger, M.-O. (**1996**). "Cooperation of regularization and speech heuristics to control automatic formant tracking," Speech Commun. **19**, 255–270.
Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**, 1683–1692.
Liu, S. (**1996**). "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am. **100**, 3417–3430.
Lotto, A., Kluender, K., and Holt, L. (**1998**). "Depolarizing the perceptual magnet effect," J. Acoust. Soc. Am. **103**, 3648–3655.
Ohde, R., and Sharf, D. (**1977**). "Order effect of acoustic segments of VC and CV syllables on stop and vowel identification," J. Speech Hear. Res. **20**, 543–554.
Öhman, S. (**1966**). "Coarticulation in VCV utterances: Spectrographic measurements," J. Acoust. Soc. Am. **39**, 151–168.
Stevens, K. (**2000**). "Diverse acoustic cues at consonantal landmarks," Phonetica **57**, 139–151.
Stevens, K. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.
Suomi, K. (**1985**). "The vowel-dependence of gross spectral cues to place of articulation of stop consonants in cv syllables," J. Phonetics **13**, 267–285.
Wajskop, M. (**1979**). "Segmental durations of french intervocalic plosives," in *Frontiers of Speech Communication Research*, edited by B. Lindblöm and S. Öhman (Academic, London), pp. 109–134.
Zue, V. (**1976**). "Acoustic characteristics of stop consonants: a controlled study," MIT Technical Report No. 523.

Acoustic characteristics of English lexical stress produced by native Mandarin speakers

Yanhong Zhang
Program in Linguistics, Purdue University, West Lafayette, Indiana 47907

Shawn L. Nissen
Department of Communication Disorders, Brigham Young University, Provo, Utah 84602

Alexander L. Francis[a)]
Program in Linguistics and Department of Speech, Language and Hearing Sciences, Purdue University, West Lafayette, Indiana 47907

(Received 20 June 2007; revised 22 February 2008; accepted 27 February 2008)

Native speakers of Mandarin Chinese have difficulty producing native-like English stress contrasts. Acoustically, English lexical stress is multidimensional, involving manipulation of fundamental frequency (F0), duration, intensity and vowel quality. Errors in any or all of these correlates could interfere with perception of the stress contrast, but it is unknown which correlates are most problematic for Mandarin speakers. This study compares the use of these correlates in the production of lexical stress contrasts by 10 Mandarin and 10 native English speakers. Results showed that Mandarin speakers produced significantly less native-like stress patterns, although they did use all four acoustic correlates to distinguish stressed from unstressed syllables. Mandarin and English speakers' use of amplitude and duration were comparable for both stressed and unstressed syllables, but Mandarin speakers produced stressed syllables with a higher F0 than English speakers. There were also significant differences in formant patterns across groups, such that Mandarin speakers produced English-like vowel reduction in certain unstressed syllables, but not in others. Results suggest that Mandarin speakers' production of lexical stress contrasts in English is influenced partly by native-language experience with Mandarin lexical tones, and partly by similarities and differences between Mandarin and English vowel inventories.
 [DOI: 10.1121/1.2902165]

PACS number(s): 43.70.Fq, 43.70.Kv [AL] Pages: 4498–4513

I. INTRODUCTION

Adults who learn a second language (L2) are seldom able to speak that language without accent. Although the degree of an accent is related to many factors such as age and language environment, the primary influence on the nature of an individual's accent is the sound system of their native language (L1) (Flege and Hillenbrand, 1987; Lord, 2005; Piske *et al.*, 2001; Tahta and Wood, 1981). The interference of native phonetics and phonology on the acquisition of non-native vowels and consonants has been studied extensively, and results typically suggest that L2 learners have relatively greater difficulty perceiving and producing non-native contrasts that involve phonetic features dissimilar to those used in their native language. Similar difficulties in L2 acquisition have been identified in suprasegmental domains as well. For example, native Mandarin speakers learning English as a second language have been repeatedly shown to have difficulties producing English lexical and/or sentential stress, and it has been argued that this difficulty may result in large part from the influence of native suprasegmental (tonal) categories (Archibald, 1997; Chen *et al.*, 2001a; Juffs, 1990; Hung, 1993). However, most research in this area has focused on impressionistic observations rather than acoustic analysis (with the notable exception of Chen *et al.*, 2001a) and often confounds the phonological issue of stress placement with the phonetic problem of native-like stress production.

Here we attempt to dissociate the question of whether, or to what degree, non-native speakers are able to apply phonological rules of stress placement, in order to focus on the question of whether they are able to correctly produce the phonetic properties that correlate with the English stress contrast under conditions in which they know unambiguously where stress is to be placed. Thus, we ask whether Mandarin speakers are capable of producing native-like patterns of fundamental frequency, intensity, duration and vowel formant frequencies associated with English stressed and unstressed syllables when there is no question of their knowing *where* to place stress. An inability to correctly produce these acoustic correlates of English stress under these circumstances would suggest that their native language experience with producing (and possibly perceiving) the specific acoustic cue patterns related to Mandarin phonetic categories (tonal and/or segmental) interferes with their ability to produce qualitatively different patterns of these same cues in the service of producing English lexical stress distinctions.

[a)]Author to whom correspondence should be addressed. Electronic mail: francisa@purdue.edu

 0001-4966/2008/123(6)/4498/16/$23.00

A. English stress

A number of studies have explored the acoustic correlates of lexical stress in American English (Beckman, 1986; Bolinger, 1958; Campbell and Beckman, 1997; Fry, 1955, 1958, 1965; Lieberman, 1960, 1975; Sluijter and van Heuven, 1996; Sluijter *et al.*, 1997). Most of these studies focused on lexical stress in English disyllabic words in which the location of stress on the first or second syllable led the word to be identified as either a noun or a verb, respectively. Results of these studies consistently indicate that the acoustic correlates of average fundamental frequency (F0), intensity, syllable duration, and vowel quality are associated with the perception and production of English lexical stress: Stressed syllables have higher F0, greater intensity, and longer duration than unstressed syllables. Moreover, recent research suggests that the *alignment* of F0 events with respect to segments within a syllable may play an important role in both tonal and intonational categories (For intonation, cf. Arvaniti and Gårding, 2007; Atterer and Ladd, 2004; Grabe *et al.*, 2000; Mennen, 2004. For tone, see Xu, 1998, 1999; Xu and Liu, 2007, 2006) and it may be worth investigating this property in the production of stress as well. Although, to our knowledge, pitch peak alignment has not been implicated as a specific cue to the placement of lexical stress, misalignment of a pitch peak in a stressed syllable might contribute to the perception of non-nativeness in L2 speakers.

The precise measure of computing intensity is debated. Fry (1955, 1958) and Beckman (1986) identified average intensity over the syllable as a possible acoustic correlate of stress differences, while others (Sluijter and van Heuven, 1996; Sluijter *et al.*, 1997) have argued that spectral tilt (differences in intensity over the frequency spectrum of a given vowel) is a more appropriate measure. Since both measures are associated with increased vocal effort (Liénard and Di Benedetto, 1999; Traunmüller, 1989), it is possible that either may serve as acceptable correlates of the English stress contrast. However, since measurement of spectral tilt is highly dependent on the height or location of the first formant (F1), it is not possible to compare spectral tilt across vowels differing in quality (formant frequencies), as between reduced (unstressed) and unreduced (stressed) versions of the same vowel, so in the current study only average intensity was used.

Finally, the process of vowel reduction has been consistently identified as a correlate of the English lexical stress contrast. Although this feature has not been extensively examined in cross-language studies, many researchers have discussed its importance in general terms. For example, nonnative speakers' use of unreduced vowels in unstressed syllables has been argued to "contribute importantly to foreign accent" (Flege and Bohn, 1989) and is an "extremely typical" phenomenon in Spanish-accented English (Hammond, 1986). Fokes *et al.* (1984, 1989) and Flege and Bohn (1989) also concluded that the inability of L2 speakers to perform appropriate vowel reduction contributed to their non-native-like production of English, although the two articles differed in their assessment of the relative importance of vowel reduction in cuing the perception of native-like stress. Fokes *et al.* (1984) suggested that the inability of L2 learners to reduce the vowel in unstressed syllables could influence their ability to manipulate other phonetic correlates of English lexical stress, resulting in poorer performance on lexical stress production tasks. In contrast, Flege and Bohn (1989) argued that L2 learners of English first learn to produce stressed vs unstressed syllables contrasting in duration and intensity, and only subsequently learn (or fail to learn) to correctly reduce the vowels in unstressed syllables. Either way, vowel quality is clearly an important acoustic correlate of stress (Beckman, 1986; Fry, 1965) and failure to appropriately reduce unstressed vowels may contribute to the perception of a non-native accent (Fokes *et al.*, 1984; Flege and Bohn, 1989; Lee *et al.*, 2006).

B. Mandarin lexical tone

Unlike English, Mandarin is a tonal language. There are four lexical tones in Mandarin: tone 1 (high-level), tone 2 (high-rising), tone 3 (dipping), and tone 4 (high-falling). Tone, like stress in English, can distinguish word meaning independently of segmental properties. Some scholars have argued that Mandarin exhibits linguistic characteristics that are similar to lexical stress. For instance, syllables carrying the so-called neutral tone, which is usually found in syntactic particles within lexical units of two or more syllables, have been found to be less prominent than syllables carrying the four basic lexical tones (Chao, 1968; Chen and Xu, 2006).

Many studies have focused on the acoustic examination of Mandarin tones (Howie, 1976; Fu *et al.*, 1998; Gandour, 1978, 1983; Liu and Samuel, 2004; Whalen and Xu, 1992). In general, these studies have demonstrated that F0 is the primary acoustic cue for Mandarin tones, but that syllable duration and amplitude contour vary consistently across lexical tone categories. For example, the falling tone (tone 4) is typically much shorter than the other tones, especially the first tone (high level) which is typically quite long. Similarly, the third (dipping) tone is long, but also exhibits a mid-syllable decrease in amplitude. Perceptual research has shown that these non-pitch cues can also function as acoustic cues to Mandarin tones in the absence of F0 information (Fu *et al.*, 1998; Liu and Samuel, 2004; Whalen and Xu, 1992). Thus, based on their experience with controlling the F0, duration, and intensity of individual syllables to express lexical tone distinctions, from a purely phonetic perspective, it is possible that Mandarin speakers may be able to control these same acoustic properties to produce native English-like lexical stress contrasts.

This seems unlikely, however, as research on cross-language perception and L2 production of speech sounds clearly indicates a strong influence of the native *phonological* system on the perception and production of non-native sounds, and only some Mandarin tones map clearly onto English intonational patterns (see Francis *et al.*, 2008, for discussion of cross-language mapping between Mandarin and Cantonese tones and English intonational categories). Interestingly, the specific nature of L1 category influence on L2 perception and production (in terms of facilitation or interference) also appears to depend in large part on the relative

degree of (phonetic featural) similarity between the native and non-native categories (Best, 1995; Flege, 1995; Flege and Davidian, 1985). For example, according to Flege's Speech Learning Model (SLM), the presence of one or more native categories that are phonetically similar to a non-native category may interfere with the perception and production or acquisition of that L2 category. In contrast, Best's Perceptual Assimilation Model (PAM) would predict improved perception of an L2 contrast if each sound is sufficiently similar to a different native category. Such a situation would result in *two-category* assimilation, whereby each sound in a non-native contrast is assimilated to a different native category. Even if both sounds of the L2 contrast are assimilated to the same native category, PAM predicts improved perception of the contrast if one of the two is more successfully assimilated (a case of a *category goodness* contrast).

More interestingly, according to PAM non-native sounds that are *uncategorizable* to any native phoneme category may be easy to discriminate perceptually (perhaps even more easily than for native speakers), while still being extremely difficult to produce in a native-like manner (Best *et al.*, 2001; Best *et al.*, 1988). However, this last possibility seems unlikely in the case of F0 patterns, since these, unlike clicks (the typical example of uncategorizable sounds) can easily be recognized as speech sounds. Still, depending on which theory one adopts, and, more importantly, on the specific degree of similarity between the native and the L2 category or categories, one might expect either an increase or decrease in ease of acquisition when an L2 category is determined to be similar to a non-native one along one or more phonetic dimensions. Although the SLM and PAM have traditionally been applied to production and perception of segmental phonemes, there is nothing about the models themselves that would necessarily restrict their predictions to the segmental domain, and either may be able to account for the acquisition of suprasegmental aspects of speech, such as intonation or stress.

C. Mandarin speakers' production of English stress

There is evidence that native Mandarin speakers have difficulty producing L2 English stress contrasts in a native-like manner. While it is possible that this difficulty arises from interference from the Mandarin sentential stress (intonational) system, existing evidence currently seems to suggest a strong interference from the Mandarin *tonal* system.[1] For example, Juffs (1990) reported errors made by native Chinese speakers who were college students and had little or no experience with spoken English outside the classroom. Many of these speakers' errors consisted of mistakes in stress placement, suggesting that they simply did not know what syllables required stress in the utterances they were asked to produce. However, even when stress was produced on the appropriate syllable, they showed evidence of difficulty with the phonetic manipulation of specific correlates of stress. For example, some speakers tended to use a falling tone to signal an English stressed syllable. The use of a falling tone, with its overall lower average F0, for a stressed syllable suggests that these speakers were not aware of the general association between English stress and higher (average) F0, but may instead have been overextending the English tendency to use a sharply falling F0 contour for strongly emphatic stress (as in "*Yes*, I *do*") (Chao, 1972). Alternatively, it is possible that they were correctly recognizing that the English stressed syllable should be produced with a higher *initial* F0 value—in other words, they were focusing on the location of a pitch peak, rather than on an average syllable value (see discussion of F0 peak location, below). In contrast, other speakers did achieve an overall higher average pitch in stressed syllables, but also lengthened these syllables much more than was necessary. This suggests that these speakers simply superimposed *all* properties of the Mandarin high tone onto the English stressed syllable (including its association with very long syllable duration), rather than simply producing an overall higher average F0. Taken together, these results suggest that, even when Mandarin speakers know which syllable to stress, they may do so by transferring production patterns from their native tonal inventory.

To control for Mandarin speakers' lack of knowledge about where stress is to be placed, Chen *et al.* (2001a) examined the production of English sentence stress under conditions in which the speaker was clearly aware of the proper location of stress. They found that native Mandarin speakers employed many of the same acoustic correlates of stress as English speakers, including duration, amplitude, and fundamental frequency, but their use of these correlates was significantly different from American speakers. For example, Mandarin speakers produced stressed words with higher F0 compared to English speakers. Chen *et al.* (2001a) argued that this was a result of Mandarin speakers' native language experience, since Mandarin typically exhibits a much greater range of pitch fluctuation during the course of a sentence than does English. Thus, Mandarin speakers are used to producing high pitches at a higher point in their average pitch range than are English speakers, and this tendency transfers to the L2 as well. Although their results regarding F0, duration and intensity are very informative, Chen *et al.* (2001a) did not examine the possible influence of native phonology (whether tonal or segmental) on the production of L2 vowel quality as a cue to English stress.

The investigation of vowel quality is central to the present study, unlike previous work, because it is in this domain that we may begin to distinguish between interference that results from the fundamental difference between tone and stress *systems* and interference that arises from incomplete or inaccurate acquisition of individual lexical items. Interference of a systematic origin should be relatively uniform across lexical items, for example, leading to a uniform lack of vowel reduction or, conversely, a tendency to overgeneralize a principle of vowel reduction in unstressed syllables. Interference that arises on an item-by-item basis should, in contrast, be much more variable across items (Flege and Bohn, 1989).

The present study focused on three factors involved in the production of stress: (1) the acoustic correlates used by Mandarin and English speakers to indicate lexical stress placement in English, including F0, duration, intensity and vowel quality; (2) differences between the two groups in

terms of their use of these features; (3) the degree to which Mandarin speakers' pattern of acoustic correlate production can be explained by the structure of their native language phonology (both suprasegmental and segmental).

II. METHODS

A. Subjects

Two groups of speakers participated in this experiment: ten native speakers of American English (five women, five men) and ten native speakers of Mandarin Chinese (five women, five men). English participants ranged in age from 21 to 28 years of age ($M=25$), while Mandarin speakers were 26–35 years of age ($M=32$). The English speakers were all native residents of the United States (U.S.), while the Mandarin speakers were all originally from the People's Republic of China (PRC) and had lived in the U.S. for three to four years prior to participating in the experiment. All participants were recruited from within Purdue University community (West Lafayette, IN) and had normal hearing, speech, and language ability by self-report.

None of the Mandarin speakers had any English-immersion experience before arriving at Purdue University; all of their prior English experience was obtained in class while in China. None was enrolled in an English language department or school in China, although eight reported having had native English speakers as college English teachers at some point in their education. Since coming to the U.S., all Mandarin speakers had been exposed primarily to Midwestern dialects of American English. Of the American English speakers, seven were from the central Midwest (six from Indiana and one from Ohio, one of whom also spoke American-African English). There was also one American English speaker from each of California, New York, and Louisiana.

B. Stimuli

Seven pairs of disyllabic words were selected following the methodology of Beckman (1986) and Fry (1955, 1958). Each word pair consisted of a noun and a verb that had identical spelling forms and differed only in terms of stress placement (noun: stress on initial syllable; verb: stress on final syllable). These stimulus pairs were formed from the following corpus of word forms: *contract*, *desert*, *object*, *permit*, *rebel*, *record*, and *subject*. Each target word was elicited in isolation and in the semantically neutral frame sentence *I said __ this time* and was accompanied by associated context sentences created specifically for each word, which are shown in Table I.

Based on the work of Peterson and Barney (1952), ten familiar English words (*beat*, *bit*, *bet*, *bat*, *bought*, *father*, *bird*, *butt*, *put, boot*) were used to map English vowel spaces by native English speakers of America and by native Mandarin speakers of English. Similarly, a list of Chinese characters was selected for mapping the Mandarin speakers' Mandarin vowel space, as shown in Table II.

TABLE I. Stimuli and context sentences to aid in establishing the stressed syllable.

Target word	Noun/verb	Context sentence
Contract	noun	Mr. Smith has finally agreed to sign the new *contract*.
	verb	Will steel *contract* when it is cooled?
Desert	noun	They got lost in the *desert*.
	verb	Will he *desert* his team?
Object	noun	What is the *object* on the table?
	verb	They won't *object* to your decision.
Permit	noun	In order to park here, you need a *permit*.
	verb	Would you *permit* her request?
Rebel	noun	The *rebel* army did this.
	verb	They *rebelled* at this unwelcome suggestion.
Record	noun	Can I get a copy of my health *record*?
	verb	She *recorded* all songs her daughter sang yesterday.
Subject	noun	What is the *subject* of this sentence?
	verb	Must you *subject* me to this boring twaddle?

C. Procedure

Prior to recording, participants were asked to fill out a language background questionnaire. All recordings took place in a single-walled sound-attenuated booth and were made using a digital audio recorder (SONY DAT, TCD-D8), Studio V3 amplifier, and a unidirectional Hypercardiod dynamic microphone (Audio-Technica D1000HE).

The microphone was placed approximately 20 cm from the speaker's lips at an angle of 45° (horizontal) during recording. The speech tokens were sampled at a rate of 44.1 kHz with a quantization of 16 bits and low-pass filtered at 22.05 kHz. Each token was then saved as an individual sound file and normalized to a RMS amplitude of 70 dB using Praat 4.3 (Boersma and Weenink, 2004).

All stimuli were presented to speakers on individual file cards organized into three sets. One set of cards showed each word (target or distracter) at the top with the corresponding context sentence and frame sentence below. The second set of cards showed only target words and corresponding context sentences. The third set of cards showed only the English words and Chinese words for mapping vowel spaces.

Speakers were instructed to speak naturally at a typical rate and loudness level. Each speaker first read the first set of cards, context sentence first then the frame sentence, twice for each card. Before the next reading, the experimenter explained to the speaker the rule that stress needs to be shifted between syllables when some English words shift from noun to verb (e.g., CONtract vs conTRACT). The need for this type of stress shift to differentiate noun from verb for some English words should be familiar to the participants, because it is part of the standard middle school English class curriculum in the PRC. For the second set of recordings, speakers read only the target words in isolation. Target word pronunciation was indicated by referring to the context sentence. Again, each card was read twice. This elicitation procedure yielded 1120 tokens (14 words × 2 contexts × 2 repetitions × 20 subjects). Only the 560 tokens produced in isolation were used in subsequent analyses (both instrumental and perceptual) since each production is assumed to represent the speaker's best attempt to produce stress on the appropriate

TABLE II. All monophthong and diphthong vowel phonemes involved in this experiment, including corresponding Chinese characters and English words used in the vowel space mapping task. Note: (ü) indicates that this transcription is used when vowel is produced in isolation. Transcriptions based on those in Duanmu (2000) with the substitution of the IPA symbol [ɐ] for [A].

Mandarin vowels (IPA)	Chinese character	PinYin romanization	Tone	English vowels (IPA)	English word
i	一	yi	55	i	beat
iɛ	爷	ye	25	*	
yɛ	约	yue	55	*	
iɐ	鸭	ya	55	*	
y	迂	yu (ü)	55	*	
*	*	*	*	I	bit
ei	悲	bei	55	ei	
*	*	*	*	ɛ	bet
*	*	*	*	æ	bat
ai	哀	ai	55	ai	
au	熬	ao	25	*	
*	*	*	*	ɑ	father
*	*	*	*	ʌ	butt
*	*	*	*	ɔ	bought
*	*	*	*	ɝ	bird
ɤ	鹅	e	25	*	
o	播	bo	55	*	
ou	欧	ou	55	ou	
*	*	*	*	ʊ	put
u	屋	wu	55	u	boot
uo	窝	wo	55	*	
uɐ	挖	wa	55	*	
ɐ	阿	a	55	*	

syllable (initial for nouns, final for verbs). Moreover, one production could not be analyzed, leaving a total of 559 stress-contrasting tokens. Finally, all speakers read the list of English vowel space-mapping words, and Mandarin speakers read the list of Chinese characters.

D. Acceptability rating

Subjective ratings of acceptability or accentedness are commonly used in the evaluation of a speaker's foreign accent (Flege, 1984, 1988; Southwood and Flege, 1999). Such ratings are obtained by asking native listeners to assign a numeric value to a segment of speech based on its perceived quality (Francis and Nusbaum, 1999; Schmidt-Nielsen, 1995). To determine the acceptability of each recorded token, a listening evaluation test was conducted. Five native English-speaking graduate students in the linguistics or English as a second language program of Purdue University served as paid consultants. Linguistically trained listeners

were selected because of the increased likelihood that they would be able to focus on stress characteristics alone, ignoring other possible non-native (segmental) pronunciations in the speech samples. Each listener evaluated the acceptability of each of the 559 tokens on five separate occasions over a two-week period. Words were presented randomly but blocked by speaker gender.

For each token, listeners first heard the word and were asked to determine which word was said. Both possible choices for each word (e.g., conTRACT or CONtract) were displayed on the screen prior to playing the sound and remained on the screen until a choice was made. After listeners identified the token a new screen appeared showing their choice (e.g., either CONtract or conTRACT) and asked them to provide a rating of acceptability on a scale from 1 (poor) to 5 (excellent). The sound was repeated after this second screen was displayed, but the screen did not clear until a choice had been selected. Token presentation and data collection was carried out using *E-prime* version 1.1 (Schneider *et al.*, 2002).

E. Acoustic measurements

Using Praat acoustic analysis software (Boersma and Weenink, 2004), the following acoustic parameters were measured for each token: syllable duration (in ms); average intensity (in dB); average fundamental frequency (F0, in Hz); time of F0 peak and the first and second formant frequencies (F1 and F2, in Hz). The parameters related to intensity and F0 were measured within a syllable, and the formant frequencies were measured within the vowel. Only F1 and F2 measures were used to map the speakers' vowel space.

Syllable and vowel boundaries were segmented according to the following criteria: (1) word/syllable 1 onset: The first upward-going zero crossing at the beginning of the waveform; (2) word/syllable 2 offset: The ending point of the sound waveform at the last downward-going zero crossing; (3) syllable 1 offset/syllable 2 onset: In words with a stop consonant as the onset of the second syllable (such as *rebel, contract, object, subject, record*), this was defined at the beginning of the silence of the stop gap. In words with no medial stop consonant (*permit, desert*), then the boundary was marked as the transition between the acoustic (spectrographic) pattern of the initial consonant of the second syllable and the segment immediately preceding it. Segmentation criteria were based on both waveform and spectrogram cues as described by Peterson and Lehiste (1960). Based on these segmentations, syllable and vowel durations were calculated in millisecond increments. In addition, for diphthongal vowels (i.e., in Mandarin), formant frequencies were measured twice, once for the initial vocalic portion and once for the final portion. For this purpose, the transition point between the two vowel segments was visually identified as the midpoint of the transition between the two steady states, or the midpoint between the initial formant frequencies and the final ones, in the absence of any steady state. Average formant values were calculated between the onset of the vowel and this midpoint (for the initial vocalic portion) and between this midpoint and the end of the vowel (for the final vocalic portion).

The average intensity measure was calculated as the mean of multiple intensity values extracted and smoothed over the number of time points necessary to capture the minimum predicted pitch of each individual participant. F0 measures were measured as the average value over the entire syllable, and were computed using a Hanning analysis window and the autocorrelation method described in Boersma (1993). When measuring F0, the pitch range for female talkers was set to 100–500 Hz and 75–300 Hz for male talkers, as recommended in the Praat manual. The time of the F0 peak was identified automatically from the F0 contour, and subsequently converted to a proportion of the syllable by reference to the syllable duration. F0 was remeasured manually (as the reciprocal of each manually identified period of the syllable's acoustic waveform) when the pitch contour was absent, or displayed incompletely or intermittently through the syllable, and when displayed F0 values were suspiciously high or low compared to the rest of that talker's utterances. In most cases, these display problems were due to the presence of glottalization, especially in unstressed syllables produced by female American English and male Chinese speakers.

A linear predictive coding (LPC) based tracking algorithm was used to determine formant calculations for the entire vocalic segment of interest [as implemented in the Praat Sound to LPC (burg) method]. The LPC analysis employed a 25 ms Gaussian window with +6 dB pre-emphasis over 50 Hz. These computed formant frequencies were then averaged across the entire vowel, or, in the case of the diphthong, across the initial or final portion of the diphthong, respectively. In order to quantify the property of vowel quality, we used two measures derived from the center frequencies of the first and second formants (F1 and F2) as described by Blomgren *et al.* (1998). The statistic compact-diffuse (C-D), calculated as the difference between F1 and F2 (F2 – F1), is correlated with the phonetic property of tongue height. High vowels such as [i] and [u] typically have a relatively large C-D value, while low vowels such as [a] have a smaller C-D value. The statistic grave-acute (G-A), calculated as the arithmetic mean of F1 and F2 [(F1 +F2)/2], is correlated with the phonetic dimension of tongue advancement (front/back), such that front vowels such as [i] or [æ] typically have a relatively small value of G-A, while back vowels such as [u] or [o] typically have relatively large values.

III. RESULTS

A. Acceptability ratings

Listeners correctly identified the majority of tokens produced by both English and Mandarin speakers. The five tokens that were identified incorrectly by more than two listeners were excluded from further analysis. The mean acceptability rating for each of the remaining tokens was then calculated only across raters who correctly identified the token (all but 11 tokens were correctly identified by all listeners), as shown in Table III. Raters were relatively uniform

TABLE III. Results of perceptual evaluation of productions by American English and Mandarin speakers. Note: Accuracy = proportion of correct identifications; Avg = mean acceptability rating across 3–5 raters (see text) on a five-point scale where 1 = poor and 5 = excellent; s.d. = standard deviation for each mean rating.

	English Speakers' Productions			Mandarin Speakers' Productions		
	Identification	Rating		Identification	Rating	
Word	Accuracy	Avg	s.d.	Accuracy	Avg	s.d.
Contract N	0.99	4.60	0.18	1.00	3.47	0.38
Contract V	0.99	4.43	0.51	0.98	2.77	0.74
Desert N	1.00	4.29	0.47	0.95	2.88	0.78
Desert V	0.99	4.51	0.19	0.99	3.19	0.60
Object N	1.00	4.44	0.29	0.97	3.17	0.59
Object V	0.99	4.30	0.45	0.97	3.22	0.45
Permit N	0.94	4.15	0.59	0.94	2.72	0.62
Permit V	0.96	3.97	0.53	1.00	2.67	0.47
Rebel N	1.00	4.64	0.22	0.98	1.88	1.14
Rebel V	0.97	4.32	0.67	0.99	3.14	0.27
Record N	0.99	4.59	0.30	0.98	2.89	0.85
Record V	0.93	4.19	0.97	0.99	3.14	0.32
Subject N	0.98	4.21	0.49	0.99	3.61	0.60
Subject V	0.97	4.07	0.51	0.98	2.98	0.43
Average	0.98	4.34	0.46	0.98	2.98	0.59

in their assessment of both the English and Mandarin utterances. The mean range between the lowest and highest acceptability rating for a given word was 1.8 overall (1.6 for English productions, 1.9 for Mandarin). The mean rating score for correctly identified words produced by Mandarin speakers was 2.98(SD=0.74, Mdn=3.04), while for the American group it was 4.34 (SD=0.53, Mdn=4.49). Most Mandarin speakers' productions were rated less than 3.5 (204 out of 277 tokens), but the majority of English speakers' productions scored higher than 4 (256 out of 276 tokens). A t-test showed that the rating difference between the two language groups was statistically significant, $t(551)=24.97$, $p<0.001$.

B. Acoustic analyses

To confirm the reliability of our acoustic measurements, 10% of all tokens (56) were selected for independent reanalysis by a second judge who was naive to the purpose of the experiment. Across raters, mean formant values differed by at most 25 Hz for F2 and 12 Hz for F1, mean F0 measures differed by no more than 3 Hz, and mean vowel and syllable durations differed by no more than 16 ms. Pearson's product moment correlation analysis of the two sets of measurements showed a strong correlation of at least $r=0.95$ and $p<0.001$ for all measures except the duration of the second syllable ($r=0.77$, $p<0.001$) and the location of the F0 peak within the second syllable ($r=0.88$, $p<0.001$). The comparatively poor correlation for measures involving the duration of the second syllable appears to derive from differences in the identification of the end of the syllable in cases in which the burst release was difficult to differentiate from background noise.

Using the originally measured values for each acoustic variable, a mixed factorial analysis of variance (ANOVA) was performed with native language and gender as between-subjects variables and stress (stressed or unstressed) as the within-subjects factor. All *post hoc* (Tukey HSD) tests were performed with a critical p value of 0.05. Means for each measure for each group, gender, and stress condition are given in Table IV.

1. Average F0

Results of the analysis of average F0 showed significant main effects of stress [$F(1,16)=148.19$, $p=0.001$], native language [$F(1,16)=15.73$, $p=0.001$], and gender [$F(1,16)=164.23$, $p<0.001$]. There were significant interactions between stress and language [$F(1,16)=12.42$, $p=0.003$] and gender and stress [$F(1,16)=21.09$, $p<0.001$]. The three-way interaction was not significant [$F(1,16)=0.41$, $p=0.53$]. The significant effect of gender was expected: The mean average F0 was 229 Hz for females and 176 Hz for males. *Post hoc* (Tukey HSD) tests showed that, for each language group, the F0 of the stressed syllables, averaged across males and females, was significantly higher than that of the unstressed syllables (Mandarin: stressed=198 Hz, unstressed =163 Hz; American: stressed=164 Hz, unstressed=145 Hz). In addition, in stressed syllables Mandarin speakers produced significantly higher F0 than English speakers, but not in unstressed syllables (averaged across genders: Mandarin: stressed=198 Hz; American: stressed 164 Hz). Thus, the language-group difference (Mandarin > American English) is purely the result of Mandarin speakers producing *stressed* syllables with significantly higher F0 than do American English speakers.

TABLE IV. Mean scores and standard deviations for all acoustic parameters for English stressed and unstressed syllable produced by native Mandarin and English speakers. Note: STR = stressed, UNSTR = unstressed. Each cell contains mean value with standard deviation in parentheses.

	Mandarin				English			
	Male		Female		Male		Female	
	STR	UNSTR	STR	UNSTR	STR	UNSTR	STR	UNSTR
F0 (Hz)	145	121	252	205	122	111	206	178
	(13)	(15)	(14)	(13)	(18)	(22)	(17)	(12)
Peak F0 loc. (%)	47	38	45	30	47	45	42	39
	(4)	(8)	(6)	(2)	(8)	(7)	(6)	(4)
Intensity (dB)	65	60	65	60	68	63	65	61
	(2)	(3)	(1)	(2)	(1)	(1)	(1)	(1)
Syllable duration (ms)	337	267	365	287	291	216	367	283
	(23)	(15)	(51)	(44)	(26)	(19)	(50)	(41)

2. Peak F0 location

There were significant effects of stress [$F(1,16)=18.18$, $p<0.001$] and of gender [$F(1,16)=10.38$, $p=0.005$], but not of language [$F(1,16)=3.45$, $p=0.079$]. There was a significant interaction between stress and native language [$F(1,16)=5.09$, $p=0.038$], but not between stress and gender [$F(1,16)=0.92$, $p=0.35$], or native language and gender [$F(1,16)=0.05$, $p=0.81$], and the three-way interaction was also not significant [$F(1,16)=0.32$, $p=0.58$]. *Post hoc* tests showed that for Mandarin speakers the location of peak F0 in stressed syllables was significantly different from the location of peak F0 in unstressed syllable ($p=0.003$, with the stressed location at 46% of the syllable and unstressed at 34%). In other words, Mandarin speakers produced the F0 peak location significantly earlier in unstressed syllables than that in stressed ones. For American English speakers, the difference in F0 peak location between stressed and unstressed syllables was not significant.

In addition, the F0 peak location of the stressed syllable in trochaic (strong-weak pattern) and in iambic (weak-strong) structure was also compared, because it was shown that English speakers tended to produce the peak F0 earlier in the stressed syllable in iambic words than in trochaic words (Munson *et al.*, 2003). A mixed factorial ANOVA was performed with native language and gender as between-subjects variables and with structure (trochee or iamb) as within-subject factor, and the F0 peak location of the stressed syllable as the dependent variable. Results showed a significant effect of structure [$F(1,16)=63.93$, $p<0.001$], but no significant effect of native language [$F(1,16)=0.66$, $p=0.43$], or gender [$F(1,16)=2.31$, $p=0.15$]. There was a significant interaction between native language and structure [$F(1,16)=12.5$, $p=0.003$], as well as between gender and structure [$F(1,16)=5.591$, $p=0.03$], but there was no significant three-way interaction [$F(1,16)=0.61$, $p=0.45$]. *Post hoc* tests showed that both Mandarin and American English speakers produced the F0 peak of the stressed syllable earlier in iambic words than that in trochaic words (Mandarin: trochaic=61%, iambic=32%; English: trochaic=50%, iambic=39%).

3. Intensity

Analysis of average intensity showed a significant effect of stress [$F(1,16)=259.85$, $p<0.001$], and language group [$F(1,16)=10.19$, $p=0.006$]. Gender did not show a main effect [$F(1,16)=2.29$, $p=0.149$], and none of the interactions were shown to be significant. *Post hoc* tests showed that for both language groups, stressed syllables (Mandarin: 65 dB; American: 67 dB) had a significant higher intensity than unstressed syllables (Mandarin: 60 dB; American: 62 dB). Interestingly, although the main effect of language group was significant, indicating that the intensity of speech produced by American English speakers was, on average, two dB higher than Mandarin speakers, *post hoc* analysis showed no significant difference between the intensities of either Mandarin and American English stressed syllables or those of unstressed syllables.

4. Duration

Results of the analyses of syllable durations showed significant effects of stress [$F(1,16)=380.68$, $p<0.01$] and gender [$F(1,16)=9.2$, $p=0.008$], but no effect of language [$F(1,16)=2.48$, $p=0.135$], and no significant interactions. Men produced syllables averaging 277 ms, while women's syllables averaged 325 ms. *Post hoc* tests showed that for both language groups stressed syllables had a significantly longer duration (Mandarin: 351 ms; American: 329 ms) than unstressed syllables (Mandarin: 277 ms; American: 250 ms).

5. Vowel space

Figure 1 shows the English, Mandarin and Mandarin-English vowel spaces, averaged across both male and female talkers (only peripheral vowels are shown). Both the Mandarin English and American English vowel spaces are roughly quadrilateral, consistent with the results of Chen *et al.*, 2001b. However, there are slight differences in the location of specific vowels between the two groups of speakers. In particular, the production of English [u] by native Mandarin speakers is farther "back" (in the sense of having lower F2) compared to the American English [u]. It has been documented that the American English production of [u] is often characterized by a higher F2 than similar phoneme produc-

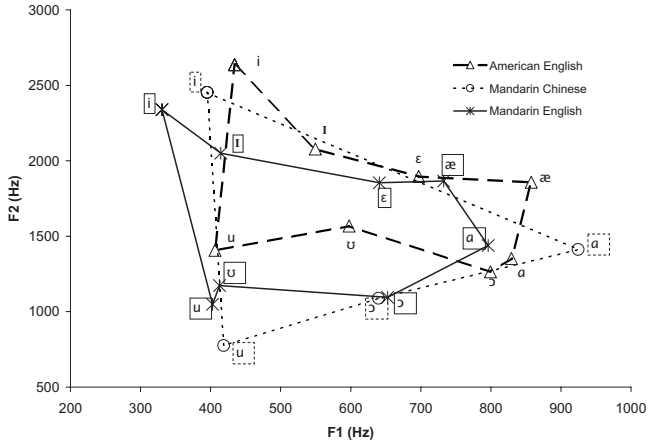


FIG. 1. Comparison of three vowel spaces of American English, Mandarin Chinese, and Mandarin English.

tions in many other languages (for examples, compare vowel charts for various languages presented in IPA, 1999), which may be the result of a more advanced tongue placement during articulation. Such production differences may be the process of historical change, as suggested by Hillenbrand *et al.*'s (1995) comparison of their data with those of Peterson and Barney (1952). This hypothesis is supported by the observation that the present measurements of the F2 of [u] are even higher in frequency (more fronted) than those of Hillenbrand *et al.* (1995), which are in turn higher than those of Peterson and Barney (1952): 1406, 1051, and 910 Hz, respectively, averaged over men and women. Of course, some of this difference may be due to the much smaller number of participants in the present study leading to greater potential influence of inter-individual differences in absolute vocal tract size, as well as possible dialect differences between the participants in the Peterson and Barney (East coast of the U.S.), Hillenbrand *et al.* (central Michigan) and current (central Indiana) studies (see also Harigawa, 1997, for data from Southern Californian English, with even more evidence of fronting of back vowels). Still, Fig. 1 shows that Mandarin speakers are attempting to approximate the more fronted American English [u] (as compared with their native [u]), although they do not achieve it perfectly. Despite these minor differences, the observation that the Mandarin English and American English vowel spaces share an overall similar structure suggests that Mandarin speakers' native vowel system does not interfere very much with production of English-like *stressed* vowels, at least when words are produced in isolation. Further analyses were carried out to investigate the production of stressed and unstressed vowels in the target disyllabic words.

6. Vowel reduction

For each syllable in each word, separate ANOVAs were conducted for both the C-D (Compact-Diffuse, related to the phonological feature contrast high/low) and G-A (Grave-Acute, related to the phonological feature contrast front/back) variables with three factors: gender, language, and stress. Significant results are shown in Table V, with a bold font indicating the significance of the *post-hoc* (Tukey HSD) test at the $p<0.05$ level. In addition, F1 and F2 values, averaged across male and female participants, are shown in Table VI.

For most syllables, stressed vowels did not show a significant difference between Mandarin and American English speakers. The exceptions were *-mit (permit)* and *re-* (*rebel*). The main differences between Mandarin and English speakers were found in their productions of unstressed syllables, and most of these differences appeared in the C-D (Compact-Diffuse) feature with the exception of the word *rebel* in which the unstressed versions of both the initial syllable *re-* and the final syllable *-bel* showed significant differences be-

TABLE V. Statistically significant pairwise comparisons between formant measures for stressed and unstressed vowels, by syllable. Note: C-D refers to the compact-diffuse dimension (F2-F1); G-A refers to grave-acute (arithmetic mean of F1 and F2). S refers to stressed syllables, U to unstressed, AE to American English speakers' productions, and M to Mandarin speakers'. Thus, for example, E<M indicates that English speakers' productions of a given syllable showed smaller mean values of a given acoustic feature than did Mandarin speakers'.

	Stressed/Unstressed				American English/Mandarin			
	AE		M		STRESSED		UNSTRESSED	
Syllable	C-D	G-A	C-D	G-A	C-D	G-A	C-D	G-A
con-	**S<U**						**AE>M**	
-tract			**S<U**					
de-	**S<U**		**S<U**				**AE<M**	
-sert								
ob-	**S<U**						**AE>M**	
-ject								
per-								
-mit			**S>U**		**AE<M**		**AE<M**	
re-		**S>U**			**AE<M**		**AE<M**	**AE<M**
-bel	**S>U**	**S>U**		**S>U**			**AE<M**	**AE<M**
re-		**S>U**	**S<U**				**AE<M**	
-cord	**S<U**	**S<U**	**S<U**				**AE>M**	
sub-								
-ject	**S<U**		**S<U**					

TABLE VI. Average F1 and F2 values in Hz across male and female native speakers of Mandarin Chinese and American English.

	Stressed				Unstressed			
	English		Mandarin		English		Mandarin	
Syllable	F1	F2	F1	F2	F1	F2	F1	F2
con-	844	1495	828	1423	564	1819	694	1519
-tract	817	1768	789	1726	773	1768	750	1756
de-	610	1807	638	1880	452	1938	412	2159
-sert	583	1691	594	1644	582	1765	599	1687
ob-	873	1396	797	1368	676	1583	706	1346
-ject	680	1886	690	1867	627	1884	625	1884
per-	710	1496	670	1462	661	1507	610	1480
-mit	633	1997	480	2325	659	1925	543	2162
re-	694	1587	583	1812	530	1632	516	1905
-bel	717	1618	725	1735	624	1235	683	1559
re-	716	1723	674	1736	524	1774	470	1898
-cord	622	1378	711	1227	617	1761	665	1471
sub-	687	1551	762	1547	617	1654	620	1528
-ject	685	1893	683	1864	619	1912	600	1908

tween Mandarin and English speakers in terms of both the C-D (Compact-Diffuse) and G-A (Grave-Acute) features. Overall, five general patterns can be distinguished:

Type 1. *Correct non-reduction*. Neither English nor Mandarin speakers reduced the vowel in the following unstressed syllables (no significant differences were found for either C-D or G-A): *per- (permit), -sert (desert), sub- (subject)*, and *-ject (object)*.

Type 2. *Unexpected reduction*. Unlike English speakers, Mandarin speakers significantly reduced unstressed vowels (in terms of either C-D or G-A) in the following words: *-tract (contract)* and*-mit (permit)*.

Type 3. *Incorrect reduction*. In these syllables, both English and Mandarin speakers showed significant differences between stressed and unstressed vowels, but the unstressed vowel used by Mandarin speakers was in each case significantly different (in terms of either C-D or G-A, or both) from its English counterpart. These syllables include: *de- (desert), -bel (rebel), re- (record)*, and *-cord (record)*.

Type 4. *Lack of reduction*. Unlike the English speakers, Mandarin speakers did not show a significant change in either the C-D or G-A features from stressed to unstressed versions of the following syllables: *con- (contract), ob- (object)*, and *re- (rebel)*.

Type 5. *Correct reduction*. The only syllable in which both American and Mandarin speakers appear to show a similar degree and quality of vowel reduction is the syllable *-ject (subject)*.

In order to evaluate possible strategies Mandarin speakers may have used in the production of the English unstressed vowels, the average formant values for each vowel were converted to Bark scale values. These values were used to compute Euclidean distances for each stressed or unstressed vowel produced in the experimental words and those from the vowel space mapping task (mapping vowels). These distance measures are shown in Tables VII–X.

Although these tables are quite complex, a few general patterns may be observed from them. Table VII shows which of the Mandarin speakers' mapping vowels are closest to the vowel in a given syllable, while Table IX does the same for English speakers. Comparing the stressed syllables in these two tables shows that Mandarin and English speakers employed approximately the same vowel categories for stressed syllables in many cases. For example, both groups' productions of the vowel in the stressed syllable *con- (contract)* and *ob- (object)* syllables were closest to their productions of [ɑ] in the mapping task, and both produced *de- (desert)* with a vowel most similar to [ɛ].

Comparison of the distance between Mandarin speakers productions and *English* speakers' mapping vowels (Table VIII) with that in Table VII also helps elucidate some more ambiguous cases, such as the nearly equivalent distance between Mandarin speakers' productions of the stressed *-ject* syllable and their mapping vowels [ɛ] and [æ]. Given the overall similarity of these vowels and the very small difference between the two distances, such productions may still be acceptable, and, indeed, as shown in Table VIII, Mandarin speakers' productions of *-ject* are clearly closest to English speakers' [ɛ] mapping vowel which suggests that this syllable is being produced with a vowel that would be clearly identifiable to English speakers as [ɛ] rather than [æ].

With respect to unstressed vowels, the situation is more complex. In some cases, such as the unstressed syllable *de- (desert)*, Mandarin speakers' productions were closest to a vowel in their own English mapping vowel productions (Table VII) that corresponded to the English speakers' mapping vowel closest to English speakers' productions of this syllable ([ɪ]). However, the Mandarin production of this vowel was significantly different from that of native English speakers as shown in Table V, column 7, suggesting that the two mapping task vowels must have been quite different (see also the greater magnitude of the distance between the Mandarin speakers' production of this syllable and the English speakers' mapping vowel [ɪ] as shown in Table VIII).

TABLE VII. Euclidean distance in F1×F2 space between Mandarin speakers' stressed and unstressed vowels in the word production task and Mandarin speakers' productions of English vowels in the vowel space mapping task. Note: Smallest distance indicated in bold.

		English vowels (Mandarin speakers)									
		i	I	ɛ	æ	ɑ	ɔ	ʌ	ɝ	ʊ	u
con-	S	5.26	4.09	2.22	1.92	**0.23**	2.09	1.07	2.13	3.52	3.90
	U	4.26	3.07	1.39	1.40	0.82	2.12	0.10	1.41	2.88	3.39
-tract	S	4.33	3.23	1.19	**0.66**	1.21	3.10	1.19	2.47	3.93	4.47
	U	4.04	2.93	0.90	**0.43**	1.36	3.14	1.15	2.37	3.81	4.38
de-	S	3.08	1.99	**0.09**	0.71	2.13	3.51	1.55	2.36	3.63	4.28
	U	0.95	**0.35**	2.20	2.81	4.11	4.89	3.34	3.40	4.01	4.71
-sert	S	3.34	2.15	**0.89**	1.36	1.75	2.65	0.94	1.40	2.69	3.33
	U	3.25	2.06	**0.71**	1.22	1.81	2.82	1.03	1.58	2.86	3.50
ob-	S	5.27	4.08	2.32	2.11	**0.32**	1.75	1.00	1.88	3.23	3.59
	U	4.89	3.69	2.17	2.17	0.78	1.36	**0.72**	1.23	2.58	2.97
-ject	S	3.46	2.39	0.38	**0.31**	1.89	3.48	1.46	2.46	3.82	4.44
	U	2.98	1.89	0.17	0.81	2.21	3.55	1.60	2.36	3.61	4.26
per-	S	4.31	3.12	1.60	1.69	0.92	1.84	**0.21**	1.12	2.58	3.09
	U	3.94	2.75	1.52	1.80	1.40	1.95	**0.59**	0.82	2.26	2.83
-mit	S	1.41	**1.03**	2.01	2.50	4.04	5.13	3.37	3.71	4.55	5.25
	U	2.01	**1.19**	1.30	1.78	3.33	4.53	2.69	3.18	4.18	4.87
re-	S	2.85	1.69	**0.48**	1.16	2.21	3.31	1.49	2.02	3.20	3.87
	U	2.20	**1.03**	1.04	1.71	2.86	3.77	2.09	2.35	3.30	4.00
-bel	S	4.33	3.23	1.19	**0.66**	1.21	3.10	1.19	2.47	3.93	4.47
	U	4.04	2.93	0.90	**0.43**	1.36	3.14	1.15	2.37	3.81	4.38
re-	S	3.60	2.45	**0.51**	0.65	1.53	2.98	0.97	1.98	3.37	3.97
	U	1.92	**0.72**	1.43	2.11	3.16	3.89	2.36	2.41	3.19	3.89
-cord	S	5.37	4.18	2.76	2.76	1.19	**0.83**	1.31	1.38	2.49	2.74
	U	4.26	3.06	1.55	1.66	0.96	1.88	**0.20**	1.11	2.58	3.09
sub-	S	4.57	3.40	1.51	1.27	**0.54**	2.36	0.60	1.89	3.36	3.84
	U	3.83	2.64	1.30	1.58	1.36	2.15	**0.52**	1.04	2.46	3.04
-ject	S	3.41	2.34	**0.32**	0.37	1.91	3.46	1.44	2.43	3.78	4.40
	U	2.77	1.67	**0.38**	1.02	2.38	3.64	1.74	2.38	3.57	4.24

In other cases, Mandarin speakers' productions of unstressed vowels did not pattern with those of native American English speakers. For example, for the unstressed *con- (contract)*, American English speakers recorded here used a vowel similar to [ɪ] as in *bit* (Table IX), but Mandarin subjects used [ʌ] as in *butt* (Table VII).[2] One possible explanation for this is that Mandarin speakers may have substituted a native short, central vowel, [ɤ], for the similar English [ʌ], and this argument is supported by the observation that, as shown in Table X, [ɤ] is indeed the closest Mandarin monophthong to the vowel in the unstressed *con-* syllable. However, this distance (1.18 Bark) is considerably larger than the distance between the vowel in the unstressed *con-* syllable and Mandarin speakers' production of English [ʌ] (0.10 Bark). This pattern of results is more consistent with the hypothesis that Mandarin speakers chose an English vowel as their target for this syllable, but, unlike the case of *de-* discussed above, the vowel that they chose was different from that chosen by the native speakers in this study (the possibility that this native production may have been nonstandard is discussed below).

Finally, sometimes Mandarin speakers seem to have tried but failed to produce sufficiently distinctive versions of stressed and unstressed vowels. For example, in the syllable *ob- (object)*, both Mandarin and American English speakers produced vowels similar to [ɑ] in stressed productions and [ʌ] in unstressed ones (Tables VII and IX). However, there was no significant difference in Mandarin speakers' stressed and unstressed vowels in terms of either the C-D or G-A dimensions (Table V). This pattern can be explained by examining the relative distance between the vowel in unstressed *ob-* and [ɑ], which was 0.78 for Mandarin speakers (Table VII), compared with 0.72 as a distance from [ʌ], and 1.51 for English speakers (Table IX), compared with 0.34 for [ʌ]. In other words, the Mandarin production of unstressed *ob-* was nearly equidistant between [ɑ] and [ʌ], while English speakers' productions were much closer to [ʌ] than to [ɑ], suggesting that Mandarin speakers were aware that they needed to produce a different vowel in the unstressed as compared to the stressed context, but were either not sure what that vowel should be or, perhaps, were simply unable to realize it to a sufficiently clear degree.

IV. DISCUSSION

Native Mandarin speakers were able to produce lexical stress contrasts that were correctly identified by linguistically trained native speakers of American English. Subsequent acoustic analyses indicated that both native English and native Mandarin speakers used the acoustic correlates of F0, intensity and duration in a similar manner: Both groups pro-

TABLE VIII. Euclidean distance in F1×F2 space between Mandarin speakers' stressed and unstressed vowels in the word production task and English speakers' productions of English vowels in the vowel space mapping task. Note: Smallest distance indicated in bold.

		English vowels (Mandarin speakers)									
		i	I	ɛ	æ	ɑ	ɔ	ʌ	ɝ	ʊ	u
con-	S	5.14	3.27	2.12	1.78	**0.33**	0.79	0.96	1.98	1.81	3.36
	U	4.26	2.38	1.48	1.76	1.23	1.41	**0.14**	0.94	0.77	2.46
-tract	S	4.00	2.20	0.91	**0.68**	1.63	2.04	0.96	1.69	1.56	3.37
	U	3.73	1.92	**0.64**	0.83	1.81	2.18	0.94	1.49	1.38	3.18
de-	S	2.84	0.98	**0.45**	1.58	2.59	2.87	1.47	1.25	1.25	2.77
	U	1.34	**1.24**	2.53	3.65	4.55	4.71	3.37	2.53	2.68	2.85
-sert	S	3.44	1.61	1.24	2.09	2.16	2.30	1.01	**0.29**	0.32	1.93
	U	3.30	1.45	1.09	1.99	2.24	2.41	1.07	**0.46**	0.49	2.06
ob-	S	5.23	3.35	2.28	2.07	**0.24**	0.51	0.96	1.90	1.73	3.15
	U	4.99	3.13	2.27	2.38	0.88	**0.78**	0.85	1.45	1.30	2.51
-ject	S	3.13	1.32	**0.11**	1.18	2.35	2.68	1.32	1.41	1.37	3.03
	U	2.75	0.89	**0.55**	1.68	2.67	2.94	1.53	1.24	1.26	2.71
per-	S	4.39	2.53	1.74	2.08	1.25	1.33	**0.44**	0.89	0.73	2.24
	U	4.13	2.31	1.78	2.35	1.72	1.74	0.81	0.51	**0.39**	1.79
-mit	S	0.93	**0.96**	2.23	3.25	4.50	4.74	3.35	2.70	2.81	3.41
	U	1.62	**0.27**	1.51	2.56	3.79	4.05	2.65	2.11	2.20	3.11
re-	S	2.81	0.96	0.93	2.02	2.66	2.87	1.49	**0.91**	0.97	2.28
	U	2.28	**0.65**	1.45	2.58	3.29	3.47	2.12	1.34	1.47	2.24
-bel	S	4.00	2.20	0.91	**0.68**	1.63	2.04	0.96	1.69	1.56	3.37
	U	3.73	1.92	**0.64**	0.83	1.81	2.18	0.94	1.49	1.38	3.18
re-	S	3.43	1.56	**0.61**	1.37	1.99	2.27	0.87	0.98	0.90	2.65
	U	2.21	**0.92**	1.85	2.98	3.58	3.72	2.42	1.53	1.68	2.07
-cord	S	5.54	3.70	2.86	2.91	1.04	**0.66**	1.45	1.94	1.81	2.68
	U	4.34	2.48	1.70	2.07	1.30	1.38	**0.44**	0.84	0.67	2.21
sub-	S	4.43	2.56	1.43	1.38	1.00	1.34	**0.37**	1.41	1.25	2.97
	U	3.96	2.13	1.55	2.16	1.73	1.81	0.70	0.41	**0.24**	1.92
-ject	S	3.10	1.28	**0.16**	1.23	2.37	2.69	1.32	1.37	1.33	2.98
	U	2.58	**0.70**	0.75	1.89	2.84	3.10	1.69	1.26	1.31	2.63

duced stressed syllables with a higher F0, longer duration and greater intensity than unstressed syllables.

However, these productions were still rated as significantly less acceptable than those of native English speakers, suggesting that the Mandarin speakers in this study produced stress contrasts with a discernable accent. Acoustically, differences between Mandarin and English speakers' production of stressed and unstressed syllables were noted, specifically in terms of the properties of average F0, F0 peak location, intensity, and vowel reduction. Mandarin speakers produced English stressed syllables with significantly higher F0 than did American speakers. Moreover, Mandarin speakers produced F0 peaks significantly earlier in the unstressed syllable than in stressed syllable, while English speakers showed no difference in F0 peak timing between stressed and unstressed syllables. In addition, Mandarin speakers were, on average, about 2 dB less intense, overall, than were English speakers, but it is unlikely that this difference, in itself, contributed significantly to the perception of non-nativeness in their production of the English stress contrast. Finally, Mandarin speakers showed a tendency to either not reduce, or incorrectly reduce vowels in unstressed syllables requiring vowel reduction. In general, these findings are consistent with the hypothesis that, although native Mandarin speakers are able to control certain acoustic correlates in an English-like manner to signal stress, they are not able to manage F0 and vowel quality in a strictly English-like manner due to interference from their native tonal system and vowel systems respectively.

With respect to the observed group differences in average F0, the present results are consistent with the results and conclusions of Chen *et al.* (2001a), who has argued that such behavior derives from tone language speakers' experience with using a larger proportion of their overall frequency range as compared to speakers of nontonal languages (Chen, 1974): Mandarin high tones are produced with an F0 at a much higher proportion of the talker's overall pitch range compared to English stress (see also Shen, 1989 and Adams and Munro 1978 for corroborative results). Therefore, although Mandarin speakers are able to transfer the use of F0 from the tonal domain to that of lexical stress, they are still strongly influenced by the native (tonal) domain within which they are used to manipulating this property. Thus, the acoustic property of F0 cannot be considered an independent feature to be manipulated at will, but rather must be controlled as part of the speakers' native language phonology.

Similarly, although analysis of the peak F0 location indicates that both American English and Mandarin speakers produced the peak F0 earlier in the stressed syllable in words with iambic stress than in words with trochaic stress, consistent with the findings of Munson *et al.* (2003), the two groups differed in terms of their location of peak F0 in

TABLE IX. Euclidean distance in F1×F2 space between English speakers' stressed and unstressed vowels in the word production task and English speakers' productions of English vowels in the vowel space mapping task. Note: Smallest distance indicated in bold.

		English vowels (Mandarin speakers)									
		i	I	ɛ	æ	ɑ	ɔ	ʌ	ɝ	ʊ	u
con-	S	4.96	3.11	1.89	1.45	**0.66**	1.13	0.95	2.00	1.83	3.49
	U	2.72	**0.90**	1.08	2.17	2.79	2.98	1.61	0.94	1.03	2.20
-tract	S	4.03	2.29	0.97	**0.43**	1.77	2.21	1.21	1.93	1.81	3.62
	U	3.81	2.02	0.72	**0.67**	1.81	2.21	1.04	1.65	1.54	3.34
de-	S	2.93	1.06	**0.74**	1.81	2.50	2.74	1.34	0.92	0.95	2.42
	U	2.06	**0.96**	2.01	3.15	3.79	3.93	2.63	1.73	1.88	2.17
-sert	S	3.23	1.40	1.17	2.11	2.35	2.50	1.19	**0.45**	0.52	1.97
	U	2.97	1.13	1.01	2.05	2.54	2.73	1.37	**0.73**	0.80	2.15
ob-	S	5.42	3.56	2.37	1.89	**0.36**	0.81	1.28	2.31	2.13	3.66
	U	3.95	2.07	1.21	1.67	1.51	1.72	**0.34**	0.77	0.61	2.40
-ject	S	3.03	1.22	**0.13**	1.26	2.45	2.77	1.40	1.42	1.39	3.02
	U	2.77	0.91	**0.54**	1.66	2.64	2.92	1.51	1.23	1.25	2.71
per-	S	4.40	2.53	1.58	1.77	1.07	1.27	**0.17**	1.09	0.91	2.56
	U	4.18	2.32	1.55	1.97	1.40	1.52	**0.39**	0.73	0.56	2.21
-mit	S	2.50	0.72	**0.60**	1.69	2.96	3.26	1.86	1.62	1.64	3.03
	U	2.82	1.01	**0.31**	1.44	2.64	2.95	1.57	1.46	1.45	2.99
re-	S	4.01	2.13	1.19	1.56	1.43	1.67	**0.26**	0.90	0.75	2.53
	U	3.31	1.62	1.66	2.60	2.58	2.65	1.49	**0.45**	0.62	1.48
-bel	S	4.03	2.29	0.97	**0.43**	1.77	2.21	1.21	1.93	1.81	3.62
	U	3.81	2.02	0.72	**0.67**	1.81	2.21	1.04	1.65	1.54	3.34
re-	S	3.66	1.81	**0.66**	1.11	1.79	2.11	0.77	1.21	1.11	2.90
	U	2.76	1.08	1.45	2.53	2.93	3.07	1.77	**0.89**	1.03	1.86
-cord	S	4.59	2.79	2.19	2.61	1.52	1.43	0.98	0.97	**0.86**	1.86
	U	3.11	1.24	0.79	1.78	2.34	2.56	1.17	**0.78**	0.79	2.35
sub-	S	4.11	2.24	1.34	1.70	1.35	1.56	**0.19**	0.86	0.70	2.44
	U	3.48	1.62	1.10	1.91	2.04	2.21	0.87	0.43	**0.39**	2.11
-ject	S	3.03	1.24	**0.09**	1.22	2.45	2.79	1.42	1.47	1.43	3.06
	U	2.66	0.79	**0.60**	1.73	2.76	3.03	1.63	1.31	1.34	2.74

stressed as compared to unstressed syllables. Mandarin speakers reached their peak F0 significantly earlier in unstressed syllables than in stressed syllables, while the American English speakers showed no difference in peak F0 timing between syllable types. Xu (1998, 1999; Xu and Liu, 2006) examined the peak F0 location in Chinese syllables across different lexical tones, finding a positive correlation between syllable duration and the location of the F0 peak. Longer syllables were found to have a later F0 peak relative to the syllable onset. In the present study, Mandarin speakers produced English unstressed syllables with significantly shorter durations than stressed syllables. This duration difference may have caused Mandarin speakers to incorrectly alter peak F0 timing.[3] Once again, it appears that, although Mandarin speakers are able to select F0 as a cue to be manipulated in the service of producing English lexical stress differences, they may only do so according to the linguistic conventions commonly used within their native language.

A. Vowel reduction

To examine vowel reduction, productions of vowels in stressed and unstressed syllables were referenced against productions of monosyllabic (stressed) vowels in the vowel space mapping task (Tables VII–X). Based on these comparisons, it appears that Mandarin speakers showed a great deal of similarity with English speakers in both their stressed and some unstressed vowel productions. In particular, for the majority of vowels used in the stressed syllable, Mandarin speakers employed approximately the same vowel categories as the English speakers. In agreement with this observation, the difference between most Mandarin and English stressed syllables was statistically insignificant (Table V, fifth and sixth columns), supporting the hypothesis that Mandarin speakers do not have significant difficulty learning to produce American English full (unreduced) monophthongal vowels.

Mandarin speakers' productions of unstressed vowels were also frequently comparable to those of English speakers. For example, in Type 1(*Correct nonreduction*) syllables such as *per- (permit)*, *-sert (desert)* and *sub- (subject)*, Mandarin speakers correctly did not reduce the vowel, just as American English speakers did not, while in Type 3(*Incorrect reduction*) syllables such as *de- (desert)* and *bel- (rebel)*, and in Type 5(*Correct reduction*) syllables such as *-ject (subject)*, Mandarin speakers reduced the vowel just as American English speakers did. However, in the Type 3(*Incorrect reduction*) cases (e.g., *de-* in the verb *desert*), although Mandarin speakers were not successful in achieving the English unstressed vowel quality, they did attain formant values that were comparable to their (accented) productions of the same vowels that were used by the native English speakers in the corresponding unstressed syllable. In other words, they ap-

TABLE X. Euclidean distance in F1×F2 space between Mandarin speakers' stressed and unstressed vowels in the word production task and Mandarin speakers' productions of Mandarin monophthongal vowels in the vowel space mapping task. Note: Smallest distance indicated in bold.

		Mandarin vowels (Mandarin speakers)					
		ɐ	o	ɣ	*i*	u	y
con-	S	**0.63**	2.19	1.94	5.02	4.90	5.00
	U	1.65	2.18	**1.18**	4.07	4.70	3.98
-tract	S	**1.61**	3.18	2.24	3.97	5.78	4.12
	U	**1.86**	3.21	2.13	3.68	5.75	3.82
de-	S	2.77	3.56	**2.15**	2.74	5.84	2.87
	U	4.85	4.88	3.30	**0.87**	6.46	0.75
-sert	S	2.56	2.68	**1.19**	3.19	4.87	3.05
	U	2.60	2.85	**1.37**	3.07	5.05	2.97
ob-	S	**0.87**	1.85	1.71	5.06	4.57	4.98
	U	1.53	1.44	**1.07**	4.76	4.08	4.58
-ject	S	2.46	3.53	**2.24**	3.08	5.93	3.26
	U	2.86	3.59	**2.15**	2.65	5.83	2.77
per-	S	1.78	1.90	**0.89**	4.16	4.39	4.01
	U	2.25	1.98	**0.58**	3.85	4.27	3.63
-mit	S	4.70	5.15	3.58	**0.85**	6.97	1.38
	U	3.97	4.56	3.02	**1.56**	6.56	1.89
re-	S	2.95	3.34	**1.83**	2.61	5.47	2.60
	U	3.61	3.78	2.21	**2.01**	5.68	1.94
-bel	S	**1.61**	3.18	2.24	3.97	5.78	4.12
	U	**1.86**	3.21	2.13	3.68	5.75	3.82
re-	S	2.21	3.03	**1.75**	3.31	5.44	3.36
	U	3.94	3.89	2.30	1.84	5.61	**1.62**
-cord	S	1.71	**0.92**	1.32	5.28	3.64	5.05
	U	1.82	1.94	**0.88**	4.11	4.42	3.96
sub-	S	**1.24**	2.43	1.67	4.30	5.07	4.31
	U	2.21	2.18	**0.80**	3.71	4.49	3.53
-ject	S	2.49	3.51	**2.21**	3.04	5.90	3.21
	U	3.06	3.67	**2.19**	2.45	5.84	2.55

peared to be aiming for the appropriate reduced vowel target, but missed producing it with the expected F1 and F2 values in the same way that they missed producing that target vowel when it was the target in a stressed monosyllable (in the vowel space mapping condition). In other words, Mandarin speakers' poor performance on vowel reduction in the present experiment appears to be due to an inability to correctly produce *specific* reduced vowels, and some of this may be related to their incorrect production of those vowels even in stressed contexts (e.g., the vowel space mapping task).

One explanation for this difficulty is interference from the native vowel system or, more properly, the *lack* of a sufficiently similar vowel in the Mandarin system leading to particularly inaccurate productions in a manner consistent with the results of Flege *et al.* (1997), who found that Mandarin speakers showed the least spectral accuracy when producing English vowels, including [ɪ], that are not found in Mandarin. Similarly, Chen *et al.* (2001b) showed that [ɪ], an "unfamiliar vowel" to Mandarin speakers, was pronounced less accurately than other vowels that were familiar to Mandarin speakers (that is, acoustically more similar to native Mandarin vowels). In particular, as in the present study, Chen *et al.* (2001b) showed that female speakers of Mandarin produced [ɪ] with a lower F1 than that of female speakers of American English, while male speakers of Mandarin produced [ɪ] with a higher F2 than that of male speakers of American English. Thus, difficulties with native-like production of [ɪ] seem to be characteristic of Mandarin speakers' production of English, in a manner independent of the issue of lexical (or sentential) stress production.

In other cases, Mandarin speakers seem to have chosen a different target vowel than did the English speakers, as in the case of unstressed *con-*, where Mandarin speakers produced a vowel very similar to their [ʌ] mapping vowel, but English speakers produced a vowel more similar to their mapping vowel productions of ([ɪ]). Since the first syllable of the verb *contract* is quite commonly produced with the vowel [ʌ] in many varieties of English, it is quite possible that the Mandarin speakers in this study were in fact successfully approximating a native-like pronunciation of this word, albeit one that differed from the native pronunciation in the local dialect. The degree to which Mandarin (or any other nonnative) speakers' perceived non-nativeness may derive from their (successfully) attaining an English target appropriate to a different English dialect than that of their listeners is an interesting and important sociolinguistic question, and deserves further exploration although it is beyond the scope of the present study.

Again, however, the fact that Mandarin speakers produced clearly different vowel qualities in the stressed and unstressed versions of the same syllable supports the hypothesis that they are capable of employing vowel change as a cue to lexical stress, at least in some cases. On the other hand, in other cases, Mandarin speakers did not appear to reduce unstressed vowels significantly, even though English speakers did show clear vowel reduction (e.g., in the syllable *ob-* in the word *object*). As described above, the behavior of this syllable can be explained in terms of Mandarin speakers' failure to correctly produce the reduced vowel [ʌ]. Examination of the two groups' vowel spaces (Fig. 1) showed that Mandarin speaker's productions of English [ɑ] and [Λ] were each quite close to the American English [ɑ] and [Λ] when producing the (stressed) words *father* and *butt*, respectively. Thus, Mandarin speakers should in principle have been able to produce both the stressed and unstressed versions of *ob-* accurately, and clearly moved in the expected (native-like) direction, but may not have managed the change with sufficient clarity. Indeed, in all cases in which American English speakers showed significant differences in formant frequencies between stressed and unstressed syllables and Mandarin speakers did not [*con-*, *ob-*, and *re-*(*bel*), see Table V], there are still some small differences observable in Mandarin speakers' productions, at least in terms of there being a difference in mapping vowel that is closest to the stressed as compared to the unstressed vowel (Table VII). The appearance of unexpected reductions (significant differences between stressed and unstressed vowel formant patterns for Mandarin but not English speakers), as in the syllables *-tract* and *-mit*, further supports the hypothesis that Mandarin speakers are aware of, and attempt to make use of, formant frequency differences to cue lexical stress differences.

V. CONCLUSION

In conclusion, it appears that Mandarin speakers are able to successfully approximate English-like patterns of duration and intensity when producing stress contrasts, as well as some of the native-like patterns of F0 production. Moreover, when their pattern of performance on these cues diverged from that of native English speakers, it did so in a manner consistent with the transfer of properties characteristic of the Mandarin tonal system. In contrast, Mandarin speakers, although clearly aware of the importance of vowel reduction as a cue to stress, had much more difficulty with this cue, but the precise pattern of difficulty was not systematic, and appeared to vary across the linguistic context or vowel category. This observation is consistent with the proposal of Flege and Bohn (1989), who suggested that L2 learners acquire L1 stress patterns for individual words. For instance, the pattern for the noun *object* might be learned at a different time than that of the verb *object*. The present results suggest further that learners might acquire the individual *cues* to stress based on the lexical item or vowel category, at least with respect to the cue of vowel reduction. Since Mandarin speakers were successful at producing English-like cues for duration, intensity, and to a limited extent F0, it is difficult to determine whether they learned to produce these cues systematically, whether they have simply already learned these cues for the specific words examined here, or whether transfer from their native suprasegmental phonological system was sufficient to achieve native-like patterns in the L2. Further research is needed to investigate the contribution of the observed non-English-like F0 patterns, such as the stressed syllables produced at F0 values that are too high and with a different alignment of F0 peaks within the syllable, to the perception of foreign accent in Mandarin speakers of English. In addition, it would be of interest to examine the relative contribution of the various cues examined in this study to the perception of stress in English.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the Program in Linguistics, College of Liberal Arts, Purdue University to Y.Z. and by NIH NIDCD Grant No. R03 DC006811 to A.L. Francis. Some of the results were presented at the *4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, Honolulu, Hawaii, November 28–December 2, 2006.

[1]The study of Mandarin intonation is still in its infancy, and is complicated by its interaction with tone. While the general consensus seems to be that Mandarin does possess at least a minimal set of intonational patterns that are independent from, but interact with, the tonal properties of a given utterance, there is considerable disagreement about the nature of the proposed system and the quality and degree of its interaction with lexical tone (Chao, 1968; Ho, 1977; Gårding, 1984; Kratochvil, 1998; Shen, 1990; see Schack, 2000, for review). This topic is far beyond the scope of the present article.

[2]In fact, it appears that English speakers produced a vowel in this context that is more or less equidistant from [ɪ], [ɛ], [ɝ], and [ʊ], though marginally closer to [ɪ]. The presence of the following [n] and concomitant nasalization of the preceding vowel may have complicated measurement of this vowel, and dialectal differences may have skewed these measures toward [ɪ] and away from the expected [ʌ]. However, the main point remains, namely that Mandarin speakers did not produce the same unstressed vowel as did native English speakers.

[3]It is not yet known whether this timing difference contributes to the perception of non-native accent in the Mandarin speakers' productions, though recent research on peak timing in Mandarin tone production and cross-dialectal differences in F0 peak timing suggests that it might (Arvaniti and Garding, 2007; Atterer and Ladd, 2004; Grabe *et al.*, 2000; Mennen, 2004). We are currently carrying out perceptual investigations to explore this issue.

Adams, C., and Munro, R. (**1978**). "In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterance of some native and non-native speakers of English," Phonetica **35**, 125–156.

Archibald, J. (**1997**). "The acquisition of English stress by speakers of non-accentual languages: Lexical storage versus computation of stress," Linguistics **35**, 167–181.

Arvaniti, A., and Gårding, G. (**2007**). "Dialectal variation in the rising accents of American English," in edited by J. Cole and J. Hualde *Laboratory Phonology*, (Mouton de Gruyter, Berlin), Vol. **9**.

Atterer, M., and Ladd, D. R. (**2004**). "On the phonetics and phonology of 'segmental anchoring' of F0: Evidence from German," J. Phonetics **32**, 177–197.

Beckman, M. E. (**1986**). *Stress and Non-stress Accent* (Foris, Dordrecht).

Best, C. T. (**1995**). "A direct realistic view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York, Baltimore), pp. 171–204.

Best, C. T., McRoberts, G. W., and Goodell, E. (**2001**). "American listeners' perception of nonnative consonant contrasts varying in perceptual assimilation to English phonology," J. Acoust. Soc. Am. **1097**, 775–794.

Best, C. T., McRoberts, G. W., and Sithole, N. M. (**1988**). "Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants," J. Exp. Psychol. Hum. Percept. Perform. **4**, 45–60.

Blomgren, M., Robb, M., and Chen, Y. (**1998**). "A note on vowel centralization in stuttering and nonstuttering individuals," J. Speech Lang. Hear. Res. **41**, 1042–1051.

Boersma, P. (**1993**). *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings [Proceedings of the Institute of Phonetic Sciences, Amsterdam]* **17**, 97–110. Downloaded from http://www.fon.hum.uva.nl/Proceedings/IFA-Proceedings.html. Last accessed May 3, 2007.

Boersma, P., and Weenink, D. (**2004**). http://www.fon.hum.uva.nl/praat/. Last accessed March 26, 2007.

Bolinger, D. L. (**1958**). "A theory of pitch accent in English," Word **14**, 109–119.

Campbell, N., and Beckman, M. (**1997**). "Stress, prominence, and spectral tilt," in *Proceedings of ESCA Workshop on Intonation: Theory, Models and Applications*, edited by A. Botinis, G. Kouroupetroglou, and G. Carayiannis, Athens, pp. 67–70.

Chao, Y. R. (**1968**). *A Grammar of Spoken Chinese* (University of California Press, Berkeley, CA).

Chao, Y. (**1972**). *Mandarin Primer* (Harvard University Press, Cambridge, MA).

Chen, G. T. (**1974**). "The pitch range of English and Chinese speakers," J. Chin. Linguist. **2**, 159–171.

Chen, Y., Robb, M. P., Gilbert, H. R., and Lerman, J. W. (**2001a**). "A study of sentence stress production in Mandarin speakers of American English," J. Acoust. Soc. Am. **4**, 1681–1690.

Chen, Y., Robb, M. P., Gilbert, H. R., and Lerman, J. W. (**2001b**). "Vowel production by Mandarin speakers of English," Clin. Linguist. Phonetics **6**, 427–440.

Chen, Y., and Xu, Y. (**2006**). "Production of weak elements in speech—evidence from f0 patterns of neutral tone in standard Chinese," Phonetica **63**, 47–75.

Duanmu, S. (**2000**) The Phonology of Standard Chinese, Oxford university Press, Oxford, England.

Flege, J. E. (**1984**). "The detection of French accent by American listeners," J. Acoust. Soc. Am. **3**, 692–707.

Flege, J. E. (**1988**). "Factors affecting degree of perceived foreign accent in English sentences," J. Acoust. Soc. Am. **1**, 70–79.

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, Baltimore), pp. 233–277.
Flege, J. E., and Bohn, O. S. (**1989**). "An instrumental study of vowel reduction and stress placement in Spanish-accented English," Stud. Second Lang. Acquis. **11**, 35–62.
Flege, J. E., Bohn, O. S., and Jang, S. (**1997**). "Effects of experience on non-native speakers' production and perception of English vowels," J. Phonetics **25**, 437–470.
Flege, J. E., and Davidian, R. (**1985**). "Transfer and developmental processes in adult foreign language speech production," Appl. Psycholinguist. **5**, 323–347.
Flege, J. E., and Hillenbrand, J. (**1987**). "Limits on phonetic accuracy in foreign language production," in *Interlanguae Phonology: the Acquisition of a Second Language Sound System*, edited by G. Ioup and S. Weinberger (Newbury House, Cambridge), pp. 176–201.
Fokes, J. E., Bond, Z. S., and Steinberg, M. (**1984**). "Patterns of word stress by native and non-native speakers," in *Proceedings of the Tenth International Congress of Phonetic Sciences*, edited by M. Van den Broecke and A. Cohen (Foris, Dordrecht), pp. 682–686.
Fokes, J., and Bond, Z. S. (**1989**). "The vowels of stressed and unstressed syllables in Nonnative English," Lang. Learn. **3**, 341–373.
Francis, A. L., and Nusbaum, H. C. (**1999**). "Evaluating the quality of synthetic speech," in *Human Factors and Voice Interactive Systems*, edited by D. Gardner-Bonneau (Kluwer, Boston), pp. 63–97.
Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). "Perceptual learning of Cantonese lexical tones by tonal and non-tonal language speakers," J. Phonetics, published online 13 February, 2008.
Fry, D. B. (**1955**). "Duration and intensity as physical correlates of linguistic stress," J. Acoust. Soc. Am. **27**, 765–768.
Fry, D. B. (**1958**). "Experiments in the perception of stress," Lang Speech **1**, 126–152.
Fry, D. B. (**1965**). "The dependence of stress judgments on vowel formant structure," in *Proceedings of the 5th International Congress of Phonetics Sciences*, eds.X. Zwerner, and W. Bethge, Karger: Basel, pp. 306–311.
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. **1**, 505–510.
Gandour, J. (**1978**). "The perception of tone," in *Tone: A Linguistic Survey*, edited by V. Fromkin (Academy, New York), pp. 41–76.
Gandour, J. (**1983**). "Tone perception in far eastern languages," J. Phonetics **11**, 149–175.
Gårding, Eva. (**1984**). "Chinese and Swedish in a generative model of intonation," in *Nordic Prosody III, Papers from a Symposium* edited by C. C. Elert, I. Johansson, and E. Strangert (Almqvist and Wiksell, Stockholm), pp. 79–91.
Grabe, E., Post, B., Nolan, F., and Farrar, K. (**2000**). "Pitch accent realization in four varieties of British English," J. Phonetics **28**, 161–185.
Hammond, R. H. (**1986**). "Error analysis and the natural approach to teaching foreign languages," Lenguas Modernas **13**, 129–139.
Harigawa, R. (**1997**). "Dialect variation and formant frequency: The American English vowels revisited," J. Acoust. Soc. Am. **1**, 655–658.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **5**, 3099–3111.
Ho, Aichen T. (**1977**). "Intonation variation in a Mandarin sentence for three expressions: Interrogative, exclamatory and declarative," Phonetica **34**, 446–457.
Howie, J. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge).
Hung, T. T. N. (**1993**). "The role of phonology in the teaching of pronunciation to bilingual students," Language, Culture and Curriculum **3**, 249–256.
International Phonetic Association (**1999**). *Handbook of the International Phonetic Association* (Cambridge University Press, Cambridge).
Juffs, A. (**1990**). "Tone, syllable structure and interlanguage phonology: Chinese learner's stress errors," Int. Rev. Appl. Linguistics **2**, 99–117.
Kratochvil, P. (**1998**). "Intonation in Beijing Chinese," in *Intonation Systems: A Survey of Twenty Languages*, edited by D. Hirst and A. DiCristo (Cambridge University Press, Cambridge, MA), pp. 417–431.
Lee, B., Guion, S. G., and Harada, T. (**2006**). "Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals," Stud. Second Lang. Acquis. **28**, 487–513.
Lieberman, P. (**1960**). "Some acoustic correlates of word stress in American English," J. Acoust. Soc. Am. **32**, 451–454.
Lieberman, P. (**1975**). *Intonation, Perception and Language* (M.I.T. Press, Cambridge, Massachusetts).
Liénard, J. S., and DiBenedetto, M. G. (**1999**). "Effects of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **1**, 411–422.
Liu, S., and Samuel, A. G. (**2004**). "Perception of Mandarin lexical tones when f0 information is neutralized," Lang Speech **47**, 109–138.
Lord, G. (**2005**). "(How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course," Hispania–A journal devoted to the teaching of Spanish and Portuguese **3**, 557–567.
Mennen, I. (**2004**). "Bi-directional interference in the intonation of Dutch speakers of Greek," J. Phonetics **32**, 543–563.
Munson, B., Bjorum, E. M., and Windsor, J. (**2003**). "Acoustic and perceptual correlates of stress in nonwords produced by children with suspected developmental apraxia of speech and children with phonological disorder," J. Speech Lang. Hear. Res. **46**, 189–202.
Peterson, G. E., and Barney, H. L., (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Peterson, G. E., and Lehiste, L. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.
Piske, T., MacKay, I. R. A., and Flege, J. E. (**2001**). "Factors affecting degree of foreign accent in an L2: A review," J. Phonetics **2**, 191–215.
Schack, K. (**2000**). "Comparison of intonation patterns in Mandarin and English for a particular speaker," in University of Rochester Working Papers in the Language Sciences, edited by K. M. Crosswhite and J. McDonough, Spring 2000, pp. 24–55. Available online at http://www.bcs.rochester.edu/cls/s2000n1/schack.pdf. Last accessed October 1, 2007.
Schmidt-Nielsen, A. (**1995**). "Intelligibility and acceptability testing for speech technology," in *Applied Speech Technology*, edited by A. K. Syrdal, R. W. Bennett, and S. L. Greenspan (CRC press, Boca Raton, FL), pp. 195–232.
Schneider, W., Eschman, A., and Zuccolotto, A. (**2002**). *E-Prime User's Guide*. (Psychology Software Tools Inc., Pittsburgh).
Shen, X. S. (**1989**). "Toward a register approach in teaching Mandarin tones," J. Chin. Lang. Teachers Assoc. **24**, 27–47.
Shen, X.-N. S. (**1990**). *The Prosody of Mandarin Chinese*, University of California Publications in Linguistics (University of California Press, Berkeley, CA), Vol. **118**.
Sluijter, A. M. C., and Heuven, V. J. (**1996**). "Spectral balance as an acoustic correlate of linguistic stress," J. Acoust. Soc. Am. **4**, 2471–2485.
Sluijter, A. M. C., Heuven, V. J., and Pacilly, J. J. A. (**1997**). "Spectral balance as a cue in the perception of linguistic stress," J. Acoust. Soc. Am. **1**, 503–513.
Southwood, M. H., and Flege, J. E. (**1999**). "Scaling foreign accent: Direct magnitude estimation versus interval scaling," Clin. Linguist. Phonetics **5**, 335–349.
Tahta, S., and Wood, M. (**1981**). "Foreign accents: Factors relating to transfer of accent from the first language to a second language," Lang Speech **3**, 265–272.
Traunmüller, H. (**1989**). "Articulatory dynamics of loud and normal speech," J. Acoust. Soc. Am. **85**, 295–312.
Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **1**, 25–47.
Xu, Y. (**1998**). "Consistency of tone-syllable alignment across different syllable structures and speaking rates," Phonetica **55**, 179–203.
Xu, Y. (**1999**). "Effects of tone and focus on the formation and alignment of F0 contours," J. Phonetics **27**, 55–105.
Xu, Y., and Liu, F. (2006). "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," Italian J. Ling. **18**, 125–159.

Binaural intelligibility prediction based on the speech transmission index[a)]

Sander J. van Wijngaarden and Rob Drullman
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

(Received 9 February 2007; revised 14 March 2008; accepted 14 March 2008)

Although the speech transmission index (STI) is a well-accepted and standardized method for objective prediction of speech intelligibility in a wide range of environments and applications, it is essentially a monaural model. Advantages of binaural hearing in speech intelligibility are disregarded. In specific conditions, this leads to considerable mismatches between subjective intelligibility and the STI. A binaural version of the STI was developed based on interaural cross correlograms, which shows a considerably improved correspondence with subjective intelligibility in dichotic listening conditions. The new binaural STI is designed to be a relatively simple model, which adds only few parameters to the original standardized STI and changes none of the existing model parameters. For monaural conditions, the outcome is identical to the standardized STI. The new model was validated on a set of 39 dichotic listening conditions, featuring anechoic, classroom, listening room, and strongly echoic environments. For these 39 conditions, speech intelligibility [consonant-vowel-consonant (CVC) word score] and binaural STI were measured. On the basis of these conditions, the relation between binaural STI and CVC word scores closely matches the STI reference curve (standardized relation between STI and CVC word score) for monaural listening. A better-ear STI appears to perform quite well in relation to the binaural STI model; the monaural STI performs poorly in these cases. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2905245]

PACS number(s): 43.71.An, 43.66.Pn [DOS] Pages: 4514–4523

I. INTRODUCTION

Speech intelligibility is most accurately and representatively measured by using subjective test procedures, involving panels of human test subjects. Unfortunately, subjective tests are cumbersome and expensive. For this reason, researchers, engineers, and acoustics consultants often rely on objective procedures to predict speech intelligibility. Examples of such procedures are the articulation index (Kryter, 1962), the speech intelligibility index (SII) (ANSI, 1997), and the speech transmission index (STI) (IEC, 2003; Steeneken and Houtgast, 1980). The SII and the STI are considered to represent the state of the art in intelligibility prediction. Although these models are generally successful in predicting intelligibility across a wide range of conditions, there are always conditions for which inaccurate results are obtained. An important source of prediction errors is that fact that standardized versions of the STI and SII are monaural models; they are based on single-channel (or single ear) estimates. By extending the prediction models to cover aspects of binaural hearing, their scope is extended to applications for which otherwise inaccurate results would be obtained.

This paper describes an extension of the STI model to a binaural intelligibility prediction model by adding algorithms that simulate binaural interaction. A similar approach could probably be adopted to modify the SII. The current paper focuses only on the STI largely for practical reasons: the STI is more widely used by acoustic consultants and engineers, due to the availability of measuring devices that are capable of rapidly producing STI values, through direct measurements.

The STI was originally designed to predict intelligibility in diotic listening conditions based on measurements with a single microphone. It is beyond the scope of this paper to give a complete description of the STI method. Basically, the method assumes that the intelligibility of a transmitted speech signal is related to the preservation of the original spectrotemporal differences between the speech sounds. These spectral differences may be reduced by bandpass limiting, masking noise, nonlinear distortion components, and distortion in the time domain (echoes and reverberation). The reduction of these spectral differences can be quantified by looking at the modulation transfer in a number of frequency (octave) bands. More background information on the STI can be found in the literature (e.g., Steeneken and Houtgast, 1980; IEC, 2003). Given the diotic listening conditions of the traditional STI, this means that all binaural (dichotic) intelligibility benefits are disregarded. The resulting inaccuracy may be considerable if sources of speech and interfering noise are separated spatially. Intelligibility, and hence the STI, depends on the relative positions of source (speech/noise) and listener within a certain space.

Potentially, the extension of the STI to a binaural model could reduce the general applicability; changes to the model might affect its validity in other conditions, unless specific precautions are taken. Care must be taken to ensure that the measured STI value is unchanged compared to the original model if binaural hearing is presumed not to play any role.

[a)]Part of this work was presented at the 151st ASA meeting in Providence, RI, 5–9 June 2006.

 0001-4966/2008/123(6)/4514/10/$23.00

Also, the attractive features of the STI method should be kept intact. In summary, this leads to the following requirements for the development of a binaural STI method:

(a) fast (15 s) measurements with a test signal in any environment;
(b) representative results in noise, reverberation, and nonlinear distortion;
(c) simple model, with very few model parameters, none of which are "tuned" to any specific application;
(d) feasible as an extension to current STI measuring devices.

These requirements almost naturally lead to the conclusion that a good option is to develop a model extension that allows STI measurements in the same way as currently standardized, but with two microphones (or rather an artificial head) instead of one. This should be achieved by incorporating a model of binaural listening into the STI framework.

II. BINAURAL INTELLIGIBILITY MODELLING

A. Background

The benefit of listening to speech with two ears instead of one in conditions with background babble is known as the cocktail party effect (Cherry, 1953). A significant body of scientific research on this topic (Bronkhorst, 2000), spanning half a century, provides ample resources to draw from for devising binaural intelligibility models.

Binaural speech intelligibility tends to be better than monaural intelligibility because of the contributions of two factors: head shadow and binaural interaction. Head shadow may result in an (effective) speech-to-noise ratio that is better at one ear than the other; by using the "better-ear" signal, the intelligibility is improved. This effect, based on interaural level differences, can probably be incorporated in the STI model relatively easily by using separate measurements corresponding to the left and right ears. The main question is how to choose from both ears: perhaps by selecting the best overall STI or selecting the best signal on a band-by-band basis (cf. Edmonds and Culling, 2006).

The effect of binaural interaction on speech intelligibility is primarily related to interaural time differences, although interaural decorrelation may also play a role in more reverberant environments (Bronkhorst and Plomp, 1990). The literature presents various models of binaural interaction (e.g., Stern and Trahiotis, 1995), mostly based on the concept of binaural cross correlation (Jeffress, 1948; Zwicker and Henning, 1985; Raatgever and Bilsen, 1986). Cross-correlation models of binaural processing help explain various auditory phenomena related to binaural hearing, such as lateralization, binaural pitch, and binaural masking level differences, while also appearing physiologically feasible (Colburn, 1995). Models of binaural interaction have been refined to a level at which detailed predictions can be obtained for many phenomena. The most important of these models, which are powerful but also quite complex, are the equalization-cancellation model (Durlach, 1963, 1972) and the auditory-nerve-based model (Colburn, 1973).

Interaural time differences (ITDs) and interaural level differences (ILDs) both contribute to an improvement in intelligibility over monaural listening. However, these contributions are not mutually independent. In an anechoic environment, an improvement in the signal-to-noise ratio (SNR) corresponding to 50% sentence intelligibility of up to 8 dB was found due to ILDs, while the improvement due to ITDs was up to 5 dB. However, the combined effect was at most 10 dB (Bronkhorst and Plomp, 1988).

A quantative model for predicting binaural advantages and directional effects in speech intelligibility was presented by Zurek (1993). It models speech and interference in 1/3-octave bands, accounting for the binaural interaction by using interaural level and phase differences. Zurek's model proved to give reasonably adequate predictions of existing data in a number of spatial configurations. However, the model is restricted to include masked speech in an anechoic environment only. Reverberation (for both speech and interference) is not incorporated, which makes this model not very suitable for typical STI applications (Houtgast and Steeneken, 2002).

Recently, Beutelmann and Brand (2006) presented a binaural intelligibility prediction model based on an extended equalization-cancellation process and the SII. They used three different acoustic environments (anechoic, office room, and cafeteria) to measure the speech reception threshold (SRT) with normal-hearing and hearing-impaired listeners. The overall correlation between predicted and observed SRTs proved to be quite high (0.95). Although, in principle, capable of handling reverberation, their model was only tested for near-field speech. Beutelmann and Brand (2006) proposed to use the STI instead of or as a correction on the SII. However, their binaural processing is quite complex, which we consider a drawback for application with the STI.

B. Incorporating binaural effects in the STI

Over the past decades, the STI model has gradually evolved from a very simple procedure suitable for a limited set of applications to a widely applicable model that is representative for most practical situations in which speech communication occurs. Features have been added to the model. For example, the current version of the STI incorporates the effects of mutual dependence between frequency bands and also the dependence of auditory masking curves on the absolute level. Whenever the model was enhanced or modified, care was taken to adhere to the following principles.

(a) The relation between STI and subjective intelligibility must remain unchanged after modification (i.e., the new version of the STI must exactly replicate results obtained with past versions, except in those cases where the "old" model was proven inaccurate).
(b) The model parameters of the STI are never tuned to a specific application. There is but one universal set of STI model parameters.
(c) STI improvements are always aimed at improving the accuracy for certain conditions. However, this always

makes the model more complex. The added complexity of a model modification must be proportional to the achieved accuracy improvement.

By sticking to the principles given above, the STI model has over the last years improved significantly without losing touch with engineers and consultants who already used it. Especially, the last principle on the list has turned out to be of great importance for the standardization of the STI. Not all modifications to the STI, proposed in the literature, have therefore been incorporated into the International Electrotechnical Commission (IEC) standard. If a new addition to the model doubles or triples its complexity, this will clearly affect the cost of STI measuring equipment. The increase in performance should warrant such an increase in cost.

For our intended extension of the STI to binaural listening conditions, model complexity is a realistic concern. The use of a comprehensive state-of-the-art binaural interaction model would greatly increase the complexity of the entire STI model. Our conclusion is that we need to look for a simplified quantification of the effects of binaural interaction. This will be less general and probably less accurate than the state of the art in binaural modeling. However, the aim is not to minimize the resulting prediction error—just to reduce this error to the same order of magnitude as other sources of variance in the STI model. Greater accuracy of the binaural interaction model would be meaningless since the overall error in the STI would then be determined by other factors (Houtgast *et al.*, 1980).

Our current proposal is to incorporate binaural interaction through the estimation of a simple interaural cross correlogram. In this, we follow the approach by Jeffress (1948), which assumes a mechanism fundamentally related to cross correlation. It is customary to incorporate auditory filter band models and hair-cell models in such an estimation of the cross correlogram. Our current aim is to simplify this as far as possible. The basic idea is visualized in Fig. 1, which shows the way in which interaural correlograms could be represented in the context of the STI model. Signals corresponding to the left (L) and right (R) ear are measured and divided into octave bands (centered from 125 Hz to 8 kHz), as customary in the STI model. In each octave band (or at least the ones covering the approximate frequency range in which humans can analyze interaural time relations, presented in gray in Fig. 1), the interaural cross correlation is calculated. The signal is reconstructed at several "internal" time delays of up to (plus or minus) a few milliseconds. Next, these internal spectral representations are analyzed in the usual way, as if corresponding to a single-channel STI measurement. This yields a quantification of the internal modulation transfer for each octave band at each interaural delay time.

The final problem is to select the most representative internal delay time for each octave band. Assuming that human binaural processing results in a straightforward strategy of intelligibility optimization, the most likely candidate is the internal delay at which the maximum modulation transfer is observed. By using these results, an overall (binaural) STI can be calculated.

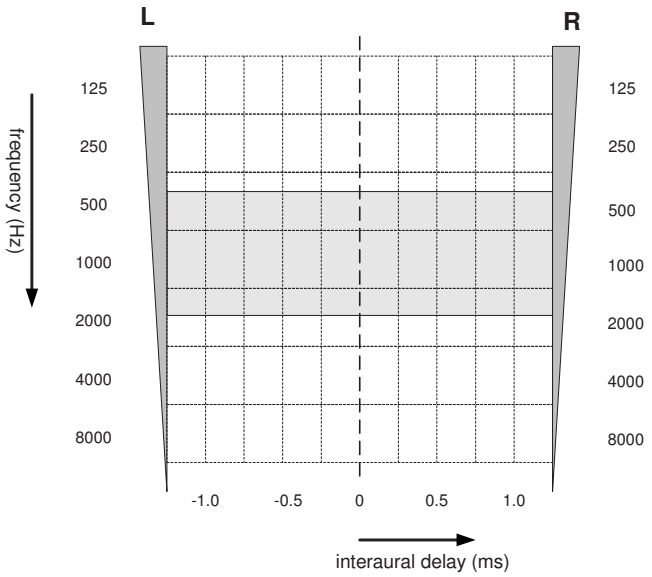


FIG. 1. Visualization of a "grid" for displaying interaural cross correlograms in the context of the STI model. Left (L) and right (R) ear signals are divided into octave bands. In applicable octave bands (gray rectangle), the interaural cross correlation is calculated and the signal is reconstructed at several "internal" time delays.

Within the framework described here, a number of different binaural STI implementations can be thought of. Construction and implementation of such a model comes down to choosing which degree of simplification is accepted and choosing model parameters. This process is outlined in the next section.

III. IMPLEMENTATION OF A BINAURAL STI MODEL: DESIGN CHOICES

A binaural STI model based on the framework described in the previous chapter was designed and implemented in MATLAB®. The measurements on which binaural STI calculations are based are straightforward extensions of the normal standardized STI measurements with the following adaptations:

(a) Each binaural STI measurement is based on a two-channel recording, obtained using an artificial head. This artificial head marks the position of the (simulated) listener.
(b) The test signal—played back at the position of a simulated talker—can be any standardized telecommunications (STI) signal, such as STITEL, STIPA, or full STI (cf. IEC, 2003). The exception is room acoustics (RASTI) , which cannot be used since the 1 kHz band, which is essential in binaural listening, is not included in the signal

A first approximation of binaural speech intelligibility is obtained by calculating the STI from the left and right ears of the artificial head separately. This can be done by using a standard equipment. By taking the highest STI (better ear), the effects of interaural level differences are taken into account. This approach can be slightly refined by taking the best signal (left-right) on a band-by-band basis. The better-

ear approximation is expected to underestimate the contribution of frequency bands in which interaural time differences can be used to (perceptually) enhance the speech signal. This is where the model could be extended.

Since the frequency range in which the most useful binaural interaction for speech intelligibility takes place extends from 500 to 1500 Hz (Zurek, 1993; Blauert, 1996), the octave bands centered at 500, 1000, and 2000 Hz in the STI analysis should be affected. For these three octave bands, interaural correlograms are calculated. The overall procedure is given below. Note that the choice of parameters is more fully explained in Sec. V, where we also discuss the optimization process to come to the final settings.

(a) The recorded signals (left and right ear of the artificial head) are analyzed in octave bands.
(b) For all bands, the modulation transfer function is calculated for the left and right ears separately.
(c) For the three frequency bands centered at 500, 1000, and 2000 Hz, an interaural correlogram is calculated. This is done in the following way:
    (1) The band-filtered signals are separated into nonoverlapping time frames of 30 ms duration and squared.
    (2) The left and right (squared) signals within each frame are cross correlated, resulting in a cross correlation of interaural delay for each frame (and for each filter).
    (3) Data for delay magnitudes >2 ms are discarded: the rest are kept.
    (4) Any offset in the cross-correlation function is subtracted so that the lowest value is set to zero.
    (5) Effectively, one interaural cross correlogram per frame is obtained for each band. The signal envelope can now be calculated for any interaural delay.
    (6) For a set of discrete interaural delays (τ) in the range $-2<\tau<2$, the signal power as a function of time is taken. This is already low pass filtered (with a cutoff frequency corresponding to $\frac{1}{2}$ the frame rate, i.e., 15 Hz). By using conventional techniques, the modulation transfer function (MTF) is calculated as a function of internal delay and frequency band (cf. Steeneken and Houtgast, 1980). The internal delay is selected at which the overall MTF contribution is highest (which leads to the highest STI, taking upward spread of masking into account as well). The MTF values for this internal delay are used.
(d) For the octave bands centered at 125 and 250 Hz and at 4000 and 8000 Hz, only the MTFs corresponding to left and right ears are considered. The highest value is taken (left ear or right ear) for each octave band separately.
(e) The selected (highest) MTF data from each of the seven octave bands are now combined to calculate an overall STI.

The rationale behind this approach is that for each separate band, the internal signal offering the most information, in terms of preservation of signal modulation, is presumed to be selected. How the described use of interaural correlograms would predict benefits related to interaural time delays is easily understood by considering the case of a single noise source and a single speaker, both in front of the listening position. If the speaker is slightly off to the left and the noise source to the right, then the maximum interaural correlation for the speech will be at a certain negative interaural delay, and the maximum interaural correlation for noise will be at a positive interaural delay. The power signals at these delays will have different modulation depths correspondingly.

This approach contains gross simplifications compared to accepted binaural models, such as the use of octave band filters instead of the much narrower auditory band filters. Also, instead of using inner-ear hair-cell models, we simply take the square of the signal amplitude. These choices were made in order to choose as simple a model as can be shown to work. However, the implementation of the binaural model, as described here, is only meaningful if it can be shown to yield sufficiently accurate predictions of speech intelligibility. To this end, equally balanced consonant-vowel-consonant (CVC) intelligibility tests were carried out in 39 binaural listening conditions. The validation carried out with the results from these listening tests is described below.

IV. VALIDATION OF THE BINAURAL STI

A. Speech material

The preferred method for subjective measurement of speech intelligibility in relation to the STI makes use of CVC words (Steeneken, 1992). This method uses simple nonsense words, embedded in carrier phrases, which were recorded digitally under good laboratory conditions (high quality microphones and no ambient noise). The recorded material consists of speech by eight speakers (four males and four females). Sequences of CVC test words were combined to obtain word lists of 51 words each. The source was digitally transferred to a computer, resampled to 22 kHz, and stored with 16 bit resolution. All CVC scores given in this report are the so-called *equally balanced* CVC scores. Since all phonemes have the same frequency of occurrence in the corpus of the test stimuli, the CVC score is by definition equally balanced.

The material was filtered with anechoic binaural impulse responses recorded with a Head Acoustics HMS III.2 dummy head (zero elevation and different azimuths) and with binaural impulse responses of environments (listening room, class room, and Grundtvigs cathedral) simulated in the ODEON® 7.0 software (Christensen, 2003). Speech shaped noise was also filtered with these binaural impulse responses.

B. Experimental design

The anechoic conditions all had a talker in front of the listening position and an interfering noise source (at signal-to-noise ratios of −3 and −6 dB) at positions around the head (0°, 30°, 60°, 90°, and 150°). In addition, conditions were created in which noise (correlated and uncorrelated between the ears) was directly added to the speech signals.

In the various (simulated) listening environments, realistic source and receiver positions were defined. Noise sources were also included in the simulation; Head-related transfer functions were included in the binaural room impulse responses yielded by Odeon. The overall impression

when listening to speech processed in this way was that a high degree of face value was offered by the simulations. Some conditions were included in which noise was added without binaural processing (diotic and dichotic/uncorrelated noise conditions). The speech material and noise files—a single speech source and a single noise source from various directions—were mixed electronically in different SNRs. This resulted in a set of 39 conditions. A survey of the 39 binaural conditions is given in the Appendix. The full STI signal was used for the speech source, i.e., 14 modulated signals per octave band with 7 simultaneous modulations. Noises were used as given in Table I.

The currently standardized version of the (monaural) STI has been validated by analyzing third-order polynomial fits through CVC data points (Steeneken, 1992). The same approach was now followed; however, instead of fitting a new polynomial through the data, the average monaural polynomial is plotted in each figure for comparison. Given our goal to have the binaural STI yield results that can be interpreted numerically in the same way as the existing STI versions, this seems to be a more appropriate choice.

To verify the validity of the monaural STI-CVC reference curve, a CVC test in a standard set of 40 representative monaural listening conditions was also carried out with 4 listeners, and the associated STI was calculated. Since exactly the same paradigm was used for the binaural conditions (except for the difference between diotic and binaural listening), a good correspondence between data from this monaural experiment and the reference curve serves to validate the applied implementation of the CVC test and the STI measurement. A survey of the monaural conditions is given in the Appendix.

C. Subjects

A total of seven young normal-hearing subjects (five males and two females, age range of 19–23 years) participated in the listening tests. All seven participated in the test with binaurally processed CVCs; four subjects participated in the test with monaurally processed CVCs. They were paid for their services.

D. Procedure

The processed lists were balanced for conditions and speakers and presented over headphones in a paced open-response test to the listeners, who were asked to respond by typing the perceived syllable on a computer keyboard. The individual responses were manually checked for typographic errors and inconsistencies, and then automatically processed. Hence, each data point consists of 56 speaker-listener pairs (8 speakers $\times$ 7 listeners). After the processing of the individual results, the mean equally balanced CVC score was calculated for each condition.

E. Results and discussion

The results show that the experiment covers almost the entire range of possible CVC scores, from 10% to 90% correct. The data are nicely spread between the minimum and maximum values. Results relating CVC scores and STI of the monaural conditions are given in Fig. 2. All STI values reported in this paper are obtained through full STI measurements (7 octaves and 14 modulation frequencies for each octave band).

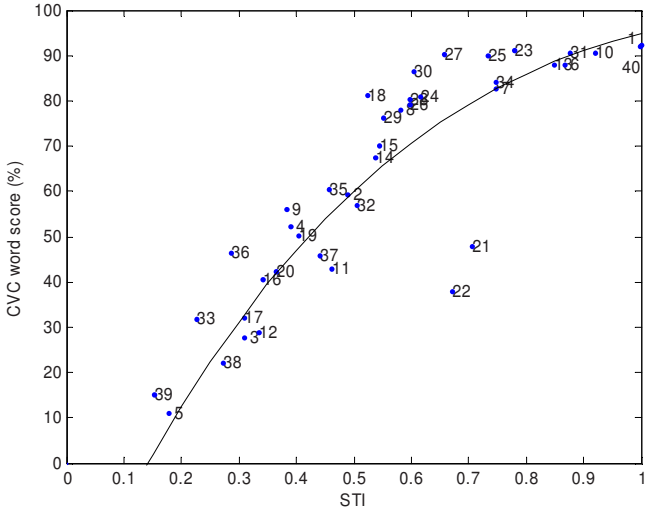


FIG. 2. (Color online) Validation of the STI vs CVC reference curve, using a set of 40 monaural reference conditions featuring noise and bandwidth limiting (1–14), nonlinear distortion (15–22), echoes (23–30), and reverberation (31–39). Condition 40 is a condition without any type of signal degradation.

Figure 2 shows that except for conditions 21 and 22, the relation between CVC and STI in monaural conditions is—on the whole—adequately described by the reference curve. Conditions 21 and 22 are center clipping conditions, for which the STI is known to overestimate intelligibility. Center clipping is nowadays rarely found in practice; it occurs with old-fashioned carbon microphones and poorly aligned push-pull amplifiers. Another noticeable deviation from the reference curve is seen at CVC scores above STI $=0.55$, where the scores appear to be approaching the saturation level more quickly. The standard deviation (or rather rms deviation), representing the vertical spread around the reference curve (cf. Steeneken, 1992), is 11.37%. This is similar to the original data set on which the reference curve is based.[1]

The relation between CVC scores and the binaural STI, in the binaural conditions described above, is shown in the top panel of Fig. 3. For comparison, the mean-ear STI in these conditions, averaged between both ears of the artificial head, is given in the middle panel, and the better-ear STI (i.e., the highest STI value of either left or right ear, as processed across all octave bands) in the bottom panel. Figure 3 shows that the relation between the binaural STI and the CVC word score comes quite close to the reference curve; the standard deviation is 9.2%, which is even smaller than for the monaural conditions of Fig. 2. For most conditions, the binaural STI seems to underestimate the intelligibility somewhat, with the exception of a cluster of data points in the cathedral environment (18–21) for which the STI is overestimated. The mean-ear STI (middle panel of Fig. 3) clearly underestimates the intelligibility in these binaural conditions with a standard deviation of 28.3%. The better-ear STI as in Fig. 3 also underestimates the binaural intelligibility. The

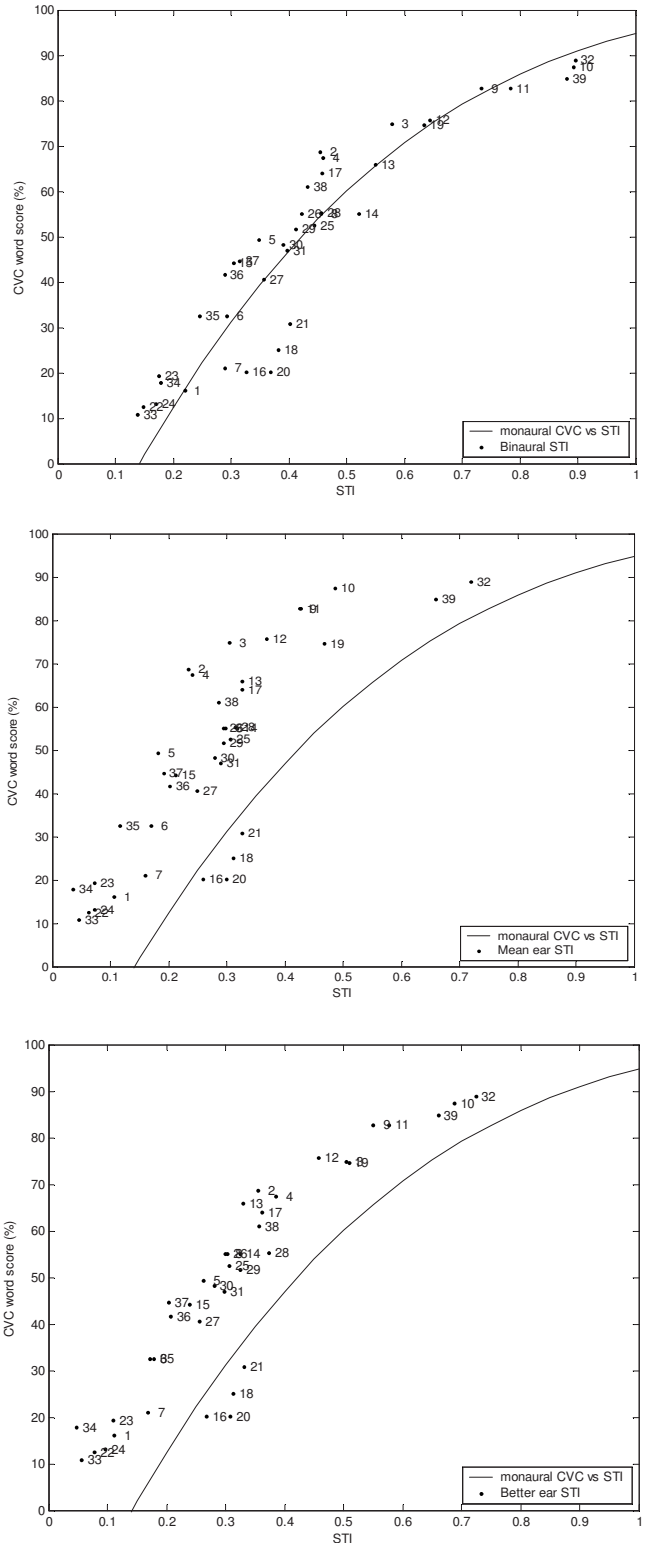


FIG. 3. CVC word score (seven subjects) as a function of the binaural STI (top panel), mean-ear STI, averaged between both ears of the artificial head (middle panel), and better-ear STI (bottom panel). Binaural conditions include anechoic conditions (1–14), a cathedral environment (15–21), a classroom (22–32), and a listening room (33–39).

standard deviation is 21.2%, which is considerably worse than for the binaural model. The better-ear STI does not take the ITD effect into account. In general, we can estimate the ITD effect to a maximum of 3 dB (difference between diotic/correlated and dichotic/uncorrelated interference). By taking the better-ear STI with a 3 dB "correction"—corresponding to a horizontal right-hand shift of 0.1 STI in Fig. 3—the data points come closer to the monaural reference curve. As a consequence, the standard deviation decreases significantly to 10.6%, quite close to the 9.2% of our binaural model.

To investigate the data further, the relation between binaural STI and CVC is given in Fig. 4 for separate categories of conditions. In various environments (anechoic, classroom, and listening room), conditions were included for which the noise presented diotically was either identical (maximum correlation) or uncorrelated. These data points are presented in a separate curve in Fig. 4.

In the anechoic conditions, the binaural STI is slightly underestimated at lower speech-to-noise ratios (−6 dB). In the cathedral environment, the binaural STI performs poorly in some cases, as observed before. This turns out to be at very large source-receiver distances (>30 m). Fortunately, such conditions are quite rare in real life. Here, more accurate estimates are actually obtained by taking the better-ear STI.

For the correlated/identical noise conditions, one would expect a difference between the anechoic conditions and to the conditions in simulated acoustic environments. Identical noise at both ears creates a clear "peak" in the interaural correlogram around an internal delay of 0 ms; uncorrelated noise contributes more or less equally at all internal delays. In an anechoic environment, speech originated from a source azimuth of 0°, straight in front of the listener position. Hence, the interaural correlation is optimal for an internal delay of 0 ms and diotic noise is expected to be a more effective masker than uncorrelated noise. However, in reverberant environments, speech signal contributions are spread out across a range of internal delays. In this case, diotic noise is expected to be less effective since the listener can "listen around" the noise peak at 0 ms (in terms of our model, the maximum STI is realized at internal delays other than 0 ms). In summary, when we subtract the intelligibility for uncorrelated noise from that for diotic noise, a positive value is expected in reverberant environments and a negative value in the anechoic environment. This is also the result found in the CVC experiment: The binaural STI correctly predicts a positive difference in reverberant conditions (+4.5% CVC difference for +0.034 STI difference) and a negative difference in the anechoic conditions (−11% CVC difference for −0.021 STI difference). However, the magnitudes of the differences are not predicted very well.

V. GENERAL DISCUSSION

Overall, the results presented in Figs. 3 and 4 appear satisfactory. However, an important question is to which degree the results presented here are influenced by the choice for the model parameters. In the proposed version, the binaural STI model has only a few free parameters, which are (a) octave bands to include in the binaural interaction model (500, 1000, and 2000 Hz), (b) range of internal delays to consider (−2–2 ms), (c) operator used to selected MTF contributions (maximum STI contribution), and (d) frame rate.

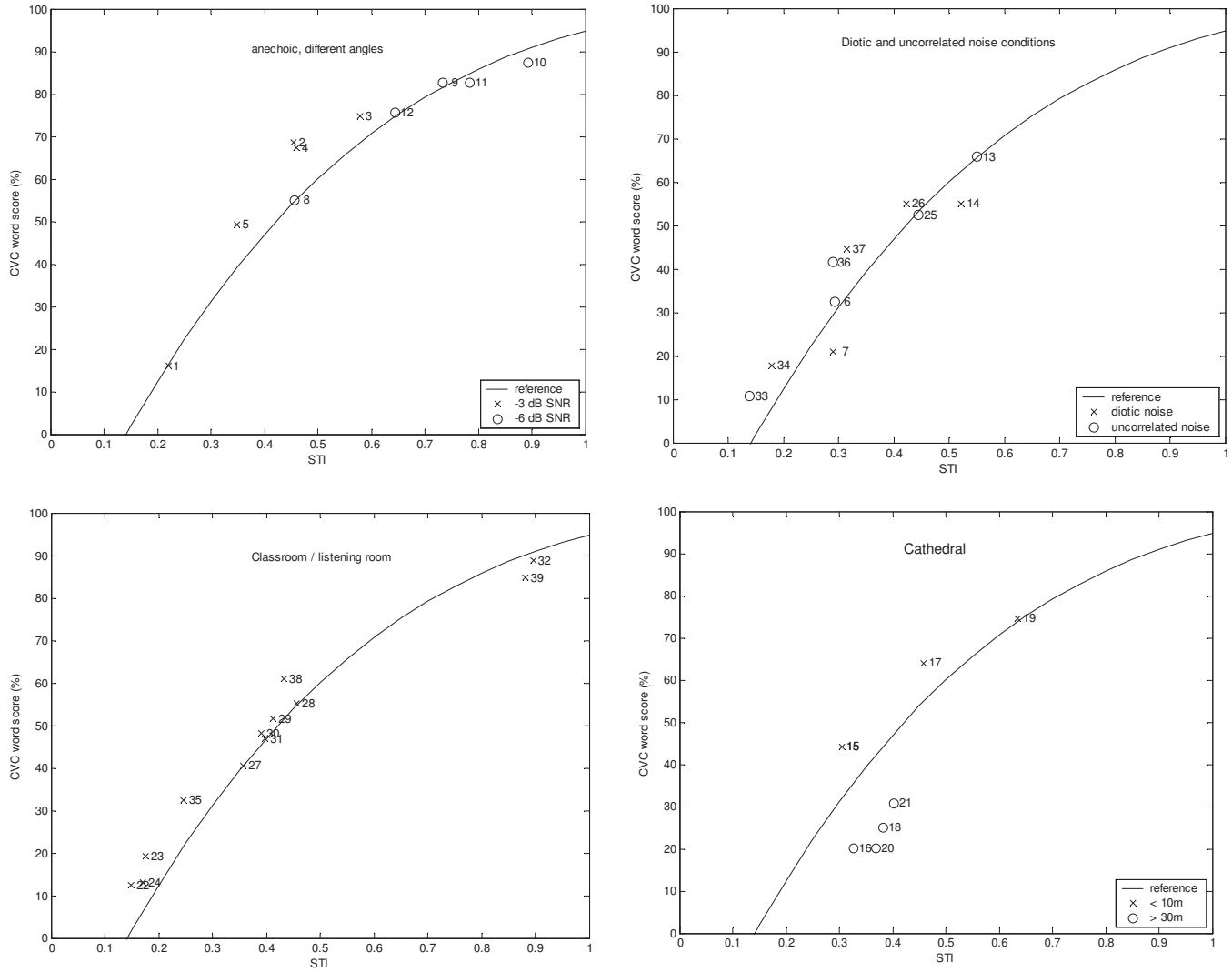


FIG. 4. Relation between CVC and binaural STI, shown separately for subsets of the binaural listening conditions (see Table I for the ID per condition). The anechoic conditions varied with respect to the noise source position. The conditions labeled "diotic noise" and "uncorrelated noise" represent various acoustic environments but with a noise signal added to both ears that is either the same or uncorrelated (no convolution with binaural impulse responses). The classroom and listening room conditions represent various source and listener positions. The cathedral conditions differed mainly with respect to the distance between the source and the receiver (here grouped in two distinct categories).

A. Octave bands

The choice which octave bands are included in the binaural interaction analysis follows from known limits of the binaural system reported in the literature. In particular, the choice to include the 2 kHz octave may be considered questionable since binaural interaction is normally presumed relatively ineffective at these frequencies, although certainly present in the lower half of the octave band. Introduction of a frequency weighting mechanism, which could be used to solve this dilemma, will only be considered as a last resort since it adds free (tunable) parameters to the model. Figure 5 shows that leaving out the 2 kHz band only slightly affects the results. Leaving out the 2 kHz *and* the 500 Hz bands clearly leads to less accurate results.

B. Internal delays

To investigate the effect of the range of internal delays taken into account, STI calculations were performed for various choices of this range. Theory predicts that interaural de-

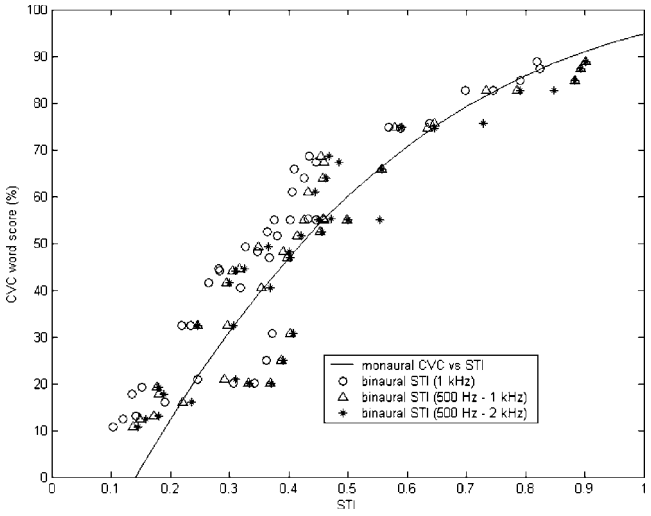


FIG. 5. CVC word score (seven subjects) as a function of the binaural STI for which binaural interaction is taken into account for one (1 kHz), two (500 Hz–1 kHz), or three (500 Hz–2 kHz) octave bands.

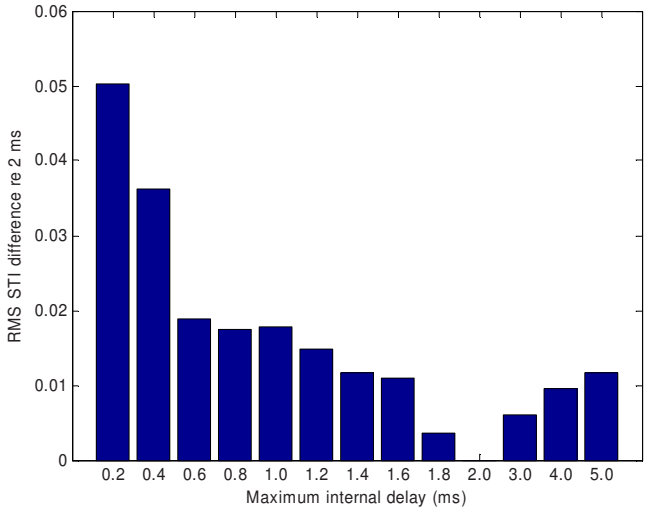


FIG. 6. (Color online) Mean absolute difference between binaural STI with a maximum internal delay of 2 ms (default) and other maximum internal delay settings. The mean is taken across the entire set of binaural conditions.

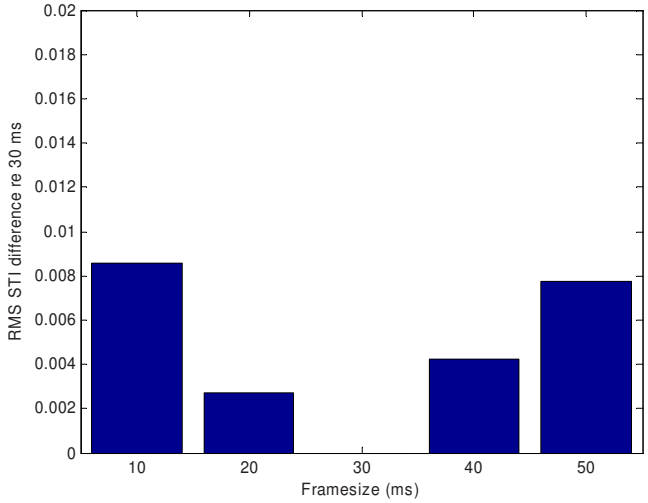


FIG. 7. (Color online) Mean absolute difference between binaural STI with frame size of 30 ms and other frame sizes. The mean is taken across the entire set of binaural conditions.

lays greater than 1 ms cannot be used effectively to enhance our internal representation of the signal (e.g., Raatgever and Bilsen, 1986). Maximum interaural delays occur for sound sources that are on the horizontal plane at an azimuth of 90° or −90°. The approximate interaural time difference is then calculated by the distance between the ears and the speed of sound; this is approximately 0.5 ms. Assuming that intelligibility benefits due to binaural interaction are limited to ecologically feasible interaural time differences, including internal delays greater than, say, 1 ms, should not result in an increased binaural STI. Figure 6 shows that this is exactly how the model behaves. Across the entire set of binaural conditions, the mean absolute difference was calculated between the binaural STI with our default internal delay range (±2 ms) and the binaural STI at various other internal delay ranges (±0.1–5 ms).

Keeping in mind that differences up to 0.03 STI occur "naturally" in monaural STI measurements due to the normal measurement error, Fig. 6 shows that it does not make a great difference whether internal delays are taken into account up to 1, 2, or 3 ms. However, if the range of internal delays is limited to a smaller maximum than, say, 1 ms, the calculated binaural STI becomes somewhat less accurate. The standard deviation relative to the monaural reference curve is, as stated above, 9.2% for the default internal delay setting (2 ms). For 0.4 ms, this standard deviation increases to 11.1%, for 0.2 ms to 12.1%, and for 0.1 ms to 21.0%.

C. MTF contributions

The choice to take the maximum of the MTF for any internal delay (instead of, for instance, the mean or median) results from the hypothesis that our binaural system selectively tunes into areas of the binaural correlogram where most information is available.

D. Frame rate

The frame rate is a parameter that may appear to be freely adjustable but for which the choices are limited for computational reasons. The STI method measures modulation frequencies up to 12.5 Hz. This means that the signal envelopes extracted from the sequence of binaural correlograms must be reliable up to this frequency, imposing a minimum frame rate of 25 Hz. Thus, the frame size must be less than 40 ms. On the other hand, the frame size must not be too small to prevent loss of accuracy in determining the interaural delays. The STI method uses a decimating filter bank to filter the signal into octave bands. This means that the signal in the 500 Hz band is sampled at 2756 Hz if the original sampling frequency is 44 100 Hz. When working with a frame size of 30 ms (our default), then this comes down to 83 samples per frame, which turns out to produce cross-correlation functions of acceptable accuracy. Shorter frames will lead to less accurate estimates of the cross-correlation function.

To determine the effect of frame size on the binaural STI, calculations were carried out similar to Fig. 3 but at frame sizes of 10, 20, 40, and 50 ms instead of 30 ms.

The effect on the calculated STI appears to be small. STI values computed on the basis of 10, 20, 40, or 50 ms are virtually identical to the values computed with a 30 ms frame size. To show this more clearly, the mean absolute difference was calculated (across the entire set of binaural conditions) between the binaural STI calculated with various frame sizes and the default frame size of 30 ms. Results are shown in Fig. 7. The mean absolute difference between monaural across STI measurements in the same condition is normally, due to measurement error alone, around 0.03. In this light, the effect of frame size is relatively minor. So, it seems fair to conclude that the model is not overly sensitive to the choice of the frame size (or the corresponding frame rate).

VI. CONCLUSIONS

When using the standard speech transmission index to predict speech intelligibility in binaural listening conditions, the intelligibility is underestimated. Significant improvement is already obtained by simply doing a two-channel STI mea-

TABLE I. Survey of the 39 binaural conditions in different environments, speech and noise positions, and signal-to-noise ratios.

Condition	ID	Speech azimuth at distance	Noise azimuth at distance	SNR (dB)
Anechoic	1, 8	0° at 1.1 m	0° at 1.1 m	−6, −3
	2, 9	0° at 1.1 m	150° at 1.1 m	−6, −3
	3, 10	0° at 1.1 m	30° at 1.1 m	−6, −3
	4, 11	0° at 1.1 m	60° at 1.1 m	−6, −3
	5, 12	0° at 1.1 m	90° at 1.1 m	−6, −3
	6, 13	0° at 1.1 m	Dichotic	−6, −3
	7, 14	0° at 1.1 m	Diotic	−6, −3
Listening room (T30≈0.4 s)	33, 36	0° at 2.6 m	Dichotic	−6, −3
	34, 37	0° at 2.6 m	Diotic	−6, −3
	35, 38	0° at 2.6 m	90° at 0.8 m	−6, −3
	39	0° at 2.6 m	No noise	∞
Classroom (T30≈0.5–1 s)	22	300° at 4.4 m	302° at 2.9 m	−6
	23	300° at 4.4 m	0° at 7.5 m	−6
	24	0° at 7.5 m	300° at 4.4 m	−6
	25, 26	0° at 2.0 m	Dichotic, diotic	−3
	27	0° at 2.0 m	0° at 2.0 m	−3
	28	0° at 2.0 m	320° at 4.2 m	−3
	29	0° at 2.0 m	230° at 3.4 m	−3
	30	0° at 2.0 m	180° at 2.2 m	−3
	31	0° at 2.0 m	140° at 2.9 m	−3
	32	0° at 2.0 m	No noise	∞
Cathedral (T30≈1.5–14 s)	15, 17	260° at 7 m	355° at 38 m	0, +3
	16, 18	345° at 38 m	270° at 7 m	0, +3
	19	345° at 38 m	No noise	∞
	20	5° at 31 m	No noise	∞
	21	330° at 33 m	No noise	∞

TABLE II. Survey of the 40 monaural conditions with different bandpass, nonlinear (peak and center clipping), and echo conditions in various signal-to-noise ratios. The peak clip level is −24 dB below the 1% speech peak level. Center clipping conditions 21 and 22 have clip levels −24 and −21 dB below the 1% speech peak level, respectively.

Condition	ID	Bandwidth (Hz)	Noise type	SNR (dB)	Echo (ms)	RT60 (ms)
Unprocessed	40	10–16 000	⋯	∞	⋯	⋯
Bandpass only	1	50–10 500	⋯	∞	⋯	⋯
	2, 3	50–10 500	White	0, −8	⋯	⋯
	4, 5	50–10 500	Pink	0, −8	⋯	⋯
	6, 7	50–10 500	Low	3, −3	⋯	⋯
	8, 9	50–10 500	Speech	3, −3	⋯	⋯
	10	300–3 400	⋯	∞	⋯	⋯
	11	300–3 400	White	0	⋯	⋯
	12	300–3 400	Pink	0	⋯	⋯
	13	300–3 400	Low	3	⋯	⋯
	14	300–3 400	Speech	3	⋯	⋯
Peak clip (+bandpass)	15	50–10 500	⋯	∞	⋯	⋯
	16	50–10 500	White	6	⋯	⋯
	17	50–10 500	Speech	3	⋯	⋯
	18	300–3 400	⋯	∞	⋯	⋯
	19	300–3 400	White	6	⋯	⋯
	20	300–3 400	Speech	6	⋯	⋯
Center clip	21,22	50–10 500	⋯	∞	⋯	⋯
Echo (+bandpass)	23	50–10 500	⋯	∞	50	⋯
	24	50–10 500	Speech	6	50	⋯
	25	50–10 500	⋯	∞	100	⋯
	26	50–10 500	Speech	6	100	⋯
	27	50–10 500	⋯	∞	200	⋯
	28	50–10 500	Speech	12	200	⋯
	29	50–10 500	Speech	6	200	⋯
	30	300–3 400	⋯	∞	200	⋯
Reverberation	31	50–10 500	⋯	∞	⋯	200
	32,33	50–10 500	Speech	6, −3	⋯	200
	34	50–10 500	⋯	∞	⋯	500
	35,36	50–10 500	Speech	6, 0	⋯	500
	37	50–10 500	⋯	∞	⋯	2000
	38,39	50–10 500	Speech	6, 0	⋯	2000

surement using an artificial head and working with the better-ear STI. However, in some conditions, this simple approach still considerably underestimates the actual intelligibility.

On the basis of the 39 binaural conditions tested in this paper, the proposed binaural STI model is capable of predicting binaural speech intelligibility with the same approximate accuracy offered by the traditional STI in monaural listening conditions. We also found that, overall, a simple better-ear STI appears to perform quite well in relation to the binaural STI model. The attractiveness of this particular binaural STI lies in a few features.

(a) The model is motivated by the existing binaural theory, considerably simplified.
(b) There are only a few "free" model parameters (frequency range and internal delay range).
(c) Changing these model parameters within reasonable bounds has little effect on the outcome of the model.
(d) The model is simple and computationally inexpensive.
(e) Known subjective binaural intelligibility data are accurately predicted by the model.

The fact that the model is relatively insensitive to changes in the model parameter values increases confidence in the strength of the model itself; it reduces that likelihood that the correspondence between subjective data and predicted intelligibility is the result of "fitting" rather than "modeling."

ACKNOWLEDGMENTS

This research was supported by grants from the European Union FP6, Project No. 004171 HEARCOM. The authors wish to thank Bastiaan van Gils for conducting the CVC experiments and Claus Lynge from Ørsted DTU, Acoustic Technology, Technical University of Denmark for providing us with the binaural impulse responses of the listening room, classroom, and cathedral.

APPENDIX: SURVEYS OF THE CONDITIONS USED IN THE EVALUATION

Tables I and II above give a survey of the binaural and

monaural signal processing conditions used in the CVC evaluation experiments described in Sec. IV.

[1]The actual standard deviations reported by Steeneken (1992) were somewhat lower (up to 8%), but these were calculated separately by category of distortions; the reference curve was fitted individually to each category. Also, the center clipping points were excluded from the standard deviation calculation. When calculated in the same straightforward way applied here, the standard deviation for Steeneken's data is about 12%.

ANSI (**1997**). "Methods for calculation of the speech intelligibility index," ANSI Report No. S3.5-1997, American National Standards Institute, New York.

Beutelmann, R., and Brand, T., (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331–342.

Blauert, J., (**1996**). *Spatial Hearing* (MIT, Cambridge, MA), Chap. 4, pp. 313–324.

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.

Bronkhorst, A. W., and Plomp, R., (**1988**). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am. **83**, 1508–1516.

Bronkhorst, A. W., and Plomp, R., (**1990**). "A clinical test for the assessment of binaural speech perception in noise," Audiology **29**, 275–285.

Cherry, E. C., (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Christensen, C. L., (**2003**). *Odeon Room Acoustics Program Version 6.5 User Manual* (Technical University of Denmark, Lyngby).

Colburn, H. S., (**1973**). "Theory of binaural detection based on auditory-nerve data. General strategy and preliminary results on interaural discrimination," J. Acoust. Soc. Am. **54**, 1458–1470.

Colburn, H. S., (**1995**). "Computational models in binaural processing," in *Auditory Computation*, edited by H. Hawkins and T. McMullin (Springer, New York).

Durlach, N. I., (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**, 1206–1218.

Durlach, N. I., (**1972**). "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, edited by J. V. Tobias (Academic, New York), pp. 369–462.

Edmonds, B. A., and Culling, J. F., (**2006**). "The spatial unmasking of speech: Evidence for better-ear listening," J. Acoust. Soc. Am. **120**, 1539–1545.

Houtgast, T., and Steeneken, H. J. M., (**2002**). "The roots of the STI approach," in *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg).

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.

IEC (**2003**). "Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index," IEC Standard 60268-16 (3rd edition), International Electrotechnical Commission, Geneva Switzerland.

Jeffress, L. A., (**1948**). "A place theory of sound localization," J. Comp. Physiol. Psychol. **41**, 35–39.

Kryter, K. D., (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Raatgever, J., and Bilsen, F. A., (**1986**). "A central spectrum theory of binaural processing. Evidence from dichotic pitch," J. Acoust. Soc. Am. **80**, 429–441.

Steeneken, H. J. M., (**1992**). Ph.D. thesis, University of Amsterdam, Amsterdam.

Steeneken, H. J. M., and Houtgast, T., (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Stern, R. M., and Trahiotis, C., (**1995**). "Models of binaural interaction," in *Hearing*, edited by B. C. J. Moore (Academic, London), pp. 347–386.

Zurek, P. M., (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, London), Chap. 15, pp. 255–276.

Zwicker, E., and Henning, G. B., (**1985**). "The four factors leading to binaural masking-level differences," Hear. Res. **19**, 29–47.

Identification and discrimination of bilingual talkers across languages[a)]

Stephen J. Winters,[b)] Susannah V. Levi,[c)] and David B. Pisoni[d)]
Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana 47405

(Received 9 May 2007; revised 2 April 2008; accepted 2 April 2008)

This study investigated the extent to which language familiarity affects the perception of the indexical properties of speech by testing listeners' identification and discrimination of bilingual talkers across two different languages. In one experiment, listeners were trained to identify bilingual talkers speaking in only one language and were then tested on their ability to identify the same talkers speaking in another language. In the second experiment, listeners discriminated between bilingual talkers across languages in an *AX* discrimination paradigm. The results of these experiments indicate that there is sufficient language-independent indexical information in speech for listeners to generalize knowledge of talkers' voices across languages and to successfully discriminate between bilingual talkers regardless of the language they are speaking. However, the results of these studies also revealed that listeners do not solely rely on language-independent information when performing these tasks. Listeners use language-dependent indexical cues to identify talkers who are speaking a familiar language. Moreover, the tendency to perceive two talkers as the "same" or "different" depends on whether the talkers are speaking in the same language. The combined results of these experiments thus suggest that indexical processing relies on both language-dependent and language-independent information in the speech signal.
 [DOI: 10.1121/1.2913046]

PACS number(s): 43.71.Bp, 43.71.Hw [PEI] Pages: 4524–4538

I. INTRODUCTION

Speech researchers have traditionally distinguished between the *linguistic* and the *indexical* properties of speech (Abercrombie, 1967). Linguistic properties of speech provide information about the message that the speaker is trying to convey, while indexical properties provide cues to personal characteristics of the speaker—such as age, gender, sociolinguistic background, or emotional state. While both indexical and linguistic information are simultaneously transmitted to listeners in the same speech signal, it is an open question as to what extent these properties of speech may interact with each other in perception. One possibility is that listeners process the indexical and linguistic properties of speech independently of one another, while another possibility is that the indexical and linguistic properties of speech are inextricably bound to one another in the speech signal and therefore interact in both language processing linguistic and indexical tasks.

In this study, we investigated the extent to which linguistic and indexical properties interact in speech perception by testing the ability of listeners to identify and discriminate bilingual talkers across languages. We first trained listeners to identify the voices of bilingual talkers from speech samples produced in one language only. We then tested their ability to identify or discriminate those same talkers' voices while they were speaking both the training language and a novel language. Since talker identification and discrimination are both indexical processing tasks, this study investigated the possible interaction of linguistic and indexical properties in speech perception by solely looking at the effects of language on indexical processing. If linguistic and indexical information do interact in speech perception, then performance in both tasks should depend on the particular language that the talkers are speaking. If, on the other hand, linguistic and indexical properties do not interact, then listeners' ability to identify and discriminate between talkers' voices should be independent of the language in which those talkers are speaking.

Paradoxically, the existing research literature provides evidence which suggests that the linguistic and indexical properties of speech interact in perception, and also that they may be independently processed. Evidence for the independent processing of linguistic and indexical information in speech comes from behavioral and neurological studies which indicate that listeners can successfully perform linguistic and indexical tasks when they do not have recourse to the other kind of information in the signal. For example, several studies have shown that listeners can identify talkers from time-reversed samples of speech, the linguistic content of which is unintelligible (Bricker and Pruzanksy, 1966; Clarke *et al.*, 1966; Williams, 1964). The same independence of talker and linguistic information was also found to a lesser

[a)]Portions of this work were presented at the 80th Annual Meeting of the Linguistic Society of America in Albuquerque, NM and LabPhon10 in Paris, France.
[b)]Present address: Department of Linguistics, University of Calgary. Electronic mail: swinters@ucalgary.ca
[c)]Present address: Department of Linguistics, University of Michigan. Electronic mail: svlevi@umich.edu
[d)]Electronic mail: pisoni@indiana.edu

 0001-4966/2008/123(6)/4524/15/$23.00

extent in filtered speech (Compton, 1963; Pollack *et al.*, 1954) and whispered speech (Pollack *et al.*, 1954; Williams, 1964). Phonagnosia, a phenomenon in which neurologically impaired listeners lose the ability to identify the voices of familiar talkers even though they can still comprehend spoken utterances in a familiar language, provides evidence that the linguistic processing of speech can also take place independently of talker recognition (Van Lancker *et al.*, 1988). Neurological research has also shown that indexical and linguistic information are processed in different parts of the brain. Landis *et al.* (1982) found hemispheric specialization for linguistic but not indexical information, while more recent studies isolated indexical processing to specific brain regions. Glisky *et al.* (1995) found that listeners with low frontal lobe function exhibited impaired indexical processing, while listeners with low medial temporal lobe function exhibited impaired linguistic processing. Additionally, Stevens (2004) found that voice discrimination primarily resulted in activation in the right frontoparietal area, whereas lexical discrimination was associated with the left frontal and bilateral parietal areas. Taken together, these behavioral and neurological findings suggest a double dissociation between linguistic comprehension and talker recognition: both processes can operate independently of one another.

In contrast, several studies have shown that the linguistic and indexical properties of speech do interact in perception. Furthermore, this interaction is bidirectional: indexical properties can affect linguistic processing, and linguistic knowledge can affect the processing of indexical information. The dependence of linguistic processing on indexical information has been documented in several studies that systematically varied the number and type of voices presented in linguistic processing tasks (Mullennix and Pisoni, 1990; Goldinger, 1996). Varying indexical information in this way consistently resulted in worse performance on these linguistic tasks. Other studies have also shown that the indexical and linguistic properties of speech are encoded and stored together in representations of spoken words in memory, thus facilitating the linguistic processing of messages spoken by familiar talkers (Goldinger *et al.*, 1991; Schacter and Church, 1992; Palmeri *et al.*, 1993; Nygaard *et al.*, 1994; Nygaard and Pisoni, 1998).

The results of these studies were particularly influential in the development of exemplar-based theories of speech perception (Johnson, 1997; Goldinger, 1998; Pierrehumbert, 2001). These theories hold that listeners store individual experiences of speech—in a relatively unanalyzed form—in memory. Representations of linguistic categories thus consist of memory traces of particular words, spoken by particular talkers, in particular contexts, and at specific places and times. In these models, the categorization of new speech experiences operates on based on the combined, similarity-based activation response of all stored exemplars to incoming speech tokens. Since both indexical and linguistic information are stored together in the speech exemplars in memory, both of these properties may interact with each other in the processing of incoming speech tokens. In particular, speech produced either in a familiar language or by familiar talkers will be facilitated through activation of similar exemplars stored in memory.

A facilitating influence of linguistic knowledge on indexical processing was established by a variety of studies showing that talker identification is improved when listeners understand the language that is being spoken. For example, Thompson (1987) used a voice line-up task to test native English listeners' ability to identify talkers speaking in either English, Spanish, or Spanish-accented English. Thompson found that listeners could identify native English talkers best, followed by Spanish-accented English talkers, and, finally, Spanish talkers worst. Goggin *et al.*, (1991) followed up on this study by presenting Spanish and English stimuli to both monolingual English listeners and monolingual Spanish listeners in a similar testing paradigm. They found that both of the groups of listeners were poorer at identifying the voices of target talkers who were speaking an unfamiliar language.

This facilitatory effect of language familiarity on indexical processing extends to non-native languages (Schiller and Köster, 1996; Köster and Schiller, 1997; Sullivan and Schlichting, 2000). Schiller and Köster (1996) also found that native German listeners and non-native learners of German do not significantly differ in their ability to identify German talkers. Likewise, the extent to which listeners are familiar with a second language does not affect their ability to identify talkers as long as they have some knowledge of the language (Sullivan and Schlichting, 2000). The facilitatory effect of language knowledge on talker identification disappears, however, when the linguistic content of the speech is eliminated, as in reiterant speech (which replaces all syllables of a spoken message with the syllable [ma] but maintains the global prosodic patterns; see Schiller *et al.*, 1997).

One confound which is inherent to all of the studies that showed a facilitatory effect of language knowledge on talker identification accuracy, however, is that they consistently changed talkers between language conditions. It is therefore unclear whether the diminished performance of listeners in an unfamiliar language is due to the properties of the unfamiliar language itself or to the particular qualities of the talkers' voices that were presented in the unfamiliar language condition. The current study dissociates these effects by testing the ability of listeners to identify and discriminate the *same* group of talkers in two different languages. Any change that listeners exhibit in talker identification or discrimination accuracy between language conditions would thus be due to the change in language rather than to any change in the specific talkers producing the stimuli. By separating the contributions of the language and talker to the spoken test materials in this way, this experimental paradigm provides a stronger test of the extent to which the linguistic and indexical properties of speech interact in speech perception.

It is currently unknown whether listeners can generalize knowledge of talkers' voices across languages. Presumably, they could only do so if particular acoustic cues to individual talkers' voices are shared across languages. We will refer to such (potential) cues as *language-independent* cues to talker identity. The use of this term is not meant to imply that such cues are language universal; it is possible that a talker's

voice could be distinguished by cues that are shared across two phonologically similar languages (such as English and German) but would not necessarily be found in all languages that the talker is capable of speaking. Such cues might include, for instance, typical formant values for lax vowels, which are not often found in non-Germanic languages. Other more general language-independent cues to talker identity might include physical characteristics (such as the size and shape of the talker's vocal tract, nasal cavities, and vocal folds), age, or sex (Abercrombie, 1967). Nagao (2006) found that a talker's age can be reliably identified in both a known and an unknown language. Listeners were also shown to identify a talker's sex with a high degree of accuracy within a language, although it is unknown to what degree this ability carries over across languages (Lass *et al.*, 1976). The data we will present indicate that a talker's sex is identifiable even in an unknown language.

There are also potential language-dependent cues to talker identity. Abercrombie (1967) listed "group membership properties," such as regional or social markers, as indexical properties of speech. Such sociolinguistic markers might help identify a talker within one language but are unlikely to transfer over to another language. Other language-dependent indexical properties may overlap to some extent with a talker's physical characteristics. Todaka (1993), for instance, showed that Japanese–English bilinguals use different laryngeal settings in Japanese and English. Johnson (2005) also showed that gender-based properties of speech may change from language to language independently of a talker's sex and Nagao (2006) found that listeners can more accurately identify the age of talkers when they are speaking in a familiar language.

The extent to which knowledge of talkers' voices may generalize across languages depends not only on whether there is language-independent indexical information available to listeners in speech but also on whether listeners attend to that information when learning to identify voices. If listeners do identify voices by solely relying on language-independent information in speech, then the language that a talker is speaking should not affect voice identification accuracy. Listeners who identify talkers from these cues should be able to generalize voice knowledge without loss from one language to another. On the other hand, listeners who solely rely on language-dependent indexical cues to identify voices speaking in a particular language should not be able to generalize knowledge of those voices to a different language. Attending to such language-dependent cues to talker identity, however, should make it easier for those listeners to identify talkers in a familiar language than in an unfamiliar language.

These perceptual possibilities are not mutually exclusive. If indexical processing relies on both language-dependent *and* language-independent information in the speech signal, then some but not all of the listeners' knowledge of the talkers' voices should generalize across languages. In this case, the listeners' ability to identify a known set of talkers in an unfamiliar language would be better than their ability to identify a novel set of talkers in a familiar language. However, the same listeners should be more accurate when identifying known talkers in a familiar language than in an unfamiliar language.

II. EXPERIMENT 1: BILINGUAL TALKER IDENTIFICATION

A. Methods

1. Stimulus materials

Twelve female and ten male German L1/English L2 speakers who were living in Bloomington, IN, were recorded in a sound-attenuated IAC booth at the Speech Research Laboratory at Indiana University. Productions were recorded using a high quality recording equipment and immediately digitized into 16 bit stereo recordings via a Tucker–Davis Technologies System II hardware at 22 050 Hz and directly saved to an IBM-PC Pentium I computer. Recordings were made of each speaker producing a single repetition of 360 English words and 360 German words. Each word was of the form consonant-vowel-consonant (CVC) and was selected from the CELEX English and German databases (Baayen *et al.*, 1995). German was selected as the second language in the experiment because it had a sufficient number of CVC words with the same phonotactic structure as the English CVC words and also because uniformly calculated frequency counts for both the English and German words were available in the CELEX database.

During recording, speakers read one-word prompts off of a computer screen while sitting in the sound-attenuated booth. These words were presented to speakers in randomized order and blocked by language. Any words that speakers produced incorrectly or too quietly were noted and rerecorded in the same manner following each recording block. An automated recording process yielded sound files that were 2000 ms long for each word. The silent portions in these sound files were later removed by hand using PRAAT sound editing software, and the resulting tokens were normalized to have a uniform rms amplitude of 66.499 dB. The total recording time for each language block was approximately 1 h for each speaker. All speakers recorded both language blocks in a single session and were paid $10/h for their time.

Words from both languages varied in frequency based on counts from the CELEX database. Words varying in frequency of occurrence were included in the stimulus materials because listeners can identify high frequency words more quickly and from less acoustic information than low frequency words (Grosjean, 1980). We expected listeners to pay more attention to the acoustic-phonetic details of the low frequency words and therefore develop a more robust mental representation of the acoustic-phonetic characteristics of the various talkers' voices from these tokens. For the purpose of analysis, words in both language blocks were divided into three equally sized groups of varying frequency (high, mid, low).

Ten speakers were selected as the training voices based on their native language background and perceived nativeness in English. Speakers with southern German ($N=2$), Austrian ($N=3$), and Romanian German ($N=1$) dialects were excluded from the set of training voices, along with speakers

TABLE I. Summary of stimuli and tasks used during each phase in all of the training sessions in experiment 1.

	Training session		
	Phase	Stimuli	Task
Training block I	Familiarization	Same five words produced by all ten talkers (500 ms ISI)	Listen and attend to voice/name pair
	Refamiliarization	Same one word produced by all ten talkers	Listen and attend to voice/name pair
	Recognition	Sets of five different words produced by each talker, presented twice in random order	Identify speaker of each word (feedback provided)
Training block II	Familiarization	Same procedure as above	Same procedure as above
	Refamiliarization	Same procedure as above	Same procedure as above
	Recognition	Same procedure as above	Same procedure as above
	Evaluation	Sets of ten different words produced by each talker, presented once in random order	Identify speaker of each word (no feedback provided)

with self-reported speech or hearing disorders ($N=2$) and one speaker who did not finish the recording session. Of the remaining speakers, only the five male and five female speakers who were rated as having the least foreign accent were used in the talker identification task (Levi *et al.*, 2007b).

Accent ratings for each talker were taken from a previous study in which individual words, which were spoken by the various talkers in the bilingual database, were rated on a Likert scale from 0 ("no foreign accent") to 6 ("most foreign accent") by native English listeners who had no familiarity with the German language (for more details, see Levi *et al.*, 2007b). Although we only included the bilingual talkers with the lowest accent ratings in experiments 1 and 2, these talkers were not "accentless." The average, z-score-transformed accent ratings for the female talkers used in this study ranged from −0.27 to 0.22, while the corresponding scores for the male talkers ranged from 0.02 to 0.69—with the higher scores indicating a higher perceived degree of foreign accent. These talkers made few phoneme substitution errors in their word productions (most commonly coda voicing substitutions, such as "news" for "noose"), so these accent ratings probably reflected more subtle phonetic distinctions in their speech or perhaps lack of nativelike articulatory speed and fluency. Native English speakers who were rated in the same study were also not rated entirely accentless; they earned a z-score range of accentedness from −0.52 to −0.09.

2. Listeners

All listeners were native English-speaking students at Indiana University in Bloomington, IN. None reported any knowledge of German prior participation in the study. None of the listeners had ever lived in Germany or had any German-speaking friends or family members. All were right handed and reported no known speech or hearing impairments at the time of the study. A total of 54 listeners participated in the study and were paid $10/h for their participation. Half of these listeners were trained on English stimuli, and half were trained on German stimuli.

The response data from only 40 of these listeners were included in the statistical analysis of the results. Two of the listeners in the English training condition and four listeners in the German training condition did not complete the experiment. The data from the listeners who did not correctly identify at least 40% of the talkers in four or more evaluation phases during training were also excluded from analysis. We considered 40% correct identification accuracy to be a reasonable level of performance for establishing that the listeners had learned the talkers' voices during training since 30% correct was significantly better than chance performance in each evaluation phase (excluding cross-gender confusions). Four participants did not meet this criterion in the English language group and two did not meet this criterion in the German language group. The last listener to complete the experiment in each of the two training conditions was also excluded from the statistical analysis, resulting in 20 listeners per group.

3. Procedure

Participants were trained and tested in a quiet room. All stimuli were presented to participants over Beyer Dynamic DT-100 headphones by a customized SuperCard (version 4.1.1) stack, running on a PowerMac G4.

(*a*) *Training*. Participants were trained to identify the ten different bilingual voices by name in eight training sessions spanning 4 days. The methodology used in these training sessions closely followed the methodology developed by Nygaard *et al.* (1994). The individual training sessions consisted of seven distinct phases, which are summarized in Table I.

In the familiarization phase, listeners heard the same sequence of five words produced by each of the ten talkers,

with an interstimulus interval (ISI) of 500 ms. As each word was presented to the listener, the name of the talker who had produced that word was shown on the computer screen. Each talker's name was a common male or female name in both English and German and was presented in a unique and consistent color in a unique and consistent position on the screen. During this phase of the training, participants did not respond to what they heard but were instructed to pay attention to the names on the computer screen and listen to the sound of each talker's voice. The refamiliarization phase was identical to the familiarization phase except that listeners heard only one word produced by each of the ten talkers.

After familiarization, listeners completed a recognition task in which they heard five different words, presented twice, from all of the ten talkers.[1] These stimuli were presented in a different random order for each participant. After the presentation of each word, listeners identified the talker of that word by clicking an on-screen button next to the appropriate talker's name. After the participants registered their responses, they received feedback by hearing the stimulus token again while the name of only the correct talker appeared on the computer screen. This portion of training was self-paced.

After completing two blocks of the familiarization, refamiliarization, and recognition phases, listeners completed an evaluation task. This evaluation phase was identical to the recognition task except that listeners did not receive feedback on their responses, and they heard ten different words from each talker, without any repetitions of the same word token. Each training session (consisting of two training blocks plus the evaluation phase) took approximately 30 min to complete. The participants completed two training sessions per day for 4 days and were required to take a short (approximately 5 min) break between consecutive sessions on each day of the training. For each participant, no more than 2 days intervened between any successive training days or the generalization test.

(*b*) *Generalization*. After eight training sessions, all listeners completed a generalization test on the final day of the experiment. The generalization testing began with a shortened familiarization phase, in which the listeners heard the same three words produced by all of the ten talkers followed by a refamiliarization phase. All of the words that were presented to listeners in these familiarization phases were spoken in the same language that listeners had heard during training. After familiarization, listeners identified talkers in two testing phases. These testing phases were identical to the evaluation phase at the end of each training session except that the stimuli were presented in different languages in each phase. In one phase, the listeners heard novel words spoken in the same language they had been trained on while in the other phase, they heard words spoken in the language they had *not* been trained on. Before the generalization, the listeners were informed that the talkers might be speaking in an unfamiliar language. The order in which trained and untrained language blocks were presented in these two generalization phases was counterbalanced across participants.

4. Stimulus selection

The stimuli presented during the training and generalization were independently selected for each listener from the larger set of individual word tokens in the bilingual talker database. For each listener, 100 words—balanced for lexical frequency in each language—were randomly selected for use in the generalization blocks. These 100 words consisted of ten different words spoken by each of the ten talkers for both language blocks. Of the remaining 260 words in the database, 100—counterbalanced for lexical frequency—were randomly selected for use in the familiarization phases during the training. The remaining 160 words in the talker database were exclusively used during the evaluation and recognition phases of the training. In both the recognition and evaluation phases, all stimuli were presented in random order to the listeners. While different sets of words were selected for each talker in these phases, it was possible for there to be an overlap between the sets of words produced by each talker in the recognition phases. No individual word was ever presented twice on consecutive trials in recognition or evaluation testing.

B. Results

1. Training

A two-way, repeated measure analysis of variance (ANOVA) was conducted on the response data from the evaluation phases of the eight training sessions. This ANOVA investigated the effects that training session (1–8)—a within-subject factor—and training language (English, German)— a between-subject factor—had on the percentage of talkers correctly identified in each testing phase. The ANOVA revealed a significant main effect of training session [$F(7,32)=61.637$; $p<0.001$] but no effect (at the $p<0.05$ level) of training language nor any interaction between the training session and training language.

Figure 1 shows the percentage of talkers that were correctly identified in the evaluation phases of each training session. *Post hoc*, paired sample *t* tests indicated that both of the groups of listeners consistently improved in identification accuracy over the duration of the training. This improvement occurred in a stepwise fashion. Identification accuracy was significantly higher in training session 2 than in training session 1 ($p<0.001$). Accuracy was also significantly higher in training session 3 than in training session 2 ($p=0.002$). After session 3, significant increases in identification accuracy only occurred between separate days of training [i.e., between sessions 4 and 5 ($p<0.001$) and between sessions 6 and 7 ($p=0.007$)]. Interestingly, this pattern of learning suggests that consolidation of learning took place only during sleep after the first day of training (Fenn *et al.*, 2003). More generally, these results indicate that the listeners were able to identify the voices of the bilingual talkers, and that all of the listeners learned to identify the talkers at the same rate regardless of the language in which they were trained.

2. Generalization

A three-way, repeated measure ANOVA was run on the response data from the generalization blocks on the final day

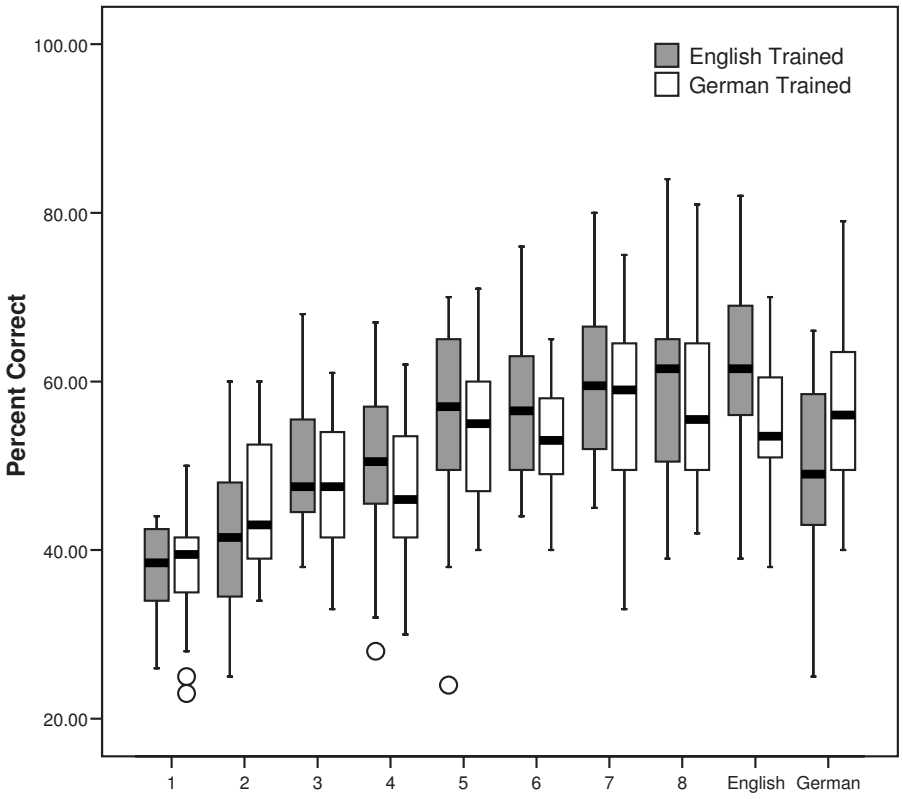


FIG. 1. Box plot of the percentage of talkers correctly identified by each group of listeners in the evaluation phase of each training session and in both generalization language blocks in experiment 1. Means are indicated by a dark line in each box, and the length of each box represents 50% of the data. Whiskers extend to the largest and smallest values for each score; circles represent outliers.

of the experiment, with testing language (English, German) as a within-subject variable and both training language (English, German) and block order (training language first, training language second) as between-subject variables. This ANOVA revealed a significant main effect of testing language [$F(1,36)=27.687$; $p<0.001$], where accuracy was significantly better for English stimuli (58.6%) than that for German stimuli (53.2%). There was also a significant interaction between the testing language and training language [$F(1,36)=47.864$; $p<0.001$]. No other main effects or interactions were significant.

Figure 1 also shows the percentage of talkers correctly identified by each group of listeners in the two generalization testing phases. *Post hoc* analysis of the significant testing language by training language interaction indicated that the English-trained listeners demonstrated significantly higher talker identification accuracy on the English generalization block than on the German generalization block ($p<0.001$). The German-trained listeners, on the other hand, did not perform significantly differently on the English and German generalization blocks ($p=0.0281$). In comparing results across the listener groups, *post hoc* tests revealed that the German-trained group performed better than the English-trained group on the German generalization block,[2] while the English-trained group performed significantly better on the English generalization block ($p=0.016$).

3. Combined data

In order to assess the extent of generalization from training to novel stimuli, paired sample *t* tests were conducted by comparing the listeners' level of performance between each training session and the two generalization testing phases. For the English-trained listeners, there were no significant differences in talker identification accuracy between the English generalization block and the evaluation sessions on the final day of training ($p=0.095$ for session 7 and $p=0.071$ for session 8). These listeners' performance on the English generalization block was, however, significantly better than their performance on the first six training sessions ($p>0.01$ in all cases). In contrast, their performance on the German generalization block was not significantly different from their performance on the third and fourth evaluation sessions, both of which took place on day 2 of the training ($p=0.779$ and $p=0.826$). Their accuracy in identifying talkers from novel German stimuli was significantly better than their identification accuracy on day 1 of the training ($p<0.01$ for both sessions) but significantly worse than their identification accuracy on days 3 and 4 of the training ($p<0.01$ in all cases). This pattern of results indicates that the English-trained listeners were able to successfully generalize some of their knowledge of the talkers' voices to German since their ability to identify the speakers of novel German tokens was significantly better than their performance on the English tokens in the first training session. These data also show that this generalization to German language stimuli was not complete; the listeners' performance in generalization was significantly better for novel English stimuli than that for novel German stimuli.

The paired-sample *t* tests also showed that the percentage of talkers correctly identified by the German-trained listeners in both generalization blocks was not significantly different than the percentage of talkers they correctly identified in training sessions 5–8.[3] However, their performance in *both* generalization blocks was significantly better than that

in all of the evaluation phases on the first two days of training (all $p < 0.001$). Thus, unlike the English-trained listeners, the German-trained listeners were able to generalize their knowledge of the talkers' voices—without a significant loss in identification accuracy—to a language they had not heard in training.

The words that the English-trained listeners heard during the evaluation phases of each training session were evenly split into three groups based on lexical frequency. The paired sample t tests revealed that the lexical frequency of the words presented during the evaluation phases of each training session did not significantly affect the listeners' ability to identify the talkers ($p > 0.08$ in all cases). The percentage of talkers correctly identified from low frequency words was 62.5%, for the mid-frequency, it was 59.4%, and for the high frequency, it was 64.2%.

C. Discussion

The results of this experiment show that there is sufficient language-independent information in speech to make the identification of familiar talkers across languages possible. Listeners steadily improved in their ability to identify talkers in both languages. Improvement in talker identification accuracy largely manifested itself between training days and did not significantly differ between the two training groups. In generalization, both of the groups of listeners identified familiar talkers better in the untrained language than they had identified those same talkers on the first day of the training. The gains in identification accuracy that listeners made in training thus carried over, in part, to stimuli produced by those same talkers speaking in a different language.

The extent to which the two groups of listeners could generalize their knowledge of the talkers' voices across languages depended, however, on the language in which they had been trained. The German-trained listeners exhibited a complete generalization of talker knowledge across languages—identifying talkers in English as well as they had identified talkers in German. The English-trained listeners, on the other hand, showed incomplete generalization of talker knowledge across languages, identifying talkers from novel English stimuli significantly better than they identified talkers from novel German stimuli. This pattern of results suggests that the two groups of the listeners processed the indexical cues in the speech tokens in different ways. The German-trained listeners apparently identified talkers by relying on language-independent indexical information in the signal. The representations of the talkers' voices that they developed in training therefore consisted of language-independent information that could be applied, without loss, to the identification of the same talkers speaking in a different language. The English-trained listeners, on the other hand, evidently learned to identify the bilingual talkers from language-dependent cues since they could not identify talkers in German as well as they had identified talkers in English.

The results of this experiment thus provide evidence for both language-dependent and language-independent indexical processings. Interestingly, listeners appeared to process indexical information in a language-dependent way when they could understand the language that was spoken; otherwise, they identified talkers' voices from only language-independent indexical information in the signal. This pattern of perceptual tendencies may provide insight into the apparently conflicting evidence for both views of speech perception in the existing research literature. The basic pattern in both previous research and this study seems to be that language-independent processing takes place only when the signal lacks linguistic information in some way (filtered speech, talkers speaking in an unfamiliar language, etc.) or when listeners are not capable of processing both kinds of information (phonagnosic listeners). When normal-hearing listeners receive an undegraded signal in a language that they can understand, however, they make use of all the information that is available to them, and the linguistic and indexical properties then interact. For this reason, *language-independent indexical processing* was exhibited by the German-trained listeners in this study, who learned to identify talkers speaking in a language they could not understand, while the English-trained listeners exhibited *language-dependent indexical processing* by relying on English-specific indexical cues that did not generalize to German.

Previous evidence for language-dependent indexical processing has been positive in nature, showing that listeners identify talkers better when they are speaking in familiar languages. Language-dependent indexical processing provided the English-trained listeners in this study with some processing benefits as well in both training and generalization. In generalization, the English-trained listeners not only performed significantly better on English stimuli than on German stimuli but they also performed better on the English stimuli than their German counterparts. That is, even though the German-trained listeners showed a "complete transfer" of their knowledge of the talkers' voices to English-language stimuli, they still could not identify talkers speaking in English as well as the English-trained listeners, and there were also subtle differences between the two listener groups in the patterns of improvement made during the training. There were no significant differences in identification accuracy between the two training groups for any given training session; however, the English-trained listeners performed better on the English stimuli in generalization than they had on all of the training sessions on the first three days of the training. In contrast, the German-trained group did not perform significantly better in generalization than they had on either the fifth or sixth training sessions (on the third day of the training). Thus, the English-trained listeners were somewhat more successful than the German-trained listeners at attaining an increasingly higher level of talker identification accuracy when generalizing to novel words spoken in the language presented during training.

The results of this study show that language-dependent indexical processing has negative implications as well. Relying on language-dependent information to identify talkers' voices in one language made it more difficult for the English-trained listeners to generalize their knowledge of the talkers' voices to a new language. Interestingly, this decrease

in talker identification accuracy occurred even though there is no a priori reason to believe that the English-trained listeners could not identify talkers from strictly language-independent indexical cues. However, the perceptual integration of linguistic and indexical information may automatically occur when listeners can understand the language that is being spoken. If so, such an automatic perceptual process offers only small gains in indexical processing within a known language at the expense of developing representations of talkers' voices which are more perceptually robust and generalizable to new languages.

III. EXPERIMENT 2: CROSS-LANGUAGE VOICE DISCRIMINATION

Experiment 2 further investigated the influence of language on indexical processing by testing the listeners' ability to discriminate voices both within and across languages. For this task, the listeners were asked to judge whether two speech stimuli were produced by the same talker or by two different talkers. The stimuli consisted of monosyllabic words that were presented in either matched-language (i.e., both English or both German) or mixed-language (i.e., English–German or German–English) pairs.

If indexical processing is language independent, then listeners should be able to discriminate talkers regardless of the language in which they are speaking. If indexical processing is language dependent, however, then language could affect performance in the voice discrimination task in at least two different ways. Listeners might, for instance, discriminate voices better when they are speaking in a familiar language. For native English listeners, voice discrimination would therefore be facilitated for English stimuli, yielding the best performance in English–English pairs, followed by the English–German and German–English pairs, and worst for the German–German pairs. Alternatively, language could affect voice discrimination performance in a different way if listeners attend to language-dependent indexical properties of speech regardless of the language they are listening to. In this case, performance in a discrimination task should be better in matched-language conditions (English–English and German–German) than in mixed-language conditions (English–German and German–English). Mixed-language conditions would force listeners to recalibrate their perceptual orientations between stimuli in order to attend to different sets of indexical properties in different languages. Not having to perform a similar recalibration between languages should facilitate discrimination accuracy in the matched-language conditions.

Not all of the listeners need to perform the voice discrimination task in the same way, of course. Some might process indexical information in a language-independent fashion while others might discriminate voices on the basis of language-dependent indexical cues. The results of experiment 1 provided evidence for both language-independent and language-dependent indexical processing depending on the language in which listeners were trained to identify voices. For this reason, both the English-trained and German-trained listeners from experiment 1 were brought back to participate in experiment 2 to determine if the perceptual proclivities they had developed in learning to identify voices transferred to a voice discrimination task. A group of untrained English listeners also participated in experiment 2 in order to determine whether listeners without any experience with the particular bilingual talkers' voices would perform the discrimination task in a language-dependent or language-independent fashion. Comparing the discrimination performance of these listeners to that of the trained listeners also provided a means of determining how much voice identification training improved the ability of the listeners to discriminate between voices.

A. Methods

1. Stimulus materials

The stimuli for experiment 2 were produced by the same set of bilingual talkers that produced the stimuli for experiment 1.

2. Listeners

Three groups of listeners participated in experiment 2: English-trained, German-trained, and untrained listeners. The trained listener groups included 15 of the 20 listeners from each training group in experiment 1. These trained listeners were paid $10 for their participation. Twenty-three additional listeners participated as untrained listeners. These listeners were students in undergraduate psychology courses at Indiana University who received partial course credit for their participation. Two untrained listeners were eliminated due to experimenter error and one was eliminated because of previous experience with German, resulting in 20 listeners in the untrained group. The remaining untrained listeners met the same qualifications as the trained participants: they were right handed, had no previous experience with German, had no history of a speech or hearing disorder, and were 18–25 years of age.

3. Procedure

Participants were tested in a quiet room. A customized software running on PCs presented the word pair stimuli to the listeners over Beyer Dynamic DT-100 headphones at a comfortable listening level. The participants were instructed to judge whether the two words in each pair were spoken by the same talker or by two different talkers. Participants registered their responses by pressing one of the two buttons on a custom-made button box; the right button registered "same" responses while the left button registered "different" responses. Participants were instructed to keep fingers on both buttons at all times during the experiment. They were also informed that the words they would be hearing might be spoken in an unfamiliar language.

Testing consisted of two blocks of 320 trials each. The stimuli in each block were evenly split between same-talker and different-talker trials. For the same-talker trials, word pairs were constructed by matching four pairs of different words produced by the same talker in four different language conditions: English–English, English–German, German–English, and German–German. For the different-talker trials, each talker was presented in combination with every other

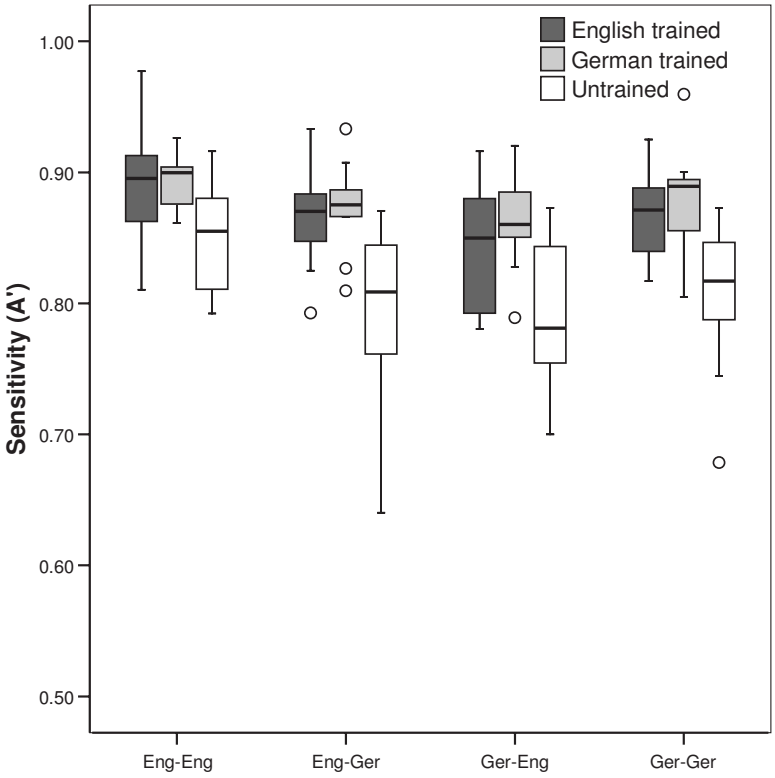


FIG. 2. Box plot of sensitivity (A') values for all of the three listener groups in each language pair condition in experiment 2. Means are indicated by a dark line in each box, and the length of each box represents 50% of the data. Whiskers extend to the largest and smallest values for each score; circles represent outliers.

same-gender talker twice—counterbalanced for order—in each of the four different language conditions. Cross-gender pairs were not presented to listeners as exceedingly few cross-gender confusions had been made in experiment 1 (approximately 1 out of every 300 identification trials). The two target words in the matched-language conditions were always different lexical items. For each listener, a different set of word pairs was selected from the database and trials were presented in a uniquely random order.

The words in each stimulus pair were separated by a 750 ms ISI. Listeners were instructed to respond as quickly as possible to each stimulus pair while still remaining accurate. Each block of trials began with a short sequence of four practice trials. After each of the trial in both the practice and testing sessions, the participants were informed whether their response was correct by a color-coded message (red for incorrect, blue for correct) presented on the computer screen. In testing, this message also informed the participants whether the pair had been a same-talker trial or a different-talker trial, along with the cumulative percentage of correct responses. Testing was self-paced, but participants generally completed each block of trials within 30 min. Participants were required to take a short break between blocks.

4. Data analysis

The same/different responses given by each listener were converted into nonparametric measures of sensitivity (A') and bias (B'') (Grier, 1971). Both of these measures are based on the proportion of "hits" and "false alarms" given by the listeners. Hits and false alarms were defined with respect to the same-talker trials; a hit was a same response to a same-talker trial, while a false alarm was a same response to a different-talker trial. A' yields a measure of listener sensitivity to the same-talker/different-talker distinction which ranges from 0.0 to 1.0, where a value of 1.0 indicates perfect discrimination and a value of 0.5 reflects chance performance on the discrimination task. B'' yields a measure of listener bias toward one response option or another. This measure ranges from −1.0 to 1.0, where negative values indicate a tendency to give same responses, while positive values reflect a tendency to give different responses. *A* B'' value of 0 indicates lack of bias. Separate A' and B'' values were calculated for the responses given by each listener in each of the four different language pair conditions.

B. Results

One-sample t tests revealed that the sensitivity measures for all of the listener groups in all language pair conditions were significantly above 0.5, the level of chance performance of A' [for English-trained listeners, all $t(14)>28.1$, $p<0.001$; for German-trained listeners, all $t(14)>36.6$, $p<0.001$; for untrained listeners, all $t(19)>24.1$, $p<0.001$]. The average and range of the sensitivity values for all of the participants in each listener group are shown in Figure 2.

Repeated measure ANOVAs with language pair (English–English, German–German, English–German, German–English) as a within-subject factor and listener group (English–trained, German-trained, untrained) as a between-subject factor were conducted on both the sensitivity (A') and bias (B'') measures. The sensitivity ANOVA yielded main effects of listener group [$F(2,47)=17.81$, $p<0.001$] and language pair [$F(3,141)=20.28$, $p<0.001$], with no significant interaction between the two factors. *Post hoc* Tukey tests of the main effect of listener group revealed that both trained groups had significantly higher sensitivity than the untrained group ($p<0.001$), but that there was no

TABLE II. Mean sensitivity (A') for the four listening conditions for all of the listener groups in experiment 2.

		Language pair				
		FE	EG	GE	GG	Mean
Listener group	English trained	0.891	0.864	0.843	0.867	0.866
	German trained	0.892	0.870	0.865	0.877	0.876
	Untrained	0.850	0.800	0.795	0.812	0.814
	Mean	0.878	0.845	0.834	0.852	0.852

significant difference in sensitivity between the two groups of trained listeners.

Post hoc paired sample t tests of the main effect of language pair revealed that listeners were better able to discriminate talkers in the English–English condition than in all other conditions (all $p<0.001$). A significant difference between the German–German ($A'=0.85$) and German–English ($A'=0.83$) conditions was also found ($p=0.007$), indicating that listeners could better discriminate talkers in the German–German condition. Table II provides the mean sensitivity values for each listener group in each of the language pair conditions.

The ANOVA on B'' yielded significant main effects of listener group [$F(2,47)=4.056$, $p=0.024$] and language pair [$F(3,141)=46.32$, $p<0.001$], as well as a significant interaction between the two factors [$F(6,141)=6.90$, $p<0.001$]. *Post hoc* Tukey tests of the main effect of listener group revealed a significant difference between the biases of the German-trained listeners and the untrained listeners ($p=0.021$). Untrained listeners were more biased to give same responses ($B''=-0.123$) than the German-trained listeners ($B''=-0.016$). Paired-sample t tests of the main effect of language pair revealed that all listeners were more likely to give same responses in the German–German condition than in the other three conditions (EE, $p=0.009$; EG, $p<0.001$; GE, $p<0.001$). The listeners were also more likely to give same responses in the English–English condition than in the two mismatched-language conditions (both $p<0.001$). Response bias did not differ significantly between the two mismatched-language conditions. Table III lists the mean response bias for each listener group in each of the language pair conditions.

The significant interaction between language pair and listener group on response bias is illustrated in Figure 3. *Post hoc* analyses revealed that this interaction was due to differences in response bias between the German-trained listeners and the other listener groups in both the English–German and German–German conditions. For these language pairs, the German-trained listeners were less likely to give same responses ($p\leq 0.005$ in the English–German condition, and $p=0.002$ in the German–German condition).

C. Discussion

The sensitivity values from the different language-pair conditions in experiment 2 indicate that the listeners performed the voice discrimination task by relying on both language-dependent and language-independent indexical information in the speech signal. For all of the three listener groups in all four language pair conditions, discrimination accuracy was significantly better than chance. Thus, listeners can accurately discriminate voices regardless of the language in which those voices were speaking. Voice discrimination was significantly better than chance even in the mismatched-language conditions (English–German and German–English). Since all listeners—including the untrained participants—exhibited this robust pattern, discrimination ability cannot solely depend on experience with either a particular language or a particular talker's voice.

The sensitivity values also indicate, however, that the language in which stimuli were presented did have some effect on the listeners' ability to discriminate talkers' voices. The listeners were better at discriminating voices in the matched-language conditions—especially the English–English pairs—than in the mixed language conditions. This pattern of language-dependent effects on sensitivity largely supports the hypothesis that listeners attend to different language-dependent indexical properties in different languages and must therefore recalibrate their perceptual orientations when listening to mixed language pairs. That the listeners showed better discrimination accuracy in the English–English condition than in the German–German condition is also consistent with the hypothesis that listeners can process language-dependent indexical information better in a familiar language.

The effects of language on sensitivity did not interact with the listeners' previous experience with the talkers; both of the English- and German-trained listeners performed sig-

TABLE III. Mean bias (B'') for the four listening conditions for all of the three listener groups in experiment 2.

		Language pair				
		FE	EG	GE	GG	Mean
Listener group	English trained	−0.093	−0.067	0.056	−0.291	−0.098
	German trained	−0.162	0.113	0.043	−0.056	−0.016
	Untrained	−0.144	−0.047	−0.024	−0.277	−0.123
	Mean	−0.133	0.000	0.025	−0.208	−0.079

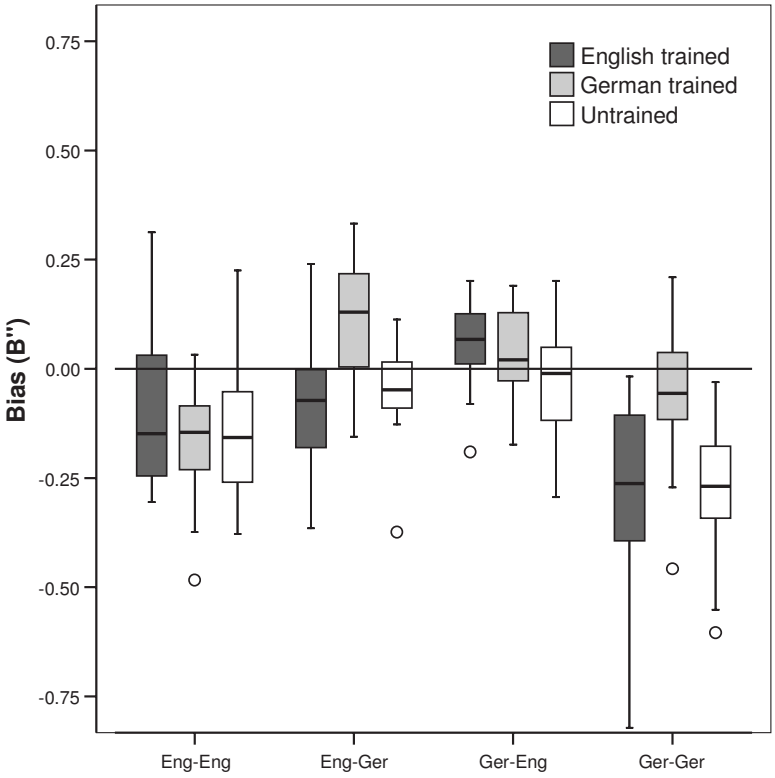


FIG. 3. Box plots of bias (B'') values for all three listener groups in each language pair condition in experiment 2. Negative values reflect a bias toward same responses; positive values reflect a bias toward different responses. Means are indicated by a dark line in each box, and the length of each box represents 50% of the data. Whiskers extend to the largest and smallest values for each score; circles represent outliers.

nificantly better on the voice discrimination task than the untrained listeners. These trained listeners were therefore able to successfully transfer their knowledge of the talkers' voices across experimental tasks. Interestingly, there were no significant differences in sensitivity between the German- and English-trained listeners in any of the language pair conditions even though the generalization data from experiment 1 suggested that the German-trained listeners had processed indexical information in a more language-independent fashion than the English-trained listeners.

This absence of an interaction between training and language condition implies that the German-trained listeners' experience with hearing the bilingual talkers speaking in German in experiment 1 did not provide them with any additional advantage over the other two groups of listeners in processing indexical information in German. The analysis of the response bias revealed a different pattern of results between the listener groups, however. The listeners from all of the three groups were more biased to give same responses to matched-language pairs than to mixed-language pairs. In other words, the listeners were more biased to respond same when both stimulus words were spoken in the same language. This pattern indicates that listeners perceptually conflated linguistic and indexical information by interpreting words in the same language as coming from the same talker even when they were spoken by two different people. Although listeners were supposed to be performing a strictly indexical task, these bias measures indicate that they based their responses in part on the linguistic information in the signal regardless of the indexical content of the stimuli.

Within the matched-language conditions, the same response bias was stronger for the German–German pairs than for the English–English pairs. All of the listeners in this experiment were thus more likely to conflate same-language information with same-voice responses when the talkers were speaking in German. The German-trained listeners, however, showed significantly reduced bias toward same responses in the German–German and English–German conditions than did the English-trained and untrained listeners. The tendency to perceptually conflate linguistic and indexical information was therefore less pronounced in the German-trained listeners. Training in English, on the other hand, did not change the listeners' response biases from their untrained counterparts. Both the English-trained and untrained listeners showed a strong same bias for the same language pairs and less same bias for the different language pairs. Thus, learning to identify voices in English improved the listeners' sensitivity to distinctions between voices but did not alter their perceptual biases; training in German, however, both improved the listeners' sensitivity and changed their response biases, such that these listeners showed a significantly reduced tendency to conflate linguistic and indexical information in perception.

It is interesting to note that the reduced same response bias by the German-trained listeners was limited to those language pairs which ended with German words. It is not entirely clear why this is the case. Had training in German simply enabled the listeners to better separate linguistic and indexical information in German words, a significant change in bias should have occurred for the German-trained listeners in the German–English language condition as well. Evidently, presenting the final word to the listeners in German triggered whatever language-independent processing abilities they had developed in training. Hearing the final word in English, on the other hand, caused them to revert to a more nativelike, language-dependent processing mode. The trans-

formation of these listeners' perceptual orientations through training in German was not, therefore, complete but rather continued to depend—to some extent—on the language of the stimuli that they heard.

IV. GENERAL DISCUSSION

One of the primary findings of these studies is that there is sufficient language-independent information in the speech signal to reliably identify and discriminate bilingual talkers' voices across two different languages. The presence of this information enabled the listeners to generalize their knowledge of talkers' voices across languages in experiment 1 and to accurately discriminate voices across languages in experiment 2. These findings support the results of previous research showing that the linguistic and indexical properties of speech are processed independently of one another. However, the results of this study also revealed that, in addition to the language-independent indexical properties of speech, listeners rely on language-dependent information to perform indexical tasks such as voice identification or discrimination. The listeners who were trained to identify talkers in English failed to generalize all of their knowledge of the talkers' voices from English to German in experiment 1; moreover, the discrimination performance of all of the groups of listeners was affected by the language in which stimuli were presented in experiment 2. These influences of language on indexical processing support the results of previous studies which have shown that listeners process the linguistic and indexical properties of speech in an integrated manner in perception. The combined results of the two experiments in this study thus validate both sides of the paradoxical findings of previous research—listeners apparently process the indexical properties of speech in both a language-dependent and a language-independent manner.

An exhaustive analysis of the acoustic properties that listeners used to identify and discriminate voices is beyond the scope of this paper, but the data do provide some clues as to what the most salient language-independent and language-dependent indexical cues were. One potential acoustic cue to talker identity that participants in the identification task consistently mentioned that they listened for was the "pitch" of each talker's voice. This was not a misguided strategy as most of the talkers used a characteristic *F*0 pattern when recording the words for the bilingual talker database. Some of the talkers used a consistently high or low average *F*0, while others used a characteristic up-sloping or down-sloping intonation. One talker produced each item with a very short duration. These suprasegmental patterns were largely unique to each talker and were not determined by the phonological or semantic content of any of the words they produced. As such, they are not language-dependent properties of speech but rather reflect a pattern of articulatory choices made by each talker during the production of items in the recording session. For that reason, these talker-specific patterns carried over, to a large extent, across languages, making them a potential language-independent cue to talker identity which is not simply based on each talker's vocal tract physiology.

TABLE IV. Percentage of talkers correctly identified by the English listeners in the German generalization testing by German word type.

Word type	Examples	Total correct	Total heard	% correct
/x/	*Bach, doch*	91	176	51.7
/ts/	*Zeit, zahm*	55	108	50.9
Final /l/	*null, Ball*	131	260	50.4
English words	*Pein, nun*	198	401	49.4
English nonwords	*heiss, lahm*	385	791	48.7
Final /r/	*nur, Tier*	112	230	48.7
/pf/	*Pfeil, Kopf*	29	66	43.9
Initial /r/	*Rausch, Reim*	81	203	39.9
/o/, /e/	*Tod, Weg*	55	141	39.0
Front rounded vowels	*schoen, kuehl*	44	124	35.5

An examination of identification accuracy in the German generalization test of experiment 1 suggests that English-trained listeners identified talkers on the basis of language-specific segmental cues as well. Table IV presents the English-trained listeners' response accuracy in German generalization for the various phonetic word types in the German database; these word types are listed in descending order of talker identification accuracy. Word types include two categories for German words that are phonetically similar to English words—including both "words," such as "mein" ("mine") and "nun" ("noon"), and nonwords, such as "heiss" [haɪs] and "lahm" [lam]. There were also eight different word types for German words that included phonetic content not found in (General American) English: words with the velar/palatal fricative /x/, the "clear" /l/, front rounded vowels (/y/ and /ø/), initial /ɾ/, final /r/, the affricates /pf/ and /ts/, and the mid- to high monophthongs /e/ and /o/.

It is difficult to draw clear conclusions about the data from this small sample set due to differences in the number of presentations for each word type, but a few general patterns emerge among the percentages. First, listeners had the most difficulty in identifying talkers from words that included vowels not found in English. Identification accuracy for words with front rounded vowels was only 35.5%, and identification accuracy for words with high to midmonophthongs was only 39%. The English-trained listeners were considerably less troubled by exotic consonant sounds, such as the velar fricative (51.7%) or the clear German /l/ (50.4%). Interestingly, the potential English word status of the German tokens did not seem to matter much to talker identification accuracy; possible English words such as "nun" and "mein" induced an accuracy level of 49.4%, while the talkers of possible English nonwords (heiss, lahm) were correctly identified 48.7% of the time. Numbers such as these indicate that, no matter how much English-trained listeners might have relied on language-independent cues such as patterns in *F*0 to identify talkers, their talker identification responses were still sensitive to language-specific segmental information in the speech signal. In particular, the unfamiliar vowels of the German language caused problems for the listeners in generalization testing, suggesting that the English-trained listeners may have heavily relied on the acoustic information in language-specific vowel categories to identify

talkers during training. The interaction of language with indexical processing at this segmental level corroborates the hypothesis of Remez *et al.*, (1997) that segmental phonetic content subserves both talker identification and word identification and is the locus for potential interactions between the two types of processing.

In postexperiment debriefing sessions, the English-trained listeners also indicated that they consciously attended to global indexical properties which would not be likely to transfer across languages. For instance, many listeners cited the relative accentedness of each talker's speech as a feature they listened to in trying to identify talkers during training. Another listener claimed that she could reliably identify one talker by the way that talker had "overexaggerated" the pronunciation of each word—a tendency, in other words, to "hyperarticulate" (Lindblom, 1990). Another listener reported that she could consistently identify one of the male speakers by the fact that he sounded "gay." Even if attending to such stylistic and sociophonetic markers of identity helped these listeners learn to identify the individual talkers' voices during training (there is no evidence for any such facilitation in the training scores), it is unlikely that any of these indexical properties would be available to listeners in a different language. Knowing how much of an "accent" a non-native talker has while they are speaking English is useless information to have when trying to identify the same talker speaking German. A tendency to attend to language-dependent global and segmental properties of speech may therefore account for why the English-trained listeners' identification accuracy dropped when generalizing across languages in experiment 1.

In fact, a closer examination of the results of both studies suggests that the English-trained listeners—whether they consciously intended to or not—attended more closely to language-dependent indexical cues than the German-trained listeners in both tasks. The German-trained listeners were more successful in generalizing their knowledge of the talkers' voices across languages in experiment 1 and also showed a reduced bias toward perceptually conflating linguistic and indexical information in experiment 2. It is likely that the German-trained listeners learned to more heavily rely on language-independent information to perform indexical tasks simply because they were exposed to bilingual talkers speaking a language that they did not understand. Under these conditions, listeners could not be distracted by the linguistic content of the utterance and therefore learned to primarily attend to the language-independent indexical information in the signal. It is also possible that the German-trained listeners developed more abstract, language-independent representations of the talkers' voices simply because they were exposed to a new source of variability in speech (i.e., the German language). Effects of high stimulus variability in training that lead to better generalization performance in testing have been documented in a variety of earlier studies (Bradlow *et al.*, 1997; Clopper, 2004; Greenspan *et al.*, 1988; Iverson *et al.*, 2005, Logan *et al.*, 1991). While training in German did not necessarily involve more variability than training in English, it did present listeners with a different kind of acoustic-phonetic variability from their pre-experiment experiences. Learning to process indexical information in this new linguistic form may therefore have better guided the listeners' perceptual systems in attending to the information in the signal which was most relevant to the indexical processing task rather than to any distracting linguistic properties of the signal.

The poorer performance of the English-trained listeners on the cross-linguistic generalization task in experiment 1 suggests that processing indexical information in a language-dependent fashion diverts attention and makes the generalization of indexical knowledge across languages more difficult. If this is the case, then why did the listeners do it? One answer is that language-dependent indexical cues may provide listeners with some indexical processing advantages, as it did for all of the groups of listeners in the English–English condition of experiment 2. Another answer may be that the listeners simply cannot help themselves—the language-dependent interpretation of indexical information in speech may be an automatic perceptual process, with negative processing consequences that are only made apparent when listeners are introduced to new types of linguistic variability in the signal. Therefore, listeners may automatically attend to language-dependent indexical cues in speech when they are listening to a language that they understand. Interestingly, the results of this study show that this process may take place even though reliance on language-dependent cues is not necessary for either talker identification or discrimination. There is sufficient language-independent information in the signal to support both talker identification and discrimination across different languages.

The different behaviors of the English-trained and German-trained listeners in these experiments have several implications for exemplar theories of speech perception, which originally drew inspiration from findings that listeners store in memory all tokens of experienced speech, without segregating indexical and linguistic properties in long-term representations. The results of this study suggest that not all exemplars are created equal. The English-trained listeners exhibited language-dependent processing in both experiments, indicating that they had developed integrated indexical and linguistic representations in memory. The German-trained listeners, on the other hand, seemed to develop largely language-independent representations of talkers' voices. If so, these listeners would not be expected to show an interaction between linguistic and indexical properties when processing English words spoken by familiar bilingual talkers. An advantage for identifying English words spoken by familiar talkers has been shown for English-trained listeners (Nygaard *et al.*, 1994), but preliminary evidence from our laboratory indicates that German-trained listeners do not exhibit such a familiar talker advantage across languages (Levi *et al.*, 2007a). The integrated representations in memory that are generally assumed by exemplar theory therefore only seem to emerge when listeners know how to interpret both the linguistic and indexical information in the signal.

There are mechanisms in place in some exemplar-based models of speech perception which can account for this pattern of findings. Pierrehumbert (2002), for instance, suggests that exemplar-based storage requires knowing how to assign

the different kinds of variability found in speech to specific category labels. Unassigned variation is simply discarded from memory. The German-trained listeners in this study who lacked category labels for German words and German-specific phonemic categories may thus have simply excluded from memory the variation in the signal that would be relevant to German language processing. Alternatively, the two groups of listeners may have differed in the amount of attention they devoted to the linguistic properties of the signal in perception. Computationally based models of exemplar categorization, such as Kruschke's ALCOVE (1992) or Johnson's XMOD (1997), typically incorporate attention weights into a front-end processing system to support exemplar storage in memory. In speech perception, particular linguistic properties may be enhanced or compressed in the stored exemplar representation according to the amount of attention listeners pay to them (e.g., Escudero, 2005; Iverson *et al.*, 2005; Kruschke, 1992; Nosofsky, 1986). In the case of the current study, the English-trained listeners may simply have paid more attention to the language-specific properties in the signal, while the German-trained listeners focused more on the language-independent properties of the talkers' voices. The different attention weights thus led to different representations in the stored exemplars themselves.

It should be noted here that the evidence for exemplar theories of speech perception largely (if not exclusively) come from studies in which listeners were presented with familiar language stimuli. It is not necessarily true that listeners process unfamiliar languages in the same way. The results of this study suggest that more consideration ought to be given to how exemplars are encoded in perception and what, exactly, is stored in memory. The findings of these experiments suggest that investigating the potential interactions of indexical and linguistic information across languages—rather than just in English—offers a promising avenue for future research on this increasingly influential theory.

V. CONCLUSION

The combined results of the two experiments reported here suggest that the extent to which indexical processing relies on language-independent information depends on the listeners' knowledge of the language presented in the speech signal. Listeners process indexical information in a language-dependent fashion when they hear a language that they know; otherwise, they perform indexical tasks by more heavily relying on language-independent information in the signal. This general perceptual pattern suggests that listeners shift their focus of attention, in fundamental ways, to adapt to the linguistic structure of the speech signals they receive and the specific processing demands of the perceptual task.

The finding that the human perception of speech can rapidly adapt to changing linguistic conditions in this way may reconcile the apparently conflicting evidence for both the language-dependent and language-independent views of speech perception that were presented in the introduction. Indexical processing operates in a language-dependent manner whenever listeners understand the language that is being spoken. In this case, they can use integrated linguistic and indexical information in the speech signal to identify or discriminate voices. For this reason, listeners identify voices better when they are speaking in a familiar language, but they also have difficulty generalizing their knowledge of voices from a familiar to an unfamiliar language. On the other hand, when linguistic information is removed from the speech signal—as in filtered speech or phonagnosia or by presenting stimuli in an unfamiliar language—the perceptual system accordingly adjusts by relying on the language-independent information in the signal to perform indexical perception tasks. The perception of the indexical properties of speech may thus be either language-dependent or language-independent depending on the context in which listeners operate. The evidence in favor of one view of speech perception does not necessarily invalidate evidence for the other, therefore, as long as the kind of information which is available to listeners in the speech signal is taken into account.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant No. DC-00012 and NIH-NIDCD Research Grant No. R01 DC-00111). The authors would like to thank Christina Fonte, Jen Karpicke, and Melissa Troyer for their help in running the subjects and editing stimuli. The authors would also like to thank two anonymous reviewers for their comments on an earlier draft of this paper.

[1]Talker-specific word tokens were presented more than once during these recognition phases because it has been found that feedback does not facilitate perceptual learning unless the stimulus items in a training paradigm are presented to them more than once (Winters *et al.*, 2005).

[2]The p-value ($p=0.049$) approaches significance with a Bonferroni correction to $p=0.0125$.

[3]This finding includes a Bonferroni correction to $p=0.006\,25$ for the eight comparisons made on these data. For sessions 5, 7, and 8, $p>0.10$. For session 6, $p=0.038$ for the German generalization and $p>0.10$ for the English generalization.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Edinburgh University, Edinburgh).

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (**1995**). The CELEX Lexical Database, Release 2 (CD-ROM), Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bradlow, A. R., Pisoni, D. P., Akahane-Yamada, R., 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.

Bricker, P. D., and Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. **40**, 1441–1449.

Clarke, F. R., Becker, R. W., and Nixon, J. C. (**1966**). "Characteristics that determine speaker recognition," Report No. ESD-TR-66-638, Electronic Systems Division, Air Force Systems Command, pp. 1–65, Hanscom Field, MA.

Clopper, C. G. (**2004**). "Linguistic experience and the perceptual classification of dialect variation," Ph.D. thesis, Indiana University, Bloomington, IN.

Compton, A. J. (**1963**). "Effects of filtering and vocal duration upon the identification of speakers, aurally," J. Acoust. Soc. Am. **53**, 1741–1743.

Escudero, P. (**2005**). "Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization," Ph.D. thesis, Utrecht University, Utrecht, the Netherlands.

Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (**2003**). "Consolidation

during sleep of perceptual learning of spoken language," Nature (London) **425**, 614–616.
Glisky, E. L., Polster, M. R., and Routhieaux, B. C. (**1995**). "Double dissociation between item and source memory," Neuropsychology **9**, 229–235.
Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Mem. Cognit. **195**, 448–458.
Goldinger, S. D. (**1996**). "Words and Voices: Episodic traces in spoken word identification and recognition memory," J. Exp. Psychol. Learn. Mem. Cogn. **22**, 1166–1183.
Goldinger, S. D. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psychol. Rev. **105**, 251–279.
Goldinger, S. D., Pisoni, D. B., and Logan, J. S. (**1991**). "On the locus of talker variability effects in recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **17**, 152–162.
Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. (**1988**). "Perceptual learning of synthetic speech produced by rule," J. Exp. Psychol. Learn. Mem. Cogn. **14**, 421–433.
Grier, J. B. (**1971**). "Nonparametric indexes for sensitivity and bias: Computing formulas," Psychol. Bull. **75**, 424–429.
Grosjean, F. (**1980**). "Spoken word recognition processes and the gating paradigm," Percept. Psychophys. **28**, 267–283.
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.
Johnson, K. (**2005**). "Speaker normalization in speech perception," *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. Remez (Blackwell, Oxford), pp. 363–389.
Johnson, K. A. (**1997**). "Speech perception without speaker normalization," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. Mullennix (Academic, San Diego), pp. 145–165.
Köster, O., and Schiller, N. O. (**1997**). "Different influences of the native language of a listener on speaker recognition," Forensic Linguistics **4**, 18–28.
Kruschke, J. (**1992**). "ALCOVE: An exemplar-based connectionist model of category learning," Psychol. Rev. **99**, 22–44.
Landis, T., Buttet, J., Assal, G., and Graves, R. (**1982**). "Dissociation of ear preference in monaural word and voice recognition," Neuropsychology **20**, 501–504.
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.
Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2007a**). "A cross-language familiar talker advantage?," *Research on Speech Perception Progress Report No.* 28, Speech Research Laboratory, Indiana University, Bloomington, IN, pp. 369–383.
Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2007b**). "Speaker-independent factors affecting the degree of perceived foreign accent in a second language," J. Acoust. Soc. Am. **121**, 2327–2338.
Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H&H theory," *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403–439.
Logan, J. D., 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.
Mullennix, J. W., and Pisoni, D. B. (**1990**). "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**, 379–390.
Nagao, K. (**2006**). "Cross-language study of age perception," Ph.D. thesis, Indiana University, Bloomington, IN.
Nosofsky, R. M. (**1986**). "Attention, similarity, and the identification-categorization relationship," J. Exp. Psychol. Gen. **115**, 39–57.
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.
Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (**1993**). "Episodic encoding of voice attributes and recognition memory for spoken words," J. Exp. Psychol. Learn. Mem. Cogn. **19**, 309–328.
Pierrehumbert, J. (**2001**). "Exemplar dynamics: Word frequency, lenition, and contrast," in *Frequency effects and the Emergence of Lexical Structure*, edited by J. Bybee and P. Hopper (Benjamins, Amsterdam), pp. 137–157.
Pierrehumbert, J. (**2002**). "Word-specific phonetics," in *Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin), pp. 101–140.
Pollack, I., Pickett, J. M., and Sumby, W. H. (**1954**). "On the identification of speakers by voice," J. Acoust. Soc. Am. **26**, 403–406.
Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.
Schacter, D. L., and Church, B. A. (**1992**). "Auditory priming: Implicit and explicit memory for words and voices," J. Exp. Psychol. Learn. Mem. Cogn. **18**, 915–930.
Schiller, N. O., and Köster, O. (**1996**). "Evaluation of a foreign speaker in forensic phonetic: A report," Forensic Linguistics **3**, 176–185.
Schiller, N. O., Köster, O., and Duckworth, M. (**1997**). "The effect of removing linguistic information upon identifying speakers of a foreign language," Forensic Linguistics **4**, 1–17.
Stevens, A. A. (**2004**). "Dissociating the cortical basis of memory for voices, words and tones," Brain Res. Cognit. Brain Res. **18**, 162–171.
Sullivan, K. P. H., and Schlichting, F. (**2000**). "Speaker discrimination in a foreign language: First language environment, second language learners," Forensic Linguistics **7**, 95–111.
Thompson, C. P. (**1987**). "A language effect in voice identification," Appl. Cognit. Psychol. **1**, 121–131.
Todaka, Y. (**1993**). "Japanese students' English intonation," Bulletin of Miyazaki Municipal University **1**, 23–47.
Van Lancker, D. R., Cummings, J. L., Kreiman, J., and Dobkin, B. H. (**1988**). "Phonagnosia: A dissociation between familiar and unfamiliar voices," Cortex **24**, 195–209.
Williams, C. E. (**1964**). "The effects of selected factors on the aural identification of speakers," *Report No.* ESD-TDR-65-153, Electronic Systems Division, Air Force Systems Command, Hanscom Field, MA.
Winters, S. J., Levi, S. V., and Pisoni, D. B. (**2005**). "When and why feedback matters in the perceptual learning of the visual properties of speech," *Research on Speech Perception Progress Report No.* 27, Speech Research Laboratory, Indiana University, Bloomington, IN, pp. 107–132.

Categorical dependence of vowel detection in long-term speech-shaped noise[a)]

Chang Liu[b)]
Department of Communication Sciences and Disorders, University of Texas at Austin, 1 University Station A1100, Austin, Texas 78712

David A. Eddins[c)]
Department of Otolaryngology, University of Rochester, 601 Elmwood Avenue, Box 609, Rochester, New York 14642 and International Center for Hearing and Speech Research, 52 Lomb Memorial Drive, Lyndon Baines Johnson Building, Room 2620, Rochester, New York 14623

(Received 25 July 2007; revised 23 January 2008; accepted 7 March 2008)

The goal of this study was to measure detection thresholds for 12 isolated American English vowels naturally spoken by three male and three female talkers for young normal-hearing listeners in the presence of a long-term speech-shaped (LTSS) noise, which was presented at 70 dB sound pressure level. The vowel duration was equalized to 170 ms and the spectrum of the LTSS noise was identical to the long-term average spectrum of 12-talker babble. Given the same duration, detection thresholds for vowels differed by 19 dB across the 72 vowels. Thresholds for vowel detection showed a roughly U-shaped pattern as a function of the vowel category across talkers with lowest thresholds at /i/ and /æ/ vowels and highest thresholds at /u/ vowel in general. Both vowel category and talker had a significant effect on vowel detectability. Detection thresholds predicted from three excitation pattern metrics by using a simulation model were well matched with thresholds obtained from human listeners, suggesting that listeners could use a constant metric in the excitation pattern of the vowel to detect the signal in noise independent of the vowel category and talker. Application of the simulation model to predict thresholds of vowel detection in noise was also discussed.
 [DOI: 10.1121/1.2903867]

PACS number(s): 43.71.Es, 43.71.An, 43.66.Ba [MSS] Pages: 4539–4546

I. INTRODUCTION

The accurate perception of speech sounds in noisy environments is a common challenge for listeners. While a number of studies have investigated identification and recognition of speech sounds in noise at suprathreshold speech levels, few studies have focused on the detection of isolated phonemes in noise (Tolhurst, 1949; Turner *et al.*, 1992). The determination of phoneme detection thresholds in noise is a critical step in the measurement of speech perception in noise if a goal is to evaluate equally audible speech stimuli. In measures of speech perception in noise, the stimulus level is often indexed by the root-mean-squared (rms) level of the digital or electrical signal and the overall sound pressure level of the acoustic signal for specific speech sounds. Interestingly, vowel stimuli with the same sound pressure level often differ both in terms of loudness, when presented at suprathreshold levels, and audibility, when masked detection thresholds for the vowel stimuli are measured. As a result, multiple stimuli presented at a fixed signal-to-noise ratio, for example, may differ dramatically in terms of their audibility, with some sounds being clearly audible while others are inaudible. The purpose of the present study was to determine the minimum sound pressure level required for the detection of isolated vowels in noise corresponding to 70.7% correct detection.

Several studies have measured thresholds of vowel detection in quiet (Fletcher, 1929; Tiffany, 1953; Kewley-Port, 1991). Fletcher (1929) estimated detection thresholds for a 13-vowel set and reported that /ɔ/ was the most audible and /i/ was the least audible, with an 11 dB threshold range across vowels. Tiffany (1953) employed a method of limits to measure detection thresholds for nine recorded vowels. In agreement with Fletcher (1929), the highest threshold was for the /i/ vowel and the lowest threshold was for /ɔ/. Examination of correlations between the thresholds and several acoustic measurements indicated the highest correlation between the vowel detection threshold and formant location relative to 1000 Hz. Kewley-Port (1991) assessed the detectability of ten isolated vowels produced by three talkers (a female, a male, and synthetic vowels) with vowel durations ranging from 20 to 160 ms. Vowel detection thresholds differed significantly across vowel types, duration, and talkers. For a given duration, detection thresholds varied as much as 22 dB across the 30 vowels.

The three studies cited above indicated that vowel detection thresholds in quiet vary substantially across vowel stimuli. Other studies of vowel perception also indicate variations in the detectability of isolated vowels in noise. For example, Pickett (1957) reported the identification of 12 English vowels embedded in phonetic-balanced word lists and

[a)]Portions of the data were presented at the 151st meeting of the Acoustical Society of America.
[b)]Electronic mail: changliu@mail.utexas.edu
[c)]Electronic mail: david_eddins@urmsc.rochester.edu

in syllable lists presented in noises of various spectra. For a given signal-to-noise ratio, the identification score was highest for /e/, /a/, and /ɔ/ and lowest for /u/ and /U/, which is grossly consistent with the vowel detection thresholds reported by Tiffany (1953).

Studies of the detectability and intelligibility of vowels presented in quiet or in noise at a fixed signal-to-noise ratio indicate that certain vowels are more easily detected and identified than other vowels (Kewley-Port, 1991; Pickett, 1957; Tiffany, 1953). As a result, at low signal-to-noise ratios, some vowel stimuli may be easily identified, while others may not even be detected. To ensure that all stimuli are audible, accurate estimates of vowel detection thresholds in noise are required. Knowledge of vowel detection thresholds in noise, and the associated acoustic characteristics of both the vowels and the noise, is important to the understanding of auditory perception of speech. Interestingly, reports of vowel detectability in noise have not been published. Thus, a primary goal of this study was to assess detection thresholds in long-term speech-shaped noise for six sets of isolated vowels spoken by three female and three male talkers. Kewley-Port (1991) reported that the high variability associated with vowel detection thresholds in quiet corresponded closely with the average excitation across audio frequency based on excitation pattern computations (Moore and Glasberg, 1987). Based on these results, a secondary goal was to determine whether or not vowel detection thresholds in noise can be predicted on the basis of a constant change in excitation estimated from excitation pattern computations. If so, then studies of speech perception in noise may control stimulus audibility by first estimating the stimulus threshold via computation rather than via time consuming empirical methods with human listeners.

II. METHODS

A. Listeners

Seven adult native speakers of American English between the ages of 20 and 32 years served as listeners and were paid for participation in this study. All listeners had normal-hearing sensitivity corresponding to pure-tone thresholds ≤15 dB Hearing Level (HL; ANSI, 1996) at octave intervals between 250 and 8000 Hz.

B. Stimuli

Detection thresholds were measured for 12 American English vowels /i, ɪ, e, ɛ, æ, ʌ, ɜ, ɑ, ɔ, o, u, U/. Vowel stimuli, which are taken from the database reported by Hillenbrand *et al.* (1995), were recorded in the syllable context of /hVd/ spoken by three female and three male talkers. Only isolated vowels were selected for threshold measurements in the present study. These isolated vowels were edited by deleting the formant transition at the beginning and end of the syllable such that only the relatively steady-state vowel nucleus remained with a duration of 170 ms. In order to minimize the effect of the fundamental frequency on vowel detection, the *F*0 contour of all 12 vowels of a given talker (including the /æ/ vowel itself) was equalized to the *F*0 contour of /æ/ vowel of that talker using a modified version of STRAIGHT (Kawahara *et al.*, 1999).

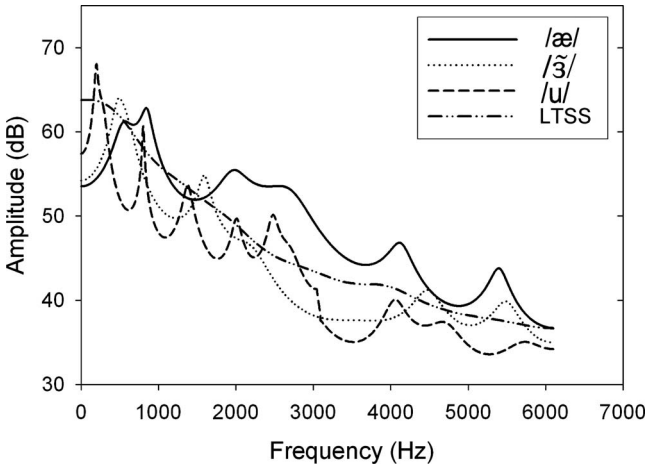


FIG. 1. Magnitude spectra for three vowels /æ,ɜ,u/ from female talker 1 and the LTSS noise. Stimuli were scaled to 70 dB SPL and sampled at 12 207 Hz. Magnitude spectra were obtained by using LPC analysis with 16 coefficients.

C. Noise

Long-term speech-shaped (LTSS) noise was chosen as the masker because LTSS noise is similar to white noise in terms of amplitude fluctuations but is a more efficient masker due to its similarity to the spectra of speech signals. The LTSS noise is presented at 70 dB sound pressure level (SPL), was generated from a Gaussian noise that was shaped by a filter with an average spectrum of 12-talker babble (Kalikow *et al.*, 1977). Figure 1 shows the linear predictive coding (LPC) spectra of the LTSS noise and three vowels, /æ, ɜ, u/, from female talker 1 with signal and noise levels equal to 70 dB SPL.

D. Stimulus generation

Digital stimuli were sampled at 12 207 Hz by a two-channel 20 bit digital-to-analog converter (TDT RP2.1), scaled to the appropriate presentation level by programable attenuators (TDT PA5), summed (TDT SM5), routed to a headphone buffer (TDT HB7), and delivered to an insert earphone (Etymotic ER-2) inserted into the right ear of the listener who was seated in a sound attenuation chamber (IAC). Noise and speech sound pressure levels were verified at the output of the insert earphones via G.R.A.S. IEC 126 2 cc coupler mated to the microphone of a Quest (model 2700) sound level meter set to the linear weighting scale.

E. Procedures

Vowel detection thresholds were measured using a two-interval, two-alternative, forced-choice (2AFC) procedure with an adaptive, two-down, one-up, tracking algorithm, estimating the signal level corresponding to 70.7% correct detection (Levitt, 1971). On each trial, independent samples of LTSS noise were presented in each of two presentation intervals. The vowel stimulus was randomly presented in one of the two presentation intervals. The task required the listener

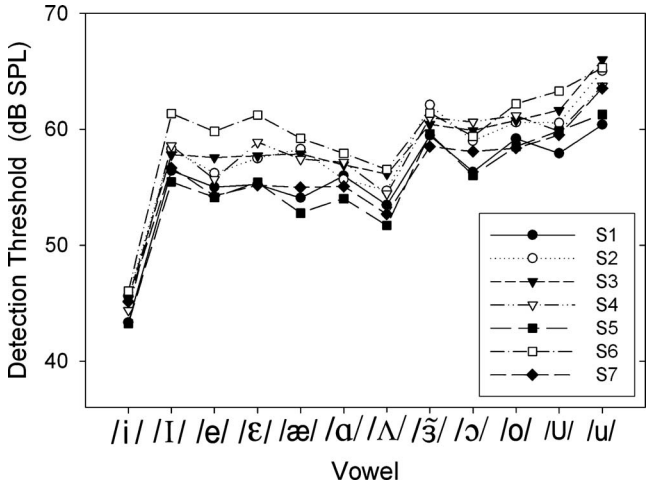


FIG. 2. Detection thresholds for vowels produced by male talker 1 in the presence of LTSS noise at 70 dB SPL as a function of the 12 vowel categories. Thresholds are shown for the seven individual listeners, as indicated in the legend.

to indicate the interval corresponding to the vowel sound via button press on a liquid crystal display monitor with an image that simulated a hand-held button box. Intervals were marked visually by a light above the corresponding interval button that was illuminated simultaneously with the masker noise burst. Vowel stimuli were temporally centered in 400 ms LTSS noise bursts. Presentation intervals were separated by a 400 ms silent period. Threshold for a given condition was taken as the average threshold based on two 60-trial blocks unless threshold for the two blocks differed by more than 3 dB, in which case a third block was completed and averaged with the first two. For each block, the signal level was adjusted in 5 dB steps for the first three reversals in the adaptive track and in 1 dB steps thereafter. Threshold for a single block of trials was based on the average level corresponding to the last even number of level reversals in the adaptive track, excluding the first three. Short breaks were provided between blocks and all threshold measures were completed in ten sessions with each session lasting about 1.5 h. Vowel detection thresholds were measured for the three female talkers first and then for the three male talkers. Within each talker gender group, the order of the experimental conditions was randomized across talkers and vowel categories. Prior to data collection, listeners were given three practice blocks in which the signal was the /æ/ vowel spoken by a female talker. The experimental design, stimulus generation, response collection, and data storage was controlled by the SYKOFIZX® *v*2.0 software application (TDT).

III. RESULTS

Figure 2 illustrates thresholds for each of the seven listeners as a function of vowel category for male talker 1. For this talker, the average threshold across listeners ranged 18.9 dB from a minimum of 44.7 dB SPL to a maximum of 63.6 dB SPL. The pattern of thresholds is quite similar across individual listeners, differing primarily in terms of a constant listener-specific threshold shift. Such a shift may be conceptualized as a shift in the operating characteristic or processing efficiency (cf. Patterson, 1976) for each listener, irrespective of vowel category. In other words, normalized thresholds across listeners would reveal seven largely overlapping functions. Data for the other talkers produced similar patterns of results.

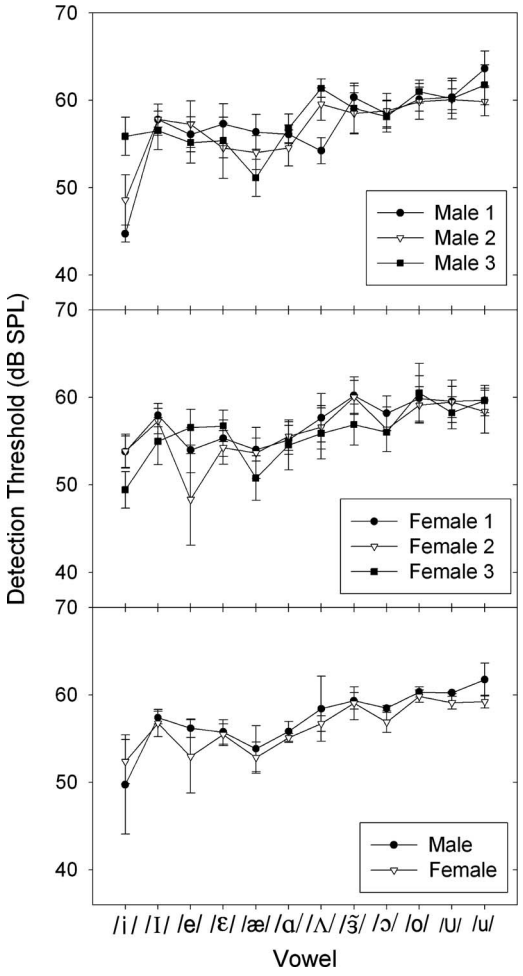


FIG. 3. Detection thresholds for vowels presented in LTSS noise (70 dB SPL) as a function of 12 vowel categories. Symbols and error bars represent the average and standard deviation, respectively, across seven listeners. Data are for the three male talkers (upper panel), three female talkers (middle panel), and average thresholds across the three male and three female talkers (lower panel).

Thresholds for each vowel stimulus, averaged across the seven listeners, are shown in Fig. 3 for the three male talkers (upper panel) and the three female talkers (middle panel). A two-factor (vowel X talker) repeated-measure analysis of variance with detection threshold as the dependent variable indicated a significant effect of vowel category [$F(11,66) = 155.355$, $p < 0.001$] and talker [$F(5,20) = 9.124$, $p < 0.001$] on vowel detection thresholds, as well as a significant vowel category-talker interaction [$F(55,330) = 18.045$, $p < 0.001$]. The effects of vowel category and talker are individually described in the next two subsections.

A. Effects of vowel category

Vowels are usually categorized along the vowel $F1 \times F2$ triangle such that the abscissa labeled "vowel" in Fig. 3

shows the order from front to back. The pattern of thresholds across vowel categories shown in Fig. 3, termed vowel detection pattern hereafter, is similar in form to the vowel detection patterns in quiet previously reported (Fletcher, 1929; Tiffany, 1953; Kewley-Port, 1991). Vowel detectability differed across categories by as much as 18.9 dB for male talker 1 and as little as 6.4 dB for female talker 1, with threshold ranges for the other four talkers falling in between these extremes. Thresholds averaged across talkers varied over a range of 9.4 dB across the 12 vowel categories, comparable to the ranges previously reported by Fletcher (1929) and Tiffany (1953) but smaller than the range reported by Kewley-Port (1991). In general, the front vowels /i/ and /æ/ had the lowest masked thresholds, while the back vowel /u/ and the central vowel /ɜ/ had the highest masked thresholds. This pattern is somewhat different from vowel detection in quiet, in which the threshold for /ɑ/ was lowest and that for /u/ was highest (Fletcher, 1929; Tiffany, 1953; Kewley-Port, 1991). The difference among vowel detection patterns likely reflects the effect of LTSS noise in the present study.

B. Effects of talker

As noted above, there was a significant dependence of vowel detectability on the talker. A *post hoc* Tukey test significantly indicated higher detection thresholds for male talker 3 than for female talker 3 ($p<0.05$), whereas differences among other talkers were not significant ($p>0.05$). A planned comparison between male and female talkers revealed no significant talker gender differences ($p>0.05$) (see lower panel of Fig. 3). Although there was a significant threshold difference among talkers, the average thresholds over the 12 vowels differed by only 1.8 dB across talkers ranging from 55.9 to 57.7 dB SPL. The significant interaction between talkers and vowel categories indicated that the vowel detection patterns were talker dependent, as seen in Fig. 3.

These results demonstrate that vowel detectability in noise depends on both the vowel category and talker, similar to the data for vowel detection in quiet previously reported. Kewley-Port (1991) reported that the pattern of vowel detection thresholds across vowel category in quiet closely corresponded to a single-valued excitation level computed as the average excitation level across the entire audio-frequency range. To determine whether or not a similar relation exists for vowel detection in noise, excitation pattern computations were carried out for the stimuli used here, based on the excitation pattern model proposed and modified by Moore and Glasberg (1987, 2004) in the context of their loudness model.

IV. MODELING VOWEL DETECTION IN NOISE

A. Description of the excitation pattern model

The excitation pattern model described by Moore and Glasberg (1987, 2004) assumes that the auditory periphery can be represented by a bank of overlapping bandpass filters, the characteristics of which were determined on the basis of the results of several investigations of auditory filter shape using the notched noise masking paradigm (e.g., Patterson, 1976). The model includes the following stages. The first stage includes a transfer function that represents the transmission of the stimulus through the outer ear from either the free field or earphones to the eardrum (a transfer function for the ER-2 insert earphone was used here). The second stage includes a transfer function accounting for the transmission of the stimulus through the middle ear. Together, the outer and middle ears provide greater transmission for middle frequencies than for low or high frequencies. In the third stage, the excitation level is computed as a function of audio frequency based on psychoacoustic estimates of auditory filter characteristics and is conceptualized as the distribution of excitation level along the basilar membrane (Moore and Glasberg, 1983). The resulting excitation is expressed in terms of equivalent rectangular bandwidth number (ERB_N for normal-hearing listeners) representing adjacent rectangular filters with bandwidths that correspond to the output of Roex filters following the procedure of Glasberg and Moore (1990).

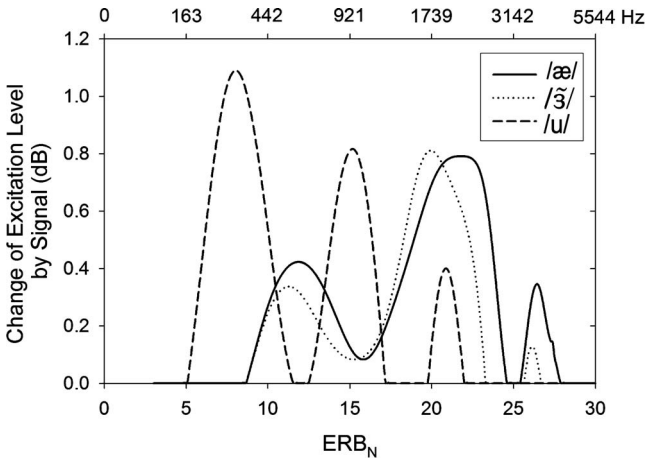


FIG. 4. Change in excitation level as a function of ERB number (lower *x* axis) and audio frequency (upper *x* axis) evoked by three vowel stimuli /æ, ɜ, u/ at threshold (averaged over the seven young normal-hearing listeners) for female talker 1 in the LTSS noise.

The excitation pattern corresponding to the masked vowel signal was calculated as the difference between the excitation pattern for the noise-only interval and the excitation pattern for the vowel-plus-noise interval in decibels. Because independent noise samples were used for each condition, random variations in the excitation patterns are expected. Excitation pattern differences computed for a single trial are shown in Fig. 4 for the vowels /æ, ʌ, u/ presented at the detection threshold in noise. In each case, the formant peaks corresponding to $F1$, $F2$, and $F3$ are clearly represented. For example, the back vowel /u/ resulted in an excitation pattern difference of 1.1 dB near $ERB_N=8$, indicating that the vowel produced a 1.1 dB increase in excitation at the corresponding audio frequency relative to the LTSS noise alone.

The primary question addressed here is whether or not a single-valued metric based on excitation pattern computations can account for differences in vowel detection thresholds across vowel categories and talkers, independent of acoustic features (e.g., fundamental frequency or spectral

shape) and/or listening conditions (quiet or noise). One approach is to simulate threshold by presenting noise alone and signal-plus-noise stimuli to the excitation pattern model and to use an adaptive up-down rule and a decision statistic to determine thresholds for each condition.

The selection of decision statistics was based on the hypothesis that vowel detection depends on a change in the excitation pattern produced by the vowel signal added to the noise background. Three metrics were evaluated including the change in the excitation pattern averaged across the entire spectrum (EP all), the average change over a limited frequency region (<3 kHz: EP 3 kHz) spanning the most salient spectral cues including $F1$ and $F2$, and the maximal excitation level in the excitation pattern (EP max) representing the greatest local change over the entire frequency. For example, consider the excitation pattern difference shown by the solid line in Fig. 4 representing the /æ/ vowel at threshold computed for a single trial. The average excitation level across the entire spectrum is 0.204 dB, whereas the average excitation level below 3 kHz (25.6 ERB_N) is 0.277 dB, and the maximal excitation level is 0.790 dB centered at 22.0 ERB_N.

B. Prediction of vowel detection thresholds

Vowel detection thresholds in LTSS noise were simulated using a two-interval 2AFC procedure with an adaptive algorithm identical to that used in the behavioral measurements of the main experiment. The success of each metric will be evaluated by two criteria: a high correlation and a low variance between predicted thresholds and thresholds measured in human listeners. These two evaluation methods are described as follows. Because predicted thresholds represent thresholds for the group of normal-hearing listeners without including listener variability, behavioral thresholds used in this section for comparison were averaged over all seven normal-hearing listeners.

Vowel detection thresholds were computed by using the following simulation technique executed in the MATLAB® software package. Stimulus generation, presentation, and adaptive tracking methods were identical to the behavioral experiment described above. By following stimulus generation, the excitation pattern was computed over the central 160 ms portion of the stimulus for each presentation interval (noise alone and noise plus vowel) of the two-interval trial. The signal interval was chosen as the interval having the greater value of the excitation pattern metric under consideration. The predicted threshold for each of the 72 experimental conditions (12 vowel categories ×6 talkers) was obtained by averaging the thresholds obtained for three 60-trial blocks. Three sets of thresholds were computed, one based on each of the metrics reported above.

Figure 5 shows a comparison between the predicted thresholds (dotted/dashed lines) for each of the three metrics and the behavioral thresholds (solid line) obtained in the main experiment, averaged across six talkers and seven listeners. The patterns of behavioral thresholds and all three sets of predicted thresholds are quite similar. The predicted thresholds were highly correlated with the behavioral thresholds, as shown in column 2 of Table I, with all correlations ≥0.88. As shown in Fig. 5, the major difference between the behavioral thresholds (solid circles) and the three sets of predicted thresholds is a constant function-specific shift in level. By following the logic used to explain differences among individual listeners (see Fig. 2), one may reason that this shift in level results from the fact that the models cannot capture listener-specific differences in processing efficiency. As will be shown in the next section, estimating this constant level difference is a key element in the application of the model with novel listeners and stimuli. To assess the goodness of fit, the rms error was then computed by averaging the squared deviations between the predicted and behavioral thresholds across the 12 vowels and 6 talkers. The rms errors are shown in Table I (column 3). Thresholds predicted on the basis of the average excitation level below 3 kHz are slightly better when considering both correlation with the behavioral thresholds and goodness of fit (i.e., rms error). These simulations indicate that the variability in vowel detection thresholds associated with different talkers and different vowel categories can be explained by a constant change in excitation level independent of vowel category and talker.

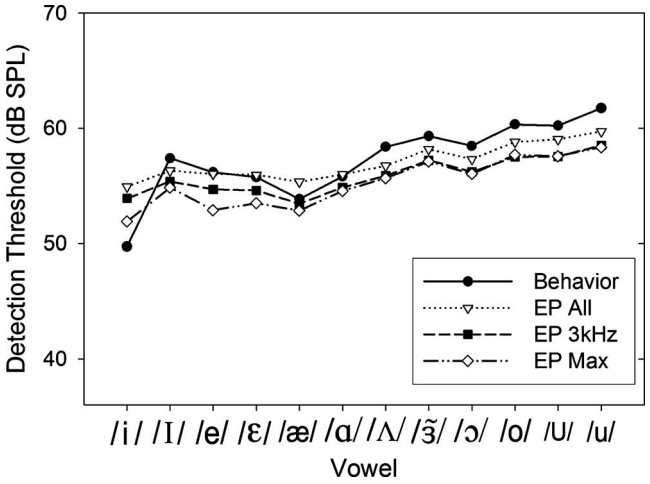


FIG. 5. Comparison between thresholds measured from human listeners and thresholds predicted from the three auditory metrics for the main detection experiment. The behavioral thresholds were averaged over the six talkers and seven listeners, while the predicted thresholds were averaged over the six talkers. EP=excitation pattern.

C. Application of the simulation model

One goal of this study was to determine if our simulation model could successfully predict vowel detection thresh-

TABLE I. Correlation between behavioral and predicted thresholds for the detection of 12 isolated vowels presented in long-term speech-shaped noise. Behavioral thresholds represent the average across seven young normal-hearing listeners and six talkers, as shown by the solid circles in Fig. 5. The predicted thresholds obtained using three excitation pattern metrics (rows, see text) were averaged across six talkers and are shown as dashed/dotted lines in Fig. 5. Corresponding rms error values are also shown.

Auditory metrics	Correlation	rms error (dB)
EP all	0.88	3.7
EP 3 kHz	0.96	3.0
EP max	0.90	3.0

TABLE II. Comparison of behavioral and predicted thresholds for each of three excitation pattern metrics (columns) as a function of vowel category (rows) for the main detection experiment. Behavioral thresholds were averaged across the seven listeners. The rms error (dB) was computed separately for each vowel category based on the behavioral and predicted thresholds obtained for the six talkers. The /æ/ vowel, which is highlighted in bold font, showed the lowest rms error among the 12 vowel categories for both the EP all and the EP max excitation pattern metrics.

Vowels	EP all	EP 3 Hz	EP max
/i/	4.0	3.0	2.8
/I/	4.2	4.0	2.9
/e/	3.6	2.6	2.7
/ɛ/	3.0	2.1	2.6
/æ/	**2.7**	**2.2**	**1.8**
/ɑ/	3.5	3.3	4.0
/ʌ/	3.6	3.9	3.8
/ɜ/	3.6	2.7	3.2
/ɔ/	3.1	2.6	3.6
/o/	4.1	3.5	2.8
/U/	3.4	3.9	4.1
/u/	5.2	3.9	3.5

olds for different talkers such that one would not need to conduct a preliminary detection in noise experiment in order to equalize audibility and/or achieve a specific sensational level for all speech stimuli used in a given study. As shown in Fig. 5, the patterns of thresholds across vowel category were very similar for the predicted and measured thresholds. Thus, the present data indicate that such a simulation can be used to predict the variations in audibility across vowel categories and talkers. An additional challenge of our simulation model is to quantify the listener-specific threshold shift in behavioral thresholds for a given talker, as shown in Fig. 2. Although the threshold shift is specific to the listener and the talker, it is fairly constant across the vowel category (see Fig. 2). Therefore, the level shift can be estimated on the basis of the threshold for a single vowel for each talker. The /æ/ vowel was chosen as the best stimulus for this task since the rms error between the predicted and measured thresholds, which was computed for the 6 talkers, 12 vowels, and 3 excitation pattern metrics of the main experiment reported above, was lowest for this stimulus, as shown in Table II.

The proposed procedure for the estimation of vowel detection thresholds for a given talker is described as follows. First, talker-specific predicted detection thresholds are obtained via threshold simulations as described above. Second, threshold for the /æ/ vowel is measured for each human listener. Third, the level difference between the predicted and measured detection thresholds for the /æ/ vowel is computed for each listener. Fourth, the predicted thresholds from the threshold simulations in the first step are scaled by the level difference obtained in step 3 for the /æ/ vowel. Fifth, scaled predicted thresholds are taken as the estimated vowel detection thresholds for a given listener and a given talker. Finally, steps 1–5 are repeated to obtain detection thresholds for other talkers.

To examine the success of this vowel detection threshold estimation procedure, detection thresholds for 12 vowels, which are spoken by 2 new talkers (1 female and 1 male),

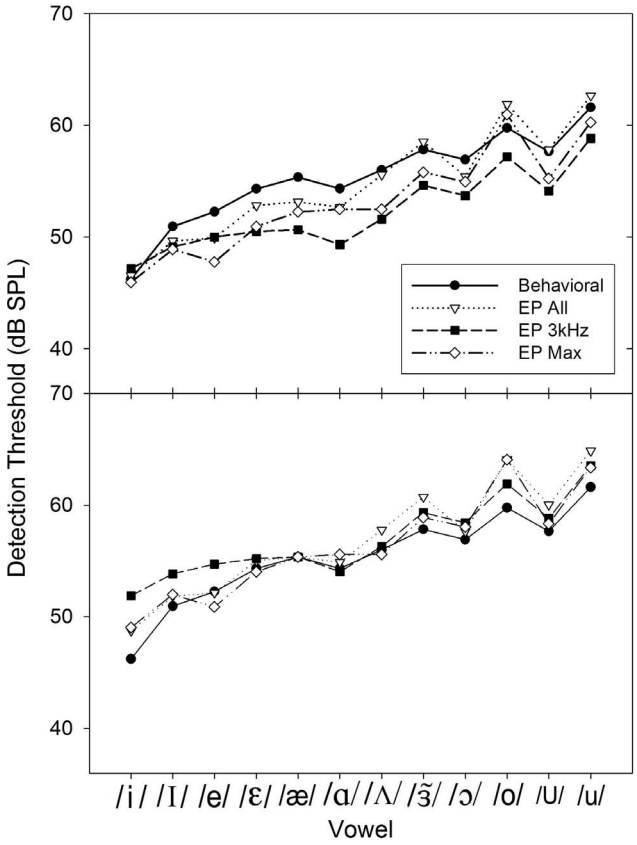


FIG. 6. Comparison of thresholds measured for human listeners and thresholds predicted from the three auditory metrics for the new detection experiment. Behavioral and predicted thresholds were obtained from the average over the two new talkers and three new listeners, while predicted thresholds were shown before (upper panel) and after (lower panel) being scaled by the detection thresholds for the /æ/ vowel. EP=excitation pattern.

presented in LTSS noise were predicted by using the procedure above and were behaviorally measured for three young normal-hearing listeners who did not participate in the main detection experiment. Figure 6 shows the predicted and measured thresholds for the three excitation pattern metrics for the average over the two new talkers and three new listeners before (upper panel) and after (lower panel) being scaled by the threshold of /æ/ vowel, while Table III provides the rms errors and correlation values used to determine the accuracy

TABLE III. Correlations between behavioral and predicted thresholds for the detection of 12 isolated vowels presented in long-term speech-shaped noise for the new detection experiment before and after being scaled by the /æ/ vowel thresholds. Behavioral thresholds represent the average across three young normal-hearing listeners and two new talkers, as shown by the solid line in Fig. 6. Predicted thresholds averaged over the two new talkers were obtained by using three excitation pattern metrics (rows, see text) and are shown as dashed/dotted lines in Fig. 6. Corresponding rms error values also are shown before and after scaling.

	Correlation		rms error (dB)	
Auditory metrics	Before scaling	After scaling	Before scaling	After scaling
EP all	0.97	0.97	2.1	2.0
EP 3 kHz	0.93	0.93	3.4	2.3
EP max	0.94	0.94	2.6	1.8

of the predictions. The high correlations (≥0.93) between the predicted and measured thresholds, remaining the same before and after the scaling, indicate that the simulation model successfully predicted the pattern of vowel detection thresholds for these talkers. The rms errors associated with the predicted thresholds ranged from 1.8 to 2.3 dB for the three excitation pattern metrics after being scaled, slightly lower than the rms errors before being scaled with the lowest rms error corresponding to the EP max. These values are consistent with the results of Skovenborg and Nielson (2004), who evaluated the relative success of several models of loudness perception in predicting the loudness perception of complex sounds. They reported rms errors between predicted and measured data that ranged from 0.8 to 3.4 dB across the various loudness models. Though the excitation pattern model on which our predictions are based (Moore and Glasberg, 1987; 2004) was not used by Skovenborg and Nielson (2004), they did include the loudness model of Zwicker and Fastl (1999) that has a similar representation of peripheral excitation, both being based in part on the model of Zwicker (1958) and Zwicker and Scharf (1965). Skovenborg and Nielson (2004) showed that the rms error between predicted and measured data produced by the Zwicker and Fastl (1999) model was 2.1 dB, which is comparable to the rms errors (see Table III) reported in the present study.

Altogether, the vowel detection threshold estimation procedure described here predicted not only the pattern of thresholds across the vowel category but also threshold differences across talkers within a single listener-specific level difference that may be likened to a listener operating characteristic or an index of efficiency for signal detection in noise. It should be noted that the current estimation procedure did not include any temporal features of the signal or masker stimuli, which may limit the generalization of this simulation model depending on the potential application. Given that the stimuli in this study were steady-state vowels with little temporal variation, the excitation pattern model successfully explained the variance in detection thresholds associated with vowel category and talker. With more dynamic stimuli, successful prediction of detection by thresholds may need to incorporate the temporal features of the stimuli by using a model such as the one described by Glasberg and Moore (2005). Likewise, stimulus duration was constant in the present experiment. Large variations in stimulus duration across a set of stimuli may require that the estimation procedure include duration specific parameters. Despite these potential limitations, the procedure described has the potential to save an enormous amount of time if the goal of an experiment is to use equally audible stimuli that are similar in duration and relatively steady state, such as vowels (Watson and Gengel, 1969).

V. SUMMARY AND CONCLUSIONS

The goal of this study was to estimate vowel detection thresholds in noise for a 12-vowel set produced by 6 different talkers. Vowel detectability in noise showed great variability across the vowel categories. The degree of threshold variability across the vowel category was talker specific ranging from 6.4 to 18.9 dB. This threshold variability is in general agreement with the range of vowel detection thresholds in quiet reported previously (Fletcher, 1929; Tiffany, 1953). In general, the vowels /æ/ and /i/ had the lowest thresholds, while the vowel /u/ had the highest masked threshold, Although vowel detection thresholds significantly differed across talkers, when averaged across vowel categories, the average difference across talkers was less than 2 dB. Considering that the pattern of vowel detection thresholds did not differ between the female and male talker groups, the difference in vowel detectability across talkers is attributed primarily to differences in the vowel spectra rather than differences related to the fundamental frequency.

In order to account for the variability in detection thresholds, three metrics derived from an excitation pattern model (Moore and Glasberg, 1987; 2004) were examined. The three different metrics used in the excitation pattern analyses differed only in terms of the frequency range over which the excitation level was estimated (the entire spectrum, below 3000 Hz, or at the excitation pattern peak). The results showed that thresholds predicted from all the three metrics closely matched the thresholds obtained from human listeners, indicating that a single-valued metric based on the excitation pattern can be used to describe vowel detection thresholds independent of vowel category and talker. The large variability in behavioral detection thresholds across vowel category reported here indicates that among a set of vowels scaled to have the same rms sound pressure level, some vowels may be inaudible while others may be clearly audible and identifiable when presented at relatively low signal-to-noise ratios.

If the primary goal of an experiment were to measure vowel identification in noise for the same set of 12 vowels spoken by six talkers, and the experimenters wanted to normalize audibility across vowel categories and talkers, then the behavioral measurements required to do so would take approximately 15 h per listener, as outlined in Sec. II. The present results suggest that to avoid this time consuming task, one might rescale the vowel level depending on excitation patterns of the vowel stimuli in noise following the measurement of a single vowel detection threshold for each listener.

ACKNOWLEDGMENTS

The authors are thankful to two anonymous reviewers and Mitchell Sommers for their constructive suggestions and comments on early drafts of this manuscript.

ANSI (**2004**). "Specification for Audiometers," ANSI Report No. S3.6-2004, (ANSI, New York).
Fletcher, H. (**1929**). *Speech and Hearing* (Van Nostrand, New York).
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Glasberg, B. R., and Moore, B. C. J. (**2005**). "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," J. Audio Eng. Soc. **53**, 906–918.
Hillenbrand, J., Getty, L. J., Clark, M. J., and Weeler, K. (**1995**). "Acoustic characteristics of American English Vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Kalikow, D. N., Stevens, K. M., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with con-

trolled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.
Kawahara, H., Masuda-Kastuse, I., and Cheveigne, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.
Kewley-Port, D. (**1991**). "Detection thresholds for isolated vowels," J. Acoust. Soc. Am. **89**, 820–829.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Moore, B. C. J., and Glasberg, B. R. (**1987**). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Hear. Res. **28**, 209–225.
Moore, B. C. J., and Glasberg, B. R. (**2004**). "A revised model of loudness perception applied to cochlear hearing loss," Hear. Res. **188**, 70–88.
Skovenborg, E., and Nielson, S. H. (**2004**). "Evaluation of different loudness models with music and speech materials," in *Proc. 117th Audio Eng. Soc. Conv.*, San Francisco, CA **2004**.
Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am. **59**, 640–654.
Pickett, J. (**1957**). "Perception of vowels heard in noises of various spectra," J. Acoust. Soc. Am. **29**, 613–620.
Tiffany, W. (**1953**). "The threshold reliability of recorded sustained vowels," J. Speech Hear Disord. **18**, 379–385.
Tolhurst, G. C. (**1949**). "Audibility of the voiceless consonants as a function of intensity," J. Speech Hear Disord. **14**, 210–215.
Turner, C. W., Fabry, D. A., Barrett, S., and Horwitz, A. R. (**1992**). "Detection and recognition of stop consonants by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **35**, 942–949.
Watson, C., and Gengel, R. (**1969**). "Signal duration and signal frequency in relation to auditory sensitivity," J. Acoust. Soc. Am. **46**, 989–997.
Zwicker, E. (**1958**). "Uber psychologische und methodische Grundlagen der Lautheit," Acustica **8**, 237–258.
Zwicker, E., and Fastl, H. (**1999**). *Psychacoustics: Facts and Models*, Springer Series in Information Sciences, Vol. **22**, 2nd ed. (Springer-Verlag, Berlin).
Zwicker, E., and Scharf, B. (**1965**). "A model of summation," Psychol. Rev. **72**, 3–26.

On the robustness of overall F0-only modifications to the perception of emotions in speech

Murtaza Bulut[a] and Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory, http://sail.usc.edu, Electrical Engineering Department, University of Southern California, Los Angeles, California 90089

(Received 26 October 2006; revised 24 March 2008; accepted 24 March 2008)

Emotional information in speech is commonly described in terms of prosody features such as F0, duration, and energy. In this paper, the focus is on how F0 characteristics can be used to effectively parametrize emotional quality in speech signals. Using an analysis-by-synthesis approach, F0 mean, range, and shape properties of emotional utterances are systematically modified. The results show the aspects of the F0 parameter that can be modified without causing any significant changes in the perception of emotions. To model this behavior the concept of emotional regions is introduced. Emotional regions represent the variability present in the emotional speech and provide a new procedure for studying speech cues for judgments of emotion. The method is applied to F0 but can be also used on other aspects of prosody such as duration or loudness. Statistical analysis of the factors affecting the emotional regions, and discussion of the effects of F0 modifications on the emotion and speech quality perception are also presented. The results show that F0 range is more important than F0 mean for emotion expression.
 [DOI: 10.1121/1.2909562]

PACS number(s): 43.72.Ar [DOS] Pages: 4547–4558

I. INTRODUCTION

Studies of emotional speech have shown that emotion change can be associated with changes in the prosodic and spectral characteristics of speech signals (Bulut *et al.*, 2005, 2002; Burkhardt and Sendlmeier, 2000; Cahn, 1990; Montero *et al.*, 1999). The focus has been mainly on prosody parameters, such as F0, duration, and energy. Among these, significant attention has been paid to F0 contour modulations occurring as a result of emotion change (Cowie *et al.*, 2001; Murray and Arnott, 1993; Scherer, 2003).

Acoustic analyses of angry or happy speech show that, in general, their F0 mean, median, range, and variance values are larger than their neutral speech counterparts, which are larger than sad emotion F0 values (Davitz, 1964; Iida *et al.*, 2003; Murray and Arnott, 1993). The F0 contours of happy and angry speech, in most cases, are more variable than neutral speech, showing fast and irregular up and down movements, while the sad speech F0 contours show smaller variation and downward inflections (Davitz, 1964; Murray and Arnott, 1993). Although these findings are fairly consistent across different studies, differences are not uncommon. For instance, in Yildirim *et al.* (2004) sad speech had a higher F0 mean than neutral speech.

Despite having a powerful descriptive value, the aforementioned technique for studying emotions has several limitations. For example, its implementation in emotional speech synthesis is limited (Cowie *et al.*, 2001) because it does not specifically account for the variability present in the natural speech (Braun *et al.*, 2006; Chu *et al.*, 2006; Pell, 2001). In the traditional analysis, an emotional utterance is represented as a point in the parameter space. We suggest a new model where each utterance is represented by an "*emotional region*" in the parameter space. The proposed model is a new procedure for studying speech cues for judgments of emotion. In this paper, the method was applied to F0 but it can also be used on other aspects of prosody such as duration or loudness.

In this paper, we show how F0 mean, range, and shape characteristics can vary in emotional utterances, and how these variations can be modeled using the emotional regions. Statistical analyses of the factors that cause the variability are presented. In addition, the effects of different F0 modifications on emotional content perception are also investigated. These results show that F0 range is more important than F0 mean for emotional expression.

The concept of the variability of prosodic patterns was studied in a database composed of two repetitions of 1000 sentences recorded with six months separation by Chu *et al.* (2006). The results showed wide variations in F0 values, sometimes corresponding to 50% of the dynamical range of the speaker. In an another study (Braun *et al.*, 2006) iterative mimicry was employed to observe whether F0 contours converge to specific English intonation patterns, referred to as "attractors." It was only after several iterations that F0 branching (i.e., clustering) patterns were seen. However, even then the variability of F0 contours was noticeable. This was due to the fact that "human variability places a lower limit on the width of the branches" (Braun *et al.*, 2006).

In this paper the concept of emotional regions is introduced to model the variability in the F0 characteristics of emotional utterances. A model based on F0 mean, range, and standard deviation statistics is proposed. Following an analysis-by-synthesis approach it is shown that the proposed

[a]Author to whom correspondence should be addressed. Electronic mail: murtaza.bulut@philips.com

model gives reliable estimation of how F0 contours of emotional utterances can be modified without significantly affecting the perceived emotional content and speech quality. This representation is helpful to better assess the role of F0 contours in emotion perception. Also, when applied together with the F0 generation models such as ToBI (Silverman *et al.*, 1992) or Tilt (Taylor, 2000), it can be used to better predict the intonation events in emotional speech synthesis.

Emotion perception is a result of the interplay between acoustical, lexical, and environmental factors (Traunmuller, 2005). These factors can be expected to have an effect on the emotional regions. We show how the speaker and utterance characteristics affect the emotional regions, and analyze the interaction between different factors using statistical methods.

The effects of modifying F0 contour shape, F0 range, and voice quality characteristics of emotional utterances (in German) were statistically analyzed by Ladd *et al.* (1985) for arousal related (relaxed/aroused, open/deceitful, annoyed/content, insecure/arrogant, and indifferent/involved) and cognition related (emphasis, cooperativeness, contradiction, surprise, and reproach) emotions. The results showed that *text* (i.e., sentence content) had a significant effect on listener judgment. Similarly, the *speaker* factor (i.e., who uttered the utterance) also had significant effect for all emotion categories, except *arrogant*. The results also showed that modifying F0 range had a significant (and *continuous*) effect on emotion perception, especially on speaker arousal. The effects of contour changes were less prominent than range modifications. Similarly, in this paper, we also analyze the effects of F0 range and contour modifications, and also consider the *sentence* and *speaker* as independent factors. However, our focus, in contrast, is on how F0 acoustic features interact with *sentence*, *speaker*, and *emotion* factors in influencing the perception of emotion and speech quality. We use *angry*, *happy*, *sad*, and *neutral* labels, which are a subset of the emotional labels suggested by Ekman and Friesen (1977), to describe emotions.

In this paper, the results were also analyzed from the emotional speech synthesis perspective (Burkhardt and Sendlmeier (2000); Cahn (1990); Raux and Black (2003); Schroder (2001)). It was observed that F0 modification caused the perception of sad and neutral emotions to increase, and angry and happy emotions to decrease. The effects of F0 range modifications were *continuous* (Ladd *et al.* (1985)) and more significant than F0 mean modifications. F0 contour shape modifications were also effective but only when performed in large semitone scales. It was also observed that the listeners were still able to perceive the emotions in a manner similar to that of the unmodified natural utterances even when the speech quality was distorted.

In the next sections we first describe the performed F0 mean, range, and shape modifications (Sec. II) and how they were evaluated (Sec. III). The concept of emotional regions is introduced in Sec. IV and the statistical analyses results are presented in Sec. V. The effects of F0 modifications on emotional content are presented in Sec. VI. The discussion and conclusion follow in Sec. VII and Sec. VIII, respectively.

II. DATA PREPARATION

In this section the emotional data collection and the F0 modifications are explained.

A. Data collection

Two sentences, "*She told me what you did*." (sentence 1) and "*This hat makes me look like an aardvark*." (sentence 2) were recorded by a female speaker (speaker 1) and a male speaker (speaker 2). Both speakers were in their late 20s. Speaker 1 had some professional acting experience, while speaker 2 did not.

The speakers were instructed to utter the two sentences in *angry*, *happy*, *sad*, and *neutral* (i.e., no particular emotion) emotion styles, resulting in a total of 16 utterances (see Fig. 1). However, no specific instructions were given on how the emotions should be expressed. In other words, the interpretation and expression of emotions was left to the speakers themselves. The speech was recorded in a quiet room at 48 kHz sampling rate using unidirectional head-worn dynamic Shure brand (model SM10) microphones. Later the speech was down sampled to 16 kHz. Listening tests were conducted, afterwards, to evaluate the success of emotion production. The results showed that human listeners were able to correctly identify (with approximately 80% success on average) the emotions expressed by the speakers.

B. F0 modifications

Several modifications manipulating the mean, range, and shape of the natural F0 contours were applied to all recorded emotional utterances (which will be also referred to as *original* utterances). The F0 mean, range, and shape modifications were performed using the Time Domain Pitch Synchronous Overlap and Addition (TD-PSOLA) algorithm (Moulines and Charpentier, 1990) as implemented in the Praat software (Boersma and Weenink, 2007).

The applied modifications can be categorized into three groups: Mean, range, and stylization modifications (summarized in Table I).

Modifications in F0 mean: The mean was modified by shifting the F0 contour up or down. The following modifications were applied: (1) Increasing/decreasing the original F0 mean by 10%, 15%, 25%, and 50%, (2) Making the F0 mean equal to 50, 100, 150, 200, 250, and 300 (Hz).

Modifications in F0 range: The range was modified by multiplying the F0 contour with a constant and then shifting the contour up or down so that the mean will be the same as the original mean value. The following modifications were applied: (1) Scaling the range by 0.5, 0.75, 1.5, and 2, (2) Making the F0 range equal to 10, 30, 50, 80, 110, and 150 (Hz).

Stylization modifications: The shape of the F0 contour of the utterances was altered by stylizing the F0 contour. The following modifications were applied: Stylizing the F0 contour by a 2, 5, 10, 15, and 40 semitone frequency resolution.

Stylization of the F0 contour was performed using the Praat software. The logic behind the stylization algorithm is to try to represent the F0 contour using linear segments. The length of the linear segments was determined by the fre-

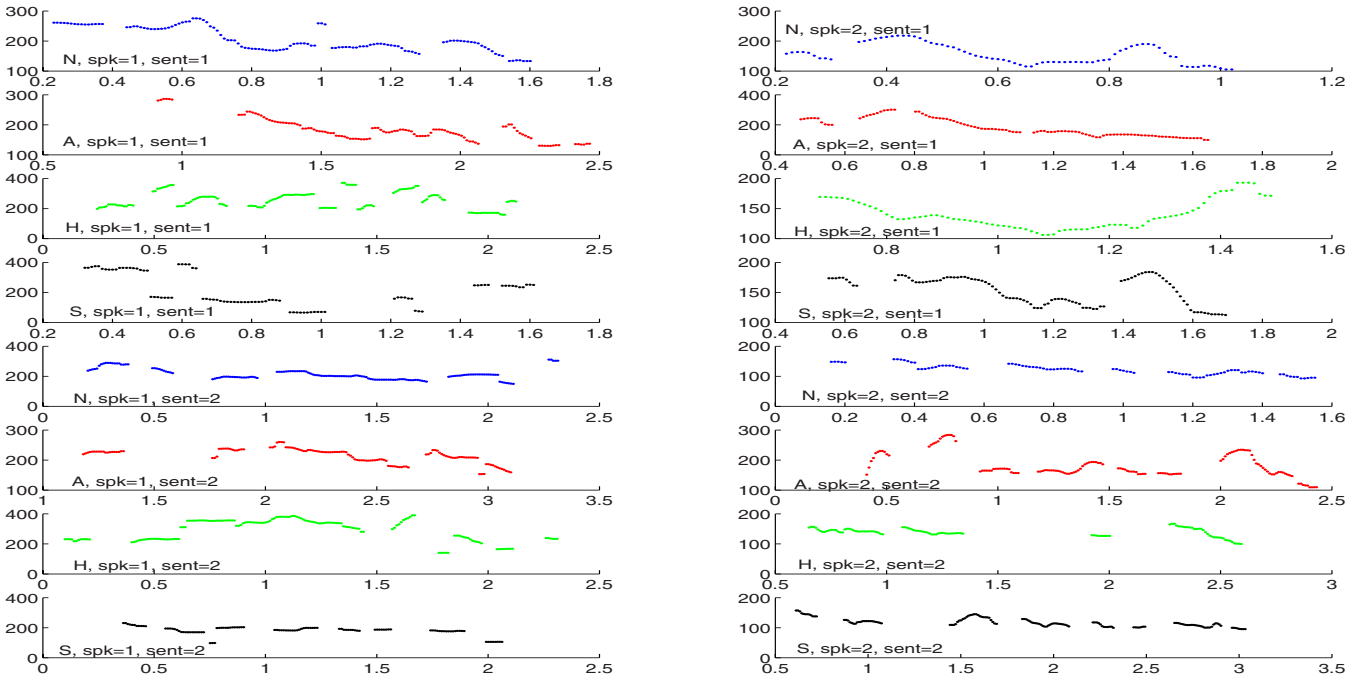


FIG. 1. (Color online) F0 contours of all 16 utterances that were recorded. *H*, *A*, *N*, *S*, *spk*, and *sent* denote happy, angry, neutral, sad, speaker, and sentence, respectively.

quency resolution component. For instance, while a 2 semitone resolution corresponds to fairly short linear segments, thus preserving the general contour shape, 40 semitone resolution may cause the whole utterance F0 contour to be a line (see Fig. 2 for an example).

As a result of applying the aforementioned modifications, 29 utterances, all having exactly the same duration as the original utterance but different F0 contours, were resynthesized for each of the original (i.e., recorded, natural) utterances. In total, including the original utterances, there were $(30\times 16=)$ 480 utterances.

III. LISTENING TESTS

All natural and resynthesized utterances were evaluated by listening experiments with naive listeners that included both native and non-native American English speakers. Before evaluation, all speech files were normalized so that the maximum digitized waveform amplitude was 1. In the listening tests—conducted in a quiet room, using headphones and with a single rater at a time—first the speech file was presented and then the raters were asked to choose among the following options: *Happy*, *angry*, *sad*, *neutral*, and *other*. The raters were particularly instructed to choose *other* if their choice of emotion was not listed or if they could not decide on the emotional content, or if the speech sounded to them as a mixture of several emotions. They were allowed to listen to each utterance as many times as they liked before making their decision. After the raters had chosen the emotion, they were asked to rate the naturalness (i.e., speech quality) of the utterance on a scale from 1 to 5, with 5 cor-

TABLE I. Summary of the performed F0 contour modifications. The values for mean and range are in Hz and the values for stylization are in semitone.

F0	Mean	Range	Stylization
Increase	m1: +10%	r3: +50%	
	m2: +15%	r4: +100%	
	m3: +25%		
	m4: +50%		
Decrease	m5: −10%	r1: −50%	
	m6: −15%	r2: −25%	
	m7: −25%		
	m8: −50%		
Set value	m9: =50	r5:= 10	s1:= 2
	m10:= 100	r6:= 30	s2:= 5
	m11:= 150	r7:= 50	s3:= 10
	m12:= 200	r8:= 80	s4:= 15
	m13:= 250	r9:= 110	s5:= 40
	m14:= 300	r10:= 150	

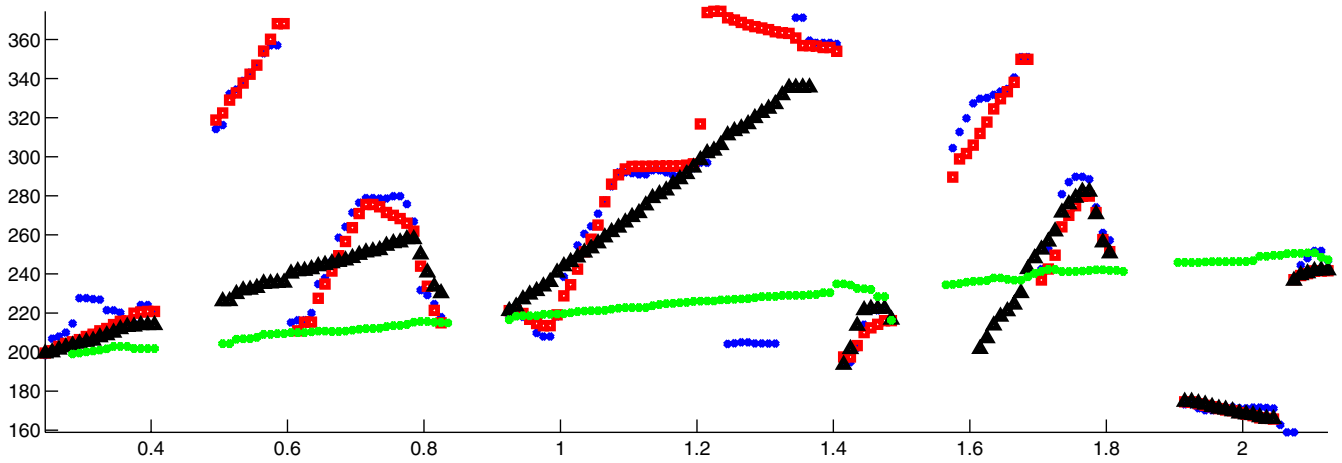


FIG. 2. (Color online) Stylization example for the happy utterance, speaker=1, sentence=1. Circles=original F0 contour, Squares=2 semitones stylization, Triangles=10 semitones stylization, Dots=40 semitones stylization.

responding to the most natural. They were specifically instructed to give low values if the speech was perceived to be different from natural human speech in terms of quality. Again, the raters were able to listen to the speech as many times as they liked. The files were presented in a different random order for each rater.

From the preliminary listening tests it was observed that long lasting tests were tiring for listeners (which may negatively affect their judgment abilities). Because of that, the average test time was limited to 20 min. per set. In order to limit the time of any single test to around 20 min., the test set was divided into ten groups of 48 utterances, each consisting or three variations of the 16 original utterances (which were chosen randomly). After the completion of a set, listeners were given the opportunity to rest (or to continue some other time), or to continue with a different test set.

The average number of raters per set was 9.2. In total, there were 14 different people that participated. Of these, seven people (three female, and four male listeners) evaluated all utterances. Most of the raters were graduate students in their mid to late 20s.

IV. EMOTIONAL REGIONS IN F0 MEAN-RANGE SPACE

One of the basic characteristics of natural speech is its variability (Braun *et al.*, 2006; Chu *et al.*, 2006). In order to generate models for speech production, synthesis, and perception this variability should be accounted for appropriately. In this section, we show examples of the variability present in emotional speech and propose a model to parametrize it.

For each of the resynthesized utterances, the F0 contour was calculated using the Praat software. After removing the outliers and smoothing using a median filter of length 3, F0 mean, F0 range [=(0.975 quantile)-(0.025 quantile)], and F0 standard deviation (std) statistics were calculated.

Based on the results of the listening tests, all resynthesized utterances were assigned an emotional label using majority voting. Then, each of these utterances was grouped together with its original version (i.e., the utterance from which it was resynthesized) only if its emotion was the same as the emotion of the original utterance. As a result, 16 (=2 speakers×2 sentences×4 emotions) groups (one for each original utterance) were generated. The utterances in these groups were used to construct the emotional regions.

We introduce the idea of emotional regions to model the variability in the F0 parameter values of emotional utterances. Using emotional regions one can theoretically represent how the F0 contour of an utterance can be modified without significantly affecting its emotion and speech quality. Note that, the dimension of these regions is dependent on the number of parameters that are used. In this paper, for easy visualization we worked with two-dimensional (2D) regions, which were estimated based on the F0 mean and range values. If F0 contour shape was also considered as a factor, the emotional regions would be three dimensional.

Grouping the utterances into 16 groups based on speaker, sentence, and emotion, as explained above, for each group, the group mean vector and covariance matrix were calculated and constant Mahalanobis distance contours—equal to 3—were determined. The center and shape of these contours are determined by the mean vectors and by the covariance matrices, respectively. The contours are ellipses (Fig. 3) and they represent the equal probability density Gaussians (Duda *et al.*, 2001). The Mahalanobis distance was set to 3 as a result of experiments that showed that these contours were reliable estimates for the distribution of resynthesized utterances as can be seen in Fig. 4.

Each of these Gaussian emotional regions, shown in Fig. 3, represent a subset of possible F0 values with which a given original utterance can be modified to maintain the same emotion perception by the majority of the listeners. Note that the Gaussian emotional regions in Fig. 3 are considered a subset of the true emotional regions because they were estimated based on a limited set of modifications (listed in Table I).

Speech quality can be also included as one of the factors determining the emotional regions. In this case, in addition to the requirement that the utterances need to be perceived as conveying a certain emotion, they are also required to be perceived with a certain minimum average speech quality.

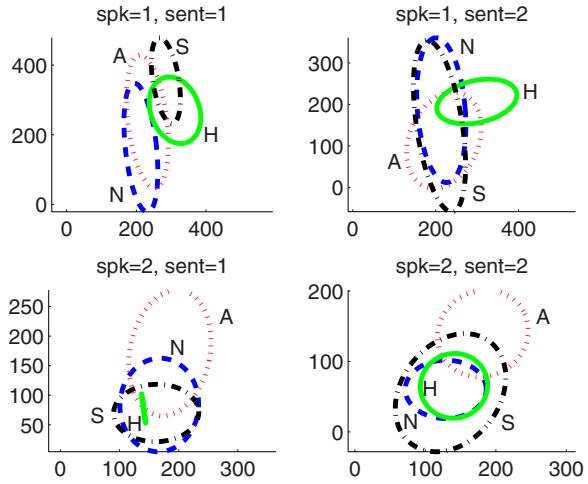


FIG. 3. (Color online) The Gaussian emotional regions for each emotion, speaker, and sentence. x axis=F0 mean (Hz), y axis=F0 range (Hz).

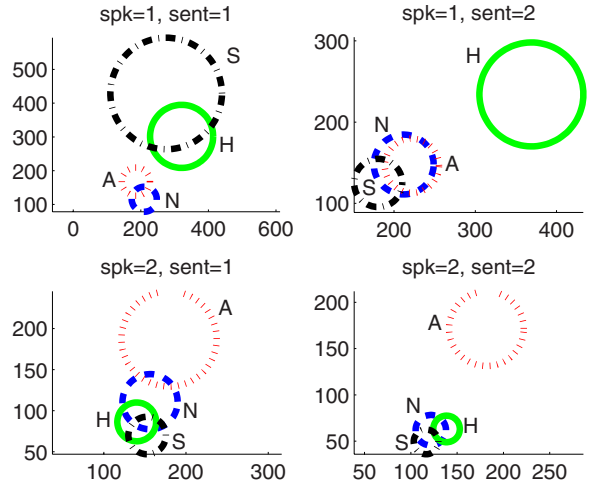


FIG. 5. (Color online) Euclidean emotional regions estimated from the F0 contours of original utterances. x axis=F0 mean (Hz), y axis=F0 range (Hz).

For example, denoting the average speech quality by ρ, it may be required for each of them to satisfy $\rho \geq 3$, $\rho \geq 3.5$, $\rho \geq 4$, or $\rho \geq 4.5$ conditions. Under these requirements, the area of emotional regions can be expected to decrease as quality requirements increase. An example is shown in Fig. 4, which shows the emotional regions for angry utterances. Although not shown, when higher quality conditions were applied, the size of the emotional regions (shown in Fig. 3) decreased in a similar manner for the other emotions as well.

The emotional regions shown in Fig. 3 and Fig. 4 were estimated using the resynthesized utterances. In order for them to be used in real life applications they need to be estimated automatically for individual utterance F0 contours. For that purpose, F0 mean, range, and std values can be used. For a given utterance, representing the center by [F0 mean, F0 range], and the radius by F0 std, (circular) Euclidean emotional regions can be constructed. As shown in Fig. 5, contours of one std Euclidean distance from the [F0 mean, F0 range] point were plotted and considered as Euclidean emotional regions for a given utterance.

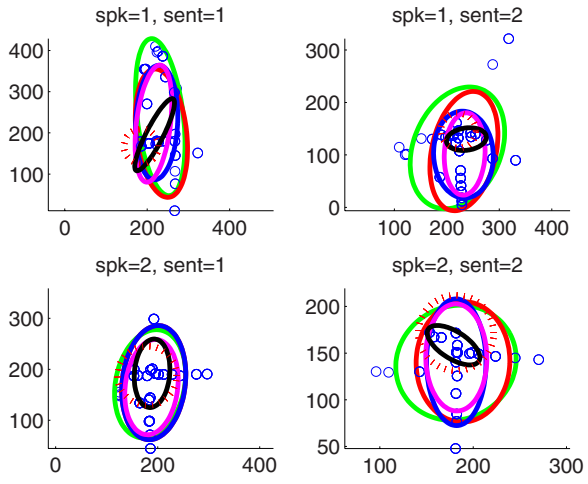


FIG. 4. (Color online) Emotional regions for different speech quality (ρ) requirements for angry emotion. The areas of emotional regions decrease as quality requirements increase. Displayed are the following quality conditions: (1) No restriction (same as Fig. 3), (2) $\rho \geq 3$, (3) $\rho \geq 3.5$, (4) $\rho \geq 4$, (5) $\rho \geq 4.5$. Small circles show the resynthesized utterances. x axis=F0 mean (Hz), y axis=F0 range (Hz).

In order to determine how well the Euclidean regions can approximate the Gaussian regions, they were plotted together in Fig. 6. Comparing the two regions we note that for most of the groups Euclidean regions lie inside the Gaussian regions. This shows that the regions estimated by the Euclidean method are reasonable and accurate, representative of a subset of Gaussian regions. This is more clearly seen in Fig. 4 where the Gaussian regions for different quality conditions were plotted together with the Euclidean regions (shown as dotted circles). While observing the plots, note the similarity between the Euclidean emotional regions and higher quality ($\rho \geq 4.5$) Gaussian regions. Figure 4 shows the results for angry utterances only, but the results were similar for other analyzed emotions.

From the figures it can be seen that the emotional regions were different for different speakers. For example in Fig. 3, note that for speaker 1, the happy region did not lie inside sad or neutral region, while for speaker 2 it did. Also note that for speaker 2, the intersection between angry and neutral regions was smaller in comparison to their intersection for speaker 1. In addition to the speaker related differences, differences due to sentences were also observed. For instance, for speaker 1, the neutral region for sentence 2 was inside the sad region, while for sentence 1 it was not.

The differences between the emotional regions can be attributed to the differences in the factors—such as sentence, speaker, and emotion—that affect the F0 contour characteristics (see Fig. 1). The effects of these factors are examined in detail in the next section (Sec. V) where statistical analyses results are presented.

It is important to note that the present emotional regions are proposed as models to represent the variability of single utterance F0 parameter values, and therefore they are specific to the utterance itself. They show the ranges within which the utterance F0 parameters can be modified without affecting its emotional and speech quality. However, they do not

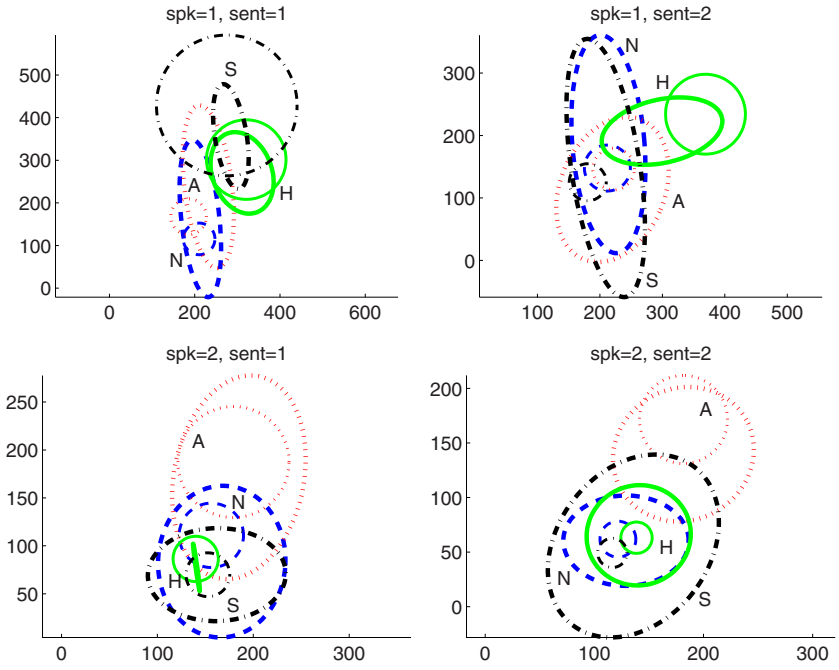


FIG. 6. (Color online) The perceived Gaussian emotional regions and estimated Euclidean emotional regions, x axis=F0 mean (Hz), y axis=F0 range (Hz).

necessarily show how these parameters should be modified to synthesize speech with a new emotion. For example, if a happy utterance is modified so that its new F0 values fall outside the happy emotional region, it is known that it will not be perceived as happy anymore. However, it is not necessarily true that if the new point is in the neutral (or any other) region then the utterance will be perceived as neutral. This is due to the fact that the perception of emotions is based not only on F0, but on the combined effects of prosodic, spectral, and linguistic factors. Therefore only when all of these factors are used to construct multidimensional regions one can predict the emotion only from the region itself. Note that the present emotional regions can be considered as projections of the hypothetical multidimensional regions on the linguistic (sentence), spectral (speaker), and F0 planes.

As shown in Fig. 4, in general, the Euclidean emotional regions can be considered to be reasonable approximations to the high quality Gaussian regions. As seen in Fig. 5, in the Euclidean method the assumption is that the variations in F0 mean and F0 range directions are equal. If needed, another model which estimates possible F0 range and F0 mean variations separately can be also constructed. For example, for an utterance, if in addition to the F0 contour, word (and/or syllable) boundaries are also known, one can calculate the F0 mean and range values for each word (syllable). Then, these two vectors (the vectors of F0 means and F0 ranges) can be used to calculate the covariance matrix, which can be used to form the new emotional regions (which will be ellipsoidal and not circular) for the utterance.

V. STATISTICAL ANALYSIS OF EMOTION AND SPEECH QUALITY PERCEPTION

In order to examine the effects of utterance emotion, speaker, sentence, and modification factors on emotion and quality perception a four-way ($4\times2\times2\times30$) repeated measures analysis of variance (ANOVA) model was designed. The model consisted of four independent variables [*original utterance emotion (4)*, *speaker (2)*, *sentence (2)*, and *modification (30)*] used as repeated measures, and two dependent variables (*emotion* and *speech quality*). The model was fully counterbalanced across seven subjects (i.e., listeners) that evaluated all 480 utterances (which correspond to all possible combinations of the independent variables).

Of the within-subject independent variables, *utterance emotion type* has four levels that reflect the emotions intended by the speakers. These levels correspond to happy, angry, sad, and neutral emotions. The *speaker* variable has two levels, corresponding to speaker 1 (who was a female) and speaker 2 (who was a male). The *sentence* variable has two levels, sentence 1 (She told me what you did.), and sentence 2 (This hat makes me look like an aardvark.). And finally, the *modification* variable has 30 levels that correspond to all performed F0 modifications cases (29), plus the no modification (i.e., original) case (see Table I for the complete list of the modifications).

There are two dependent variables. Of these, *emotion selection* is a nominal variable that was defined as a dichotomous outcome reflecting whether the emotion selected by a listener for a resynthesized utterance was the same as the emotion of the original utterance. If they were the same the variable was set to 1, if they were different it was set to 0. A dichotomous variable was used, because the purpose of the experiment was to investigate specifically the role of the F0 component in the perception of the original emotion.

The *quality* dependent variable was used as a measure of the perceived speech quality that was evaluated on a five point scale as explained in Sec. III.

TABLE II. Cochran's Q statistics calculated for *emotion selection* dependent variable. Significant results are in *italic* form.

Modification		Mean		Range		Stylization	
Emotion	Spk./Sent.	Sent1	Sent2	Sent1	Sent2	Sent1	Sent2
Happy	spk1	Q(14)=16.26 p=0.298	*Q(14)=27.06 p=0.019*	*Q(10)=20.00 p=0.029*	Q(10)=17.61 p=0.062	*Q(5)=19.52 p=0.002*	Q(5)=10.77 p=0.056
	spk2	*Q(14)=29.63 p=0.009*	*Q(14)=32.36 p=0.004*	Q(10)=16.79 p=0.079	Q(10)=9.00 p=0.532	Q(5)=10.91 p=0.053	Q(5)=4.00 p=0.549
Angry	spk1	*Q(14)=28.24 p=0.013*	Q(14)=11.41 p=0.654	*Q(10)=38.31 p<0.001*	Q(10)=10.00 p=0.440	Q(5)=5.00 p=0.416	Q(5)=10 p=0.075
	spk2	Q(14)=14.25 p=0.431	*Q(14)=30.84 p=0.006*	*Q(10)=33.08 p<0.001*	Q(10)=14.70 p=0.144	Q(5)=3.46 p=0.629	*Q(5)=18.10 p=0.003*
Sad	spk1	Q(14)=11.17 p=0.673	*Q(14)=31.31 p=0.005*	Q(10)=7.78 p=0.651	Q(10)=4.24 p=0.936	Q(5)=3.40 p=0.639	Q(5)=5.56 p=0.352
	spk2	*Q(14)=32.62 p=0.003*	Q(14)=16.63 p=0.277	Q(10)=15.22 p=0.124	Q(10)=7.14 p=0.712	Q(5)=8.23 p=0.144	*Q(5)=15.00 p=0.010*
Neutral	spk1	*Q(14)=29.38 p=0.009*	*Q(14)=30.92 p=0.006*	*Q(10)=24.00 p=0.008*	*Q(10)=29.34 p=0.001*	*Q(5)=15.85 p=0.007*	*Q(5)=21.07 p=0.001*
	spk2	Q(14)=21.33 p=0.093	*Q(14)=26.80 p=0.020*	Q(10)=11.72 p=0.304	Q(10)=10.00 p=0.440	*Q(5)=15.29 p=0.009*	Q(5)=6.30 p=0.278

A. Factors influencing emotion perception

The null hypothesis tested was the following: *The probability of intended (i.e., original) emotions correctly perceived by listeners is equal across different variants in a group*. The variants in this case, as explained above, consisted of all possible combinations of independent variables. There were 480 different variants in total, which were grouped based on the sentence, speaker, emotion, and modification (mean, range, or stylization) factors, resulting in 48 ($=2\times2\times4\times3$) groups.

In our experimental setup each of the seven listeners (i.e., subjects) evaluated all of the utterances. Thus, the subjects were treated as related samples. Therefore, to test the null hypothesis, Cochran's Q test was used. The required condition for the application of Cochran's Q test, that the number of the conditions (K) and number of the listeners (N) are such that $KN>30$, was satisfied for all of the analyzed groups. The results of these tests are shown in Table II. The statistically significant ($p<0.05$) results are shown in *italics* for ease of differentiation.

Note that, since the purpose of this analysis was to investigate whether F0 modifications are sufficient to alter the emotional content of original utterances, for each of the cases compared in Table II, the original (i.e., unmodified) utterances were also included. For example, the results reported in the lower right corner (of Table II) are for the group consisting of neutral sentence 2 recorded by speaker 2 and its stylization modifications, in total 6 utterances (5 modified and one original). The size of the groups comparing mean, range, and stylization modifications were 15, 11, and 6, respectively.

From the results it is observed that the effects of F0 modifications on emotion perception were dependent on *emotion*, *speaker*, and *sentence* factors, showing the complex interactions between these parameters. For example, it is seen that sentence 2 uttered in angry emotion was not significantly influenced by the range modifications, while in contrast, the same modifications caused the perceived emotions for angry sentence 1 to be significantly different than its original. Note, however, that when sentence 2 uttered by speaker 1 in neutral style was modified by the same F0 range modifications, the perceived emotions were different than the emotions perceived for the original utterance. Other similar observations can be made from the results in Table II.

Also it is notable that especially for speaker 1 modifying the F0 characteristics of neutral utterances caused the perception of new emotional nuances. In contrast, this result was less common for angry and happy emotions, and the least common for sad emotion.

B. Factors influencing speech quality perception

The null hypothesis tested was the following: *The mean of the perceived quality is the same under different conditions*. The repeated measures ANOVA results are reported in Table III (the significant results are shown in *italic*). Shown in the tables are the F values calculated from Greenhouse–Geisser tests. This test was preferred because it accounts—by adjusting the degrees of freedom—for the violations of sphericity condition.

The results show that the main effects of emotion, speaker, and sentence factors were insignificant, while the main effect of modification was significant (see Table III). Interesting results were found from the interaction analysis of the within-subject factors. Note, for instance, that the effect of F0 modifications (on the perceived speech quality) was significantly dependent on emotion, speaker, and sentence variables. Also note that the effect of speakers was significantly dependent on emotion, but not on sentence.

In order to analyze the effects of F0 modifications, speaker and sentence factors for different emotion conditions, statistical analyses were performed separately for different emotions. These results are shown in Table IV. For all emotions, it is seen that the main effect of speaker was not statistically significant. In contrast, the main effect of modi-

TABLE III. Repeated measures ANOVA statistics calculated for *quality* dependent variable. The reported are the F values for Greenhouse–Geisser tests.

Factor	Greenhouse–Geisser statistics
Emotion	$F(1.72,10.30)=1.87$, $p=0.203$
Speaker	$F(1,6)=0.96$, $p=0.366$
Sentence	$F(1,6)=0.02$, $p=0.890$
Modification	$F(3.84,23.02)=28.27$, $p<0.001$
*Emo*Spk*	$F(2.50,15.03)=3.64$, $p=0.043$
*Emo*Sent*	$F(1.20,7.21)=7.70$, $p=0.024$
*Emo*Modif*	$F(5.40,32.40)=6.07$, $p<0.001$
Spk*Sent	$F(1,6)=3.144$, $p=0.127$
*Spk*Modif*	$F(4.24,25.45)=13.25$, $p<0.001$
*Sent*Modif*	$F(4.12,24.74)=5.16$, $p=0.003$
*Emo*Spk*Sent*	$F(1.22,7.30)=7.21$, $p=0.027$
*Emo*Spk*Modif*	$F(5.43,32.60)=2.64$, $P=0.037$
*Emo*Sent*Modif*	$F(5.20,31.22)=2.76$, $p=0.034$
*Spk*Sent*Modif*	$F(4.67,28.06)=3.36$, $p=0.019$
*Emo*Spk*Sent*Modif*	$F(4.89,29.35)=2.84$, $P=0.034$

fication was significant in all cases. Interestingly, we also observe that the effect of sentence was significant for angry and neutral emotions, but not for happy and sad. In fact, note that the patterns of significant results were the same for happy and sad emotions and somewhat similar between angry and neutral emotions.

VI. EFFECTS OF F0 MODIFICATIONS ON EMOTIONAL CONTENT

In this section, the effects of F0 modifications are compared in terms of emotional content that was perceived. Responses from all 14 listeners were included in these evaluations.

In Fig. 7 the changes in the emotion recognition percentages observed after each modification are shown. The change was defined as the difference between recognition percentages of unmodified and modified utterances. Chi-square tests with 95% confidence interval were used to calculate whether or not the change was significant. The discussions below focus mainly on the significant modifications.

The mean modifications that caused significant ($p<0.05$) emotion perception changes for speaker 1 were m4, m8, m9, m10, m11 [Fig. 7(a)]. These results show that speaker 1 was quite robust against the F0 mean modifications. It was only when the F0 mean was changed by $\pm 50\%$ significant changes were observed. Increasing F0 mean caused the *neutral* and *angry* recognition percentages to drop, and *sad* and *other* recognition percentages to increase. Interestingly, adjusting the mean to be in 50–150 Hz range caused increase in *happy* and *other* responses. Note that in all these instances the speech quality degraded significantly [Fig. 7(e)].

For speaker 2—as seen with speaker 1—increasing or decreasing F0 mean by 50% caused an increase in *sad* and *other* perception percentages [Fig. 7(c)]. It was also observed that some of the modifications caused an increase in the *neutral* and *other* responses, but not in the *happy* or *angry* responses. The statistically significant modifications in this case were m1, m4, m7, m8, m9, m10, m13, and m14. All of

TABLE IV. Repeated ANOVA statistics calculated for quality dependent variable. The reported are the F values for Greenhouse–Geisser tests. Significant results are shown in *italic* for easy differentiation.

Factor	Happy	Angry	Sad	Neutral
Speaker	$F(1,6)=1.66$	$F(1,6)=0.75$	$F(1,6)=2.13$	$F(1,6)=0.03$
	$p=0.245$	$p=0.421$	$p=0.195$	$p=0.864$
Sentence	$F(1,6)=0.85$	$F(1,6)=11.97$	$F(1,6)=0.55$	$F(1,6)=21.26$
	$p=0.392$	$p=0.013$	$p=0.486$	$p=0.004$
Modification	$F(3.68,22.06)=16.84$	$F(4.23,25.38)=26.93$	$F(4.63,27.79)=7.97$	$F(4.52,27,12)=24.5$
	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
Spk*Sent	$F(1,6)=1.92$	$F(1,6)=33.51$	$F(1,6)=3.67$	$F(1,6)=9.72$
	$p=0.215$	$p=0.001$	$p=0.104$	$p=0.021$
Spk*Modif	$F(4.43,26.56)=6.98$	$F(4.21,25.27)=8.46$	$F(4.72,28.33)=3.05$	$F(5.01,30.07)=7.59$
	$p<0.001$	$p<0.001$	$p=0.027$	$p<0.001$
Sent*Modif	$F(4.75,28.50)=2.12$	$F(4.69,28.12)=8.36$	$F(4.72,28.29)=1.87$	$F(5.02,30.13)=2.35$
	$p=0.095$	$p<0.001$	$p=0.135$	$p=0.065$
Spk*Sent*Modif	$F(4.14,24.82)=1.55$	$F(4.41,26.44)=2.64$	$F(4.51,27.03)=2.57$	$F(4.74,28.46)=5.23$
	$p=0.217$	$p=0.051$	$p=0.055$	$p=0.002$

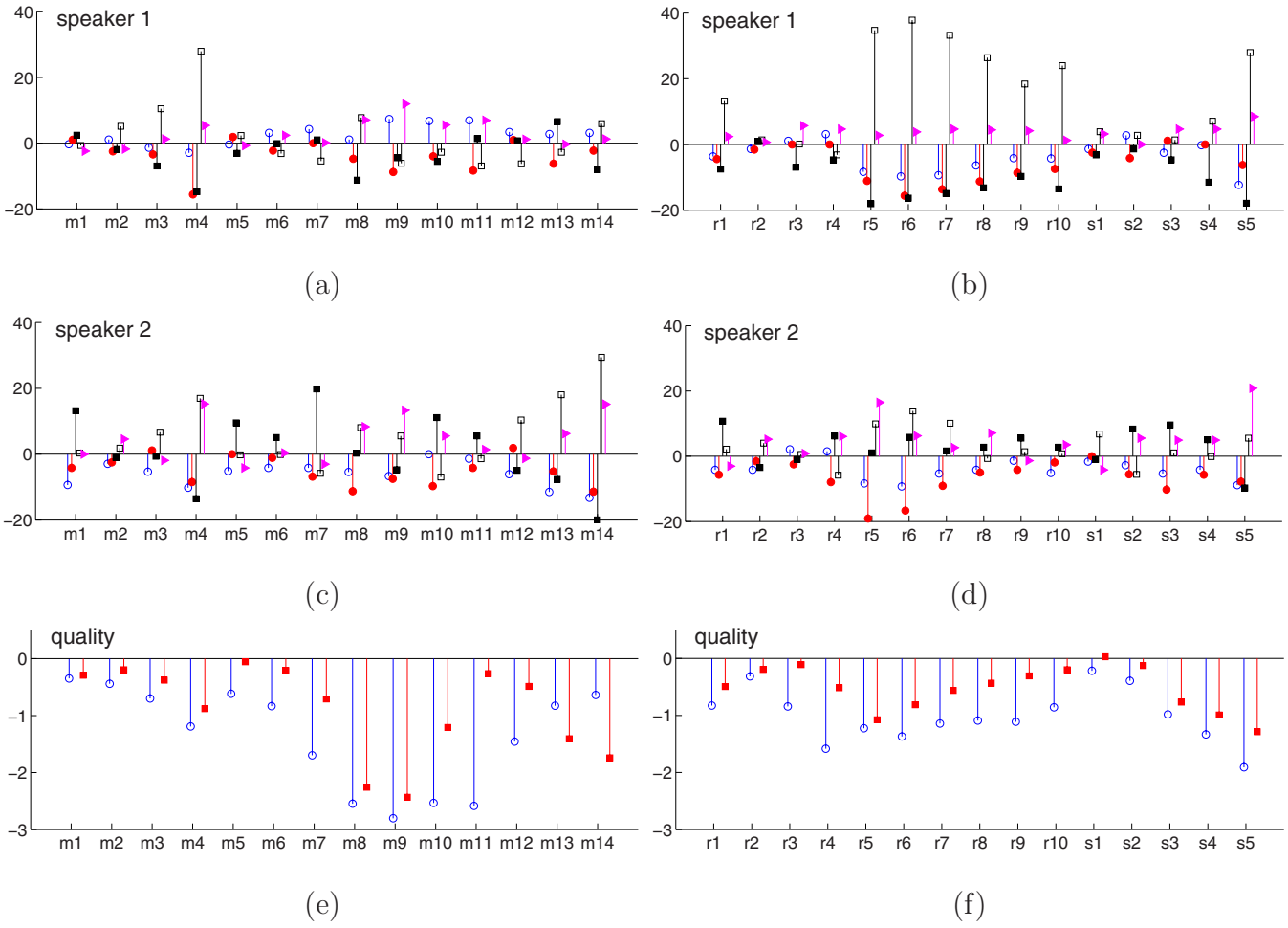


FIG. 7. (Color online) Figures (a), (b), (c), (d): The differences between the emotion recognition percentages of original and modified utterances. Happy=open circle, Angry=filled circle, Sad=open square, Neutral=filled square, Other=filled triangle. Figures (e), (f): The differences between the average speech qualities (5=excellent, 4=good, 3=fair, 2=poor, 1=bad) of original and modified utterances. Speaker 1=open circle, Speaker 2=filled square.

these (except m1) caused a significant drop in the speech quality.

The effects of F0 range modifications were more prominent than F0 mean. For speaker 1, decreasing the F0 range by more than 50% caused a significant increase in *sad* responses (r1, r5, r6, r7, r8, r9, r10). The effect of the F0 range on the sad emotion percentages was continuous (Ladd *et al.*, 1985) and it could be easily parametrized [Fig. 7(b)]. The drop in the perceived speech quality was less severe than F0 mean modifications, suggesting that one should perform range and not mean modifications during the synthesis of emotional speech.

The effects of range modifications on speaker 2 were also significant, however not as strong as they were for speaker 1 [Fig. 7(d)]. This can be attributed to the lower F0 range of this speaker. The modifications that caused significant emotion perception difference were r4, r5, r6, r7, r8. These modifications increased *sad* perception, and decreased *happy* and *angry* perception. In contrast to speaker 1, some of them (r1, r6) also increased the *neutral* perception.

Interesting results were observed for stylization modifications. For speaker 1 [Fig. 7(b)], only s4 and s5 caused significantly different results. An increase in the *sad* and *other* responses and drop in quality was seen for these cases. These results show that eliminating the small prosodic variations (s1, s2, s3) in the F0 contour shape did not significantly decrease the perception of the original emotions. It was only when the F0 contour at the sentence level was fully linearized (as seen in Fig. 2)—eliminating any accents and foot patterns (Klabbers and van Santen, 2004)—the percentages of *happy* and *angry* emotions started to decrease. In these cases the utterances were mostly perceived as *sad* or *other*.

This is an important result which has implications for emotional speech synthesis. As shown in our previous work (Bulut *et al.*, 2005, 2002), for synthesis of emotions such as anger and happiness, in addition to prosody, spectral characteristics also play an important role. Therefore, during synthesis of these emotions one needs to concentrate more on the overall F0 contour shape, F0 range, and spectral characteristics and can ignore the small prosodic variations in the F0 contour shape. As we show later, these small prosodic variations were more important for high quality perception than emotion perception.

The arguments above were also valid for the speaker 2, for whom only some particular stylization modifications (s2, s3, s5) caused significant changes [Fig. 7(d)], with minimal degradation in quality [Fig. 7(f)]. In these cases increase in the *other* responses was accompanied either by increase in

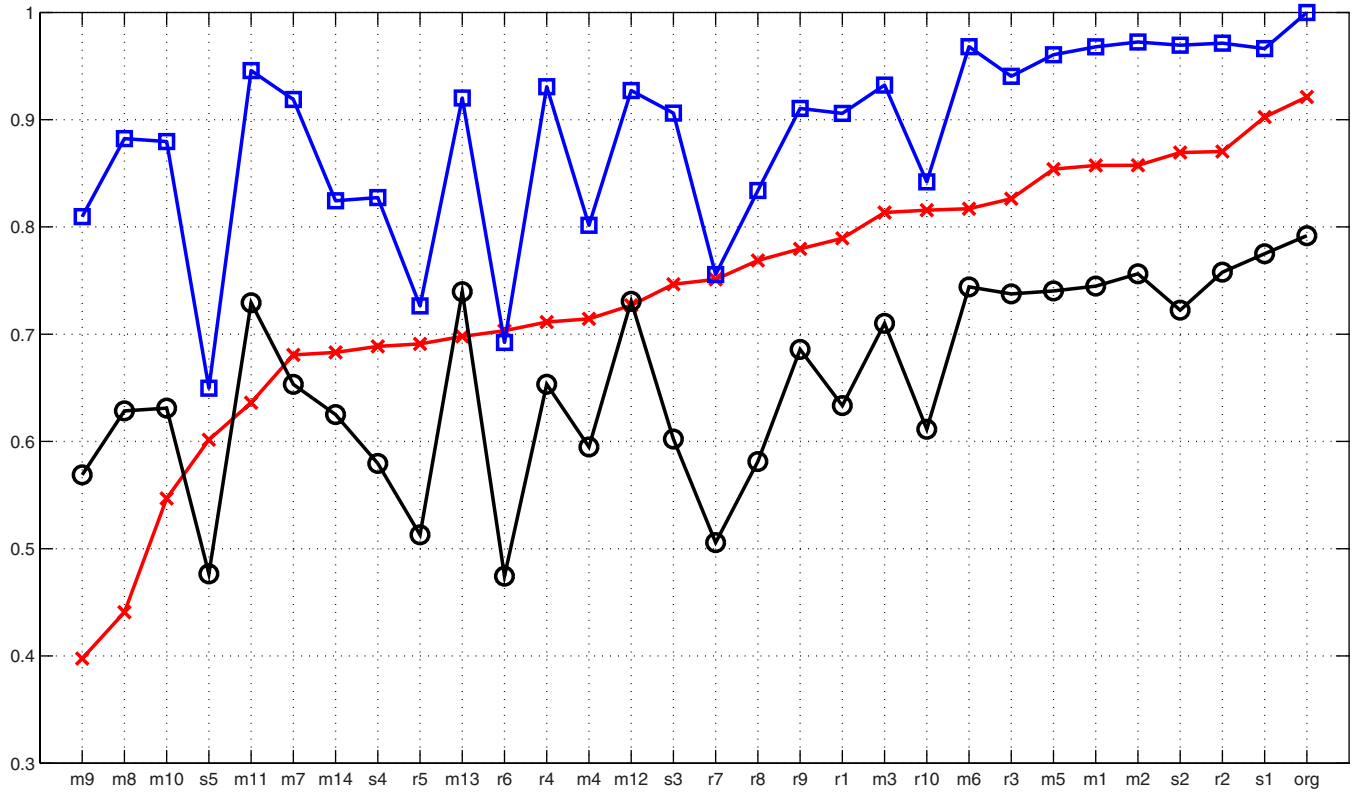


FIG. 8. (Color online) Relation between *average quality*, *similarity*, and *percentage* parameters. Note that the quality is normalized: 1=Excellent, 0.8=Good, 0.6=Fair. 0.4=Poor. 0.2=Bad. Squares (□) are used for *similarity*, circles (○) for *percentage*, and (*x*) for *quality* variables.

sad or *neutral* responses. It was particularly interesting to note that both s1 and s4, which caused significantly different F0 contour shapes, did not cause any significant emotion changes. That the s4 modification was not significant, while s3 was, was unexpected and it can be attributed to the variability inherent in the subjective nature of the listening tests.

As expected, the modifications that received high quality responses were the ones that did not cause any significant changes in the emotional content. Significant emotion change was in general accompanied with the degradation in quality. However in some instances, especially when the F0 range of speaker 2 was modified, despite significant emotion change the quality was not affected.

In almost all instances, the original emotions were correctly recognized by the majority (i.e., 50% or more) of the listeners. This shows that despite the quality distortions the original emotions were still well perceived.

In order to visualize these relations between quality and emotion perception, we define two new variables, *percentage* and *similarity* (see Fig. 8). The percentage variable represents the percentage of listeners that perceived the same emotion as the original emotion. The similarity variable is a measure that is the cosine of the angle between two vectors and has a large value (i.e., close to 1) when the vectors point in the same direction (Duda *et al.*, 2001). It is calculated using Eq. (1), where x and y are vectors, of size $[5\times1]$, showing the fractions of perceived emotions, for an original (x) and a modified (y) utterance, respectively. For example, a vector $y=[0.5\ 0.3\ 0.1\ 0\ 0.1]$ was used for an utterance that was perceived as happy, angry, sad, neutral, and other by 50%, 30%, 10%, 0%, and 10% of human raters, respectively.

$$s(x,y)=\frac{x^t y}{\|x\|\|y\|}. \tag{1}$$

In summary, the effects of F0 range modifications were more significant than F0 mean modifications. Stylization modifications were also effective, but only when performed in large semitone scales. They showed that small prosodic variations in the F0 contour shape were more related to the quality of speech and not to its emotional content.

VII. DISCUSSION

In speech synthesis, it is important to study the role of the acoustic parameters in connection with the human perception of prosodic and paralinguistic features (Picard, 1997; Picard *et al.*, 2004; Roach, 2000; Traunmuller, 2005). The results of this paper show that in order to be able to describe F0 variations occurring in emotional speech *sentence, speaker*, and *emotion* factors should be considered. These are the factors that determine how emotional regions (Sec. IV) will be shaped.

Sentence (i.e., linguistic content) should be taken into account because—together with speaker and emotion characteristics—sentence structure [i.e., focus, modality, length (Pell, 2001)] determines how the pitch (and also duration, energy, formant frequencies, and meaning) will be generated.

Instead of including the linguistic content as a factor in the analysis, an alternative approach is to minimize its ef-

fects. One way to do that is by using nonsense sentences (Banziger and Scherer, 2006). This method eliminates the semantic effects, however it also may cause the acoustic parameters (e.g., F0, duration, energy) to be modulated in an unnatural fashion. Therefore, the results may not be easily generalizable to real life utterances.

Probably a better parametrization of emotions can be achieved not by restricting the variance in the different features but by restricting the emotion space itself. This may be achieved by defining more homogeneous emotion categories. One good example is given by Banziger and Scherer (2006), who, in addition to the classic categorical emotion labels, also used activation level differences (Grimm *et al.*, 2007) to describe the emotions. This suggests that in order to better relate the acoustic parameter variation to particular emotions, a hybrid labeling scheme combining categorical (Ekman and Friesen, 1977) and attribute descriptions (Schlosberg, 1954) can be utilized. For example, considering the findings showing that valence, activation, and intensity dimensions are correlated with the acoustic features of emotional speech (Grimm *et al.*, 2007; Schroder *et al.*, 2001), an angry utterance can be described as *angry*, *high (low, medium) activation*, *high (low, medium) valence*, *high (low, medium) intensity*, instead of just *angry*.

Having a better description for emotions can be expected to produce smaller emotional regions. Smaller regions can be expected to overlap less, which in turn will help to better parameterize and differentiate between different emotions in terms of their acoustic features. For example, evaluating angry speech as high or low activation anger would have created two emotional regions instead of one, which theoretically would have helped to better describe how F0 characteristics relate to the angry emotional content. As shown in this paper, the significant overlap between the regions of emotions labeled using the categorical labels indicates that a hybrid labeling technique is necessary for future research in this area.

Considering the small number of sentences and speakers that were analyzed in this study, our future plan is to perform similar studies on a larger dataset. Also, we plan to perform similar analyses for the duration and energy parameters. Increasing the number of sentences, speakers, emotions, and acoustic parameters will provide better information about how the interactions between different factors can be described and parametrized.

VIII. CONCLUSION

The variability of pitch (and therefore F0 contour) is one of the basic characteristics of natural human speech. It has been shown that the same text recorded at different times can have very different F0 characteristics. In this study using an analysis by synthesis method we showed the variability that exists in the F0 mean, range, and shape parameters of emotional speech. The results showed that even significant variation in F0 parameters did not mask the original emotion perception. It was observed that F0 modification caused *sad*, *neutral* or *other* emotion perception to increase, and *angry* or *happy* perception to decrease. The effects of F0 range modifications on emotion perception were more prominent than F0 mean modifications. Also, for F0 range modifications, the drop in the perceived speech quality was less than F0 mean modifications. These results suggest that one should focus on range and not mean modifications during the synthesis of emotional speech. The results were significantly dependent on the speaker and the original utterance characteristics.

In order to model the observed variability in the F0 contour, an emotional regions approach was introduced. The observed emotional regions derived from the data were represented as 2D Gaussian ellipses which showed the limits within which the F0 contour of a given utterance can be modified. In order to model these observed regions, Euclidean emotion regions based on F0 statistics (mean, range, std) were proposed. It was shown that the Euclidean regions can be used as reliable approximations to the high quality Gaussian emotional regions.

The emotional regions concept can be applied to the other acoustic parameters as well. If duration and spectral envelope variations are modeled together with energy and F0 variations, it will be possible to build multidimensional emotional regions for each emotion, which can then be used in emotion conversion and synthesis of speech. This is a task for our future research.

Banziger, T., and Scherer, K. R. (**2006**). "The role of intonation in emotional expressions," Speech Commun. **46**, 252–267.

Boersma, P., and Weenink, D. (**2007**). "Praat: doing phonetics by computer (version 4.5.18) [computer program]," URL http://www.fon.hum.uva.nl/praat/, last retrieved March 9, 2008.

Braun, B., Kochanski, G., Grabe, E., and Rosner, B. S. (**2006**). "Evidence for attractors in English intonation," J. Acoust. Soc. Am. **119**(6), 4006–4015.

Bulut, M., Busso, C., Yildirim, S., Kazemzadeh, A., Lee, C. M., Lee, S., and Narayanan, S. (**2005**). "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Proc. of Eurospeech, Interspeech*, Lisbon, Portugal.

Bulut, M., Narayanan, S., and Syrdal, A. K. (**2002**). "Expressive speech synthesis using a concatenative synthesizer," in *International Conference on Spoken Language Processing*, Denver.

Burkhardt, F., and Sendlmeier, W. F. (**2000**). "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Workshop on Speech and Emotion*, New Castle, Northern Ireland, UK.

Cahn, J. E. (**1990**). "The generation of affect in synthesized speech," J. Am. Voice I/O Soci. **8**, 1–19.

Chu, M., Zhao, Y., and Chang, E. (**2006**). "Modeling stylized invariance and local variability of prosody in text-to-speech synthesis," Speech Commun. **48**, 716–726.

Cowie, R., Cowie, E. D., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (**2001**). "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag. **18**(1), 32–80.

Davitz, J. R. (**1964**). *The Communication of Emotional Meaning* (MCGraw–Hill, New York).

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd ed. (Wiley-Interscience, New York).

Ekman, P., and Friesen, W. V. (**1977**). *Manual for the Facial Action Coding System* (Consulting Psychologist Press, Palo Alto).

Grimm, M., Mower, E., Kroschel, K., and Narayanan, S. (**2007**). "Primitives based estimation and evaluation of emotions in speech," Speech Commun. **49**, 787–800.

Iida, A., Campbell, N., Higuchi, F., and Yasumura, M. (**2003**). "A corpus-based speech synthesis system with emotion," Speech Commun. **40**, 161–187.

Klabbers, E., and van Santen, J. P. H. (**2004**). "Clustering of foot-based pitch contours in expressive speech," in *Proc. of the 5th ISCA Speech Synthesis Workshop*, Pittsburg.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation con-

tour type, voice quality, and F0 range in signaling speaker affect," J. Acoust. Soc. Am. **78**(2) 435–444.
Montero, J. M., Gutierrez-Arriola, J., Colas, J., Enriquez, E., and Pardo, J. M. (**1999**). "Analysis and modeling of emotional speech in Spanish," in *International Congress of Phonetic Sciences*, San Francisco.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.
Murray, I. R., and Arnott, J. L. (**1993**). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Am. **93** 1097–1108.
Pell, M. D. (**2001**). "Influence of emotion and focus location prosody in matched statements and questions," J. Acoust. Soc. Am. **109**(4), 1668–1680.
Picard, R. (**1997**). *Affective Computing* (MIT Press, Cambridge, MA).
Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., and Strohecker, C. (**2004**). "Affective learning – a manifesto," BT Technol. J. **22**(4), 253–269.
Raux, A., and Black, A. (**2003**). "A unit selection approach to F0 modeling and its application to emphasis," in *Proc. of ASRU* (St. Thomas, U.S. Virgin Islands).
Roach, P. (**2000**). "Techniques for the phonetic description of emotional speech," in *ISCA Workshop on Speech and Emotion*, Newcastle. Northern Ireland, UK.
Scherer, K. R. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**(1–2), 227–256.
Schlosberg, H. (**1954**). "Three dimensions of emotion," Psychol. Rev. **61**, 81–88.
Schroder, M. (**2001**). "Emotional speech synthesis - a review," in *Eurospeech* (Aalborg, Denmark).
Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (**2001**). "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Eurospeech* (Aalborg, Denmark).
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierre-humbert, J., and Hirschberg, J. (**1992**). "ToBI: A standard for labeling English prosody," in *International Conference on Spoken Language Processing*, Banff, Alberta, Canada, pp. 867–870.
Taylor, P. (**2000**). "Analysis and synthesis of intonation using the tilt model," J. Acoust. Soc. Am. **107**, 1697–1714.
Traunmuller, H. (**2005**). "Speech considered as modulated voice," URL http://www.ling.su.se/STAFF/hartmut/aktupub.htm, revised manuscript (last retrieved **March 9, 2008**).
Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. (**2004**). "An acoustic study of emotions expressed in speech," in *International Conference on Spoken Language Processing*, Jeju, Korea.

A spectral/temporal method for robust fundamental frequency tracking

Stephen A. Zahorian[a] and Hongbing Hu
Department of Electrical and Computer Engineering, State University of New York at Binghamton, Binghamton, New York 13902, USA

(Received 14 December 2006; revised 2 April 2008; accepted 7 April 2008)

In this paper, a fundamental frequency (F_0) tracking algorithm is presented that is extremely robust for both high quality and telephone speech, at signal to noise ratios ranging from clean speech to very noisy speech. The algorithm is named "YAAPT," for "yet another algorithm for pitch tracking." The algorithm is based on a combination of time domain processing, using the normalized cross correlation, and frequency domain processing. Major steps include processing of the original acoustic signal and a nonlinearly processed version of the signal, the use of a new method for computing a modified autocorrelation function that incorporates information from multiple spectral harmonic peaks, peak picking to select multiple F_0 candidates and associated figures of merit, and extensive use of dynamic programming to find the "best" track among the multiple F_0 candidates. The algorithm was evaluated by using three databases and compared to three other published F_0 tracking algorithms by using both high quality and telephone speech for various noise conditions. For clean speech, the error rates obtained are comparable to those obtained with the best results reported for any other algorithm; for noisy telephone speech, the error rates obtained are lower than those obtained with other methods.
 [DOI: 10.1121/1.2916590]

PACS number(s): 43.72.Ar, 43.72.Dv [DOS] Pages: 4559–4571

I. INTRODUCTION

Numerous studies show the importance of prosody for human speech recognition, but only a few automatic systems actually combine and use fundamental frequency (F_0),[1] with other acoustic features in the recognition process to significantly increase the performance of automatic speech recognition (ASR) systems (Ostendorf and Ross, 1997; Shriberg *et al.*, 1997; Ramana and Srichland, 1996; Wang and Seneff, 2000; Bagshaw *et al.*, 1993). F_0 tracking is especially important for ASR in tonal languages, such as Mandarin speech, for which pitch patterns are phonemically important (Wang and Seneff, 1998; Chang *et al.*, 2000). Other applications for accurate F_0 tracking include devices for speech analysis, transmission, synthesis, speaker recognition, speech articulation training aids for the deaf (Zahorian *et al.*, 1998), and foreign language training. Despite decades of research, automatic F_0 tracking is still not adequate for routine applications in ASR or for scientific speech measurements.

An important consideration for any speech processing algorithm is performance using telephone speech, due to the many applications of ASR in this domain. However, since the fundamental frequency is often weak or missing for telephone speech and the signal is distorted, noisy, and degraded in quality overall, pitch detection for telephone speech is especially difficult (Wang and Seneff, 2000).

A number of pitch detection algorithms have been reported by using time domain and frequency domain methods with varying degrees of accuracy (Talkin, 1995; Liu and Lin, 2001; Boersma and Weenink, 2005; de Cheveigne and Kawahara, 2002; Nakatani and Irino, 2004). Many studies have compared the robustness of pitch tracking for a variety of speech conditions (Rabiner *et al.*, 1976; Mousset *et al.*, 1996; Parsa and Jamieson, 1999). However, robust pitch tracking methods, which can easily be integrated with other speech processing steps in ASR, are not widely available. To make available a public domain algorithm for accurate and robust pitch tracking, the methods presented in this in this paper were developed.

A key component in "yet another algorithm for pitch tracking" (YAAPT) is the normalized cross correlation function (NCCF) as used in the "robust algorithm for pitch tracking" (RAPT) (Talkin, 1995). However, in early pilot testing, the NCCF alone did not reliably give good F_0 tracks, especially for noisy and/or telephone speech. Frequently, the NCCF method alone resulted in gross F_0 errors (especially F_0 doubling for telephone speech) that could easily be spotted by overlaying obtained F_0 tracks with the low frequency part of a spectrogram. YAAPT is the result of efforts to incorporate this observation in a formal algorithm.

In this paper, we describe methods for enhancing and extracting spectrographic information and combining it with F_0 estimates from correlation methods to create a more robust overall F_0 track. Another innovation is to separately compute F_0 candidates from both the original speech signal and a nonlinearly processed version of the signal and then to find the "lowest cost" track among the candidates by using dynamic programming. The basic elements of YAAPT were first given in the work of Kasi and Zahorian (2002) and modifications were described in the work of Zahorian *et al.* (2006). In this paper, we give a comprehensive description of

[a]Author to whom correspondence should be addressed. Tel.: (607) 777-4846. FAX: (607) 777-4464. Electronic mail: zahorian@binghamton.edu.

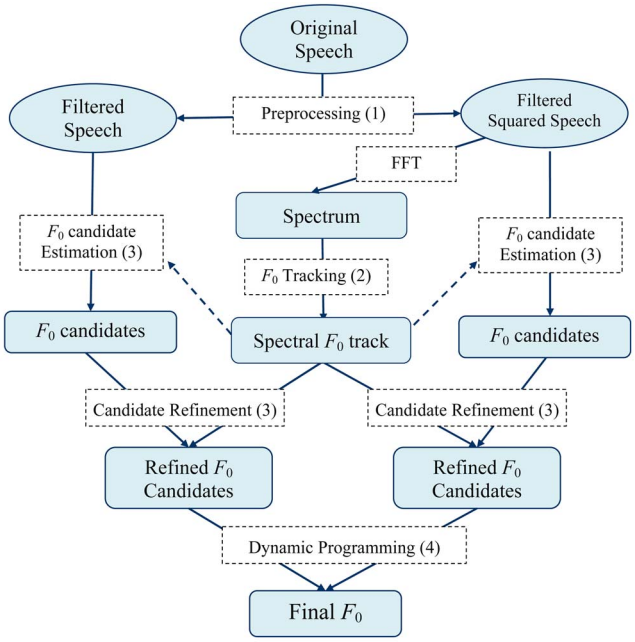


FIG. 1. (Color online) Flow chart of YAAPT. Numbers in parentheses correspond to the steps listed in Sec. II A.

the complete algorithm and extensive formal evaluation results.

II. THE ALGORITHM

A. Algorithm overview

The F_0 tracking algorithm presented in this paper performs F_0 tracking in both the time domain and frequency domain. As summarized in the flow chart in Fig. 1, the algorithm can be loosely divided into four main steps:

(1) Preprocessing: Multiple versions of the signal are created via nonlinear processing (Sec. II B).
(2) F_0 track calculation from the spectrogram of the nonlinearly processed signal: An approximate F_0 track is estimated by using a spectral harmonics correlation (SHC) technique and dynamic programming. The normalized low frequency energy ratio (NLFER) is also computed from the spectrogram as an aid for F_0 tracking (Sec. II C).
(3) F_0 candidate estimation based on the NCCF: Candidates are extracted from both the original and nonlinearly processed signals with further candidate refinement based on the spectral F_0 track estimated in step 2 (Sec. II D).
(4) Final F_0 determination: Dynamic programming is applied to the information from steps 2 and 3 to arrive at a final F_0 track, including voiced/unvoiced decisions (Sec. II E).

The algorithm incorporates several experimentally determined parameters, such as F_0 search ranges, thresholds for peak picking, filter bandwidths, and dynamic programming weights. These parameters are listed in Table I along with values used for experimental results reported in this paper. Similarly, to aid in the explanation of the algorithm and the error measures used for evaluation, primary variables used in this paper are given in Table II. The algorithm is frame based by using overlapping frames with frame lengths and frame spacings as given in Table I.

B. Preprocessing

Preprocessing consists of creating multiple versions of the signal, as shown in the block diagram of Fig. 1. The key idea is to create two versions of the signal: bandpass filtered versions of both the original and nonlinearly processed signals. The bandwidths (50–1500 Hz) and orders (150 points) of the bandpass finite impulse response (FIR) filters were empirically determined by inspection of many signals in time and frequency and also by overall F_0 tracking accuracy. These two signals are then independently processed to obtain F_0 candidates by using the time domain NCCF algorithm, as discussed in Sec. II D.

1. Nonlinear processing

Nonlinear processing of a signal creates sum and difference frequencies, which can be used to partially restore a missing fundamental. Two types of nonlinear processing, the absolute value of the signal and squared value of the signal, were considered. Since experimental evaluations indicated slightly better F_0 tracking accuracy by using the squared value, the squared value was used for the primary experimental results reported in this paper. The general idea of using nonlinearities such as center clipping to emphasize F_0 has long been known (see the work of Hess, 1983 for an extensive discussion) but appears not to be used in most of the pitch detectors developed since about 1990. For example, the pitch detectors YIN (de Cheveigne and Kawahara, 2002) and DASH (Nakatani and Irino, 2004) do not make use of nonlinearities. Of the seven pitch detectors evaluated by Parsa and Jamieson (1999), only one used a nonlinearity (center clipping). Most previous use of nonlinearities in F_0 detection algorithms was aimed at spectral flattening or reducing formant strength, rather than restoring a missing fundamental (for example, the work of Rabiner and Schafer, 1978).

As shown in the work of Zahorian *et al.* (2006), the fundamental frequency (F_0) reappears by squaring the signal in which the fundamental is either very weak or absent, such as telephone speech. The restoration of the fundamental by using the squaring operation is also illustrated by using spectrograms in Fig. 2. The top panel depicts the spectrogram of a studio quality version of a speech signal, for which the fundamental frequency is clearly apparent. The middle panel shows the spectrogram of the telephone version of the same speech sample, for which the fundamental frequency below 200 Hz is largely missing. In contrast, the fundamental frequency is more clearly apparent in the spectrogram of the nonlinearly processed telephone signal shown in the bottom panel. A bandpass filter (50–1500 Hz) was used after the nonlinearity to reduce the magnitude of the dc component. This same effect was observed for many other examples.

TABLE I. Primary parameters used to configure YAAPT. Value 1 numbers are used to minimize gross errors; value 2 numbers are used to minimize big errors.

Parameter	Meaning	Value 1	Value 2
F_0_min	Minimum F_0 searched (Hz)	60	60
F_0_max	Maximum F_0 searched (Hz)	400	400
Frame_length	Length of each analysis frame (ms)	35	25
Frame_space	Spacing between analysis frames (ms)	10	10
FFT_length	FFT length	8192	8192
BP_low	Low frequency of bandpass filter passband (Hz)	50	50
BP_high	High frequency of bandpass filter passband (Hz)	1500	1500
BP_order	Order of bandpass filter	150	150
Max_cand	Maximum number of F_0 candidates per frame	6	6
NLFER_Thresh1	NLFER boundary for voiced/unvoiced decisions, used in spectral F_0 tracking	0.75	0.75
NLFER_Thresh2	Threshold for definitely unvoiced using NLFER	0.0	0.1
N_H	Number of harmonics in SHC calculation	3	3
WL	SHC window length (Hz)	40	40
SHC_thresh	Threshold for SHC peak picking	0.2	0.2
F_0_mid	F_0 doubling/halving decision threshold (Hz)	150	150
NCCF_Thresh1	Threshold for considering a peak in NCCF	0.25	0.25
NCCF_Thresh2	Threshold for terminating search in NCCF	0.85	0.90
Merit_extra	Merit assigned to extra candidates in reducing F_0 doubling and halving logic	0.4	0.4
Merit_pivot	Merit assigned to unvoiced candidates in definitely unvoiced frames	0.99	0.99
W_1	DP weight factor for V-V transitions	0.15	0.15
W_2	DP weight factor for V-UV or VU-V transitions	0.5	0.5
W_3	DP weight factor for UV-UV transitions	100	0.1
W_4	Overall weight factor for local costs relative to transition costs	0.07	0.9

TABLE II. Variable used in YAAPT on for evaluation of F_0 tracking.

Variable	Meaning
s	Speech signal in a frame
S	Magnitude spectrum of speech signal
n	Time sample index within a frame
t	Time in terms of frame index
f	Frequency in Hz
k	Lag index used in NCCF calculations
i,j	Indices uses used for F_0 candidates within a frame
T	Number of signal frames
SHC	Spectral harmonics correlation
F_0_spec	Spectarl F_0 track, all voiced
F_0_avg	Average of spectral F_0 track
F_0_std	Standard deviation of F_0 computed from spectral F_0 track
NLFER	Normilized low frequency energy ratio
merit	Figure of merit for a F_0 candidate, on a scale of 0 to 1
NCCF	Normilized cross correlation function
K_min	Longest lag evaluated for each frame
K_max	Shortest lag evaluated for each frame
F_0_mean	Arthimetic average over all frames of the highest merit nonzero F_0 candidates for each frame
BP	Back pointer array used in dynamic programming
G_err	Error rate based on large errors in all frame where reference indicates voiced speech
B_err	All large error, including those in G_err+errors of the from UV to V

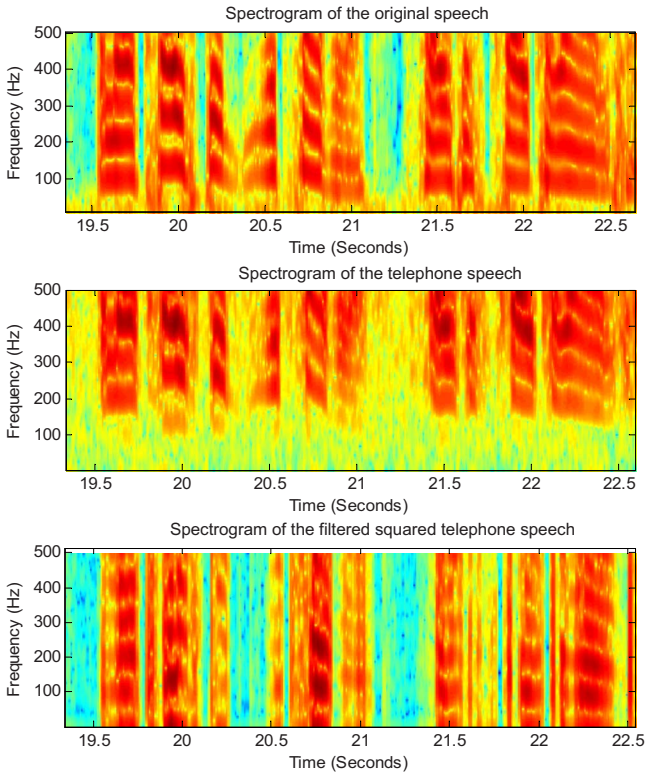


FIG. 2. (Color online) Illustration of the effects of nonlinear processing of the speech signal. The spectrogram of a studio quality speech signal is shown in the top panel, the spectrogram of the telephone version of the signal is shown in the middle panel, and the spectrogram of the squared telephone signal is shown in the bottom panel.

C. Spectrally based F_0 track

One of the key features of YAAPT is the use of spectral information to guide F_0 tracking. Spectral F_0 tracks can be derived by using the spectral peaks which occur at the fundamental frequency and its harmonics. In this paper, it is experimentally shown that the F_0 track obtained from the spectrogram is useful for refining the F_0 candidates estimated from the acoustic waveform, especially in the case of noisy telephone speech. The spectral F_0 track is computed by using the nonlinearly processed speech only.

The initial motivation for exploring the use of spectral F_0 tracks was that the examination of the low frequency parts of spectrograms revealed clear but smoothed F_0 tracks, even for noisy speech. The resolution of the spectral F_0 track depends on the frequency resolution of the spectral analysis, which, in turn, depends on both the frame length and fast Fourier transform (FFT) length used for spectral analysis. For the work reported in this paper, the values of these parameters are listed in Table I. Note that the frame lengths used (25 and 35 ms) are typical of those used in many speech processing applications. The FFT length of 8192 was chosen so that the spectrum was sampled at 2.44 Hz for a sampling rate of 20 kHz, the highest rate used for speech data evaluated in experiments reported in this paper. We hypothesized that this smoothed track could be used to guide the NCCF processing but that the NCCF processing, with a high inherent time resolution of one sampling interval, would give more accurate F_0 estimates. Ultimately, experimental evaluation is needed to check the accuracy of spectral F_0 tracking, versus NCCF-based tracking, versus a combined approach.

1. Spectral harmonics correlation

One way of determining the F_0 from the spectrum is to first locate the spectral peak at the fundamental frequency. This requires that the peak at the fundamental frequency be present and identifiable, which is often not the case, especially for noisy telephone speech. Although the nonlinear processing described in the previous section partially restores the fundamental, additional techniques are needed to obtain an even more noise robust F_0 track. Therefore, a frequency domain autocorrelation type of function, which we call SHC, is used. This method is conceptually similar to the subharmonic summation method (Hermes, 1988) and the discrete logarithmic Fourier transform (Wang and Seneff, 2000), but the details are quite different.

The spectral harmonics correlation is defined to use multiple harmonics as follows:

$$\mathrm{SHC}(t,f) = \sum_{f'=-\mathrm{WL}/2}^{\mathrm{WL}/2} \prod_{r=1}^{\mathrm{NH}+1} S(t, rf+f'),$$

where $S(t,f)$ is the magnitude spectrum for frame t at frequency f, WL is the spectral window length in frequency, and N_H is the number of harmonics. $\mathrm{SHC}(t,f)$ is then amplitude normalized so that the maximum value is 1.0 for each frame. f is a discrete variable with a spacing dependent on FFT length and sampling rate, as mentioned previously.

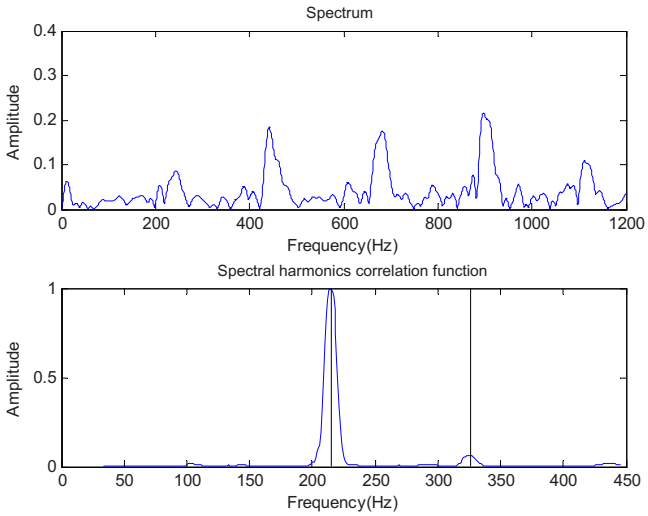


FIG. 3. (Color online) The peaks in the spectral harmonics correlation function. Compared to the small peak at the fundamental frequency of around 220 Hz in the spectrum (top), a very prominent peak is observed in the spectral harmonics correlation function (bottom).

For each frequency f, $\mathrm{SHC}(t,f)$, thus, represents the extent to which the spectrum has high amplitude at integer multiples of that f. The use of a window in frequency, empirically determined to be approximately 40 Hz, makes the calculation less sensitive to noise, while still resulting in prominent peaks for $\mathrm{SHC}(t,f)$ at the fundamental frequency. The calculation is performed only for a limited search range ($F_{0\_}\mathrm{min} \leqslant f \leqslant F_{0\_}\mathrm{max}$, with $F_{0\_}\mathrm{min}$ and $F_{0\_}\mathrm{max}$ values as given in Table I). Experiments were conducted to determine the best value for the number of harmonics. Empirically, it appeared that $N_H=3$ resulted in the most prominent peaks in $\mathrm{SHC}(t,f)$ for voiced speech and, thus, was used for the results given in this paper.

Figure 3 shows the spectrum (top panel) and the spectral harmonics correlation function (bottom panel). Compared to the small peak at the fundamental frequency of around 220 Hz in the spectrum, a very prominent peak is observed in the spectral harmonics correlation function.

2. Normalized low frequency energy ratio

Another primary use of spectral information in YAAPT is as an aid for making voicing decisions. The parameter used is referred to as the NLFER. The sum of spectral samples (the average energy per frame) over the low frequency regions is computed and then divided by the average low frequency energy per frame over the entire signal. In equation form, NLFER is given by

$$\mathrm{NLFER}(t) = \frac{\sum_{f=2\times F_{0\_}\mathrm{min}}^{F_{0\_}\mathrm{max}} S(t,f)}{\frac{1}{T}\sum_{t=1}^{T} \sum_{f=2\times F_{0\_}\mathrm{min}}^{F_{0\_}\mathrm{max}} S(t,f)},$$

where T is the total number of frames, and the frequency range, based on $F_{0\_}\mathrm{min}$ and $F_{0\_}\mathrm{max}$, was empirically chosen to correspond to the expected range of F_0. $S(t,f)$ is the spectrum of the signal for frame t and frequency f. Note that,

with this definition, the average NLFER over all frames of an utterance is 1.0. In general, NLFER is high for voiced frames and low for unvoiced frames and, thus, NLFER is used as information for voiced/unvoiced decision making. In addition, NLFER is used to guide NCCF candidate selection (Sec. II D).

3. Selection of F_0 spectral candidates and spectral F_0 tracking

Beginning with the SHC as described above, F_0 candidates were selected, concatenated, and smoothed by using the following empirically determined method and parameters. Values of the parameters used in experiments throughout this paper are listed in Table I.

(1) The frequency and amplitude of each SHC peak in each frame above threshold SHC_Thresh were selected as spectral F_0 candidates and merits, respectively. For the example shown in Fig. 3, two F_0 candidates were selected. If the merit of the highest merit F_0 candidate is less than SHC_Thresh or if the NLFER is less than NLFER_Thresh1, the frame is considered unvoiced and not considered in the following steps.
(2) To reduce F_0 doubling or halving for voiced frames (a persistent problem with pitch trackers, e.g., the work of Nakatani and Irino, 2004), an additional candidate is inserted at half the frequency of the highest merit candidate if all the candidates are above the F_0 doubling/having decision threshold F_0_mid. Similarly, if all candidates are below F_0_mid, an additional F_0 candidate is inserted at twice the frequency of the highest ranking candidate. The merit of these inserted candidates is set at the midrange value Merit_extra.
(3) All estimated voiced segments are concatenated and viewed as one continuous voiced segment. For each frame in this concatenated segment, one additional F_0 candidate is inserted as the median smoothed (seven point smoothing window) value of the highest merit candidate for each frame. This additional candidate is assigned a merit as Merit_extra.
(4) Dynamic programming, as described in Sec. II E, is used to select the lowest cost path among the candidates. This use of dynamic programming is the same as that used for final F_0 tracking, with the constants as listed in Table I. However, the transition costs involving unvoiced speech segments were relevant, since no unvoiced segments were considered.
(5) The F_0 track is then lengthened to its original length by using linear interpolation to span the sections estimated to be unvoiced from step 1 above.
(6) The result of this whole process is a smoothed F_0 track (F_0_spec) with every frame considered to be voiced. Experiments, reported in a later section, indicate that the spectral F_0 track is quite good but not quite as good as the one obtained by combining the spectral and NCCF tracks introduced in the next section.

D. F_0 candidate estimation from NCCF

F_0 candidates are computed from both the original and the nonlinearly processed signals by using a modified autocorrelation processing in the time domain. The basic idea of correlation based F_0 estimation is that the correlation signal has a peak of large magnitude at a lag corresponding to the period of F_0. This section explains the modified version used for YAAPT: the NCCF (Talkin, 1995), as well as the selection of NCCF F_0 candidates.

1. Normalized cross correlation function

The NCCF is defined as follows:[2] Given a frame of sampled speech $s(n)$, $0 \leq n \leq N-1$,

$$\mathrm{NCCF}(k) = \frac{1}{\sqrt{e_0 e_k}} \sum_{n=0}^{N-K\_\max} s(n)s(n+k),$$

where

$$e_0 = \sum_{n=0}^{n=N-K\_\max} s^2(n), \quad e_k = \sum_{n=k}^{n=k+N-K\_\max} s^2(n),$$

$$K\_\min \leq k \leq K\_\max.$$

In the equation, N is the frame length in samples and K_min and K_max are the lag values needed to accommodate the F_0 search range as described below. As with an autocorrelation, the NCCF is self-normalized for a range of $[-1,1]$ and periodic signals result in NCCF values of 1 at lag values equal to integer multiples of the period. As previously reported by Talkin (1995), the NCCF is better suited for F_0 detection than the "standard" autocorrelation function, as the peaks are better defined and less affected by rapid variations in signal amplitude. The only apparent disadvantage is the increase in computational complexity. Nevertheless, it is still possible for the largest peak to occur at double or half the correct lag value or simply at an "incorrect" value. Thus, the additional processing described below is used.

2. Selection of F_0 candidates and merits from NCCF

The following empirically determined procedure was used to create a collection of F_0 candidates and merits from the NCCF peaks:

(1) The spectral F_0 track (F_0_spec) was used to refine the search F_0 range for frame t as follows:

$$F_0\_\mathrm{search}\_\min(t) = \max[F_0\_\mathrm{spec}(t) - 2 \times \mathrm{std}, F_0\_\min],$$

$$F_0\_\mathrm{search}\_\max(t) = \min[F_0\_\mathrm{spec}(t) + 2 \times F_0\_\mathrm{std}, F_{0,}\max],$$

where F_0_std is the standard deviation of F_0 values appearing in the estimated spectral F_0 track.
(2) For each frame, all peaks found over the search range of F_0_search_min(t) to F_0_search_max(t) are located. To be a peak, a NCCF value must be at least NCCF_Thresh1

in amplitude and larger than the two values on either side of the point under consideration. If more than Max_cand/2 peaks are found, only the Max_cand/2 peaks with the highest values of NCCF are retained. Additionally, with searching beginning at a lag value corresponding to F_0_search_max(t) (shortest lag), if a peak is found with NCCF value greater than NCCF_Thresh2, peak searching is terminated. This step was empirically found to reduce F_0 halving instances. This process is repeated for all frames and for both the original and nonlinearly processed versions of the signal and the results combined for each frame. At the end of this step, up to Max_cand, F_0 candidates are found for each frame of the signal.

(3) All peaks found in step 2 are assigned a preliminary merit value equal to the amplitude of the peak. If fewer than Max_cand F_0 candidates are found in step 2, unvoiced candidates (F_0=0) are inserted, each with merit =[1−(merit of the nonzero F_0 candidate with the highest merit for that frame)]. For those frames where no peaks are found in step 2, the frame is preliminarily considered to be unvoiced; all F_0 candidates are set to 0 with merit=Merit_pivot.

(4) The initial merit values from step 3 are modified by using the spectral F_0 track, so as to increase the merits of NCCF F_0 candidates close to the spectral track. First, F_0_avg and F_0_std are computed as the average and standard deviation of F_0 from the spectral F_0 track (F_0_spec). Then, for candidates whose values are less than $5\times F_0$_std from the spectral F_0 value of that frame, the merit is changed as follows:

$$\text{merit}'(t,j) = \text{merit}(t,j) + 0.2 \times [1 - |F_0(t,j) - F_0\_\text{spec}(t)|/F_0\_\text{avg}],$$

where merit′ is the updated merit. For all other candidates, the merit is unchanged (merit′=merit). Note that j is the candidate index and t the frame index.

(5) For all frames with NLFER≤NLFER_Thresh2, the frame is considered to be definitely unvoiced and all F_0 candidates are adjusted to 0 (unvoiced) and have merits set to Merit_pivot. For all frames with NLFER >NLFER_Thresh2, the candidates are inspected to ensure that there is at least one nonzero F_0 estimate as well as an unvoiced candidate (F_0=0). If there initially was no nonzero F_0 candidate, the spectral F_0 is used as a candidate with a merit equal to half of the NLFER amplitude, if NLFER<2 or 1 if NLFER≥2. If there was initially no unvoiced ($F_0\neq 0$) candidate, the lowest merit F_0 candidate is replaced by the F_0=0 candidate, with merit=[1−(merit of the F_0 candidate with the highest meritfor that frame)], as in step 3.

E. Final F_0 determination with dynamic programming

After the processing steps mentioned above, a F_0 candidate matrix and associated merit matrix are created over the interval of a speech utterance. The F_0 candidates and the merits are used to compute transition costs, associated with every pair of F_0 candidates in successive frames, and local costs, for each candidate for each frame. In the remainder of this section, the calculation of these costs is described and the dynamic programming algorithm is summarized.

Three cases are considered for transition costs of successive F_0 candidates as follows:

(1) For each pair of successive voiced candidates (i.e., nonzero F_0 candidates),

$$\text{Cost}_{\text{transition}}(t-1,j{:}t,i) = W_1 \times |F_0(t,i) - F_0(t-1,j)|/F_0\_\text{mean}.$$

F_0_mean is the arithmetic average over all frames of the highest merit nonzero F_0 candidates for each frame. Note that the cost is for transitioning from candidate j in frame $t-1$ to candidate i in frame t.

(2) For each pair of successive candidates, only one of which is voiced,

$$\text{Cost}_{\text{transition}}(t-1,j{:}t,i) = W_2 \times [1 - \text{VCost}(t)],$$

where

$$\text{VCost}(t) = \min[1, |\text{NLFER}(t) - \text{NLFER}(t-1)|].$$

(3) For each pair of successive candidates, both of which are unvoiced,

$$\text{Cost}_{\text{transition}}(t-1,j{:}t,i) = W_3.$$

Values of W_1, W_2, and W_3 used in the experiments are given in Table I. The value of W_3 can be increased to a large value (e.g., 100) to force the dynamic programming routine to select all voiced candidates except for frames considered definitely unvoiced.

The local cost for each F_0 candidate is computed in the straightforward way,

$$\text{Cost}_{\text{local}}(t,i) = W_4 \times [1 - \text{merit}'(t,i)].$$

Thus, F_0 candidates with high merit have low local cost. W_4 is used to control the relative contribution of local costs to transition costs in the overall cost. The dynamic programming is a standard Viterbi decoding method, as described in the works of Rabiner and Juang (1993) and Duda *et al.*, (2000). The program is summarized here for completeness. Initialize:

$$\text{Cost}(1,i) = \text{Cost}_{\text{local}}(1,i), \quad 1 \leq i \leq \text{Max\_cand}.$$

Iterate: for $2 \leq t \leq T$,

for $1 \leq i \leq$ Max_cand,

$$\text{Cost}(t,i) = \text{MIN\_j}\ \{\text{Cost}(t-1,j) \times \text{Cost}_{\text{transition}}(t-1,j{:}t,i) + \text{Cost}_{\text{local}}(t,i)\},$$

$$\text{BP}(t,i) = \text{ARGMIN\_j}\ \{\text{Cost}(t-1,j) \times \text{Cost}_{\text{transition}}(t-1,j{:}t,i) + \text{Cost}_{\text{local}}(t,i)\}.$$

Max_cand and T are as, respectively, defined in Tables I and II. At the completion of the iterations over t, beginning with ARGMIN_i [Cost(T,i)], the BP array is traced back to

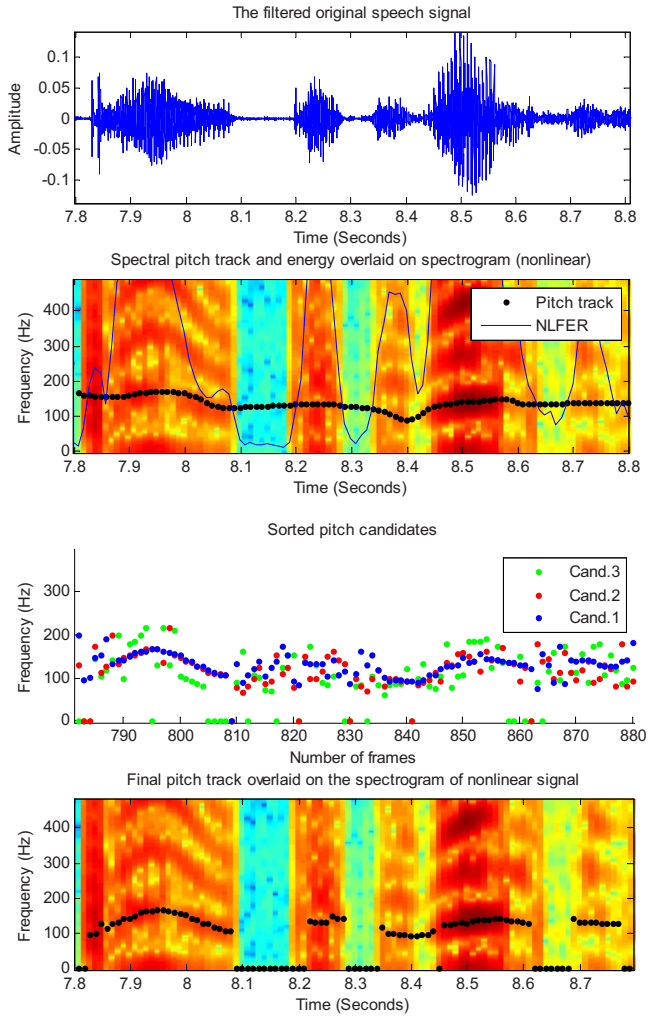


FIG. 4. (Color online) The first panel shows the time domain acoustic signal, the second panel shows the spectrogram of the signal with the low frequency energy ratio and spectral F_0 track overlaid on it, and the third panel shows multiple candidates chosen from the NCCF. The fourth panel shows the final F_0 track.

yield the overall lowest cost F_0 track. An illustration of the overall F_0 tracking algorithm is shown by the four panels in Fig. 4.

III. EXPERIMENTAL EVALUATION

A. Database description

In the F_0 estimation evaluation, performance comparison of different algorithms based on the same database are of great importance to allow better comparisons among the algorithms. Fortunately, common databases are freely provided for comparative pitch study by different research laboratories. For these databases, the laryngograph signal and/or a reference pitch are usually provided.

In our evaluation, we used the following three databases to evaluate various aspects of the YAAPT algorithm and to compare it with other algorithms:

(1) The Keele pitch database (DB1): This database consists of ten phonetically balanced sentences spoken by five male and five female English speakers (Plante *et al.*, 1995). Speech signals are studio quality speech sampled at 20 kHz. The total duration of the database is approximately 6 min. The laryngograph and the manually checked reference pitch are also provided in the database. The telephone version of the Keele database, formed by transmitting the studio quality speech signals through telephone lines and resampling at 8 kHz, was also used in experiments reported in this paper.

(2) The fundamental frequency determination algorithm evaluation database (DB2): This database is provided by the University of Edinburgh, UK (Bagshaw *et al.*, 1993). Fifty sentences are spoken by one male and one female English speaker. The total duration of the 100 sentences is about 7 min. The signal was sampled at a 20 kHz rate by using 16-bit quantization. The laryngograph and the manually checked reference pitch are also included.

(3) The Japanese database (DB3): This database consists of 30 utterances by 14 male and 14 female speakers (total of 840 utterances, total durations of 40 min, 16 kHz sampling, and 16-bit quantization). For experiments reported in this paper, 100 utterances were used, with approximately half of male speakers and half of female speakers. For this database, the reference used is the same one used in the works of de Cheveigne and Kawahara (2002) and Nakatani and Irino (2004).

B. Evaluation method

As the ground truth for pitch evaluation, the supplied reference pitches were used. These reference pitches were computed from the laryngograph signal and manually corrected. Although these references should be very accurate, by visual inspection of pitch tracks, they still appeared to have some problems with F_0 halving. Consequently, in previous studies, these references were not always used, but instead an algorithm-specific reference was computed from the laryngograph signal (for example, the work of Nakatani and Irino, 2004). Nevertheless, for experiments reported in this paper, supplied references were used for all results.[3]

To test the robustness of the algorithm, additive background noise was also used in the evaluation. The background noise consisted of two kinds of noise: white noise and babble noise. The signal-to-noise ratio (SNR) in terms of the average power ranges from infinity (that is no additional added noise or "clean") to 0 dB. The average power was calculated only from the frames whose power was more than 1/30 of the entire signal's average power (as per the work of Nakatani and Irino, 2004). Evaluations were made with two kinds of telephone speech: the actual telephone speech available in DB1 and simulated telephone speech for all three databases by using a SRAEN (300–3400 Hz 150th order FIR bandpass) filter.

C. Error measures

Errors for F_0 tracking include major errors [unvoiced (UV) frames incorrectly labeled as voiced (V), V frames incorrectly labeled as UV, and large errors in F_0 accuracy for voiced frames such as F_0 doubling or halving] and smaller errors in F_0 tracking accuracy. Of the many error measures

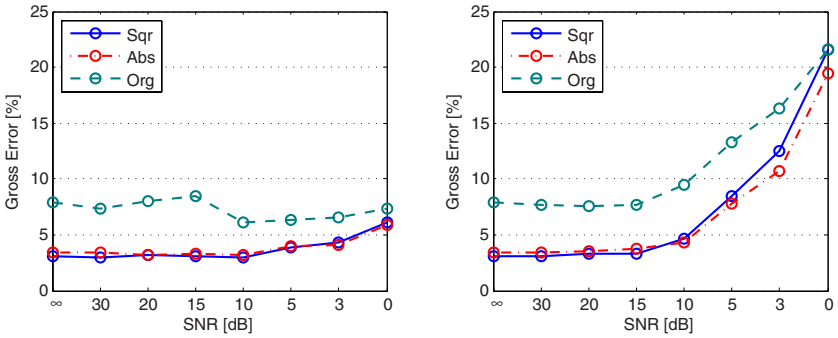


FIG. 5. (Color online) The effect of nonlinear processing for DB1 studio quality speech at various SNR white noises (left) and babble noises (right).

that can be used to quantify F_0 tracking accuracy, we used the following measures to evaluate the tracking method reported in this paper:

(1) Gross error (G_err): This is computed as the percentage of voiced frames, such that the pitch estimate of the tracker significantly deviates (20% is generally used) from the pitch estimate of the reference. The measure is based on all frames for which the reference pitch is voiced, regardless of whether the estimate is voiced or unvoiced. Thus, G_err includes V to UV errors as well as large errors in F_0 values,

$$\mathrm{G\_err} = \frac{1}{\mathrm{NVF}} \sum_{t=1}^{\mathrm{NVF}} \delta[F_0^{\mathrm{ref}}(t), F_0^{\mathrm{est}}(t)], \quad \delta(F_0^{\mathrm{ref}}, F_0^{\mathrm{est}}) = \begin{cases} 1 & |(F_0^{\mathrm{ref}} - F_0^{\mathrm{est}})/F_0^{\mathrm{ref}}| > 0.2 \\ 0 & \text{otherwise,} \end{cases}$$

where F_0^{ref} is reference F_0, F_0^{est} is estimated F_0, and NVF is the number of voiced reference frames.

(2) Big error (B_err): This error is equal to the number of voiced frames with a large error in F_0, plus the number of unvoiced frames erroneously labeled as voiced frames (UV_V_N), divided by the total number of frames T. In equation form,

$$\mathrm{B\_err} = (\mathrm{NVF} \times \mathrm{G\_err} + \mathrm{UV\_V\_N})/T.$$

Both G_err and B_err are expressed as percentages in experiments.

In the following sections of this paper, experimental results are first given to illustrate the effects of nonlinear processing and the performance of various components of YAAPT. These results are followed by a section with experiments and results based on the complete algorithm, including a comparison with three other algorithms (PRAAT, RAPT, and YIN) and a comparison with results reported in the literature using the same databases and the same error measures.

D. The effect of nonlinear processing

As described in Sec. II B, nonlinear processing could be either the absolute value or squared value, or a variety of other nonlinearities (Hess, 1983), to help restore the missing fundamental in the telephone speech. To evaluate the benefits of using this nonlinear processing, we computed the gross errors for three conditions: using the original signal only (no nonlinear processing), using absolute values as the nonlinear processing, and using the squared value as the nonlinear processing.

Figures 5 and 6, respectively, show the gross errors for studio quality speech and for telephone speech for various noise conditions using DB1. Error performance is very similar using either the absolute value or squaring operation. The nonlinear processing is quite beneficial for nearly all conditions tested, except for very high levels of additive babble noise. The most surprising result is that the nonlinear processing improves error performance even for noise-free studio quality speech.[4]

E. Evaluation of individual components of the algorithm

YAAPT computes the F_0 track by using a combination of both spectral and temporal (NCCF) information. The spectral F_0 track is used to determine the F_0 search range for the NCCF calculations and to modify the merits of the temporal F_0 candidates. It could be questioned whether or not both the temporal and spectral tracks are needed and the

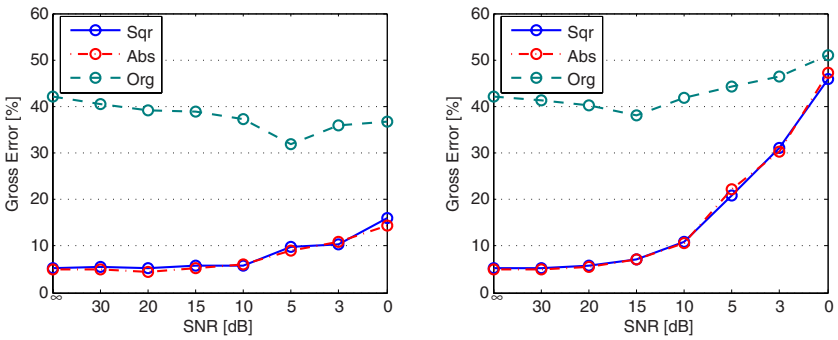


FIG. 6. (Color online) The effect of nonlinear processing for simulated DB1 telephone speech at various SNR white noises (left) and babble noises (right).

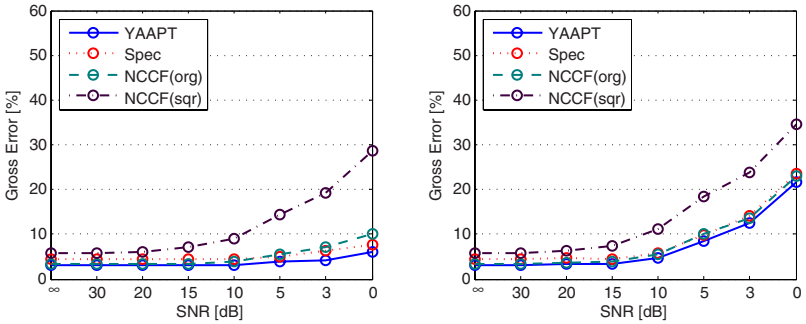


FIG. 7. (Color online) Performance based on individual components of YAAPT for DB1 studio quality speech at various SNR white noises (left) and babble noises (right).

extent to which each of these sources of information contributes to the accuracy of the F_0 tracking. Additionally, it might also be questioned whether or not the nonlinear processing is needed for the time domain F_0 candidates, especially for the case of studio quality speech. Therefore, F_0 tracking was computed by using four different approaches:

(1) using the NCCF candidates from the original signal only, with the final track determined by dynamic programming,
(2) using the NCCF candidates from the squared signal only, with the final track determined by dynamic programming,
(3) using the spectral F_0 track only, and
(4) using the entire YAAPT algorithm, combining both the temporal and spectral information.

Evaluations were conducted for each of these four methods by using both studio quality and telephone speech, and both added white and babble noises. Results are shown in Figs. 7 and 8. The combination of the temporal and spectral tracks results in better performance than using any individual component, illustrating the benefits of using both temporal and spectral information. As shown in Figs. 7 and 8, the gross error results based on the NCCF of the original signal is better than those obtained from the squared signal. For both the studio quality and telephone speech cases, the spectral F_0 tracking obtained by using the squared signal gives a very low gross error. These results, thus, show that the squared signal plays an important role in improving the performance of the entire algorithm for telephone speech.

F. Overall results

The overall evaluation of YAAPT is reported in this section, as well as a comparison with the PRAAT (Boersma and Weenink, 2005), RAPT (Talkin, 1995), and YIN (de Cheveigne and Kawahara, 2002) pitch tracking methods. The autocorrelation method described in the work of Boersma (1993) was used in PRAAT, as opposed to the cross-correlation method, as the autocorrelation option gave better results in pilot experiments. The RAPT tracker used is the MATLAB version of the Talkin algorithm. The RAPT pitch tracker was previously implemented commercially in XWAVES software and is considered to be a robust pitch tracker. More recently, the YIN tracker, which uses a modified version of the autocorrelation method, has been shown to give very high accuracy for pitch tracking for clean speech and music. The DASH and REPS trackers (Nakatani and Irino, 2004) are reported to be the most noise robust trackers developed for telephone speech.[5]

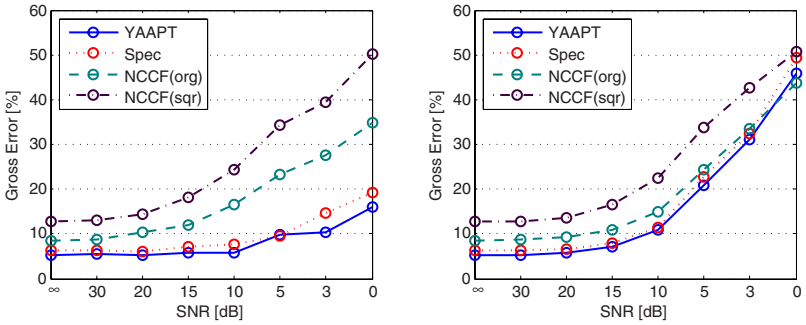


FIG. 8. (Color online) Performance based on individual components of YAAPT for DB1 telephone speech at various SNR white noises (left) and babble noises (right).

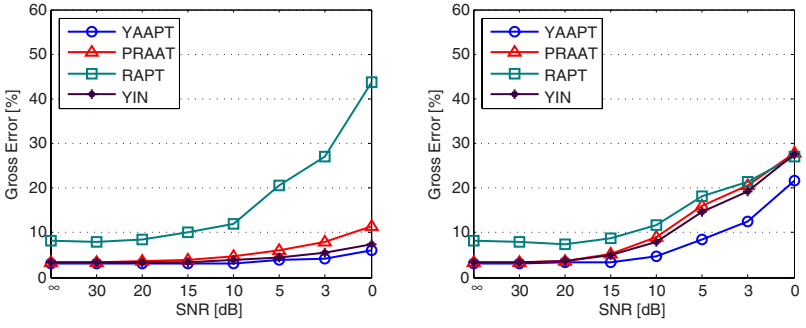


FIG. 9. (Color online) Gross errors for DB1 studio quality speech at various SNR white noises (left) and babble noises (right).

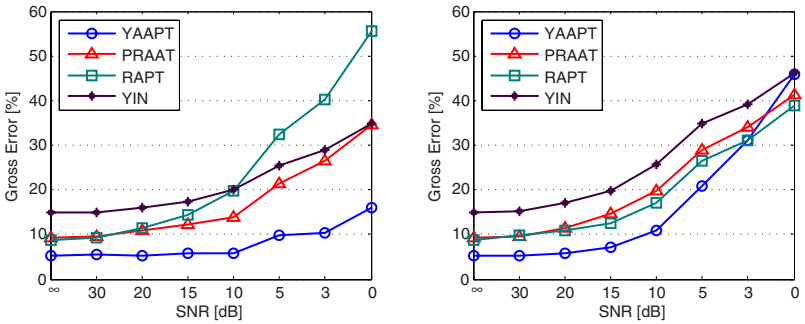


FIG. 10. (Color online) Gross errors for DB1 telephone speech at various SNR white noises (left) and babble noises (right).

1. Gross error results

Figure 9 depicts the gross F_0 errors of the studio quality speech for DB1 in the presence of additive white noise and babble noise, for the YAAPT, PRAAT, RAPT, and YIN pitch trackers. To obtain these results, the parameter values (e.g., Table I, column 1, for YAAPT) were adjusted so that nearly all frames were estimated to be voiced. Similarly, for the three control trackers, parameters were adjusted to minimize gross errors. Note that the gross F_0 errors are based on all large errors (including voiced to unvoiced errors) that a tracker makes for frames that are voiced in the reference. Figure 10 gives results of the telephone speech for the same conditions.

These results show that YAAPT has better gross error performance than the other methods, for all conditions at nearly all SNRs. The performance difference is greatest for telephone speech. The error performance of YAAPT is poor only for telephone speech with very high levels of additive babble noise (SNR$\leqslant$3 dB). It should be noted that this is very noisy speech; in informal listening tests, this speech was nearly unintelligible, with intermittent sections so noisy that the pitch was difficult to discern. Based on an inspection of F_0 candidates and the final F_0 track for YAAPT, it appeared that the final dynamic programming was unable to reliably choose the "correct" candidate for this very noisy condition.

In Table III, gross voicing error values for all three databases are listed for studio quality speech and simulated telephone speech. In this table, as well as other tables, results are given for clean speech, white noise at a 5 dB SNR (W-5), and babble noise at a 5 dB SNR (B-5). For both studio quality and telephone speech, with either no added noise or the W-5 condition, YAAPT has the best performance, sometimes dramatically better. However, for the B-5 telephone condition, YAAPT performance is sometimes worse (depending on database) than that of the other trackers. All four trackers are subject to large increases in error rates as signal degradation increases beyond a certain point.

2. Big error results

For some applications of F_0 tracking, both errors in voicing decisions and large errors in F_0 during voiced sections should be minimized. Thus, big error (B_err), as defined in Sec. III C and which includes both of these types of errors, is the most relevant measure of performance. The big error performance of YAAPT is compared only to that of the RAPT and PRAAT trackers, since the YIN tracker assumes that all frames are voiced. For all trackers, parameter settings were used that are intended to give the best accuracy with respect to big error (e.g., column 2 in Table I parameter values for YAAPT). Big error results, for studio and telephone speech, are shown in Fig. 11 as a function of SNR for added white noise. YAAPT performs better than PRAAT and RAPT for all conditions shown. The minimum big error performance of about 6% for studio quality speech is given by YAAPT. However, since most of the low frequency components are missing, higher big errors are obtained with tele-

TABLE III. Gross errors (%) for studio and simulated telephone speech for various noise conditions.

		Studio			Simulated telephone		
Database	Method	Clean	W-5	B-5	Clean	W-5	B-5
DB1	YAAPT	3.08	3.77	8.48	4.23	6.21	28.66
	PRAAT	3.35	6.91	15.98	9.91	15.72	32.56
	RAPT	8.24	21.33	18.04	9.5	18.21	29.09
	YIN	3.23	4.85	14.74	20.9	25.96	37.4
DB2	YAAPT	3.78	3.81	9.6	4.93	7.8	37.24
	PRAAT	5.96	10.5	19.61	7.81	19.03	32.53
	RAPT	14.08	30.63	23.76	14.63	30.68	30.43
	YIN	3.79	5.36	15.12	13.42	20.13	31.12
DB3	YAAPT	1.16	1.69	4.3	3.63	5.55	21.76
	PRAAT	2.02	3.97	14.75	5.35	11.93	29.7
	RAPT	5.36	12.87	13.78	3.84	13.97	24.88
	YIN	1.38	2.24	12.03	13.83	19.38	32.64

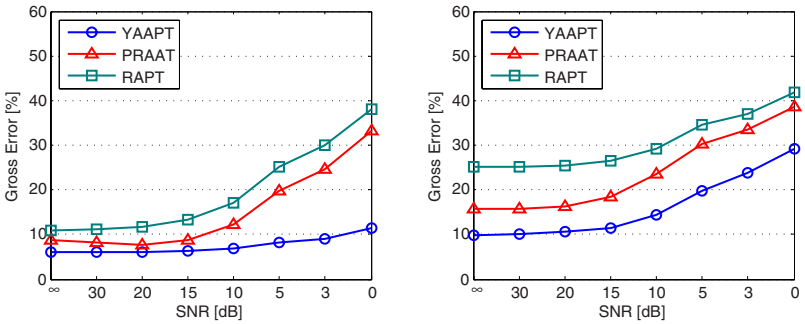


FIG. 11. (Color online) Big error for DB1 studio quality (left) and telephone (right) speech at various SNR white noises.

phone speech. In addition, high noise levels greatly affect the performance of the voiced/unvoiced determination, which, in turn, increase the big error.

A tabular presentation of big error performance is given for YAAPT, PRAAT, and RAPT in Table IV, for studio and simulated telephone speech, for the same noise conditions as used for the gross error results given in Table III. For all cases and all trackers, errors in voicing decisions (UV to V and V to UV) formed the largest portion of the big errors. For these results, YAAPT has the lowest error among the trackers for studio speech but not for the simulated telephone speech. However, as indicated by the results shown in Fig. 11, YAAPT does have the best big error performance for actual telephone speech.

3. Results with telephone speech

To examine results in more detail for real telephone speech, both gross error results and big error results are given in Table V, for the same noise conditions as used in Tables III and IV. YAAPT is compared to PRAAT, RAPT, and YIN for gross errors but to only PRAAT and RAPT for big errors. YAAPT has lower gross and big errors than PRAAT, RAPT, and YIN for the no added noise and W-5 conditions; for big errors in the B-5 condition, YAAPT has similar (poor) performance to PRAAT and RAPT.

G. Comparison of results with other published results

Selected results for gross errors obtained with YAAPT and YIN in this study are tabulated in Table VI along with previously reported results for YIN, DASH, and REPS and for all three databases used in this study. Although test conditions and parameter settings are intended to be identical, clearly, there are differences since the results obtained with YIN in this study and those obtained with YIN in these previous studies are significantly different. There may have been some differences in the reference pitch used, method for simulating telephone speech, methods for adding noise, parameter settings, or even versions of the code used. Nevertheless, the conditions are reasonably close and general comparisons can be made. Overall, the previously reported gross error results for DASH are the lowest. The previously reported gross error rates for YIN are very low for clean studio speech and very high for noisy telephone speech, as compared to the two other trackers.

TABLE IV. Big errors (%) for studio and simulated telephone speech for various noise conditions.

		Studio			Simulated telephone		
Database	Method	Clean	W-5	B-5	Clean	W-5	B-5
DB1	YAAPT	6.09	7.99	22.44	14.07	16.89	44.39
	PRAAT	8.64	19.87	34.9	12.83	20.12	46.9
	RAPT	10.99	25.35	34.28	11.69	21.81	45.6
DB2	YAAPT	7.66	8.41	26.75	8.39	12.72	47.96
	PRAAT	10.43	16.28	37.96	9.98	16.01	45.21
	RAPT	13.54	21.42	35.81	14.44	19.05	39.75
DB3	YAAPT	4.98	7.46	15.05	12.23	16.59	35.54
	PRAAT	8.42	18.9	31.55	9.5	21.99	43.92
	RAPT	11.72	25.26	32.09	11.44	25.04	40.3

TABLE V. Gross and big errors for telephone speech using DB1 for various noise conditions.

		Gross errors (%)			Big errors (%)		
Database	Method	Clean	W-5	B-5	Clean	W-5	B-5
DB1	YAAPT	5.3	8.86	20.93	9.93	19.83	46.22
	PRAAT	9.33	21.5	28.96	15.78	30.38	44.97
	RAPT	8.8	34.29	26.53	25.09	34.3	38.07
	YIN	14.85	25.63	34.95	⋯	⋯	⋯

TABLE VI. Comparison of gross errors for YAAPT, YIN, DASH, and REPS. The "*" indicates the results reported by Nakatani and Irino (2004).

		Studio			Simulated telephone		
Database	Method	Clean	W-5	B-5	Clean	W-5	B-5
DB1	YAAPT	3.08	3.77	8.48	4.23	6.21	28.66
	YIN	3.23	4.85	14.74	20.9	25.96	37.4
	DASH*	2.81	3.32	16.5	3.73	4.15	20.0
	REPS*	2.68	2.98	12.3	6.91	8.49	26.2
	YIN*	2.57	7.22	31.0	7.55	14.6	40.0
DB2	YAAPT	3.78	3.81	9.6	4.93	7.8	37.24
	YIN	3.79	5.36	15.12	13.42	20.13	31.12
	DASH*	0.42	1.34	14.6	0.63	0.97	15.1
	REPS*	0.68	1.05	11.1	2.11	3.25	18.9
	YIN*	1.3	4.38	33.5	5.53	10.3	35
DB3	YAAPT	1.16	1.69	4.3	3.63	5.55	21.76
	YIN	1.38	2.24	12.03	13.83	19.38	32.64
	DASH*	0.3	0.43	8.82	0.73	1.55	14.1
	REPS*	0.26	0.29	4.9	2.11	2.67	12.7
	YIN*	0.44	2.1	28.4	3.27	7.32	34.6

No similar comparisons can be given for big errors, since big error results are not reported for these databases. The focus for the YIN, DASH, and REPS trackers was tracking for the purpose of prosodic modeling, thus eliminating the need for voiced/unvoiced decision making. Consequently, results were only reported for gross errors, the large errors which occur in the clearly voiced (as per the reference) sections of speech.

IV. CONCLUSION

In this paper, a new F_0 tracking algorithm has been developed which combines multiple information sources to enable accurate robust F_0 tracking. The multiple information sources include F_0 candidates selected from the normalized cross correlation of both the original and squared signals and smoothed F_0 tracks obtained from spectral information. Although methods similar to all the individual components of YAAPT have been used to some extent in previous F_0 trackers, these components have been implemented and integrated in a unique fashion in the current algorithm. The resulting information sources are combined by using experimentally determined heuristics and dynamic programming to create a noise robust F_0 tracker. An analysis of errors indicates that YAAPT compares favorably with other reported pitch tracking methods, especially for moderately noisy telephone speech. The entire YAAPT algorithm is available from Zahorian as MATLAB functions.

Except for different settings used to evaluate gross error and big error, all parameter values used in the results reported in this paper were the same for all conditions tested. These conditions span three databases for two languages (English and Japanese), both studio quality and telephone speech, and noise conditions ranging from no added noise to 0 dB SNR with added white and babble noises. Over this wide range of conditions, F_0 tracking accuracy with YAAPT is better, or at least comparable, to the best accuracy achievable with other reported trackers.

From a computational perspective, YAAPT is quite demanding due to the variety of signal processing approaches used and then combined in the complete algorithm. For applications such as prosodic modeling where the voicing decision may not be needed, a very good voiced-only pitch track can be obtained by using the spectral pitch track method described in this paper, with greatly reduced computational overhead and only slight degradation in performance.

ACKNOWLEDGMENTS

This work was partially supported by JWFC 900 and NSF Grant No. BES-9977260. We would like to thank A. de Cheveigne, T. Nakatani, and T. Nearey for access to databases and control F_0 trackers. We also thank the anonymous reviewers for their detailed and helpful comments.

[1]In this paper, we use the terms F_0 and pitch interchangeably, although technically, pitch is a perceptual attribute, whereas F_0 is an acoustic property, generally considered to be the primary cue for pitch.

[2]This implementation of NCCF is slightly different from the one used in the second pass of RAPT, in that RAPT includes a small positive constant inside the radical, to reduce the magnitude of peaks in low amplitude regions of speech. Based on pilot testing, this constant did not improve F_0 tracking accuracy for YAAPT, so it was not used.

[3]Based on experimental testing, the patterns of error results obtained with supplied references and algorithm generated ones are very similar, except that the errors obtained with algorithm generated references are usually 1%–2% lower than those obtained with supplied references. This difference in performance is, thus, significant for clean studio speech but not significant for noisy telephone speech.

[4]It is quite likely that some modifications and changing of parameter values would have resulted in better performance of YAAPT without nonlinear processing, for studio speech. However, the experimental results shown were obtained without changing the algorithm or parameter values, except for changes in the nonlinear signal processing.

[5]The DASH and REPS trackers are proprietary code and were not available for comparison testing.

Bagshaw, P. C., Miller, S. M., and Jack, M. A. (**1993**). "Enhanced pitch tracking and the processing of the F_0 contours for computer aided intonation teaching," in *Proceedings of EUROSPEECH*, Berlin, Germany, 1003–1006. The database used in this paper is also available at http://www.cstr.ed.ac.uk/research/projects/fda (last accessed 4/1/2008).

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, Vol. **17**, 97–110.

Boersma, P., and Weenink, D. (**2005**). "PRAAT: Doing phonetics by computer," version 4.3.14, Institute of Phonetic Sciences, http://www.praat.org (retrieved 5/26/2005).

Chang, E., Zhou, J., Di, S., Huang, C., and Lee, K. F. (**2000**). "Large vocabulary mandarin speech recognition with different approaches in modeling tones," in *Proceedings of the Sixth International Conference of Spoken Language Processing*, Interspeech 2000—ICSLP, Beijing, China, 983–986.

de Cheveigne, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.

Duda, R., Hart, P., and Stork, D. (**2000**). *Pattern Classification* (Wiley-Interscience, New York), pp. 128–137.

Hermes, D. J. (**1988**). "Measurement of pitch by subharmonic summation," J. Acoust. Soc. Am. **83**, 257–264.

Hess, W. (**1983**). *Pitch Determination of Speech Signals* (Springer-Verlag, Berlin), pp. 310–355.

Kasi, K., and Zahorian, S. A. (**2002**). "Yet another algorithm for pitch tracking," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, Orlando, Florida, 361–364.

Liu, D. J., and Lin, C. T. (**2001**). "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," IEEE Trans. Speech Audio Process. **9**, 609–621.

Mousset, E., Ainsworth, W. A., and Fonollosa, J. A. R. (**1996**). "A comparison of several recent methods of fundamental frequency and voicing decision estimation," in *Proceedings of the Fourth International Conference on Spoken Language Processing*, ICSLP, Philadelphia, Pennsylvania, 1273–1276.

Nakatani, T., and Irino, T. (**2004**). "Robust and accurate fundamental frequency estimation based on dominant harmonic components," J. Acoust. Soc. Am. **116**, 3690–3700.

Ostendorf, M., and Ross, K. (**1997**). "A multi-level model for recognition of intonation labels," in *Computing Prosody*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi (Springer-Verlag, New York), pp. 291–308.

Parsa, V., and Jamieson, D. G. (**1999**). "A comparison of high precision F_0 extraction algorithms for sustained vowels," J. Speech Lang. Hear. Res. **42**, 112–126.

Plante, F., Meyer, G., and Ainsworth, W. A. (**1995**). "A pitch extraction reference database," in *Proceedings of the Fourth European Conference on Speech Communication and Technology*, EUROSPEECH, Madrid, Spain, 837–840; Information about this database is available at http://www.liv.ac.uk/psychology/hmp/projects/pitch.html (last accessed 4/1/2008).

Rabiner, L., Cheng, M., Rosenberg, A., and McGonegal, C. (**1976**). "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-24**, 399–418.

Rabiner, L., and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ), pp. 204–208.

Rabiner, L., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ), pp. 150–157.

Ramana, R., and Srichand, J. (**1996**). "Word boundary detection using pitch variations," in *Proceedings of the Fourth International Conference on Spoken Language Processing*, ICSLP, Philadelphia, Pennsylvania, 813–816.

Shriberg, E., Bates, R., and Stolcke, A. (**1997**). "A prosody-only decision-tree model for disfluency detection," in *Proceedings of the Fifth European Conference on Speech Communication and Technology*, EUROSPEECH, Rhodes, Greece, 2383–2386.

Talkin, D. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier Science, New York), pp. 495–518.

Wang, C., and Seneff, S. (**1998**). "A study of tones and tempo in continuous mandarin digit strings and their application in telephone quality speech recognition," in *Proceedings of the Fifth International Conference on Spoken Language Processing*, ICSLP, Sydney, Australia, 635–638.

Wang, C., and Seneff, S. (**2000**). "Robust pitch tracking for prosodic modeling in telephone speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, Istanbul, Turkey, 1143–1146.

Zahorian, S. A., Dikshit, P., and Hu, H. (**2006**). "A spectral-temporal method for pitch tracking," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, Interspeech 2006—ICSLP, Pittsburgh, Pennsylvania, 1710–1713.

Zahorian, S. A., Zimmer, A., and Dai, B. (**1998**). "Personal computer software vowel training aid for the hearing impaired," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, Seattle, Washington, Vol. **6**, 3625–3628.

Phonatory characteristics of excised pig, sheep, and cow larynges

Fariborz Alipour[a] and Sanyukta Jaiswal
Department of Speech Pathology and Audiology, The University of Iowa, Iowa City, Iowa 52242

(Received 26 July 2007; revised 13 March 2008; accepted 19 March 2008)

The purpose of this study was to examine the phonatory characteristics of pig, sheep, and cow excised larynges and to find out which of these animal species is the best model for human phonation. Excised pig, sheep, and cow larynges were prepared and mounted over a tapered tube on the excised bench that supplied pressurized, heated, and humidified air in a manner similar to that for excised canine models. Each excised larynx was subjected to a series of pressure-flow experiments with adduction as major control parameter. The subglottal pressure, electroglottograph (EGG), mean flow rate, audio signal, and sound pressure level were recorded during each experiment. EGG signal was used to extract the fundamental frequency. It was found that pressure-frequency relations were nonlinear for these species with large rate of frequency changes for the pig. The average oscillation frequencies for these species were 220 ± 57 Hz for the pig, 102 ± 33 Hz for the sheep, and 73 ± 10 Hz for the cow. The average phonation threshold pressure for the pig was 7.4 ± 2.0 cm H_2O, 6.9 ± 2.9 cm H_2O for the sheep, and 4.4 ± 2.3 cm H_2O for the cow. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2908289]

PACS number(s): 43.80.Ka, 43.70.Aj, 43.70.Bk [AL] Pages: 4572–4581

INTRODUCTION

Phonatory models are important tools that help in the systematic study of phonation. Biomechanical aspects of phonatory models are best described by animal larynges, whether *in vivo* or when mounted and oscillated on the laboratory bench. The major advantages of these models are their capability for self-oscillation and easy access for manipulation of the control parameters. The closer the geometric and morphological similarities between these animal models and the human larynx, the better the results can be applied and interpreted for human phonation. A widely used animal model has been the canine larynx because of its geometric similarities to human larynx. It has been extensively used to study many aspects of phonation either *in vivo* or with excised models. For example, Koyama *et al.* (1969, 1971) used *in vivo* canine models to investigate the mechanics of voice production including regulation of pitch and intensity. Berke *et al.*, (1989a, 1989b) also used *in vivo* canine models to study the effect of recurrent laryngeal nerve and superior laryngeal nerve stimulation on the phonatory characteristics. Others have used excised canine larynges to study the phonatory physiology. Slavit *et al.* (1990) studied the effects of vocal fold tension and elongation on glottographic waveforms by using the excised canine larynx model. Alipour *et al.* (1997) used excised canine larynges to study the effects of adduction on laryngeal aerodynamics. Also, recently, the excised canine models have helped to quantify the aerodynamic and acoustic effects of the false folds and epiglottis [Alipour *et al.* (2007)] and pressure-frequency relations during phonation [Alipour and Scherer (2007)].

Besides the canine larynx, pig, sheep, and cow larynges may serve as other alternatives for phonatory models. Comparative anatomy of canine, pig, sheep, and human larynges [Kurita *et al.* (1983); Jiang *et al.* (2001); Hahn *et al.* (2005); Hahn *et al.*, (2006a, 2000b)] has shown that the three animal species have certain similarities in their basic structures to human larynx as well as differences that would account for the differences in phonatory productions.

Jiang *et al.* (2001) reported anatomical measurements from pig, deer, dog, and human excised larynges. The pig larynges had larger thyroid cartilages than the dogs, but the size of the cricothyroid muscle was greater in the dogs than the pigs. They defined the vocal fold height as the perpendicular distance from the axis of rotation of cricothyroid (CT) joint to the longitudinal axis of the vocal fold, which they also referred to as an index of the mechanical advantage of the thyroid cartilage for the purpose of phonation. This vocal fold height and the range of rotation of the CT joint were found to be comparable in all the three species, and the authors concluded that the mechanical advantage for vocal fold lengthening was similar in these species. They concluded that from the structural perspective, the pig larynx is a superior model.

Kim *et al.* (2004) compared the laryngeal dimensions of sheep, dog, and human excised larynges. The ovine laryngeal height was significantly greater than those in dogs and humans due to its vertically elongated thyroid cartilage. In sheep, the CT gap was considerably smaller, suggesting a smaller range of fundamental frequency control in sheep (as the cricoid-thyroid rotation angle would be smaller as well because of the larger cricoid dimensions). The sheep larynx also lacked well-defined ventricles and vocal fold boundaries. They asserted that while canine larynx measurements were within the range of those of human larynges, the ovine larynx showed distinctive differences.

[a]Author to whom correspondence should be addressed. Tel.: (319) 335-8694. FAX: (319) 335-8851. Electronic mail: alipour@iowa.uiowa.edu

From a histological point of view, the animal species had a two-layered lamina propria structure unlike the three-layered structure in the humans [Kurita *et al.* 1983]. The pig vocal fold was assumed to have a greater resemblance to the human vocal folds [Jiang *et al.* (2001)] because of the similarity in mucosal thickness (0.9 mm compared to 1.1 mm in the human). Also, intrinsic laryngeal muscles of the pig were found to be similar to human in terms of their origins and insertions [Knight *et al.* (2005)]. Sheep laryngeal muscles had fiber types and composition comparable to humans [Happak *et al.* (1989); Zrunek *et al.* (1988)] and were therefore deemed as one of the models for physiological studies of phonation.

The geometrical differences between porcine, bovine, and ovine larynx compared to the canine and human larynx were not only in the laryngeal and vocal fold dimensions but also in certain protective modifications in the supraglottic structures. Unlike the canine, these species are herbivorous and bear a large epiglottis and arytenoids with high lateral walls that enable them to breathe and swallow at the same time [Harrison (1995)]. Consequently, there exists a long supraglottic duct about 1–2 in. long, that is formed between their larger arytenoids and supraglottic thyroid wall. The wall of this duct in these three species participates in the vocal fold oscillations in certain conditions and has influence on the overall acoustics of phonation. This functional difference may cause some limitation to the applicability and comparison of their phonatory data to human phonation.

There were some similarities and differences in the molecular composition and stiffness characteristics of the vocal folds as well. In the human vocal folds, the density of the collagen fibers progressively increased in the direction of the vocalis muscle laterally. The density of elastic fibers was greatest in the intermediate layer of the lamina propria and gradually decreased in the superficial and deep layers. Kurita *et al.* (1983) did a cross-species comparison of thickness of the mucosa and the density of collagen and elastic fibers in the lamina propria of the vocal folds. The vocal folds from larynges of dogs, pigs, sheep, and humans were sectioned, stained, and examined under a microscope. In contrast to the canine and sheep vocal folds, the pig larynx demonstrated the humanlike superficial layer with sparse fibrous components, with increasing collagenous fibers and decreasing elastic fibers close to the vocalis muscle. Hahn *et al.* (2005, 2006a, and 2006b) did a more quantitative immunohistochemical analyses to examine the collagen, elastin, and proteoglycan (PG) and associated glycosaminoglycan (GAG) content of the midmembrane vocal fold lamina propria (LP) in humans, dogs, and pigs. The comparison of LP distribution of specific PG/GAG indicated that the canine LP subdivisions did not fit into the laminar arrangement characteristic of the human LP which may be indicative of the differences in the functional requirements.

Phonatory models require vocal fold dimensions and their viscoelastic properties for comparison to human subjects. Majority of these data were collected in the past on the canine tissues. Our preliminary studies of the vocal fold elastic properties of the pig, sheep, and cow [Alipour and Jaiswal (2006)] indicated significant differences between elastic properties among these species. The mean longitudinal Young's modulus for the superior vocal folds of the porcine larynx was 20.2 kPa that was significantly larger than its inferior vocal fold Young's modulus of 16.9 kPa (four samples, 18 cases, $p=0.022$). The bovine samples had a higher modulus value of 29.9 kPa (three samples, nine cases) but were also longer and thicker. The sheep vocal folds were more pliable with a value of 17.6 kPa (four samples, 27 cases). Although the number of samples were small, these preliminary data suggest that the fundamental frequency of canine vocal fold with low strain Young's modulus of about 42 kPa has higher frequency sensitivity to elongation than pig vocal folds. Further investigation into the elastic properties of these vocal folds may be needed for a better comparison.

Interspecies comparison of phonatory characteristics is of interest as the similarities and differences across species allow for greater understanding of phonatory physiology. In addition, the elucidation of phonatory characteristics helps in the selection of appropriate model for experimentation. It also furthers the knowledge regarding phonatory modifications and its evolutionary bases. Despite the wide usage of pig, sheep, and cow larynges in laryngeal muscle physiology and comparative histology studies [Kurita *et al.* (1983); Happak *et al.* (1989); Zrunek *et al.* (1988)], these species have not been extensively used in phonation study. Scherer *et al.* (1985) used the bovine excised larynx model to measure contact pressure during phonation because of its large vocal folds and its stable oscillations. In a previous study, Alipour and Jaiswal (in press) quantified the glottal flow resistance in the pig, sheep, and cow excised larynges. They found that unlike the canine larynges, these species had nonlinear pressure-flow relations and oscillated in different ranges of frequencies. Cow larynx with its larger dimensions had the lowest oscillation range and the pig larynx had the highest range. Thus, this paper is a follow-up to the earlier work. The purpose of this study was to investigate phonatory characteristics of these species, to examine the relation between subglottal pressure and the fundamental frequency of vocal fold vibration, and to find out which of these animal species is the best model for human phonation. This was accomplished by measuring and reporting pressure-flow sweep data and comparing vocal fold vibration characteristics across the species.

METHODOLOGY

Larynges of pig ($N=8$), sheep ($N=8$), and cow ($N=6$) were procured from a local butcher shop. They were cleaned, and extraneous tissue and muscles were dissected and removed. Since the larynges were initially slow frozen at the butcher shop, the prepared tissue was slow frozen following the cleaning. They were packaged in plastic bags and stored in the freezer at −20 °C for a few days to a few weeks. Prior to the experiment, the specimen was thawed overnight in saline solution and mounted on a base (by the trachea) for better handling. The epiglottis was dissected out, and the resulting free ends of the lateral wall were sutured to the thyroid cartilage along its superior margin to prevent any interference with the experimental conditions. Sutures were

placed on the larynges to stabilize the structures and simulate different degrees of adduction. Electrode plates from an electroglottograph (EGG) device (Synchrovoice) were placed on either side of the thyroid laminae to obtain the EGG signal during phonation. The EGG signal was later used to extract fundamental frequency.

Air from the building pipeline first entered a desiccating air filter (Devilbiss DEVDAD500) to remove dirt and water content, then passed through an in-line flow meter (Gilmont rotameter model J197) for airflow rate monitoring and a pneumatic flow meter (Rudolph 4700). The filtered air was then heated and humidified to about 37 °C and 100% humidity (ConchaTherm III, Hudson RCI) and entered the larynx via appropriately tapered tubing. A pressure tap located about 10 cm below the larynx was used to monitor the subglottal pressure through a well-type manometer (Dwyer model 1230-8).

Adduction was manipulated by simulating the action of the lateral cricoarytenoid muscle with sutures that medialized the vocal processes of the arytenoid cartilages. The sutures were placed on the muscular processes on either side, coursed parallel to the wall lateral to the vocal folds, and anteriorly emerging through the thyroid lamina. When the sutures were anteriorly pulled by addition of graded weights, the muscular process was rotated and consequently, the vocal processes medially moved, thereby increasing the degree of vocal fold adduction. Depending on the size of each larynx, adduction weights ranged from 20 to 200 g for the sheep, 50–300 g for the pig, and 200–1200 g for the cow larynges to provide graded low to high adductions levels. Each experiment started with one upward and one downward pressure-flow sweep of 20 s duration that determined the operational range of pressure, flow, and frequency for each larynx. The main control parameters were the levels of adduction and subglottal pressure. Due to the major differences in the dimensions and the angles of the vocal folds in these species, elongation was not considered in this study.

The time-varying subglottal pressure was recorded using a pressure transducer (Microswitch 136PC01G1) mounted in the tracheal tube across from the manometer pressure tap. The time-varying flow rate was recorded with a pneumatic flow meter and low-range pressure transducer (Validyne DP103) upstream of the humidifier. The sound intensity was measured with a sound level meter (Extec model 407738), placed 10–15 cm from the larynx. Analog signals from the EGG, microphone, pressure, and flow transducers were simultaneously recorded on a Sony SIR1000 digital tape recorder at a sampling rate of 40 kHz per channel and directly onto a computer using a DATAQ A/D converter and WINDAQ software. The EGG signal and subglottal pressure were low-pass filtered at 500 Hz and monitored on a digital oscilloscope (Tektronix, TDS2014). The oscilloscope was set to show few cycles and the signal frequency. The superior view of the vocal fold oscillation was monitored with a stroboscopic light source on a television screen while it was video taped (Phaser Strobe, Monarch Instruments).

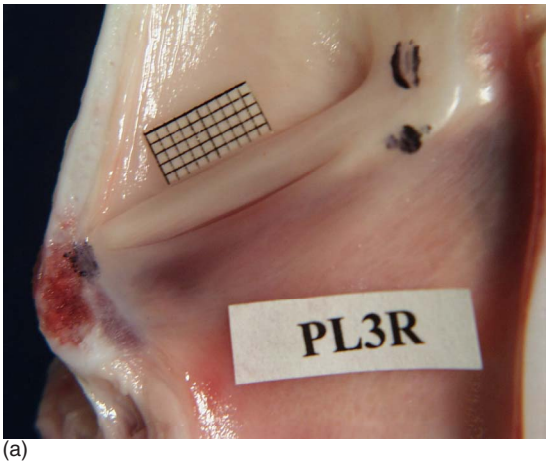
For each excised larynx, the experiment started with two pressure-flow sweeps (upward and downward) for low, medium, and high adduction levels. Then, a series of sustained oscillation runs were made within the working range of pressure and flow to record and observe oscillation of vocal folds in slow motion visualized with strobe light. For some of the larynges, instead of adding graduated weights for discrete adduction levels, the adduction suture was attached to a micrometer with an attached strain gauge to measure the adduction force (in grams). The manipulation of the micrometer toward or away from the larynx generated continuous adduction variations in the form of adduction sweeps and was measured as adduction force on the strain gauge.

The recorded signals were then converted to physical values with MATLAB™ software routines and used for the aerodynamic and acoustic analyses. Mean values of subglottal pressure, flow rate, pressure amplitude, and fundamental frequency were calculated during each sweep in the following manner. First the highest value of fundamental frequency was estimated from the spectrogram of the EGG signal. With this period a time segment that could include 20 cycles was estimated. Then, the duration of each sweep was divided into these segments, and the mean subglottal pressure and mean flow rate were calculated for each segment. To calculate the fundamental frequency, the EGG signal was low-pass filtered at 150% of the previously estimated frequency. The fundamental frequency was calculated with a zero crossing method from the filtered EGG signal. In zero crossing method, first the signal dc offset was removed and then periods of all the cycles in the selected segment were calculated from consecutive zero crossings and averaged. Due to possible variation in the fundamental frequency during a sweep, the number of cycles in each segment could vary a few cycles above or below 20.

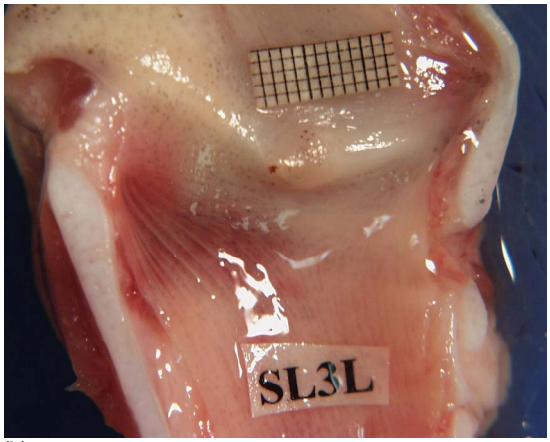
Once data were calibrated to their physical quantities in MATLAB environment, the mean values of major aerodynamic and acoustic information such as subglottal pressure, flow rate, glottal flow resistance, fundamental frequency, and pressure amplitudes were calculated and exported to EXCEL worksheets. Then statistical analysis was applied to these calculated mean values through the *t*-test for independent samples by variables in Basic Statistics of the STATISTICA package. Using a criteria of $p=0.05$, parameters were compared for significant differences. Also, whenever a correlation was established between two variables, its correlation coefficient (R^2) was calculated.

RESULTS

Figure 1(a) shows the internal cross-sectional view of the right half of a pig larynx. The pig larynx has a narrow slitlike, long ventricle that separates its vocal folds with well-differentiated boundaries. The two vocal folds are slanted at an angle of about 40° with the posterior end raised more than the anterior. The inferior vocal folds have a thinner mucosal cover with reddish muscular tissue visible underneath, especially in its inferior edge. The superior vocal folds are more distinct, usually 28 ± 2 mm long with a thickness of about 3 mm. Video observation indicates that superior vocal fold is an active oscillator, which covers the inferior folds from top view. Unlike the canine and human larynges, this oscillator appears to differently behave from



(a)



(b)

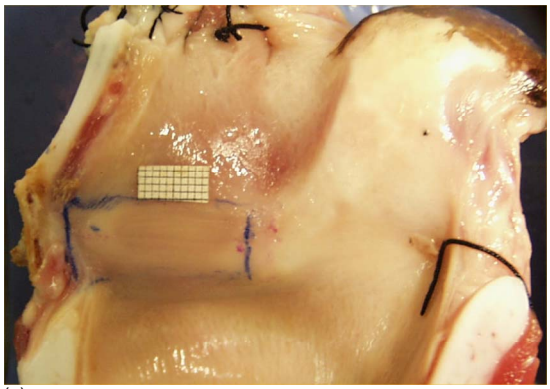

(c)

FIG. 1. (Color online) (a) Pig hemilarynx with a millimeter grid place above the superior fold. A large, modified arytenoid is attached to the superior and inferior folds on the upper right. A narrow ventricle separates the two folds. The vertically angled position of the vocal folds can be seen. (b) Sheep hemilarynx showing the padlike small vocal folds. (c) Cow hemilarynx showing the padlike large vocal folds.

the false vocal fold because of its higher Young's modulus than the inferior vocal folds. The elastic modulus for the false vocal folds measured for human larynx is reported to be much less than those of the true vocal folds [Chan *et al.* (2006)].

Figures 1(b) and 1(c) show the internal cross-sectional view of the sheep and cow larynges, respectively. Sheep vocal folds are soft and pliable with an average length of 17.2 ± 2 mm and thickness of 6 mm. Unlike the pig, sheep and cow lack a ventricle and their vocal folds are padlike with no sharp boundaries. The cow vocal folds are longer and stiffer with an average length of 37 ± 0.7 mm and a thickness of about 18 mm. Observation of the superior view of these larynges indicates that oscillation in these two species generates large mucosal waves that travel both in vertical and horizontal directions.

Figure 2 shows the beginning of typically recorded sweep for pig larynx, including from top to bottom are the following signals: EGG, subglottal pressure (P_s), and flow rate signal (F). These data correspond to oscillations at 96.3 Hz. The EGG data indicate some irregularity of oscillation that was typical of pig excised larynx. The fundamental frequency and pressure amplitude for each segment are calculated, such that each sweep generates a graph of glottal parameters such as the mean pressure, flow rate, fundamental frequency, and pressure amplitude as a function of time. The beginning upward sweep and the end of downward sweep data were used to calculate the onset and offset phonation threshold pressures (PTPs). These calculated PTP values complement their manually recorded values.

Figure 3 shows the variation of mean glottal parameters for a pig excised larynx (PL31) during an upward sweep (UP) and downward sweep (DN) for high adduction. The top graph shows the fundamental frequency (F_0), the second graph shows the subglottal pressure (P_s), and the third graph is the mean flow rate (Flow) variation with time. For this pressure-flow sweep, the onset PTP was about 9 cm H_2O and the maximum sound pressure level (SPL) at a distance of 15 cm from the larynx was 82.5 dB. During the upward sweep the mean fundamental frequency shows a sudden, rapid jump of about one octave and continues to rise to over 360 Hz. A mode change was observed at this point with an increase in audio pitch and sudden change in the oscillation amplitude observed in the video image from above. The downward sweep shows the decreasing trend across all the three parameters, though it did not reduce to the lower frequency range observed during the upward sweep. Pig larynges were usually found to oscillate with large amplitudes and loud SPLs.

In a similar experiment, a sheep excised larynx (SL22) was subjected to the pressure-flow sweeps at high adduction (Fig. 4). All three mean values of glottal parameters (F_0, P_s, and Flow) show monotonic changes with time either in upward sweep or downward sweep. This sheep larynx started to oscillate at about 6 cm H_2O and ended at about 5 cm H_2O. The maximum SPL at a distance of 15 cm from the larynx was 76.4 dB. In comparison to the previous case of pig larynx, this sheep larynx oscillated at lower ranges of subglottal pressure and fundamental frequency, but flow rates were comparable.

Figure 5 shows a similar pressure-flow sweeps in a cow excised larynx (BL17) at high adduction level. The maximum SPL at a distance of 15 cm from the larynx was 82.7 dB. The major difference observed in the cow oscillation is the smaller change in the fundamental frequency (about 9%, 4–5 Hz) from the beginning to the end of sweep. The cow excised larynx, thus, consistently showed sustained oscillation with steady frequency that did not drastically change with pressure variations. The typical frequency of the

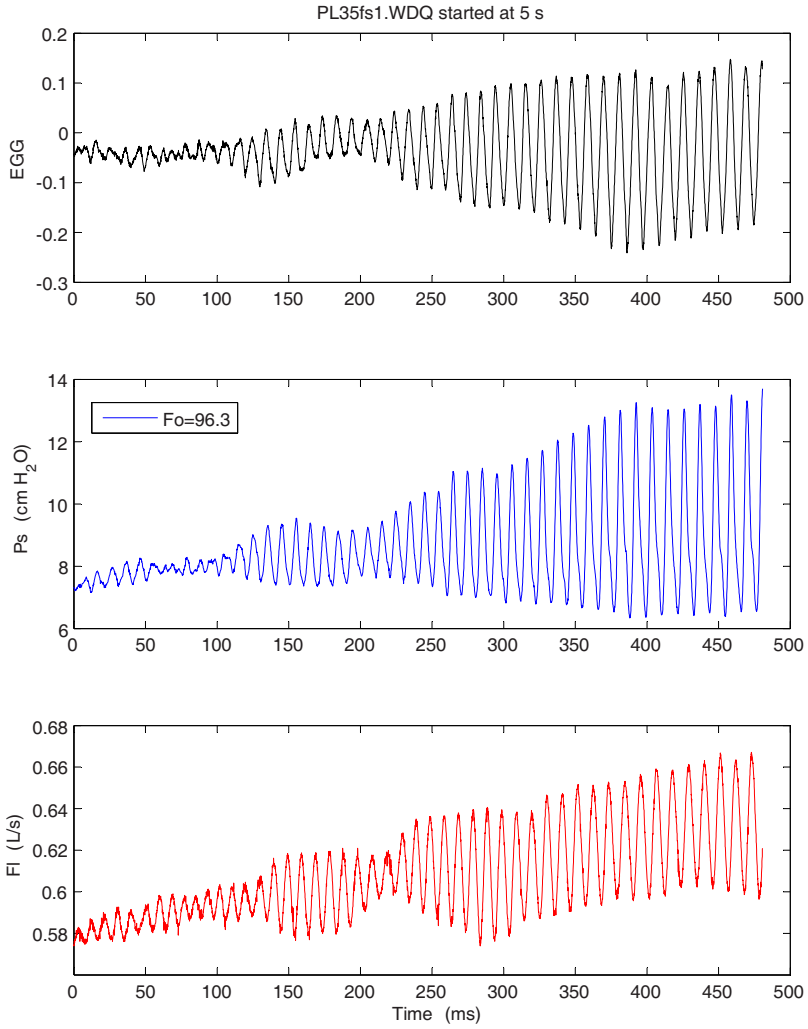


FIG. 2. (Color online) Initial portion of the pressure-flow sweep, including (from top to bottom) electroglottograph (EGG), subglottal pressure (P_s), and flow rate (Flow) signals.

cow larynx oscillation was about 74 Hz. However, there were conditions in which the larynx initially oscillated at higher rate, and once the subglottal pressure increased, the frequency dropped to a lower value. The video observations of the oscillation with stroboscopic light indicated that at higher frequencies, only a segment of total length of the vocal fold oscillated with low amplitude, but once the subglottal pressure was increased, a greater portion or even the entire length of the vocal folds was set into large amplitude oscillation, resulting in the fundamental frequency drop.

Figures 6–8 show the pressure-frequency relations for the excised pig, sheep, and cow larynges during three upward sweeps, respectively. These graphs are based on the mean values of subglottal pressure and fundamental frequency calculated from pressure-flow sweep experiments. The pig larynx (PL34, Fig. 6) showed a wider range of oscillation with frequency ranging from 160 to 300 Hz excluding the mode change and up to 500 Hz with mode change. The high adduction condition showed a mode change with a sharp increase in the fundamental frequency between 15 and 17 cm H_2O subglottal pressure. The duration of this high frequency oscillation is short, but suggests nonlinear behavior in the pig vocal folds. Nonlinearity in the pressure-frequency relations is observable even in the average curves without the mode changes.

The sheep larynx pressure-frequency relationship (SL15) showed slightly different behavior for low pressures in comparison to high pressures (Fig. 7). At higher subglottal pressure ranges, the pressure-frequency appeared to be fairly linear with very little sensitivity to the degree of adduction. However, at the lower subglottal pressure ranges, a more nonlinear relation was observed with the frequency showing minima that was also dependent on the adduction level with very low and high adduction levels showing the least drop in frequency. The negative values of slope (dF/dP) at beginning of the oscillation may be related to the decrease in the oscillation frequency with greater extent of participation of vocal fold masses for oscillation, as seen in lower levels of adduction. The average oscillation frequency in the sheep larynges was lower with smaller range than the pig larynges (65–90 Hz).

In Fig. 8, the pressure-frequency relationship for the cow larynx BL17 is presented. The graph suggests that not only the oscillation frequency range was limited (70–80 Hz); its rate of change with pressure (dF/dP) did not show a similar relationship to the adduction level as in the pig larynx. The widest range of oscillation was observed for the highest adduction level with the lowest range for the low adduction level. The fact that frequency did not show

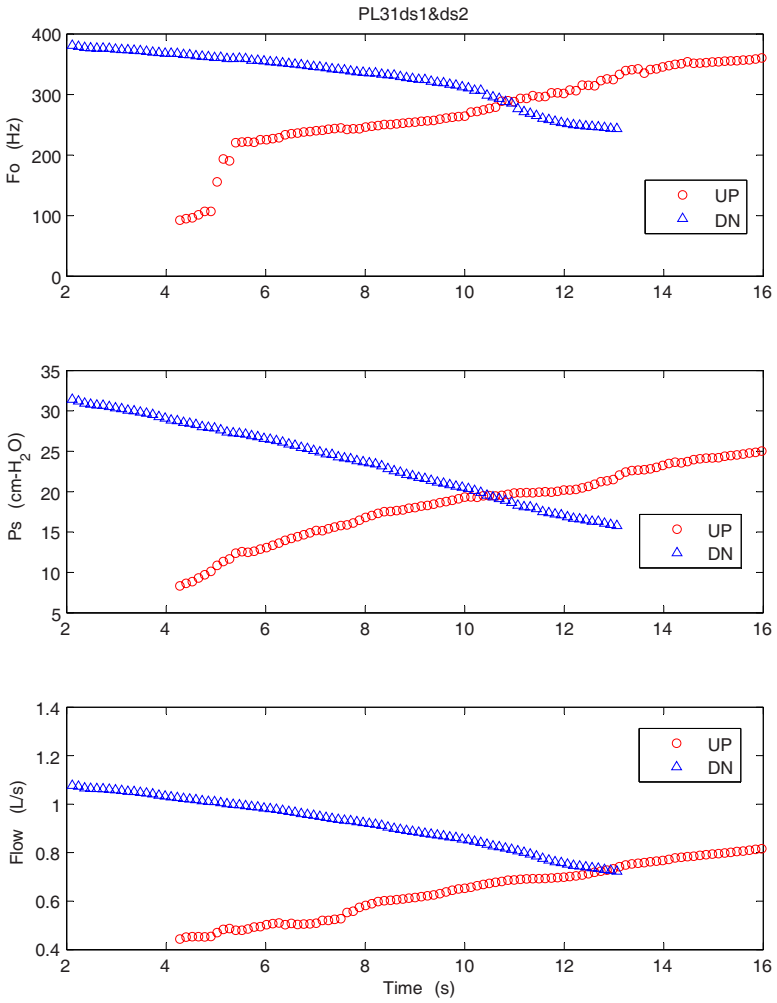


FIG. 3. (Color online) Mean values of the fundamental frequency, subglottal pressure, and flow rate for an excised pig larynx during the upward (○) and downward (△) pressure-flow sweeps at high adduction level.

any nonlinearity associated with oscillation condition makes the cow larynx a suitable oscillator for the aerodynamic and acoustic studies.

The mean glottal parameters of these larynges during the pressure-flow sweeps across different samples and at various adduction levels were calculated and statistical analysis was applied to these calculated means through the *t*-test for independent samples by variables in Basic Statistics of the STATISTICA package. Figure 9 compares the mean values of subglottal pressure (P_s), pressure amplitude (P_a), and PTP for the pig (eight samples, 38 sweep cases), sheep (eight samples, 34 sweep cases), and cow (six samples, 27 sweep cases). The operating mean subglottal pressures were in the same range, but the cow larynx had significantly larger pressure amplitude than pig and sheep ($p<0.01$). This can be related to the almost rectangular glottis in the cow larynx with large surface area and padlike vocal folds that displaced a larger volume of air during the oscillation. The cow larynx had the PTP range of 2–10 cm H_2O with the lowest average value of 4.4 ± 2.3 cm H_2O, sheep larynx had PTP range of 3–14 cm H_2O with average value of 6.9 ± 2.9 cm H_2O, and pig larynx had PTP range of 4–12.3 cm H_2O with an average of 7.4 ± 2.0 cm H_2O. The same reason for the pressure amplitudes may also explain why the cow larynx has a significantly smaller PTP than pig and sheep ($p<0.001$).

Figure 10 compares the mean values fundamental frequency (F_0), SPL, and the rate of frequency changes with pressure (dF/dP) for these larynges. The mean oscillation frequency for the pig was 220 ± 57 Hz and for the sheep was 102 ± 33 Hz, and 73 ± 10 Hz for the cow. The pig larynx has significantly higher F_0 than sheep larynx ($p<0.0001$) and sheep larynx has significantly higher F_0 than cow larynx ($p<0.0001$). The pig larynx usually generated the loudest sound ($p<0.0001$) that could reach as high as 96.1 dB with an average of 88.3 ± 4.5 dB. The sheep and cow larynges had produced sounds within similar intensity ranges. The SPL for the sheep was 78.3 ± 5.0 dB and for the cow was 79.5 ± 5.3 dB. Finally, the pig larynx had the highest dynamic range of frequency ($dF/dP=1.9-19.7$) with an average of 6.0 Hz/cm H_2O which is significantly larger than others ($p<0.0001$). The sheep larynx had a peculiar pattern with a negative region and a range of −6.5–5.5 with an average of 0.47 Hz/cm H_2O. The cow larynx had the lowest range of 0.1–0.9 with an average of 0.5 Hz/cm H_2O.

DISCUSSION

The purpose of this study was to examine the phonatory characteristics of pig, sheep, and cow excised larynges and to examine the relation between subglottal pressure and the fundamental frequency of vocal fold vibration in these spe-

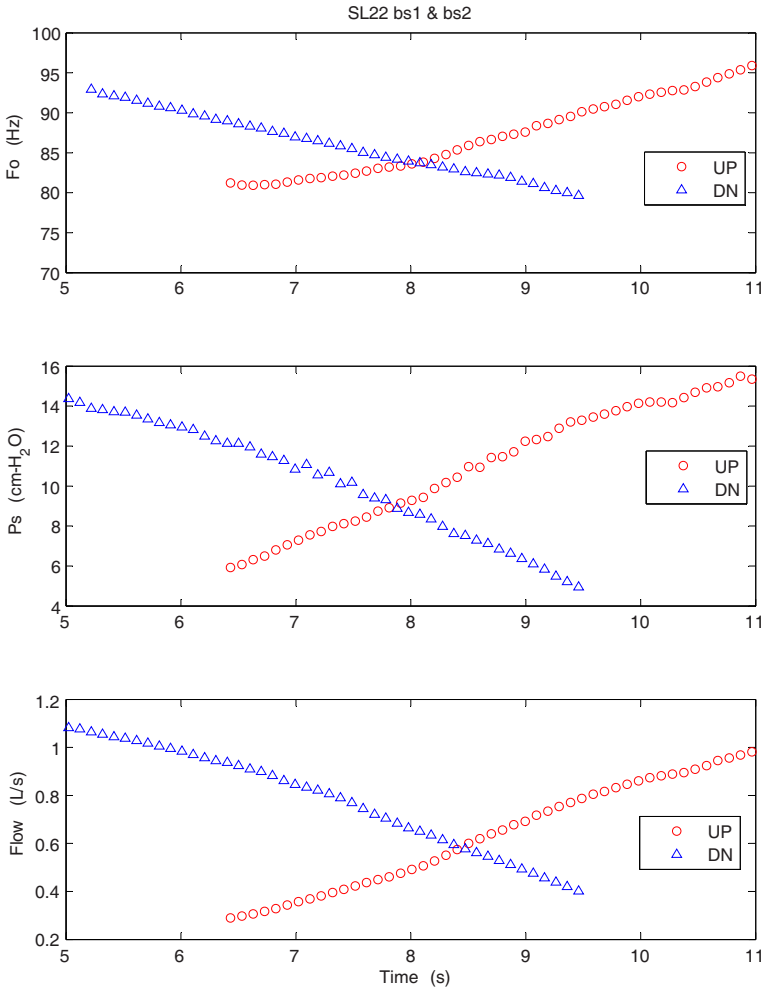


FIG. 4. (Color online) Mean values of the fundamental frequency, subglottal pressure, and flow rate for an excised sheep larynx during the upward (○) and downward (△) pressure-flow sweeps at high adduction level.

cies. Each one of these species demonstrated different oscillation ranges and pressure-frequency behavior. The presence of two oscillating vocal folds in addition to the supraglottic structure wall was probably responsible for the greater dynamic range of F_0 in pig larynx. This was possibly responsible for greater instability in oscillation as well. Mode changes in the form of sudden variations in oscillation frequency during the pressure sweep were observed in pig and in some cases in sheep larynges as well. The physiological explanation of the phenomenon can be attributed to the participation of additional masses, whether in the form of superior/inferior vocal folds and glottal wall in the pigs or supraglottic structures in the sheep in the oscillation.

One thing that was common in all three species was the existence of a supraglottic duct with an approximately 2–4 cm length. The combination of this duct and the spout formed by the large epiglottis and arytenoids in these species creates an inertive load that helps stabilize their phonation. The low frequency oscillations of the supraglottic wall could have supplemented the glottal oscillations. These three larynges considerably varied in terms of dimensions of the cartilages and the vocal folds. The lack of well-defined vocal fold boundaries in the sheep and cow larynges made it difficult to accurately measure their vocal fold dimensions.

In a previous study reported on these larynges [Alipour and Jaiswal (in press)], it was reported that these larynges operated in almost the same ranges of maximum pressure and maximum flow rate. The most obvious difference observed was their maximum frequency ranges, with the highest value for the pig and the lowest for the cow larynx. Similarly, in this study, the mean fundamental frequency is reported for each larynx during the sweep. The pig larynx with well-defined superior and inferior vocal folds and a ventricle in between had higher frequency because its narrow and stiff superior vocal folds acted as the main oscillator. This phonation type probably matched with the high frequency natural "squeal" of the pig vocalization. Participation of the supraglottic wall and other structures probably results in the lower pitched "grunt."

These pressure-frequency behaviors suggest that pig larynx with its widest operating range of frequency, large amplitude, and higher dF/dP is a good model for the study of pitch control. Pig larynx can oscillate from 100 to 300 Hz with large amplitude and loud intensity, but most often with large frequency jitter. Histological studies have suggested a closer similarity of its LP to the human vocal folds than canine. However, the steeply angled position of the pig vocal folds creates a big challenge for the image analysis and vocal amplitude measurements. Moreover, the previous histological studies seemed to have focused on the inferior folds, assuming the superior folds to be nonparticipatory as in the human and canine false folds. The phonatory data provide evidence that the superior folds are actually one of the primary vibratory sources. More information regarding the his-

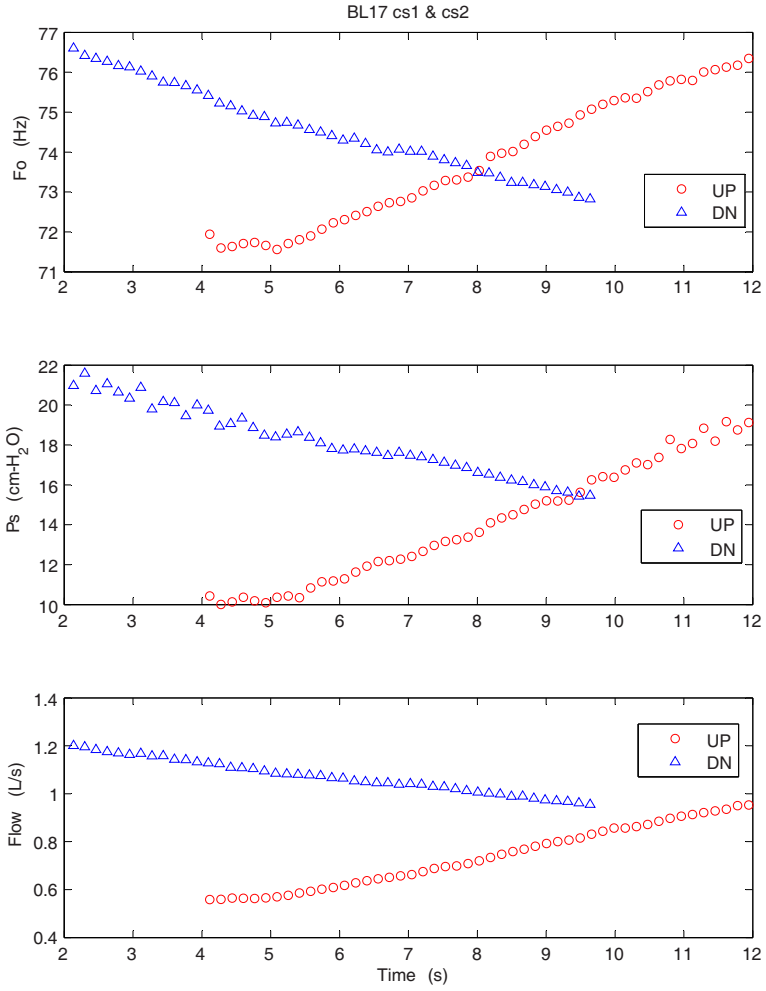


FIG. 5. (Color online) Mean values of the fundamental frequency, subglottal pressure, and flow rate for an excised cow larynx during the upward (○) and downward (△) pressure-flow sweeps at high adduction level.

tological and biomechanical nature of the superior folds is required to provide physiological basis to the acoustic output. Moreover, possible participation of the supraglottic wall can further create phonatory instability.

The cow larynx with low PTP and almost steady pitch is suitable for studies involving aerodynamic measurements. Also, the SPL values of cow larynges seem to correlate well with their pressure amplitude (R^2=0.81) and the subglottal pressure (R^2=0.64). The pressure-amplitude dependence was reported for the canine larynges by Titze (1989). Thus, cow larynges can be a good model for studies involving measurements of oscillation amplitude and intraglottal [Alipour *et al.* (2001)] pressure due to its large dimensions and steady pitch profile. The larger dimension has allowed for the insertion of contact pressure transducers in the larynx in the past [Scherer *et al.* (1985)].

The sheep larynx with a soft and pliable vocal fold tissue produced large vibrational amplitudes with big mucosal waves that could serve as a useful phonatory model as well. The similarity of its vocal fold length, laryngeal dimensions, and tissue histochemistry to human vocal folds make it a good physiological model. It would also serve as suitable phonatory model for studies requiring smaller frequency ranges.

The findings from this study suggest that the pig, sheep, and cow larynges with their distinctive aerodynamic and acoustic behaviors could be used as phonatory and aerodynamic models, especially when different frequency ranges and sizes of the larynx are crucial factors. When replacing pig larynx with canine, the angled position of the vocal folds, longer dimensions of vocal folds, smaller CT size, and lower mechanical advantage in pig larynx would need to be considered. The variations in pitch with elongation may not be as clear as in the canine larynx. The results of this study also provide some clues about the phonatory effects of evolutionary modifications of the larynx. This is especially reflected in the shapes and sizes of the supraglottal duct and its participation in phonation. Some of the variability in the data may be attributed to the use of some samples that were slow frozen, and their mechanical properties were deteriorated by the ice crystals. Future works should use fresh or fast frozen samples to avoid this problem.

SUMMARY AND CONCLUSIONS

Pressure-frequency relations were obtained for pig (eight samples), sheep (eight samples), and cow (six samples) larynges through a series of pressure-flow sweep experiments that were conducted on the excised bench. Vocal fold adduction was the major control parameter. The following was found.

(1) Pressure-frequency relations were nonlinear for these species, with large ranges in df/dp slope for the pig and

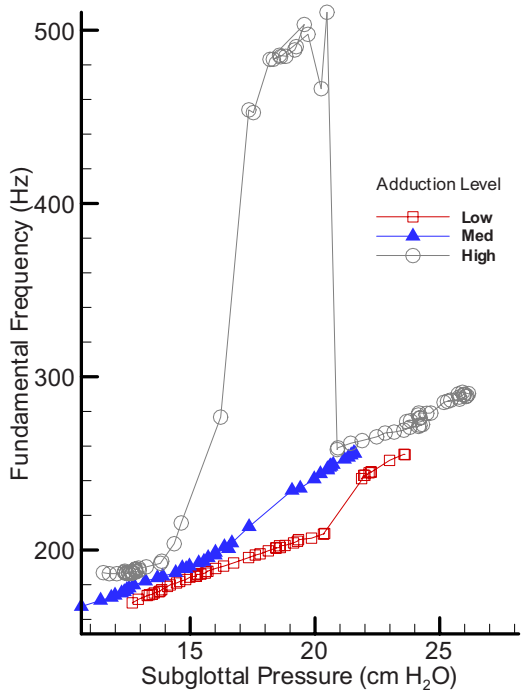


FIG. 6. (Color online) Pressure-frequency relationship for an oscillating pig larynx at three adduction levels: low, medium, and high.

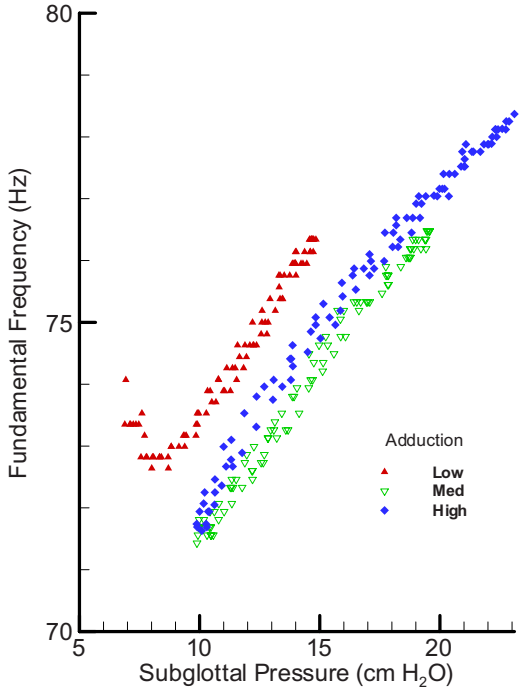


FIG. 8. (Color online) Pressure-frequency relationship for an oscillating cow larynx at three adduction levels: low, medium, and high.

sheep excised larynges. Sheep larynx showed negative slope under low pressure conditions that stabilized with increase in subglottal pressure.

(2) The average oscillation frequencies for these species were 220 ± 57 Hz for the pig, 102 ± 33 Hz for the sheep, and 73 ± 10 Hz for the cow.

(3) The average PTP for the pig was 7.4 ± 2.0 cm H_2O, 6.9 ± 2.9 cm H_2O for the sheep, and 4.4 ± 2.3 cm H_2O for the cow.

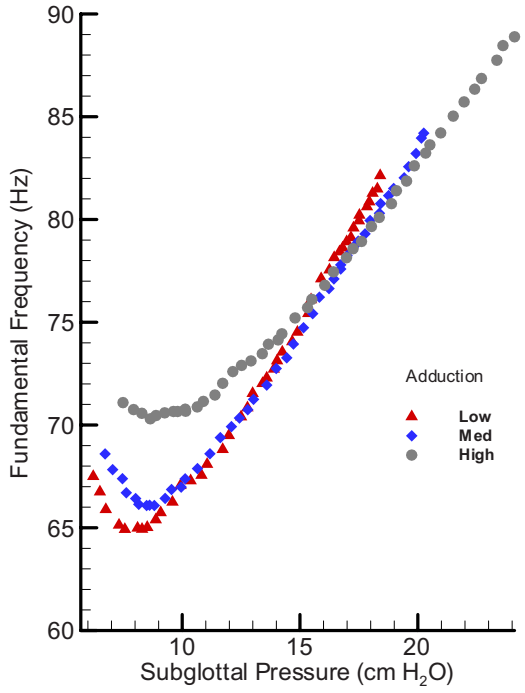


FIG. 7. (Color online) Pressure-frequency relationship for an oscillating sheep larynx at three adduction levels: low, medium, and high.

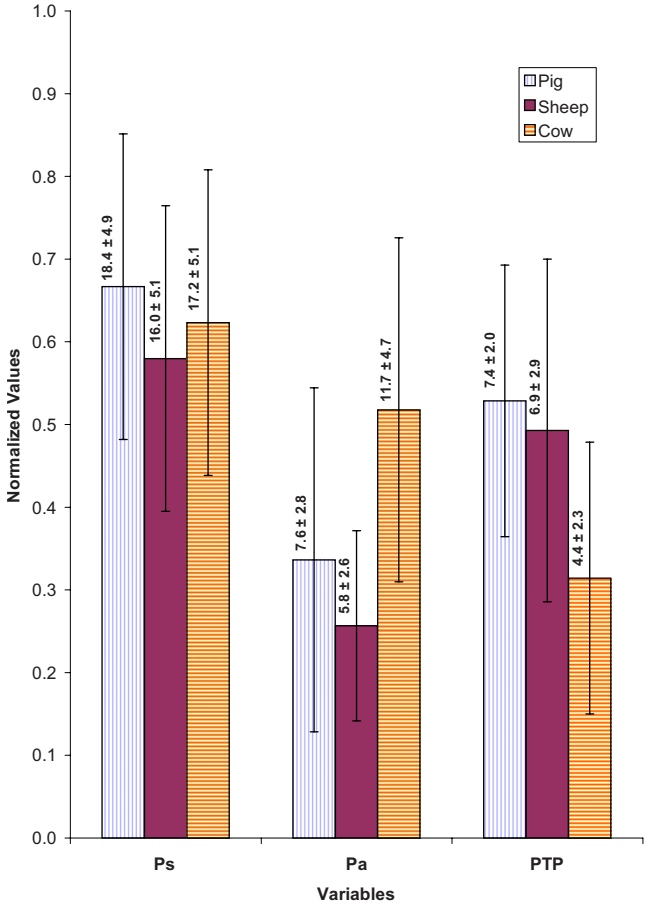


FIG. 9. (Color online) Normalized composite profile of different pressure variables during pressure-flow sweeps of pig, sheep, and cow larynges. P_s is subglottal pressure, P_a is the pressure amplitude, and PTP is phonation threshold pressure. All pressure values are in cm H_2O. The mean value for each condition is provided over the corresponding bar.

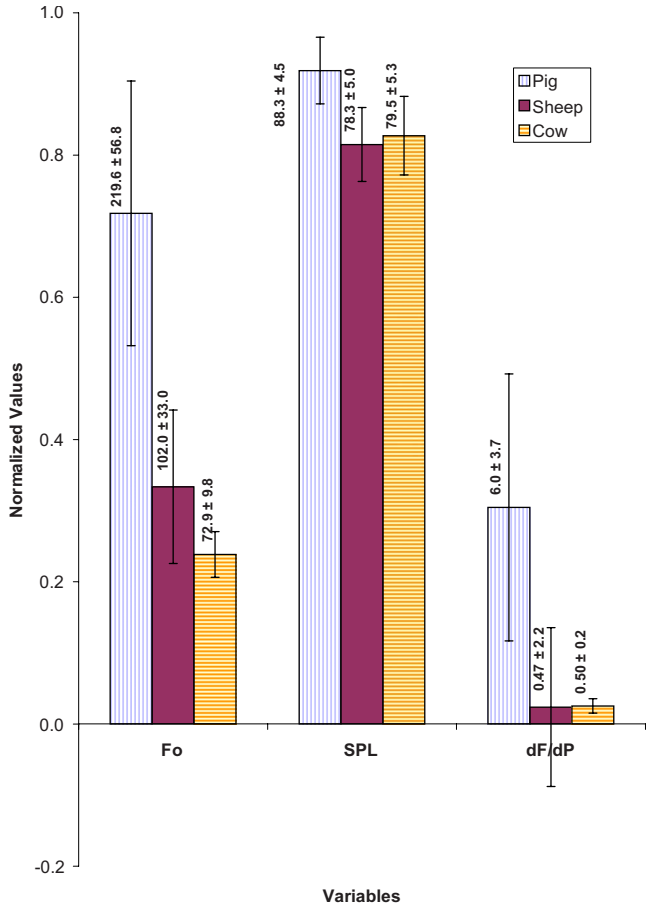


FIG. 10. (Color online) Normalized composite profile of different phonatory variables during pressure-flow sweeps of pig, sheep, and cow larynges. F_0 is the fundamental frequency (Hz), SPL is sound pressure level (dB), and dF/dP is the rate of change of frequency with pressure (Hz/cm H_2O). The mean value for each condition is provided over the corresponding bar.

ACKNOWLEDGMENTS

National Institute on Deafness and other Communication Disorders, Grant No. DC03566 supported this work. The authors would like to thank Jaclyn Curiel for assistance in data collection and Dr. Anders Löfqvist and two anonymous reviewers for their helpful comments.

Alipour, F., Scherer, R. C., and Finnegan, E. (**1997**). "Pressure-flow relationships during phonation as a function of adduction," J. Voice **11**, 187–194.

Alipour, F., Montequin, D., and Tayama, N. (**2001**). "Aerodynamic profiles of a hemilarynx with vocal tract," Ann. Otol. Rhinol. Laryngol. **110**, 550–555.

Alipour, F., and Jaiswal, S. (**2006**). "Vocal Fold Elasticity of pig, sheep and cow larynges," *Proceedings of the Fifth International Conference on Voice Physiology and Biomechanics*, Tokyo, Japan, July 12–14 pp. 28–29.

Alipour, F., Jaiswal, S., and Finnegan, E. (**2007**). "Aerodynamic and acoustic effects of false vocal folds and epiglottis in excised larynx models," Ann. Otol. Rhinol. Laryngol. **116**, 135–144.

Alipour, F., and Jaiswal, S. (**2008**). "Glottal airflow resistance in excised pig, sheep and cow larynges," J. Voice (in press).

Alipour, F., and Scherer, R. C. (**2007**). "On pressure-frequency relations in the excised larynx," J. Acoust. Soc. Am. **124**, 2296–2305.

Berke, G. S., Moore, D. M., Gerratt, B. R., Hanson, D. G., and Natividad, M. (**1989a**). "Effect of superior laryngeal nerve stimulation on phonation in an in vivo canine model," Am. J. Otolaryngol. **10**, 181–187.

Berke, G. S., Moore, D. M., Gerratt, B. R., Hanson, D. G., Bell, T. S., and Natividad, M. (**1989b**). "The effect of recurrent laryngeal nerve stimulation on phonation in an in vivo canine model," Laryngoscope **99**, 977–982.

Chan, R. W., Fu, M., and Tirunagari, N. (**2006**). "Elasticity of the human false vocal fold," Ann. Otol. Rhinol. Laryngol. **115**, 370–381.

Hahn, M. S., Kobler, J. B., Zeitels, S. M., and Langer, R. (**2005**). "Midmembranous vocal fold lamina propria proteoglycans across selected species," Ann. Otol. Rhinol. Laryngol. **114**, 451–462.

Hahn, M. S., Kobler, J. B., Starcher, B. C., Zeitels, S. M., and Langer, R. (**2006a**). "Quantitative and comparative studies of the vocal fold extracellular matrix. I: Elastic fibers and hyaluronic acid," Ann. Otol. Rhinol. Laryngol. **115**, 156–164.

Hahn, M. S., Kobler, J. B., Zeitels, S. M., and Langer, R. (**2006b**). "Quantitative and comparative studies of the vocal fold extracellular matrix II: Collagen," Ann. Otol. Rhinol. Laryngol. **115**, 225–232.

Happak, W., Zrunek, M., Pechmann, U., and Streinzer, W. (**1989**). "Comparative histochemistry of human and sheep laryngeal muscles," Acta Oto-Laryngol. **107**, 283–288.

Harrison, D. F. N. (**1995**). *The Anatomy and Physiology of the Mammalian Larynx* (Cambridge University Press, Cambridge).

Jiang, J. J., Raviv, J. R., and Hanson, D. G. (**2001**). "Comparison of the phonation related structures among pig, dog, white-tailed deer, and human larynges," Ann. Otol. Rhinol. Laryngol. **110**, 1120–1125.

Kim, M. J., Hunter, E. J., and Titze, I. R. (**2004**). "Comparison of human, canine, and ovine laryngeal dimensions," Ann. Otol. Rhinol. Laryngol. **113**, 60–68.

Knight, M. J., McDonald, S. E., and Birchall, M. A. (**2005**). "Intrinsic muscles and distribution of the recurrent laryngeal nerve in the pig larynx," Eur. Arch. Otorhinolaryngol. **262**, 281–5.

Koyama, T., Kawasaki, M., and Ogura, J. H. (**1969**). "Mechanics of voice production I. Regulation of vocal intensity," Laryngoscope **79**, 337–354.

Koyama, T., Harvey, J. E., and Ogura, J. H. (**1971**). "Mechanics of voice production II. Regulation of pitch," Laryngoscope **81**, 45–65.

Kurita, S., Nagata, K., and Hirano, H. (**1983**). "A comparative study of the layer structure of the vocal fold," in *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, edited by Bless, D. M., and Abbs, J. H. (College Hill, San Diego), pp. 3–21.

Scherer, R. C., Cooper, D., Alipour-Haghighi, F., and Titze, I. R. (**1985**). "Contact pressure between the vocal processes of an excised bovine larynx," in *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, edited by I. R. Titze, and R. C. Scherer (The Denver Center for the Performing Arts, Denver), pp. 292–303.

Slavit, D. H., Lipton, R. J., and McCaffrey, T. V. (**1990**). "Glottographic analysis of phonation in the excised canine larynx," Ann. Otol. Rhinol. Laryngol. **99**, 396–402.

Titze, I. R. (**1989**). "On the relation between subglottal pressure and fundamental frequency in phonation," J. Acoust. Soc. Am. **85**, 901–906.

Zrunek, M., Happak, W., Hermann, M., and Streinzer, W. (**1988**). "Comparative anatomy of human and sheep laryngeal skeleton," Acta Oto-Laryngol. **105**, 155–162.

Evidence for spatial representation of object shape by echolocating bats (*Eptesicus fuscus*)

Caroline M. DeLong, Rebecca Bragg, and James A. Simmons
Department of Neuroscience, Brown University, Box GL-N, Providence, Rhode Island 02912

(Received 11 September 2007; revised 21 March 2008; accepted 1 April 2008)

Big brown bats were trained in a two-choice task to locate a two-cylinder dipole object with a constant 5 cm spacing in the presence of either a one-cylinder monopole or another two-cylinder dipole with a shorter spacing. For the dipole versus monopole task, the objects were either stationary or in motion during each trial. The dipole and monopole objects varied from trial to trial in the left-right position while also roving in range (10–40 cm), cross range separation (15–40 cm), and dipole aspect angle (0°–90°). These manipulations prevented any single feature of the acoustic stimuli from being a stable indicator of which object was the correct choice. After accounting for effects of masking between echoes from pairs of cylinders at similar distances, the bats discriminated the 5 cm dipole from both the monopole and dipole alternatives with performance independent of aspect angle, implying a distal, spatial object representation rather than a proximal, acoustic object representation. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2912450]

PACS number(s): 43.80.Ka, 43.80.Lb [WWA] Pages: 4582–4598

I. INTRODUCTION

Insectivorous bats use an active mode of sound perception called *echolocation* (Griffin, 1958). They emit high-frequency sounds and perceive objects in their environment from the echoes that return to their ears. The sonar broadcasts of most bats are wideband signals covering frequencies of roughly 20–150 kHz (varying across species; Neuweiler, 2000). For example, the big brown bat (*Eptesicus fuscus*) emits frequency-modulated (FM) biosonar sounds with frequencies from 20 to 100 kHz arranged in several downward-sweeping harmonics (Hartley, 1992; Saillant *et al.*, 2007; Surlykke and Moss, 2000). These bats use echoes to navigate, to find prey, and to avoid obstacles.

Behavioral tests of object discrimination have focused on how bats employ acoustic parameters such as echo delay, echo amplitude, and echo spectrum to discriminate among objects that vary in size, shape, and distance from the bat (Grinnell, 1995; Moss and Schnitzler, 1995; Neuweiler, 2000; Simmons *et al.*, 1995). Additionally, neuroethological studies have explored how these acoustic parameters are encoded in the ear and auditory nervous system of bats (Neuweiler, 2000; Pollak and Casseday, 1989; Simmons *et al.*, 1996). However, there has been only a limited examination of the nature of echolocating bats' perception from a psychological perspective (Simmons, 1989). Specifically, we are interested in the nature of bats' representations. A cognitive representation is an internal model of a past experience (e.g., an object) that is used to guide future action and must be inferred from an organism's behavior (i.e., it cannot be investigated by examining the nervous system).

An organism's representation of objects in its environment may correspond to the parameters of the object itself (the *distal* stimulus) or to the parameters of the sensory experiences produced by the object (the *proximal* stimulus). For example, a person listening to a ball bouncing on the floor in another room could form a representation of the characteristics of the sounds produced by the ball, such as the loudness, duration, and repetition rate of the sounds. Alternatively, the person could form a representation of the characteristics of the ball itself, such as its shape, size, and structure. Similarly, echolocating bats' representations could contain the proximal stimulus or the distal stimulus. The proximal stimuli are the echoes that return to the bats' ears after sending out sonar signals. The bats could form a representation that would contain merely a list of echo acoustic features (e.g., amplitude, frequency, delay). In contrast, the bats could represent the distal stimulus, which would contain the size, distance, and spatial dimensions of the object.

Distal object representations are advantageous. For instance, they allow an organism to link different perceptual experiences of the same object. The same object representation could be accessed whether the object was observed visually or via echolocation. In addition, a distal object representation would allow an organism to identify an aspect-dependent object from different orientations relative to the object. Many objects that bats need to identify, such as prey animals, have an appearance that is aspect-dependent—that is, the object presents different-sized or different-shaped surfaces from different orientations. Consequently, the echoes the targets reflect depend on the aspect angle at the instant individual incident sounds impinge on the object to form echoes. The acoustic features of echoes returning from aspect-dependent targets vary greatly with the target's orientation with respect to the echolocating animal (e.g., Moss and Zagaeski, 1994).

The purpose of this study was to determine whether bats form proximal or distal object representations. Echolocating bats were trained to discriminate between two objects presented on flat circular surfaces located to the bat's left and right. Both objects were presented together on each trial, so the bats faced a simultaneous-discrimination two-choice test. Within the limits set by randomly introduced variations in the relative distance to each object, echoes from both objects

 0001-4966/2008/123(6)/4582/17/$23.00

reached the bat at about the same time. The bat's task was to decide which of these objects consisted of a pair of cylinders separated by 5 cm (referred to as the *dipole target*) and to move toward that object to receive a food reward. In the first experiments (experiments 1A and 1B), the other object—the unrewarded stimulus—consisted of just one cylinder (referred to as the *monopole target*). The dipole target was presented at different aspect angles from one trial to the next. Both the dipole and the monopole were also roved to different distances and different horizontal directions from trial to trial, so their specific positions and overall echo strengths could not serve as reliable cues.

In the second experiment, both the positive and the negative stimuli were dipoles. The negative stimulus had a spacing of 1.0, 1.5, 2.0, 2.5, or 3.0 cm between the cylinders (versus the 5 cm spacing for the positive stimulus). The 5 cm dipole versus shorter dipole experiment is likely to be more difficult for the bats because there are two reflecting parts to each target, so the average strength of the echoes will be the same, and both objects now contain two parts, so the simple two versus one cue is no longer present. By varying the distance, direction, and aspect angle of both dipoles from trial to trial, as well as the spacing of the shorter dipole, conditions were created for masking between reflections from individual parts of the targets at the same time at the bat's ears, thus lowering the bat's performance to the region of greatest sensitivity between 60% and 85% correct responses.

Taking the dipole versus monopole and dipole versus dipole experiments together and roving the targets to different locations and orientations between trials prevented any single feature of the acoustic stimuli from being a stable indicator of which target was the correct choice. The bat had to recognize the dipole with the 5 cm spacing without using any of the extraneous features. By reconstructing the locations of the various reflecting parts of each target on each trial, we could determine which acoustic or object features did affect the bat's performance and, thus, whether the bats were likely to have used proximal or distal object representations.

II. EXPERIMENT 1: MONOPOLE TARGET VERSUS DIPOLE TARGET

A. Method

1. Animal subjects

Five adult big brown bats, *Eptesicus fuscus*—two females ("Frodo" and "Astro") and three males ("Buddy," "Chris," and "Patrick")—were used in this experiment (for information about this species, see Kurta and Baker, 1990). They were obtained from colonies in the attics of houses in Rhode Island. The animals were kept in individual aluminum mesh cages in a temperature-controlled (72 °F) room with relative humidity maintained between 60% and 70%. The light-dark cycle of the room was 12–12 h, with lights on at 10:30 p.m. and lights off at 10:30 a.m. This cycle was implemented to enable the bats to be active during the daytime. The bats had continual access to vitamin-infused water (Poly-Vi-Sol drops) in their cages and were fed the majority of their mealworm diet during the experiments. The weights of the bats were recorded daily, and mealworm intake was adjusted according to the measured weights. During the course of experimentation, the bats' weights were kept roughly between 14 and 18 g. Notably, the big brown bats do not require extensive food deprivation to render them sensitive to food reward; bats at normal body weights in this range will respond to food without having to undergo deprivation beforehand.

2. Stimuli

Each bat was trained in a two-alternative forced choice procedure to discriminate between a dipole target consisting of two cylinders and a monopole target consisting of just one cylinder. The dipole target was always the positive stimulus. In the test condition, the cylinders making up the targets were identical; they each had a diameter of 1.59 cm and a height of 1.27 cm. They were made of a hard black plastic. The single cylinder for the monopole and the two cylinders for the dipole were mounted on thin strips of clear polyethylene (0.1 mm thick) to serve as a base and to keep the cylinders in the dipole at a fixed separation of 5.0 cm (measured from the center of the first cylinder to the center of the second cylinder). The monopole was centered on a polyethylene strip (1.91 cm × 2.54 cm) and the dipole targets were mounted on polyethylene strips (7.3 cm × 2.54 cm) such that the cylinders were 0.32 cm from the left and right edges of the strips. The targets were spray painted flat black and were presented on flat black surfaces located to the bat's left and right.

3. Experimental setup and procedure

Figure 1 shows the experimental setup. The bat was trained to sit on an elevated Y-shaped platform and broadcast its sonar sounds to the left and right to discriminate the dipole target from the monopole target. The bat was rewarded with a piece of mealworm offered in forceps following the completion of each correct response, which consisted of crawling forward onto the arm of the platform that faced the dipole target (see arrow in Fig. 1). An incorrect response consisted of crawling onto the arm of the platform that faced the monopole target and resulted in the trainer making a "shh" sound and delivering no mealworm reward. Prior experience showed that the "shh" served as a signal that the bat had made an error and it was not intrinsically aversive. Use of this cuing sound greatly speeds up training.

The experiment took place in a dark experimental chamber (5.2 m long × 3.4 m wide × 2.4 m high). The Y-shaped platform (15.5 cm in length and 20.0 cm in width) was mounted on a heavy Brunson optical tooling stand with a tripod base. A plastic dish to hold the mealworms ready for reinforcement was mounted on the same stand 30.0 cm below the platform and out of the bats' visual range. A dim light consisting of a cluster of three light emitting diodes (LEDs) was mounted above the mealworm dish, so the experimenter could see the mealworms in the dark. A ceiling mounted spotlight weakly illuminated the center of the platform (covering the bat but not the targets), so the experi-

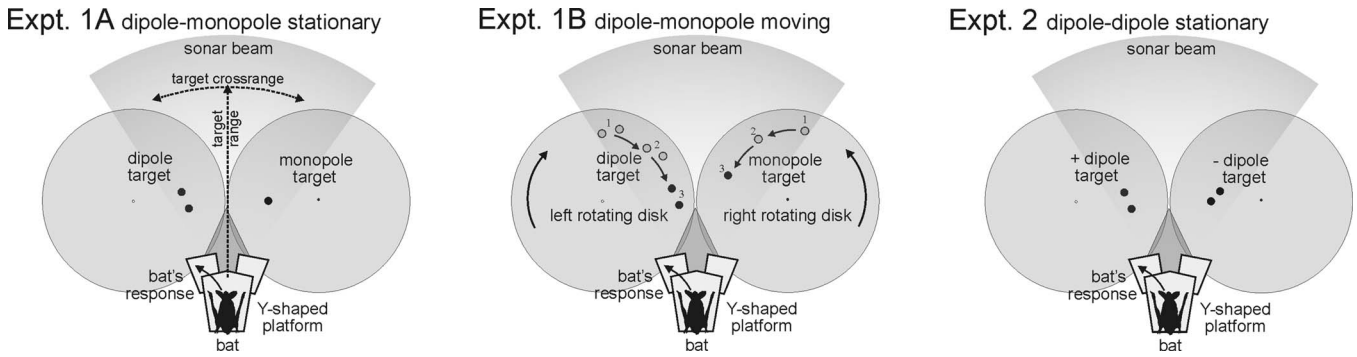


FIG. 1. Experimental setup and design of stimuli for experiments: experiment 1A, dipole versus monopole stationary; experiment 1B, dipole versus monopole moving; experiment 2, dipole versus dipole stationary. The positive stimulus is the 5 cm dipole target, and the bat's response is to move forward onto the arm of the Y-shaped platform facing this target to receive its food reward. The bat's broadcast beam is wide enough to ensonify both targets even when the bat's head is aimed at either one. In experiment 1B, the disks rotate (the left disk rotates clockwise; the right disk rotates counterclockwise) to bring the targets toward the bat during each trial. In experiment 2, the positive dipole cylinders were separated by 5 cm, whereas the negative dipole cylinders were closer together (1.0, 1.5, 2.0, 2.5, or 3.0 cm separation).

menter could see the bat but the bat was unlikely to be able to see the targets, which were flat black on a flat black surface. The stimuli were placed on two Plexiglas disks (diameter=91.5 cm, thickness=0.5 cm) controlled by a pulley system driven by a motor (Minarik Electric Co. model SH-14) that allowed the disks to rotate at various speeds. The Plexiglas disks were approximately 4 cm below the bottom of the right and left arms of the Y-shaped platform.

A downward facing black and white charge coupled device camera (DSP, Inc. model 15CB221) was mounted on the ceiling approximately 125 cm above the rotating disks. On each side of the camera were two IR lights (each an array of infrared LEDs 16.0 cm in length $\times$ 12.0 cm in width) that automatically turned on when the illumination in the room was low and trials were about to be conducted for a bat. Two Titley Electronic Ltd. ultrasonic microphones were located to the bat's left and right and 150 cm from the bats' observing position at the center of the platform. These microphones allowed for high-quality ultrasonic recordings to be made of the bat's signals during trials. Additionally, a mini-2 heterodyning bat detector (Ultra Sound Advice Ltd.) tuned to 28–29 kHz was mounted on a stand located 150 cm from the center of the platform and in front of the bat. The bat detector translated the bats' ultrasonic sounds down to the audio range, so the experimenter could listen to the bats' sonar sounds during trials via the auditory output of the bat detector. The low-frequency auditory output of the bat detector was not aversive to the bats, as there was no difference in their performance when it was turned on or off. All experimental trials were recorded on a digital video cassette recorder (Sony model GV-D800 NTSC "Video Walkman") by using Fujifilm Hi8 MP P6-120 digital video tapes. The bat detector's audio-frequency signals were stored on the sound track of this video tape to mark the time of occurrence of echolocation sounds. A Sony digital instrumentation/video recorder (model SIR-1000 W) was used to capture high-fidelity recordings of the ultrasonic sounds along with the video stream to record representative trials.

Two experimenters, a bat trainer and a recorder, were present for each test trial. At the beginning of each trial, the bat would crawl from the trainer's hand onto the back of the Y-shaped platform. Then, the bat would emit sonar signals for approximately 1–10 s and then walk forward toward one arm of the Y-shaped platform. When the bat reached the end of the arm, if it was correct (facing the dipole target), the trainer would then deliver the mealworm reward by presenting the mealworm piece in the forceps just in front of the bat's mouth. After the bat grasped the mealworm at the conclusion of a correct trial (or after the trainer said "shh" at the conclusion of an incorrect trial), the trainer would pick up the bat and hold it while it ate the mealworm for an intertrial interval of about 5 s. During this intertrial interval, the recorder would record the bat's choice on a list of trial alternations from left to right according to a pseudorandom schedule (Gellermann, 1933) and then reposition the targets for the next trial. The trainer shielded the bat from the movements of the recorder during the intertrial interval to avoid inadvertent cuing.

The bats were tested 5–6 days per week. On test days, each bat was run for a total number of trials to attain a predetermined number of correct, rewarded trials that was determined by the quantity of mealworms it could eat on that day to maintain its current body weight. Each bat typically worked through 30–60 trials per day. The first trial of the day was prompted, such that the mealworm was held in front of the arm of the platform with the correct target (the dipole). If during the test session the bat got three incorrect choices in a row, the next trial was also prompted. On all other trials, the mealworm was held out of view, so that the bat had to use its sonar sounds to discriminate between the targets. Prompted trials were not included in the data.

(a) Pretraining. Although during testing the cylinders in the monopole and dipole targets were the same in height, which is 1.27 cm, during the pretraining phase before testing, the monopole and dipole differed in size, so that the monopole was always shorter than the dipole. This added a size difference and thus a reliable echo amplitude difference to the monopole versus dipole stimuli that facilitated training the bats by a fading procedure of reducing the size difference. Figure 2(a) is a diagram of the different combinations of cylinders used for pretraining the bats. There were two different sets of training stimuli used in two different pre-

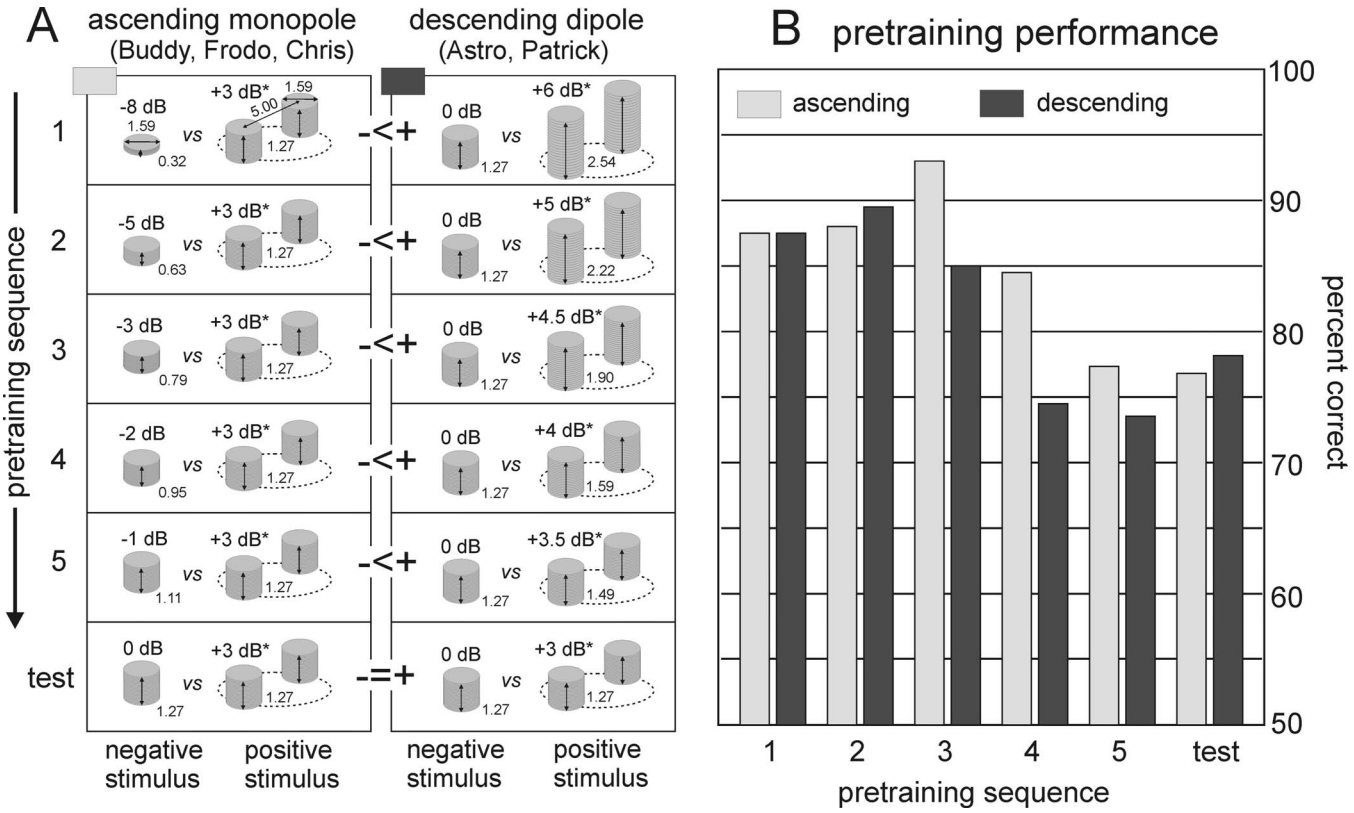


FIG. 2. (A) Arrangement of stimuli in two sequences for pretraining to bring bats to discriminate monopole and dipole with the same cylinder height (cylinder dimensions and spacing in cm). Three bats learned the test task with progressively increasing height of monopole cylinder (left panels; pretraining stages 1–5 culminating in the test task with 1.27-cm-high cylinders). Two bats learned the test task with progressively decreasing heights of dipole cylinders (right panels; pretraining stages 1–5 culminating in the same test task with 1.27-cm-high cylinders). Decibel values give target strengths of the monopole and dipole targets relative to a single 1.27-cm-high test cylinder. (* denotes an average 3 dB increase in dipole target strength relative to test-sized 1.27-cm-high monopole for most aspect angles.) (B) Mean performance of three bats in ascending pretraining sequence (light gray bars) and two bats in descending pretraining sequence (dark gray bars).

training regimes. Three bats (Buddy, Frodo, Chris) received the "ascending monopole" pretraining set, in which the cylinders in the dipole target were kept at their intended height of 1.27 cm, while the height of the cylinder in the monopole target started very short, which was at only 0.32 cm, and was increased in several steps to arrive at 1.27 cm [left side of Fig. 2(a)]. The other two bats (Patrick, Astro) received the "descending dipole" pretraining set, in which the monopole cylinder was kept at the intended height of 1.27 cm, while the cylinders in the dipole target started much higher, which were at 2.54 cm, and were decreased in several steps to arrive at 1.27 cm [right side of Fig. 2(a)]. For both pretraining sets, each bat completed a minimum of 200 trials on each target pairing and was made to achieve a mean of approximately 75%–95% correct responses (reaching asymptotic performance) before moving to the next pair in the pretraining series [in Fig. 2(a), starting with pair 1 and ending with pair 6].

(b) Experiment 1A: Stationary targets. In experiment 1A, the disks that the dipole and monopole test stimuli were placed on [see Fig. 1(a)] did not rotate during each trial. The targets were placed in different positions between trials, but they were stationary during individual test trials while the bat made its choice [Fig. 1(a)]. The targets were repositioned during each intertrial interval in three ways: (1) left/right position, (2) distance from the Y-shaped platform, and (3) dipole orientation with respect to the platform [examples shown in Fig. 3(a)]. First, the dipole target appeared on either the left or right disk as determined by a pseudorandom Gellermann series (Gellermann, 1933). Second, the dipole and monopole were placed at varying distances approximately 10–40 cm from the end of the arms on the Y-shaped platform. This resulted in three situations: the monopole was closer to the bat than the dipole [Fig. 3(A1), top], the dipole was closer to the bat than the monopole [Fig. 3(A2), middle], or the monopole and the dipole were approximately equidistant from the bat [Fig. 3(A3), bottom]. There were no three distinct positions for the targets, rather, the monopole and dipole were placed anywhere on a continuum of locations ranging from 10 to 40 cm from the platform. The cross range (left to right) distance between the dipole and monopole varied from roughly 15 to 40 cm from trial to trial. This varied the angular separation of the targets from the bat's position. Third, the orientation, or aspect angle, of the dipole was varied so that it could be in any possible position from 0° to 90° in either direction [see Fig. 3(a)]. The orientation in which the second cylinder of the dipole was directly behind the first cylinder was called 0°, and the orientation in which the two cylinders were side by side was called 90°. The positions, distances, angular separations, and orientation of the targets were semirandomly (spontaneously) determined by the recorder and altered between trials by using a small,

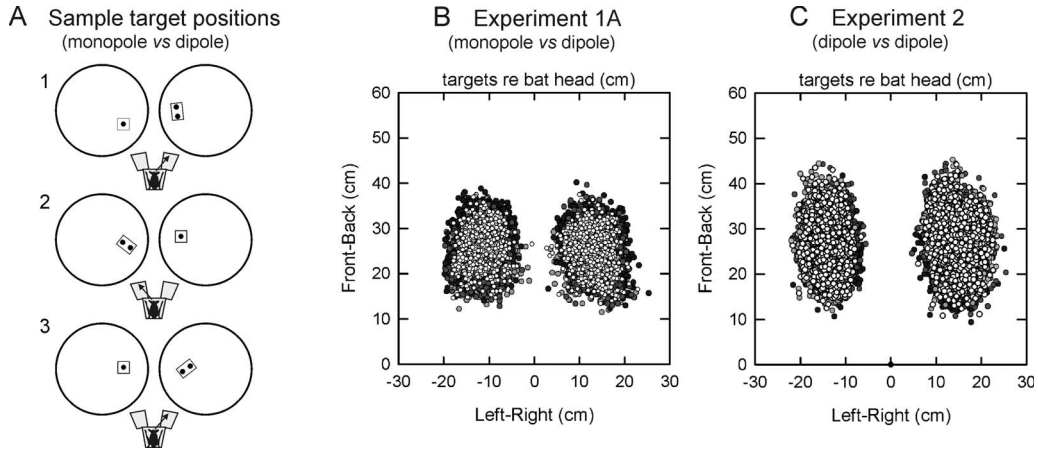


FIG. 3. (A) Three sample test trial configurations of the dipole and monopole for experiment 1A. The circles represent the disks (stimuli are shown larger than they actually were compared to the size of the disks). The Y-shaped platform with the bat is shown in front of the disks. At the top, the monopole is in front of the dipole and the dipole is oriented at approximately 5°. In the middle, the dipole is in front of the monopole and the dipole is oriented at 45°. At the bottom, the monopole and dipole are equidistant from the bat and the dipole is oriented at 45°. (B) Overlay of all positions for cylinders in the monopole and the dipole targets from experiment 1A. (C) Overlay of all positions for cylinders in the two dipole targets from experiment 2. The point (0,0) is the position of the bat's head. Each point represents the center of the each cylinder. The different cylinders (e.g., monopole, dipole near, dipole far) are plotted in different colors and overlayed. Plots B and C show that the target position varied from trial to trial by amounts that were, on average, several times larger than the target dimensions.

handheld LED flashlight while the trainer was holding the bat. The large quantity of test trials run for each bat ensured that all possible distances and orientations were included. The actual positions of the cylinders with respect to each other and to the bat were determined by digitizing the video records of all trials [see Fig. 3(b)]. Each of the five bats run in experiment 1A completed a minimum of 1000 trials.

(c) Experiment 1B: Moving targets. In experiment 1B, the large Plexiglas disks that the monopole and dipole stimuli were placed on rotated during the trials [see Fig. 1(b)]. The stimuli and the procedure were otherwise the same as in experiment 1A in that the left/right position, distance from the platform, angular separation, and orientation of the dipole target were changed during intertrial intervals. At the beginning of each trial, the dipole and monopole targets would be placed approximately 65 cm from the back of the Y-shaped platform, which is nearer the tops of the disks in Fig. 1(b) than they would have been placed for experiment 1A. During the ensuing trial, the disks would rotate inward (left disk would rotate clockwise; right disk would rotate counterclockwise), carrying the targets toward the bat just placed on the platform. As in experiment 1A, the bat would emit sonar sounds for 1–5 s and then walk down one of the arms of the platform. If the bat did not make a choice before the stimuli rotated past the platform (at a distance of about 3–6 cm from the ends of the arms), that trial was not counted as an incorrect choice. Instead, the bat was picked up and the same trial was restarted. It took about 5 s for the stimuli to reach the end of the platform as the disks rotated, but the bats made their choices before the stimuli rotated past the platform on about 85% of all trials. Two of the five bats (Chris and Astro) participated in experiment 1B. (The other three bats directly went to experiment 2.) Chris completed 1378 trials and Astro completed 693 trials with the moving stimuli.

4. Data analysis

Videotapes of all the test sessions were analyzed to determine the bats' performance in the first experiment. First, a single video frame representing the bats' choice point for each test trial was clipped from the videotape of the daily test session and then all the frames for a given bat on a given day were put together and saved in chronological order as a single video (avi) file by using PINNACLE STUDIO 9 (v 9.4.3, Pinnacle Systems Inc., 2004). The individual choice-point frame was chosen to capture the moment at which the bat started rapidly moving forward toward its final destination (either the correct arm or the incorrect arm of the platform). The bats often made their decision in the center of the platform but sometimes would walk toward one arm then change their direction and walk down the other arm. The frame that was clipped represented the bat's last turn toward its chosen target regardless of where the bat was on the platform.

The video file containing the clipped images from all the trials in a day's session was then imported into PEAK MOTUS (v 8.2, Peak Performance Technologies Inc., 2004) for analysis. In this program, six different spatial reference points were located and digitized for each test trial: (1) the bat's head (defined as the tip of the nose), (2) the left corner of the platform, (3) the right corner of the platform, (4) the center of the monopole cylinder, and [(5) and (6)] the center of each of the dipole cylinders. For each trial, the location of these points in the corresponding video image was used to calculate the distance from the bat's head to each of the cylinders and also to calculate the orientation of the dipole target.

5. Echo measurements

Echoes reflected by each of the targets were recorded using an "artificial bat" consisting of two Bruel & Kjäer model 4136 ("$\frac{1}{4}$ in.") condenser microphones separated by 3 cm to record the echoes and a centrally located custom-

built condenser loudspeaker to broadcast the test sounds [see Fig. 8(B) for the arrangement of the microphones and loudspeaker in relation to the target and the disk on which the target was rotated to different aspect angles]. The test signal consisted of a 1-ms-long FM sweep from 110 to 20 kHz that was generated by a National Instruments PCI-6111e digital-to-analog board in a Pentium-III computer. Echoes picked up by the microphones were amplified (1000×) and band limited to 15–100 kHz (Wavetek-Rockland variable bandpass filters) and digitized at a sampling rate of 500 kHz in a National Instruments PCI-6111e two-channel analog-to-digital converter board. Custom software written in LABVIEW and MATLAB provided for windowing the data and averaging the echoes reflected from each of the targets for 20 repetitions of the test signal.

To make the dipole target echo measurements, the dipole target was placed on the center of a Plexiglas rotating disk 30 cm away from the artificial bat [Fig. 8(b)]. The target was first oriented at 0° with respect to the artificial bat and ensonified for 20 repetitions of the test signal. (Each such data-collection procedure was repeated two times to check for possible acoustic transients.) Then, the circular table holding the target was rotated 2° clockwise and the same echo-measurement procedure was repeated. Following each recording of echoes, the target was again rotated 2° clockwise until echoes had been recorded form target aspect angles from 0° to 94° in increments of 2°. This provided a set of echoes that encompassed all possible orientations of the dipole. The monopole target was measured using the same setup and procedure as the dipole target, except that instead of rotating the monopole, it was measured at varying distances from the artificial bat, including 20, 25, 30, 35, and 40 cm. The set of pretraining stimuli was measured 30 cm away from the artificial bat. Individual cylinders reflected strong, short-duration specular echoes, and the peak-to-peak amplitudes of these primary reflections proved to be adequate to characterize the relative target strength for different-sized cylinders.

B. Results

1. Pretraining trials

Figure 4(a) plots the results from echo measurements of the various cylinders with different heights used in pretraining [see Fig. 2(a)]. The sequence of stimuli used in the two pretraining regimes—monopole ascending height and dipole descending height [Fig. 2(a)]—is marked on the graph in Fig. 4(a). During pretraining, the negative stimulus (monopole) always had a lower target strength than the positive stimulus (dipole), first, because the dipole returned two reflections for the monopole's one reflection and, second, because the heights of the cylinders always favored the dipole with several decibels of additional target strength during either of the pretraining approaches to the test stimulus with equal cylinder heights. As expected, differences in peak-to-peak echo amplitude remained a reliable cue during the pretraining trials, with the higher amplitude echoes always from the positive stimulus.

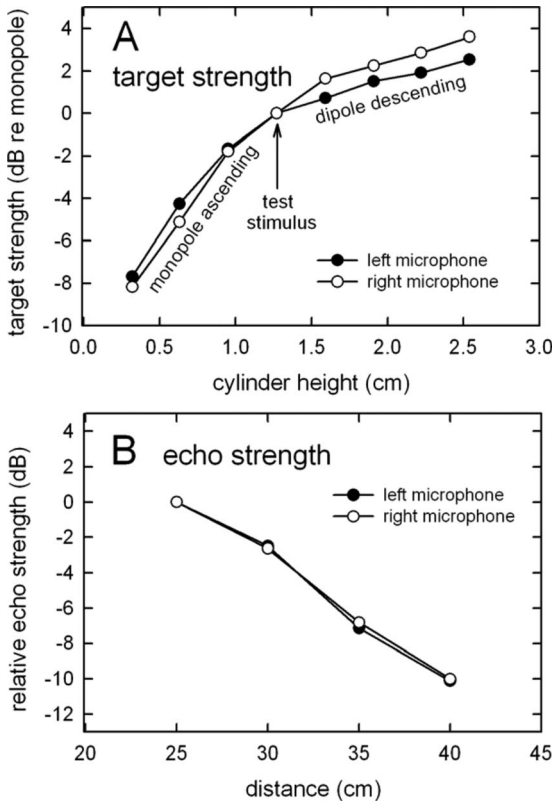


FIG. 4. (A) Graph showing peak-to-peak amplitudes of echoes recorded with two microphones from cylinders of different heights in decibels relative to peak-to-peak amplitude of echoes from 1.27-cm-high test cylinder (target distances are 30 cm; see Fig. 2 for relationship to pretraining sequences). (B) Graph showing the relationship between the target distance and relative echo strength for a 1.27-cm-high cylinder.

Figure 2(b) displays the bats' performance on the pretraining trials for the two pretraining regimes illustrated in Fig. 2(a). In general, their pretraining performance was related to the difference in height between the monopole and the dipole targets and, thus, the difference in echo strength [Fig. 4(a)]. The bats' performance was high when the difference in height between the monopole and dipole was large (stages 1–3 of either pretraining series) and lower when there was no difference in height between the dipole and the monopole [i.e., the test condition in Fig. 2(b)]. Most importantly, there was no significant difference in performance between the bats in the ascending monopole and the descending dipole pretraining regimes when they reached the final test stimuli [$t(121)=1.93$, $p>0.05$], so the data from the two groups of bats were pooled for subsequent analyses of performance.

2. Experiment 1A: Stationary targets

Figure 3(b) shows the locations of the individual cylinders in the dipole and the monopole targets on all trials of experiment 1A. Target distances roughly varied over a 2:1 range, mostly from 15 cm to almost 40 cm. Differences in range between the dipole and the monopole mostly varied over ±10 cm (see below). The angular separation of the targets varied from about 25° to 120° relative to the bat's location. Because the individual cylinders were presented at different distances from the bat, the amplitude of reflections at

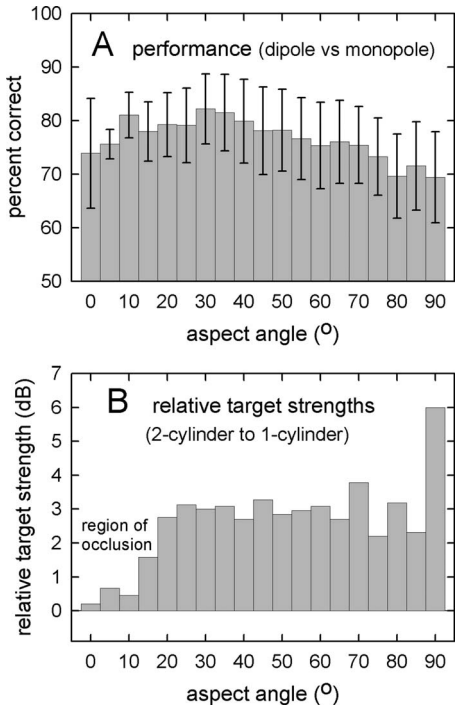


FIG. 5. (A) Performance as a function of dipole aspect angle for experiment 1A (mean performance of all five bats is shown; error bars show standard deviations). The dipole could be rotated either clockwise (e.g., +40°) or counterclockwise (e.g., –40°), but these two categories are plotted together (shown as 40°). Chance performance is 50% correct. The average performance of the bats was significantly above chance for every dipole orientation (summed binomial test, $p<0.001$). (B) Target strength of the dipole relative to the monopole for different aspect angles.

the bat's ears necessarily differed, too. Figure 4(b) shows the strength of echoes reaching the microphones from a single test cylinder at different distances. Approximately 90% of all trials in experiment 1A involved individual cylinders at distances of 25–40 cm, and the plot in Fig. 4(b) depicts the relationship between the distance and the strength of echoes acting as stimuli. Note that the spread of differences in echo amplitude as a function of distance in Fig. 4(b) is comparable to the spread of amplitudes as a function of cylinder height during the pretraining series in Fig. 4(a).

The bats were able to discriminate between the monopole and the dipole across all orientations of the dipole, as shown in Fig. 5(a). From the bat's position, by considering the dipole as a single object, 0° corresponds to an end-on view, while 90° corresponds to a side-on view. The mean choice accuracy of the five bats was significantly above chance for all dipole orientations (summed binomial test, $p<0.001$). The performance of individual bats is shown in Table I. Choice accuracy was above chance for all dipole orientations for three of the bats (Frodo, Chris, and Patrick). The other two bats' (Buddy and Astro) performance was above chance for all orientations except 80°. The bar graph in Fig. 5(a) shows only a moderate dependence of the bats' performance on the orientation of the dipole target, with differences between bats equaling or exceeding differences between orientations (see Table I). The highest mean performance was at the 30° orientation and the lowest mean performance was at the 80° orientation.

TABLE I. Performance accuracy as a function of dipole aspect angle for each of the five bats in experiment 1A. (The number of trials completed for each bat in each orientation is shown in italics under the accuracy value. Values with an asterisk are not significantly different from chance. All other values are significant at $p<0.001$.)

Dipole orientation (deg)	Bat					*M* (Total)
	Buddy	Frodo	Chris	Patrick	Astro	
0	80.3% (*61*)	79.7% (*64*)	76.6% (*47*)	82.2% (*45*)	76.7% (*60*)	79.1% (*277*)
10	68.4% (*117*)	77.4% (*133*)	77.3% (*128*)	84.7% (*85*)	73.9% (*138*)	75.9% (*601*)
20	74.5% (*153*)	83.0% (*88*)	78.2% (*87*)	84.1% (*69*)	72.8% (*103*)	77.6% (*500*)
30	79.3% (*179*)	79.6% (*98*)	84.5% (*103*)	89.1% (*137*)	76.2% (*126*)	81.6% (*643*)
40	78.9% (*208*)	73.6% (*140*)	83.6% (*134*)	83.8% (*105*)	75.0% (*128*)	78.7% (*715*)
50	78.8% (*184*)	73.8% (122)	74.2% (*124*)	84.3% (*127*)	79.2% (*130*)	78.2% (*687*)
60	79.1% (*158*)	71.8% (*117*)	80.4% (*102*)	78.4% (*116*)	68.3% (*123*)	75.6% (*616*)
70	72.0% (*118*)	68.4% (*79*)	72.2% (*72*)	77.6% (*67*)	70.6% (*92*)	72.0% (*428*)
80	*58.4% (*77*)	73.1% (*78*)	78.4% (*51*)	73.4% (*64*)	*61.8% (*68*)	68.3% (*338*)
90	75.5% (*94*)	66.3% (*101*)	67.0% (*88*)	75.6% (*119*)	69.5% (*82*)	71.1% (*484*)
M/Total	75.6% (*1349*)	74.5% (*1020*)	77.7% (*936*)	81.8% (*934*)	72.9% (*1050*)	76.3% (*5289*)

3. Experiment 1B: Moving targets

The two bats used in this experimental condition (Astro and Chris) continued to discriminate between the monopole and the dipole when the disks supporting the targets were rotating during the test trials [see Fig. 1(b)]. Both bats' mean choice accuracy in experiment 1B for moving targets was significantly above chance (summed binomial test, $p < 0.001$). Astro's performance for stationary targets was not significantly different than its performance for moving targets [stationary: 72.9% on 1050 trials, moving: 69.3% on 655 trials; $t(33)=1.66$, $p>0.05$]. Chris' performance for stationary targets was significantly higher than its performance for moving targets [stationary: 77.7% on 936 trials, moving: 73.8% on 1378 trials; $t(62)=2.39$, $p<0.05$].

C. Discussion

Figure 5(b) shows the target strength of the dipole target relative to the monopole target at the same distance but different aspect angles. At an aspect angle of 0°, the two cylinders are located one behind the other, and their reflections arrive 290 μs apart. At an aspect angle of 90°, the two cylinders are side by side, and their reflections arrive at approximately the same time (time difference of 0 μs, give or take the effect of the bat's head movements). If the bats were using echo amplitude to perform the task, the best performance should have been sharply highest at the 90° orientation since that is where the difference in amplitude between the dipole and monopole echoes is largest. Instead, the profile of performance at different dipole aspect angles does not resemble the profile of echo amplitude differences at different aspect angles.

The plot in Fig. 5(a) shows reduced performance at some dipole orientations relative to others, but the relation is not monotonic, as would be expected if the presence of two cylinders is more difficult to detect when the aspect angle is at one extreme (0°) or the other (90°). The bats' performance is somewhat lower at both extremes of the dipole angle range, at 0°–5° and at 75°–90°, than at intermediate angles. It might be expected that dipole aspect angles that provide for a greater separation of the cylinders (<60°) would be easier to detect, but performance is highest for angles around 30°–40° and gradually lower for both higher and lower angles. It thus does not seem as though the aspect angle is behaving as the critical feature of the dipole that determines its discriminability from the monopole.

There are other relations among the reflecting points—the cylinders—making up the targets that could have caused the effect shown in Fig. 5(b) without the aspect angle itself being directly responsible. Besides the dipole orientation itself, the most important additional feature of the stimulus ensemble is the placement of the monopole and dipole targets relative to each other and to the bat. Although the monopole and dipole targets are separated from each other in *direction* because they individually appear on the bats left and right, many of the trial-to-trial positions involve them being at the same *distance* from the bat. As shown in Fig. 3(b), these directions and distances vary over a large range across all trials in experiment 1A, but how does the bats' performance vary when the trials are sorted into different groups according to the relative distances of the cylinders? Figure 6(a) shows that the near cylinder of the dipole was closer to the bat than the monopole on slightly over half of the trials (black circle data points). On average, the near cylinder of the dipole was about 1 cm closer to the bat than the single cylinder of the monopole, but the spread of differences is large, over a span of ±6 cm on the majority of trials. On average, the far cylinder of the dipole was 2–3 cm farther away than the monopole on the majority of the trials (gray triangle data points), but, again, the spread of differences is large. In fact, there were several hundred trials where the far cylinder of the dipole was closer even than the monopole. Figure 6(a) also shows the distribution of differences in distance to two cylinders of the dipole (white square data points). Of course, the near cylinder of the dipole was closer to the bat than the far cylinder by definition, but the aspect angle varied over the entire 0°–90° span, so there were nearly 600 trials with both cylinders at the same distance [0 cm in Fig. 6(a)].

The performance of the bats as a function of the distance to each cylinder in the dipole target with respect to the monopole target or the cylinders in the dipole target with respect to each other is shown in Fig. 6(b). Across trials, the errors made by the bats in experiment 1A proved to be uniformly related to the relative distances from the bat to the individual cylinders. Whenever the trial-to-trial shifts in the locations of the monopole and dipole targets resulted in the monopole being at about the same distance [target range=0 in Fig. 6(b)] as either of the two cylinders of the dipole [e.g., see Fig. 3(a)], the bats' performance declined. Similarly, whenever either of the cylinders of the dipole were located at about the same distance [which occurred when the aspect angle approached 75°–90°, see Fig. 5(a)], the bats' performance declined. In all three cases, the echoes from one of the cylinders arrived at about the same delay as the echoes from the other cylinder, and a masking effect could have occurred due to a simultaneous reception of two reflections that had to be distinguished for a successful identification of the dipole versus monopole. Figure 7 illustrates the origin of this masking effect. In the diagram, going from Fig. 7(a) to Fig. 7(e), the dipole target rotates from an aspect angle of 0° (end-on view) to an angle of nearly 90° (side-on view). Each cylinder in the dipole, as well as the cylinder in the monopole, creates a trace in the bat's perception that is represented by the gray distribution located below either of the dipole cylinders or above the monopole cylinder. If these traces are separated along the horizontal echo delay or target range axis in the diagram, they are separable to the bat, so that each of the cylinders is perceived to be at its corresponding location on the horizontal axis. However, if any two of the cylinders are located at about the same distance from the bat, then their echoes arrive at about the same delay, and the gray distributions that depict those cylinders in the bat's perception overlap with each other and appear less distinguishable. The masking effect is absent in Figs. 7(a) and 7(b) but it is very strong in Fig. 7(c) and moderately strong in Figs. 7(d) and 7(e). Note that there can be interference between the monopole and either of the dipole cylinders or between the two

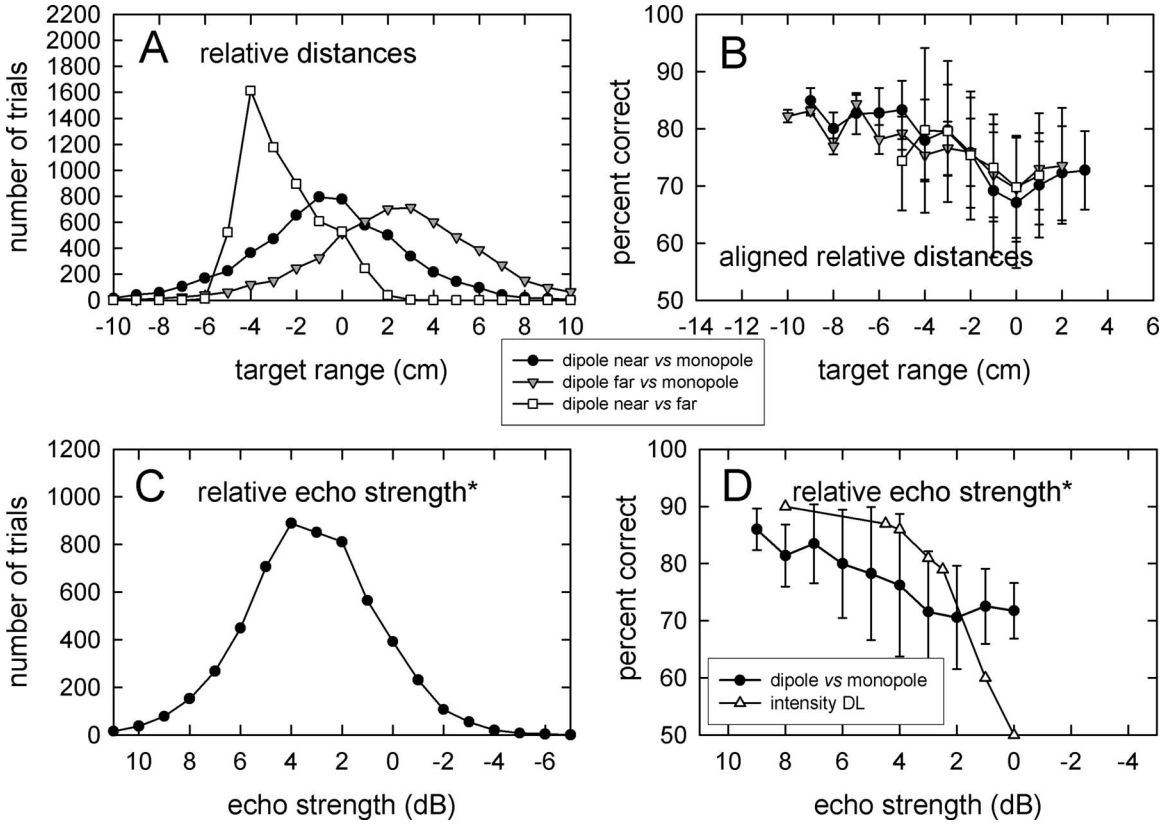


FIG. 6. (A) Distribution of relative distances to each of the cylinders in experiment. 1A for all trials of five bats. Curves show distances from the monopole to the near and far cylinders of the dipole and distances from the near to the far dipole cylinders. (B) Plots of mean performance of five bats separately sorted by trials and aligned according to the relative distance of each cylinder. (C) Distribution of overall echo strengths for echoes from the dipole relative to the monopole. (* assumes random aspect angle with an average 3 dB increase in echo strength.) (D) Plot of mean performance of five bats in experiment 1A for various differences in overall echo strength (black circles) for comparison with performance in echo intensity discrimination (white triangles; Simmons and Vernon, 1971).

dipole cylinders themselves. The occurrence of this masking is visible in the bat's performance whenever one of the cylinders is aligned at the same distance as either of the other cylinders. The curves in Fig. 6(b) are aligned to the same scale of differences in distance between pairs of cylinders. Their superimposition reveals that the masking effects among the three cylinders are of equivalent strengths. That is, the cylinders were roughly equal to each other in their perceptual salience and thus their potential for reducing performance when their reflections were simultaneous, as opposed to when their reflections arrived at somewhat different delays (as explained in Fig. 7). In Fig. 6(b), the range of differences in performance from about 85% correct responses down to 67% correct responses, depending on the alignment of the cylinders in distance, is about the same as the range of differences in performance in Fig. 5(a) as a function of dipole aspect angle, from 82% correct responses at 30° to 70% correct responses at 90°. This finding suggests that the apparent dependence of the bats' performance on aspect angle *as such* may instead be caused by the range differences between the near and far dipole cylinders relative to the bat. However, the decline in performance for aspect angles at 0°–5° in Fig. 5(a) cannot be related to simultaneity and masking of reflections from the two cylinders in the dipole because, at these angles, the dipole cylinders are about 5 cm apart from the bat's view, so their reflections arrive at or near the maximum separation of 290 μs.

Besides the direct effect that the presence of two cylinders has on the amplitude of echoes from the dipole relative to the monopole [average difference of 3 dB; see Fig. 5(b)], the dipole's spread along the range axis ensures that, on average, the nearer cylinder of the dipole will be closer to the bat than the farther cylinder, although the distribution of relative distances is broad [see Fig. 6(a)]. Apart from creating conditions for masking when the cylinders are at the same distance [Figs. 6(b) and 7], the nearer cylinder of the dipole will, on average, return a stronger reflection than the single cylinder of the monopole just because that cylinder is nearer on many trials. Therefore, its reflections undergo less spreading losses and atmospheric absorption. Figure 4(b) shows the relation between the cylinder distance and echo strength to serve as a model for estimating the strength of echoes on test trials in the experiment from the locations of the cylinders in the video reconstructions of their positions relative to the bat. From these reconstructions, Fig. 6(c) plots the relative amplitude of echoes from the dipole and the monopole targets. These estimates of echo strength are based on measurements of the amplitude of echoes from cylinders at different distances in Fig. 4(b) and assuming an additional mean difference of 3 dB due to the presence of two cylinders in the

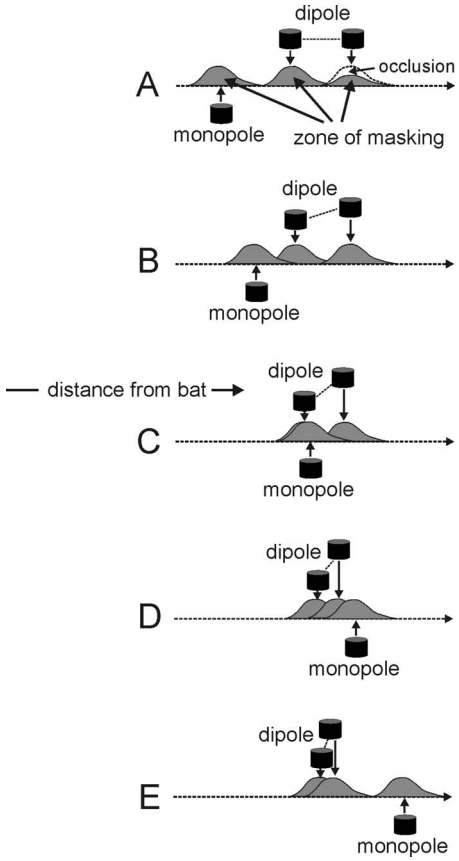


FIG. 7. Potential masking effects to be expected whenever one of the target's cylinders is aligned at the same distance from the bat as another of the cylinders. Then their echoes arrive at the same time to create mutual masking. Gray peaks represent zone of masking for each cylinder. The dipole also rotates to different aspect angles in each drawing. (A) Monopole and dipole cylinders are all located at different distances for no masking. Note that the dipole is at 0°, and the echo from the second dipole cylinder is partly occluded by the first cylinder. (B) Monopole cylinder is closer to the first dipole cylinder but masking is still largely absent. (C) Monopole cylinder is aligned at the same range as the first dipole cylinder, and strong masking occurs. The second dipole cylinder is further away and is thus unaffected, but masking of the first cylinder is sufficient by itself to conceal the dipole. (D) Monopole cylinder is aligned at a similar range to both dipole cylinders and a very strong masking occurs. The dipole aspect angle also brings both dipole cylinders into positions to mask each other. In this condition, mutual masking affects all three cylinders. (E) Monopole cylinder is farther away than either dipole cylinder, but the dipole cylinders are at a similar range and they mask each other.

dipole rather than one cylinder. For most trials in Fig. 6(c), the echoes from the dipole were 2–5 dB stronger than the echoes from the monopole due to the relative distances, although for an appreciable number of trials, the range was −1 to 8 dB. Moreover, there are about 400 trials where the echoes from the two targets were of equal strength.

Figure 6(d) plots the performance of the bats with data regrouped according to the relative strength of echoes from the dipole versus the monopole. On this scale of echo strengths, the bats' performance (black circle data points) extends from about 80% to 85% correct responses down to about 70% correct responses. Figure 6(d) also plots the only data from an experiment that directly tested echo amplitude discrimination by big brown bats (white triangle data points; from Simmons and Vernon, 1971). The curve for echo amplitude discrimination is very steep for small amplitude changes, rising from chance near 50% correct responses for 0 dB difference to 80% correct responses for 3 dB difference. In contrast, the curve showing the performance of the bats in experiment 1A as a function of differences in the strength of echoes from the dipole and monopole targets at different distances is very shallow over this same range of amplitude differences. Most notably, when the dipole versus monopole echo amplitude difference ranges from 3 dB down to 0 dB in Fig. 6(d), the bats' performance stabilizes and does not fall below 70% correct responses, whereas performance rapidly falls to 50% correct responses in the echo intensity discrimination experiment. By using the binomial distribution, the probability that 70% correct responses would occur by chance is $\ll 0.01$. The divergence of the two curves in Fig. 6(d) over this 3–0 dB range indicates that the bats did not distinguish the dipole target from the monopole target as an echo amplitude discrimination task, which is a conclusion reinforced by the difference between echo strength as a function of aspect angle [Fig. 5(b)] and the bats' performance as a function of aspect angle [Fig. 5(a)]. In neither analysis does echo amplitude betray an effect on performance compared to the definite effect shown by the congruence of the curves in Fig. 6(b). Except for the anomalous decline in performance for aspect angles of 0°–5° [Fig. 5(a)], the masking effect of one cylinder on another appears to dominate the bats' performance and is responsible for the apparent dependence of performance on aspect angle.

Figure 8 shows the results of echo measurements for the monopole target and the dipole target at different orientations but the same overall distance. Figure 8(b) is a diagram of the echo-measurement procedure, showing the placement of the target in relation to the loudspeaker and the two microphones. These echoes are illustrated by their spectrograms. The spectrogram for the reflection from the monopole in Fig. 8(a) is virtually identical to the spectrogram of the broadcast signal because the monopole echo is completely dominated by the specular return from the front face of the cylinder. This spectrogram shows the first harmonic of the test signal sweeping from 100 kHz down to 20 kHz and the weaker second harmonic sweeping from near 130 kHz down to 40 kHz. (This second harmonic is caused by distortion of the loudspeaker when it is driven to generate a FM signal with a sound pressure of 100 dB sound pressure level (SPL) at a distance of 10 cm.) There is a single, sharply defined sweep for each harmonic in the echo returned from the monopole. Figure 8(d) shows spectrograms for echoes from the dipole at orientations from 0° to 90° in 10° steps. The corresponding target outlines in Fig. 8(c) show the incident and reflected sounds in relation to the locations of the two cylinders at these aspect angles. In Fig. 8(d), at aspect angles of 0° and 10°, the dipole target returns reflections whose spectrograms resemble the spectrogram from the monopole [Fig. 8(a)] in that each spectrogram contains a single prominent FM sweep for the first harmonic and a second single sweep for the second harmonic. In Fig. 8(d), as the dipole aspect angle opens more, at 20°, 30°, 40°, and 50°, there are two different sweeps for the first harmonic and the second harmonic in the spectrograms. At these aspect angles, the range separation of the cylinders is large enough [Fig. 8(c)] for

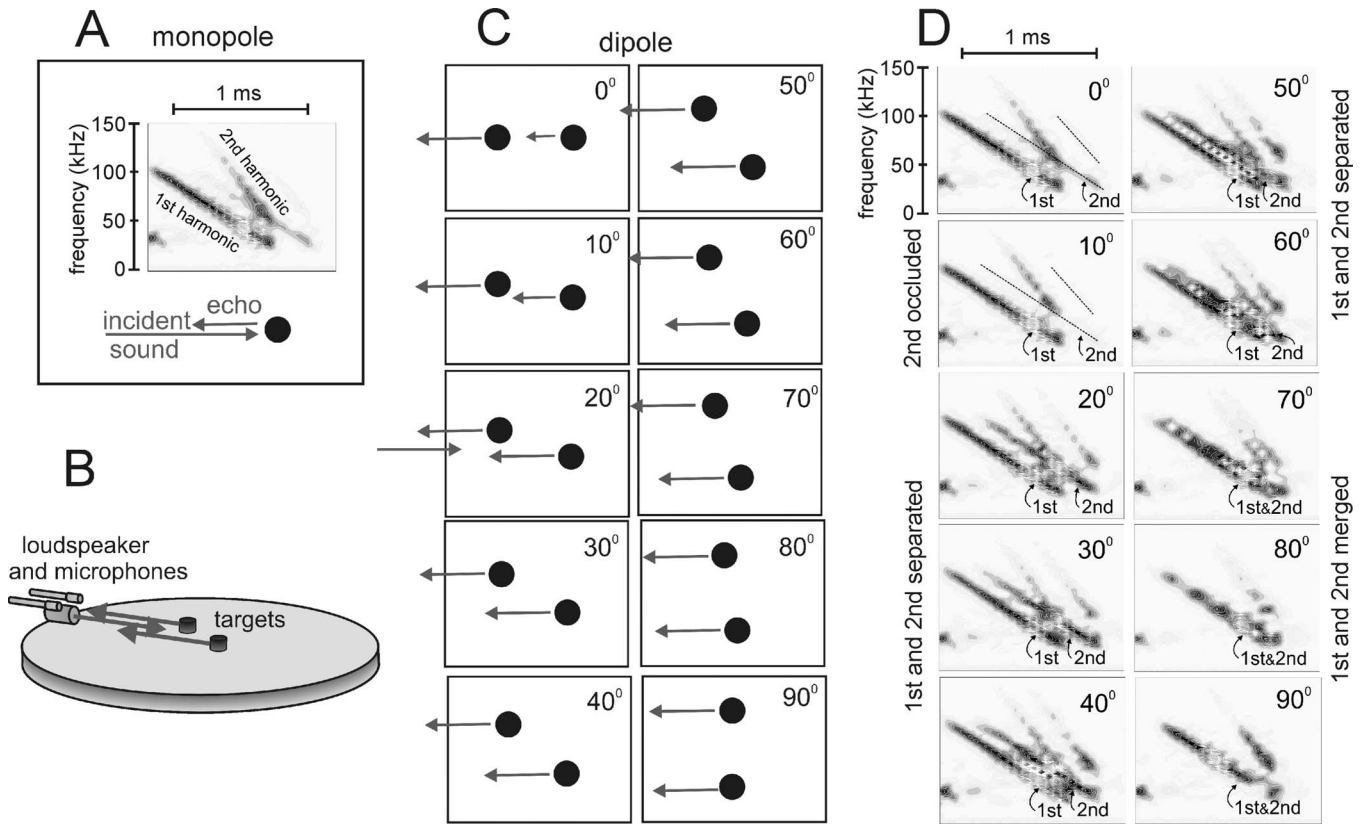


FIG. 8. Spectrograms of echoes from monopole and dipole targets. (A) Spectrogram of monopole echo. (The transmitted sound has an identical spectrogram.) The signal contains two frequency sweeps, a first harmonic sweeping down from 100 to 23 kHz and a second harmonic sweeping down from about 130 to 45 kHz. The arrows show the direction of transmitted sound and echoes. (B) Diagram showing setup for broadcasting sounds and recording echoes from two microphones (see Sec. II A). (C) Aspect angles of dipole target. (D) Spectrograms of echoes from the dipole target at different aspect angles (see text for details). The dashed lines in the spectrograms for 0° and 10° aspect angles show where the echo for the second cylinder would be if it was not occluded by the first cylinder.

each harmonic to be separately registered from each cylinder. However, at an aspect angle of 60°, the two cylinders are near enough in range that the sweeps in the spectrogram begin to merge. Finally, at aspect angles of 70°, 80°, and 90°, the spectrograms in Fig. 8(d) show overlap between the FM sweeps reflected by the two cylinders, which are now very close together in range [Fig. 8(c)]. From Fig. 8(d), the spectrograms for the echoes of the monopole and the dipole at different orientations show a shift in the nature of the information that reveals whether the target is a dipole or a monopole. It is unknown whether the spectrograms themselves give a direct representation of the acoustic information processed by the bats, but we are assuming that similar effects appear in whatever acoustic information is used by the bats.

No single acoustic feature in the time domain or the frequency domain would allow the bats to distinguish between the dipole and monopole targets from these spectrograms. At angles of 0° and 10°, the dipole's second cylinder is located behind the first cylinder and its echoes are partly occluded [dashed lines in Fig. 8(d) show expected locations of the second-cylinder echo, which is too weak to emerge in these two spectrograms]. At angles of 20°–60°, the echoes from both cylinders are strong and well separated in the spectrograms. At angles of 70°–90°, the spectrograms of the echoes from the two cylinders merge together into a single spectrogram that has a series of interference peaks and notches running along its ridges. This occurs when the time separation of the echoes becomes smaller than the horizontal width of the spectrogram ridge (spectrogram integration time). Thus, by using the criterion of two recognizable sweep traces for each harmonic, at 0°–10° and also at 90°, it is harder to determine that there are two cylinders in the dipole, while at 20°–80°, it is easier. In contrast, by using the criterion of ripples caused by the overlap of sweeps in the spectrogram, it is easier to determine which target is the dipole at 70°–80° than at other angles. Because the acoustic information for revealing the presence of two cylinders is different at 20°–60° (time domain cue of separable echoes) than that at 70°–80° (frequency domain cue of ripple pattern of interference), the bat may instead perceive the objects in a different format, which is one that combines the dual dimensions of the proximal, acoustic representation into a unitary distal representation that is more nearly spatial in nature. Use of spatially dimensioned perceptions would account for the independence of the bats' performance of the dipole aspect angle beyond the effects of masking between the individual cylinders in the targets. Further behavioral experiments are needed to shed light on this possibility.

III. EXPERIMENT 2: DIPOLE TARGET VERSUS DIPOLE TARGET

The stimuli for the dipole versus monopole experiments (1A and 1B) were intrinsically asymmetrical in several respects. Whether the bats perceived the overall echo strength or the number of reflecting sources, the dipole should, for the most part, be readily distinguished from the monopole. The analysis of results described above dissociated the bats' performance in experiment 1A from echo strength or from any one acoustic feature such as spectral ripples from interference or separate sweeps in spectrograms. Instead, this analysis tied the bats' performance to perception of the pair of cylinders, implying that the dipole was identified as having two reflecting parts and the monopole as having only one part. Variations in performance during experiment 1A appear to be caused by masking that occurs whenever any two of the three cylinders are located at about the same distance from the bat, so that their reflections coincided in time—which is a condition that has been demonstrated to cause masking in previous studies (Simmons *et al.*, 1989). However, the essential spatial asymmetry of the targets still remains with regard to the number of cylinders. To remove this asymmetry, experiment 2 replaced the monopole target with a new dipole target that differed only in having a shorter spacing between the cylinders. This transformed the bat's task from discriminating between targets that differed in type to discriminating between two similar dipole objects that differed only in one dimension, which is their length or dipole spacing. As in experiment 1, the targets were shifted in aspect angle, direction, range, and cross range separation from one trial to the next, which prevented any one acoustic cue or spatial location from serving as an indicator of the presence of the positive dipole on the bat's left or right.

A. Method

1. Animal subjects

Three of the same bats (Buddy, Frodo, and Patrick) from experiment 1 were tested.

2. Stimuli

The same dipole target from experiment 1 (two 1.59-cm-diameter cylinders separated by 5.0 cm) was the rewarded stimulus (positive dipole). There were five additional dipole targets used as negative, or unrewarded, stimuli. They were identical to the positive dipole target except for the spatial separation between the two cylinders within each target, which was 1.0, 1.5, 2.0, 2.5, or 3.0 cm at different stages of experiment 2. Each of these negative dipole targets was paired with the standard 5.0 cm dipole target in succession, starting with the shortest (1.0 cm) and ending with the longest (3.0 cm). The same negative dipole target was used for all the trials within each daily session (20–50 trials per session, where the number of trials depended on the quantity of mealworms each bat could eat on that day to maintain its current body weight). Negative dipole targets were changed between blocks of sessions. Individual bats completed a minimum of 400 trials for each negative dipole target (five to six sessions per week).

3. Experimental setup and procedure

Figure 1(c) shows the positive dipole versus negative dipole task. For each bat, experiment 2 immediately began after experiment 1 ended. During each trial, while the bat sat on the Y-shaped platform and examined the objects to make its choice, the targets remained stationary. Then, during each intertrial interval, the targets were repositioned in three ways: (1) the positive dipole appeared on the left or right as determined by a pseudorandom Gellerman sequence, (2) the positive and negative dipole targets were placed at variable distances from the bat on the platform, and (3) the orientations, or aspect angles, of both the positive and negative dipole targets were randomly changed (independently of one another). The distance of the dipole targets from the platform and the orientation of both dipole targets on each trial were spontaneously determined by the recorder.

The general setup for experiment 2 was the same as experiment 1, but the procedure was different in that the trials were run fully double blind to control for inadvertent cuing (the trainer was unaware whether the positive stimulus appeared on the left or right side until after the bat has made its choice). This precaution was taken because the targets now differed very little in structure, shape, or target strength, leaving the procedure especially sensitive to potential cues originating from the experimenters themselves. Two experimenters, a bat trainer and a recorder, were present for each test trial. At the beginning of each trial, the trainer, which was holding the bat, would be faced away from the platform. The recorder would place the stimuli in front of the Y-shaped platform. Once the stimuli were in position, the trainer then turned to face the Y-shaped platform and the bat crawled from the trainer's hand to the back of the platform. Since there were no visible lights on in the room (with the exception of the weak LED lights under the platform), the trainer was unable to see the stimuli. The bat would emit sonar signals for approximately 1–10 s and then walk forward toward one arm of the Y-shaped platform. When the bat reached the end of the arm, if it was correct (facing the dipole target), the recorder would tap the trainer on the shoulder. The trainer would then deliver the mealworm reward by presenting the mealworm piece in the forceps just in front of the bat's mouth, and then pick up the bat. If the bat was incorrect, the recorder would say "shh," which prompted the trainer to pick up the bat without delivering the mealworm reward. At the conclusion of the trial, the trainer turned away from the platform for an intertrial interval of about 5 s as the recorder recorded the trial and repositioned the stimuli for the next trial.

4. Data analysis

Videotapes of all test sessions with the 2.0, 2.5, and 3.0 cm negative dipole targets were analyzed for each bat (except for Frodo). For Frodo, 21 out of 47 sessions with the 2.0 cm target and 4 out of 51 sessions with the 3.0 cm target were not analyzed due to video errors. Videotapes of the test sessions were analyzed in the same manner as in experiment 1, except that a new digitization point was added for the second cylinder in the negative dipole.

TABLE II. Performance accuracy as a function of stimulus condition (negative dipole spatial separation) for each of the three bats in experiment 2. (The number of trials completed for each bat in each condition shown in italics under the accuracy value. Buddy did not complete the 2.5 or 3.0 cm negative dipole stimulus conditions.)

Negative dipole (cm)	Bat			*M* (Total)
	Buddy	Frodo	Patrick	
1.0	73.3% (*438*)	71.5% (*1302*)	75.8% (*1132*)	73.5% (*2872*)
1.5	75.2% (*1231*)	75.7% (*1413*)	75.5% (*1098*)	75.5% (*3742*)
2.0	67.8% (*574*)	77.1% (*1994*)	71.4% (*511*)	72.1% (*3079*)
2.5	NA	71.8% (*2143*)	71.1% (*570*)	71.4% (*2713*)
3.0	NA	63.2% (*549*)	59.8% (*433*)	61.5% (*982*)
M/Total	72.1% (*2243*)	71.9% (*7401*)	70.7% (*3744*)	70.8% (*13388*)

5. Echo measurements

Echo measurements of the five negative dipole targets were made using the same setup and procedure as those in experiment 1.

B. Results

Two bats (Frodo and Patrick) completed test sessions with all five negative targets (1.0, 1.5, 2.0, 2.5, and 3.0 cm dipoles), while a third bat (Buddy) completed test sessions with three of the targets (1.0, 1.5, and 2.0 cm dipoles). Figure 3(c) shows the locations of the individual cylinders in the positive and negative dipole targets on all trials of experiment 2. Target distances varied by about 2:1 in range, mostly from 15 cm to almost 45 cm. Differences in range between the cylinders in the dipoles mostly varied over ±15 cm (see below). The angular separation of the targets varied from about 30° to 120° relative to the bat's location. As in experiment 1A, because the individual cylinders were presented at different distances from the bat, the amplitude of reflections at the bat's ears necessarily differed [see Fig. 4(b)].

The three bats were all able to discriminate between the positive dipole and each negative dipole, as shown in Table II. Performance (percent correct responses) for the three bats was significantly above chance (50%) for all tested combinations of the positive dipole versus the negative dipole (summed binomial test, $p<0.001$). The bats' performance progressively declined as the spatial separation of the cylinders in the smaller (negative) dipole approached the spatial separation of the cylinders in the larger (positive) dipole. The highest mean performance was for the 5 cm positive dipole versus the 1.5 cm negative dipole (75.5%), and the lowest mean performance was for the 5 cm positive dipole versus the 3.0 cm negative dipole (61.5%). Across bats and conditions, performance in experiment 1A (76.3% correct responses from Table I) was slightly higher than that in experiment 2 (70.8% correct responses from Table II).

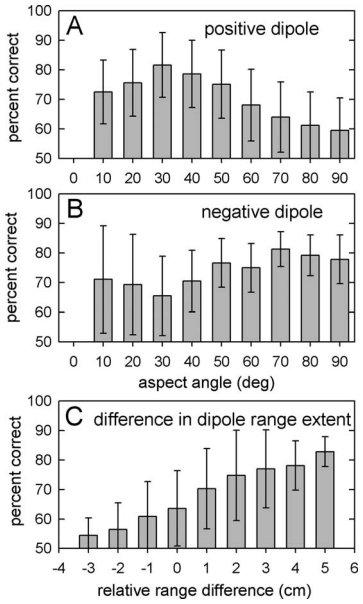


FIG. 9. (A) Graph of the performance of three bats as a function of positive dipole aspect angle in experiment 2. (B) Graph of the performance of the three bats as a function of negative dipole aspect angle in experiment 2. In plots A and B, no data are shown for the 0° orientation because there were less than 100 trials in that category. (C) Graph of performance as a function of the difference in the extent of the dipoles along the range axis in experiment 2. The absolute value of the range difference between the cylinders in the negative dipole stimulus was subtracted from the absolute value of the range difference in the positive dipole stimulus. The error bars show standard deviations.

Aside from the discriminability of the whole series of dipoles, which is established by the data in Table II, further considerations involve the relative distances to individual cylinders or the aspect angles of the dipoles with respect to simultaneity of reflections from the cylinders and consequent masking, as illustrated in Fig. 6(b) for experiment 1A. The data for all five negative dipoles were combined for these analyses of experiment 2, as discussed below.

C. Discussion

Figure 9 shows the bats' performance as a function of the orientation of the positive dipole [Fig. 9(a)] and the negative dipoles [Fig. 9(b)]. The bats' mean performance was significantly above chance for all orientations of the targets (summed binomial test, $p<0.001$). For the positive dipole, in Fig. 9(a), the bats' highest mean performance was for the 30° aspect angle (81.6%), and the lowest mean performance was for the 90° aspect angle (59.5%). This is the same pattern of performance as that obtained in experiment 1A [Fig. 5(a)]. In contrast, for the negative dipoles, in Fig. 9(b), the bats' highest mean performance was for the 70° aspect angle (81.3%), while the lowest mean performance was for the 30° aspect angle (65.5%).

The presence of two cylinders in each dipole creates an additional feature related to the targets' aspect angles. This is the relative size of each dipole along the range axis. At different orientations, the difference in distance or range between the near and far cylinders in each dipole is the target's range extent. The positive dipole has range extents from 0 to 5 cm as its aspect angle changes from 90° down to 0°.

The negative dipoles all have range extents of 0 cm at an aspect angle of 90° and range extents from 1.5 to 3 cm at an aspect angle of 0° according to their cylinder spacing. Figure 9(c) shows the mean performance (percent correct responses) for all three bats as a function of the difference in range extent between the positive dipole and all five negative dipoles at different combinations of aspect angles (see Fig. 9 caption). The bat's performance is highest when the positive dipole has the largest range extent relative to the negative dipole, while performance is lowest when the negative dipole has the largest range extent relative to the positive dipole. Moreover, this relation is linear and covers the entire span of performance from 83% correct responses, which is near the highest level achieved by any bat in experiment 2 (see Table II), down to near chance at 55% correct responses. The bats behaved as though they discriminated the dipoles by primarily choosing whichever target had the larger extent in range regardless of whether this was in fact the correct (5 cm) dipole. However, in Fig. 9(c), at a relative range extent of 0 cm (positive and negative dipoles oriented so that they have the same range extent), when the two dipole targets were of the same "size" along the single dimension of the range axis, the bats' performance of 64% correct responses was still significantly different from chance ($p<0.05$). In this condition, from the bat's vantage point, the longer (5.0 cm) dipole had a larger left-right or cross range disparity than the smaller dipole (1.5–3.0 cm). The bats thus did not solely rely on range extent to find the larger dipole; rather, they perceived that the larger dipole had a longer dipole spacing or vector length in the range/cross range plane. This is equivalent to correctly perceiving the relative dipole lengths at different aspect angles. The ability of the bats to select the longer dipole when only their relative cross range extents are available as a cue definitely invokes a binaural dimension to the perception of the dipoles. It strongly points to a spatial representation for the dipoles, but it does not prove it.

As in experiment 1A, the errors made by the bats were related to the relative distances from the bat to the individual cylinders in the dipole targets and, thus, to the potential masking effect generated by the simultaneous reception of reflections from different cylinders. However, the addition of the fourth cylinder to make the negative target a dipole, too, increased the likelihood that for the majority of trials, some degree of masking always occurred. Overall, the bats' performance in discriminating between the dipoles in experiment 2 was slightly lower than the performance in discriminating the dipole from the monopole in experiment 1A (compare the total mean of 76.3% from Table I to the total mean of 70.8% from Table II). This could be a consequence of the increased similarity of the targets in experiment 2 (both objects were dipoles), with its attendant increase in the difficulty of the task. It could also be a consequence of a decrease in the likelihood that one of the cylinders in the positive target will be nearer than either cylinder in the negative target or to an increase in the level of masking prevailing across trials because there is now an additional cylinder present. With respect to masking, on average, one of the cylinders in the positive dipole now has an increased chance of being aligned

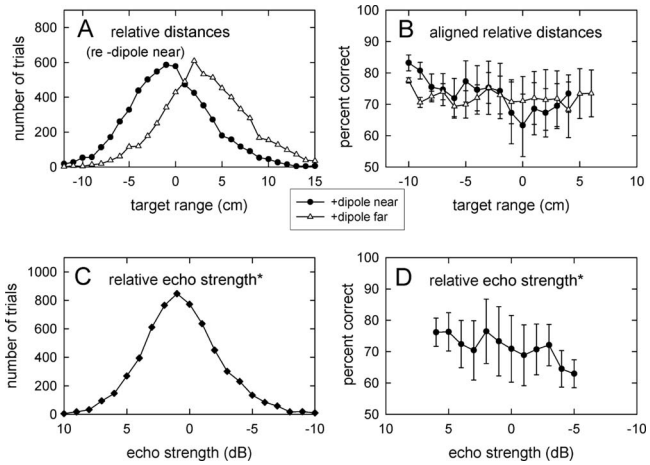


FIG. 10. (A) Distribution of distances to the near and far cylinders in the positive dipole (5 cm spacing) relative to the distance of the near cylinder in the negative dipole in experiment 2 for all trials of three bats. (B) Plots of mean performance of three bats separately sorted by trials according to the distance of the near and far cylinders in the positive dipole and of the far cylinder in the negative dipole relative to the distance to the near cylinder in the negative dipole. (C) Distribution of overall echo strengths for echoes from the positive dipole relative to the negative. (* assumes random aspect angle with an average 3 dB increase in echo strength from dipole.) (D) Plot of mean performance of three bats in experiment 2 for differences in overall echo strength.

in range with one of the cylinders in the negative dipole. Figure 10(a) shows that the near cylinder of the positive dipole was closer to the bat than the near cylinder of the negative dipole on slightly more than 50% of the trials, while the far cylinder of the positive dipole was closer to the bat than the near cylinder of the negative dipole on slightly less than 50% of the trials. However, the spread of relative distances to the near and far cylinders of both dipoles still allowed either target to be presented nearer to the bat on many trials. This provided the scope for the near and far cylinders of both dipoles to coincide in range often enough to test for the presence of masking due to the simultaneous reception of their reflections. Figure 10(b) shows that whenever the near cylinder of the positive dipole was at about the same distance from the bat as the near cylinder of the negative dipole (black circle data points), the bats' performance declined. This effect is similar to that for the cylinders in the dipole and monopole targets in experiment 1A: The curve in Fig. 10(b) comparing the relative distances of the near cylinders of both dipoles in experiment 2 has a shape and level that are similar to those of the curves in Fig. 6(b) comparing the relative distances of the near or far cylinders in the dipole relative to the monopole in experiment 1A. Thus, mutual masking of reflections from the near cylinders of the dipoles may have been a factor in lowering the bats' performance. However, in experiment 1, all three combinations of cylinder overlap and masking have about the same effect [range of performance for the three curves in Fig. 6(b) is from 85% down to 68% correct responses]. In experiment 2, only the curve for the near cylinders in both dipoles has this same range and shape [black circle data points range from 84% down to 64% in Fig. 10(b)]. In Fig. 10(b), in the curve comparing the far cylinder in the positive dipole with the near cylinder in the negative dipole (white triangle data points),

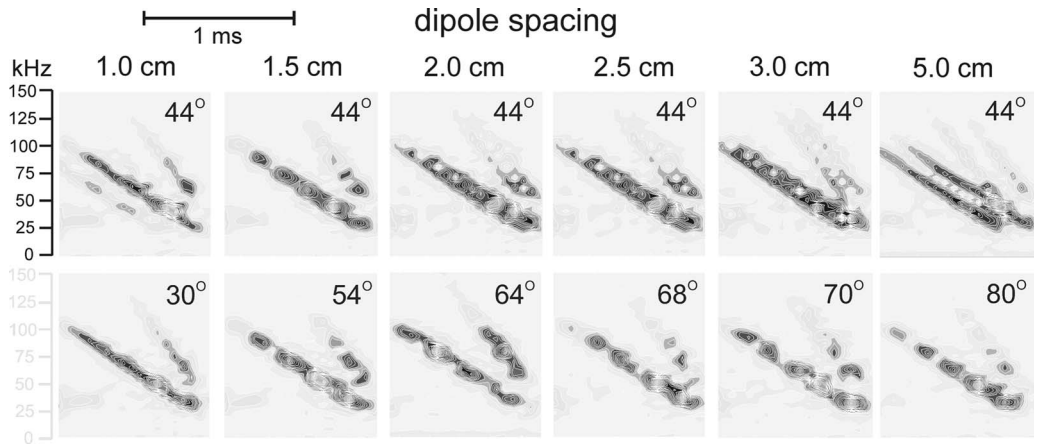


FIG. 11. Spectrograms of the six different dipole echoes (1.0–5.0 cm). The top row of the plots shows spectrograms of echoes for the five spacings of the negative dipole and for the positive dipole (rightmost spectrogram) at an aspect angle of 44°. The bottom row of the plots shows spectrograms for echoes from the six dipoles at different aspect angles.

the bats' performance is lower overall, ranging from 78% to 68% with no corresponding relation to the relative distances of the cylinders. The level of performance shown in this curve may be lowered overall due to the increased likelihood of masking from the other cylinders in both dipoles at various aspect angles, which would obscure any clear relationship to only one of the cylinders in the negative target.

In experiment 2, the echo strength of the positive dipole exceeded the echo strength of the negative dipole on slightly more than half of the trials, which is shown in Fig. 10(c). There is thus little scope for the bats to use just the echo strength to find the positive dipole. If the bats were relying on echo strength to discriminate between the dipoles regardless of which target had the larger (5 cm) spacing, their performance should have been above chance level (50% correct) when the echo strength of the positive dipole exceeded that of the negative dipole and below chance when the echo strength of the negative dipole exceeded that of the positive dipole. In effect, the bats' choices should reverse from the positive to the negative stimulus according to the echo strength. Instead, as shown in Fig. 10(d), the bats' performance remained above chance for differences in echo amplitude from −5 to 5 dB.

As in the case of the dipole versus monopole task of experiment 1 [see Fig. 8(d)], the bats' performance in the dipole versus dipole task of experiment 2 cannot be entirely explained by relying on any single feature of echoes or their spectrograms because the nature of the information the echoes contain varies with the aspect angle of the dipole. This occurs because the relative distances to the cylinders comprising the dipoles change with the aspect angle. The resulting changes in the time separation of their reflections is manifested as a change in the relative importance of the time axis or the frequency axis for spectrograms of the echoes [see Fig. 8(c)]. The top row of the spectrograms in Fig. 11 shows echoes for the five negative dipoles (spacing from 1.0 to 3.0 cm) and the positive dipole at an aspect angle of 44°. At this oblique angle, the cylinders are far enough apart in distance that the longer 5 cm positive dipole target yields two distinct frequency sweeps for the first and second harmonics of the sound in the spectrogram [see Fig. 8(d)]. At this same angle, the cylinders in all five of the negative dipoles are close enough together in distance that their reflections merge to appear as a single sweep for each harmonic in the spectrogram, with the rippling pattern characteristic of overlap and interference. However, if the aspect angle of the positive dipole increases past 50°, the two sweeps in its echoes merge together so that the spectrograms resemble the echoes from the shorter dipoles in Fig. 11. [To see this effect for the positive dipole at different aspect angles, compare the spectrogram for 50° with the spectrograms for 60°, 70°, and 80° in Fig. 8(d)]. Because the bats can distinguish the positive dipole from the negative dipoles at all the different aspect angles, they cannot be using just the separation of the sweeps for the positive dipole and the merging of the sweeps for the negative dipoles. The bottom row of the plots in Fig. 11 shows spectrograms for echoes from the six dipoles at different aspect angles for each target. These angles were chosen to illustrate how the spectrograms can appear the same for each target by selecting conditions where the difference in the arrival time of the reflections from the cylinders is the same.

IV. GENERAL DISCUSSION

By taking the dipole versus monopole (experiment 1) and dipole versus dipole (experiment 2) results together, big brown bats are capable of recognizing the dipole target with the 5 cm spacing in the presence of several competing influences that act to reduce the bat's performance. These influences are not present with the same strength in every trial due to the roving positions and aspect angles of the targets. Roving the targets to different locations and orientations between trials prevented any single feature of the acoustic stimuli from being a stable indicator of which target was the correct choice. The bat had to recognize the dipole with the 5 cm spacing without using any of the extraneous features. There is also uncertainty introduced by the roving itself, which causes the level of performance to decrease across all trials. First, in both the dipole versus monopole and dipole versus dipole experiments, alignment of the cylinders in one target with one of the cylinders in the other target creates a

masking effect because the reflections from these cylinders coincide in arrival time [Figs. 6(b) and 10(b)]. When cylinders do not align in range, masking does not occur and performance is higher. Masking interacts with the interpretation of results for other features of the stimuli. Recognition of the 5 cm dipole appears to depend on the aspect angle of that target [Figs. 5(a) and 9(a)], but the rotation of the dipole toward the 90° aspect angle brings both cylinders to similar ranges, creating a masking effect apart from the aspect angle itself. Figure 6(b) shows that the masking effect between either cylinder of the dipole and the monopole has the same strength as the putative dependence on aspect angle approaching 90°, suggesting that the effect of aspect angle is really masking between the cylinders in the dipole. When the effects of masking are taken into account, the bats in fact appear able to recognize the dipole at all aspect angles. When the task is the discrimination of two dipoles (experiment 2), the bats are able to recognize the 5 cm dipole from dipoles with shorter spacings of 1.0–3.0 cm.

The conceptual rationale for these experiments involves a fundamental distinction: The representation of targets in echolocation could be organized in terms of the proximal stimuli—the echoes, or the distal stimuli—that object themselves. That is, echolocation could be based on perception of certain acoustic features of echoes received at the ears (e.g., amplitude, frequency, delay, spectrum) whose inventory in any given circumstances is sufficient to describe the content of the bat's images. However, echolocation might instead be based on the perception of the spatial features of the objects remote from the bat that affect the echoes in the course of the process of reflection (e.g., distance, direction, size, shape). In this case, the bat's images will have an additional content beyond what can be explained by an inventory of echo features because objects rather than sounds are perceived. The bat would have to transform the acoustic information carried by echoes into information about the objects and incorporate this transformation into a representation that, in effect, projects the objects onto their locations in space. Although this transformation would necessarily be more computationally complicated than a direct perception of the sounds as surrogates for the objects, it is quite possible that in the end, it would be more parsimonious for the bat to go to the trouble of converting echoes into perceptual entities more like objects than sounds because subsequent processing in an object-oriented regime could prove to be more effective. That is, if complicated, parallel organization of neuronal computations is a salient characteristic of brain organization, it is not self-evident that perceiving objects in terms of echo features is "simpler" as an explanation than perceiving objects in terms of spatial images. Although more complicated neuronal computational transformations are required to go beyond segregated representations of echo acoustic features to create spatial representations of objects, complex computations are present in the bat's auditory midbrain and forebrain anyway (Simmons *et al.*, 1996), and it may be simpler just to use them. By rendering echo information into spatial images, bats might be able to exploit the presence of all sorts of mechanisms for spatial cognition in general rather than create a stand-alone system based on its auditory system. Besides, from a comparative perspective, auditory perception is just as much object oriented as it is acoustically oriented: Sounds are not perceived as disembodied acoustic events with characteristic features such as pitch, loudness, and timbre, they are perceived as emanating from sources at well-defined locations—sources whose structure is perceived from acoustic features along with location.

It would be advantageous for bats in flight to perceive the distal stimulus (the object) instead of the proximal stimulus (the echoes) because they must find or avoid aspect-dependent objects (e.g., insects, trees) and are constantly changing their position relative to the objects (thus receiving different echoes from the same object at different orientations). When bats maneuver through vegetation to attack insects surrounded by branches and leaves, when they dodge past rows of obstacles, or when they recognize and intercept specific targets out of a cluster of airborne objects, they evince an awareness of the locations of both the target of interest and surrounding objects that is difficult to avoid calling "spatial perception" (Simmons *et al.*, 1995; Moss and Surlykke, 2001). Another echolocating mammal, the dolphin, must also solve the problem of identifying moving aspect-dependent objects (e.g., fish) in a three-dimensional environment. The acoustic features of fish echoes vary depending on the orientation of the fish with respect to the dolphin (Au *et al.*, 2007). Thus, it would seem to be ecologically sensible for dolphins to perceive objects on the level of the distal stimulus. There is evidence that bottlenose dolphins (*Tursiops truncatus*) form distal instead of proximal representations (Harley *et al.*, 1996; Harley *et al.*, 2003; Helweg *et al.*, 1996a; Helweg *et al.*, 1996b; Herman *et al.*, 1998; Pack and Herman, 1995). This study suggests that bats, like dolphins, transform acoustic information carried by echoes into representations containing object features.

ACKNOWLEDGMENTS

This research was supported by National Institute of Mental Health Grant No. R01-MH 069633 and by Office of Naval Research Grant No. N00014-04-1-0415. The authors would like to thank Alida Kinney, Donald Mowlds, Oliver Eng, Kara Aurbach, Peter Ryg, Benjamin Wu, Dale Jun, and Anthony Petrites for their assistance in training the bats and performing some video data analyses. The authors thank Hazem Baqaen for writing the MATLAB programs used to analyze the echoes. Care and use of the animals were supervised by Brown University veterinarians and the Institutional Animal Care and Use Committee in accordance with *Principles of Animal Care* No. 86-23 (revised 1985) of the U.S. National Institutes of Health Publication.

Au, W. W. L., Benoit-Bird, K. J., and Kastelein, R. A. (**2007**). "Modeling the detection range of fish by echolocating bottlenose dolphins and harbor porpoises," J. Acoust. Soc. Am. **121**, 3954–3962.

Gellermann, L. W. (**1933**). "Chance orders of alternating stimuli in visual discrimination experiments," J. Genet. Psychol. **42**, 206–208.

Griffin, D. R. (**1958**). *Listening in the dark: The Acoustic Orientation of Bats and Men* (Yale University Press, New Haven, CT).

Grinnell, A. D. (**1995**). "Hearing in bats: An overview," in *Hearing by Bats*, edited by A. N. Popper and R. R. Fay (Springer-Verlag, New York), pp. 1–36.

Harley, H. E., Putman, E. A., and Roitblat, H. L. (**2003**). "Bottlenose dol-

phins perceive object features through echolocation," Nature (London) **424**, 667–669.
Harley, H. E., Roitblat, H. L., and Nachtigall, P. E. (**1996**). "Object representation in the bottlenose dolphin (*Tursiops truncatus*): Integration of visual and echoic information," J. Exp. Psychol. Anim. Behav. Process **22**, 164–174.
Hartley, D. J. (**1992**). "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.
Helweg, D. A., Au, W. W. L., Roitblat, H. L., and Nachtigall, P. E. (**1996a**). "Acoustic basis for recognition of aspect-dependent three-dimensional targets by an echolocating bottlenose dolphin," J. Acoust. Soc. Am. **99**, 2409–2420.
Helweg, D. A., Roitblat, H. L., Nachtigall, P. E., and Hautus, M. J. (**1996b**). "Recognition of aspect-dependent three-dimensional objects by an echolocating Atlantic bottlenose dolphin," J. Exp. Psychol. Anim. Behav. Process **22**, 19–31.
Herman, L. M., Pack, A. A., and Hoffman-Kuhnt, M. (**1998**). "Seeing through sound: Dolphins (*Tursiops truncatus*) perceive the spatial structure of objects through echolocation," J. Comp. Psychol. **112**, 292–305.
Kurta, A., and Baker, R. H. (**1990**). "*Eptesicus fuscus*," Mammalian Species **356**, 1–10.
Moss, C. F., and Schnitzler, H.-U. (**1995**). "Behavioral studies of auditory information processing," *Springer Handbook of Auditory Research: Hearing by Bats*, edited by R. Fay and A. Popper (Springer-Verlag, Berlin), pp. 87–145.
Moss, C. F., and Surlykke, A. (**2001**). "Auditory scene analysis by echolocation in bats," J. Acoust. Soc. Am. **110**, 2207–2226.
Moss, C. F., and Zagaeski, M. (**1994**). "Acoustic information available to bats using frequency-modulated sounds for the perception of insect prey," J. Acoust. Soc. Am. **95**, 2745–2756.
Neuweiler, G. (**2000**). *The Biology of Bats* (Oxford University Press, New York).
Pack, A. A., and Herman, L. (**1995**). "Sensory integration in the bottlenose dolphin: Immediate recognition of complex shapes across the senses of echolocation and vision," J. Acoust. Soc. Am. **98**, 722–733.
Pollak, G. D., and Casseday, J. H. (**1989**). *The Neural Basis of Echolocation in Bats* (Springer, New York).
Saillant, P. A., Simmons, J. A., Bouffard, F. H., Lee, D. N., and Dear, S. P. (**2007**). "Biosonar signals impinging on the target during interception by big brown bats, *Eptesicus fuscus*," J. Acoust. Soc. Am. **121**, 3001–3010.
Simmons, J. A. (**1989**). "A view of the world through the bat's ear: The formation of acoustic images in echolocation," Cognition **33**, 155–199.
Simmons, J. A., Ferragamo, M. J., Salliant, P. A., Haresign, T., Wotton, J. M., Dear, S. P., and Lee, D. N. (**1995**). "Auditory dimensions of acoustic images in echolocation," in *Hearing by Bats*, edited by A. N. Popper and R. R. Fay (Spring-Verlag, New York), pp. 146–190.
Simmons, J. A., Freedman, E. G., Stevenson, S. B., and Chen, L. (**1989**). "Clutter interference and the integration time for echoes in the bat, *Eptesicus fuscus*," J. Acoust. Soc. Am. **86**, 1318–1332.
Simmons, J. A., Saillant, P. A., Ferragamo, M. J., Haresign, T., Dear, S. P., Fritz, J. B., and McMullen, T. A. (**1996**). "Auditory computations for acoustic imaging in bat sonar," in *Auditory Computation*, edited by H. L. Hawkins, T. A. McMullen, A. N. Popper, and R. R. Fay (Spring-Verlag, New York), pp. 401–468.
Simmons, J. A., and Vernon, J. A. (**1971**). "Echolocation: Discrimination of targets by the bat, *Eptesicus fuscus*," J. Exp. Zool. **176**, 315–328.
Surlykke, A., and Moss, C. F. (**2000**). "Echolocation behavior of big brown bats, *Eptesicus fuscus*, in the field and the laboratory," J. Acoust. Soc. Am. **108**, 2419–2429.

Improved scatterer size estimation using backscatter coefficient measurements with coded excitation and pulse compression

Steven G. Kanzler[a] and Michael L. Oelze[b]
Bioacoustics Research Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

(Received 27 August 2007; revised 17 March 2008; accepted 19 March 2008)

Scatterer size estimates from ultrasonic backscatter coefficient measurements have been used to differentiate diseased tissue from normal. A low echo signal-to-noise ratio (eSNR) leads to increased bias and variance in scatterer size estimates. One way to improve the eSNR is to use coded excitation (CE). The normalized backscatter coefficient was measured from three tissue-mimicking phantoms by using CE and conventional pulsing (CP) techniques. The three phantoms contained randomly spaced glass beads with median diameters of 30, 45, and 82 μm, respectively. Measurements were made with two weakly focused, single-element transducers (f_0=5 MHz and f_0=10 MHz). For CE, a linear frequency modulated chirp with a time bandwidth product of 40 was used and pulse compression was accomplished by the use of a Wiener filter. Preliminary results indicated that improved estimation bias versus penetration depth was obtained by using CE compared to CP. The depth of penetration, where the accuracy of scatterer diameter estimates (absolute divergence $<$25%) were obtained with the 10 MHz transducer, was increased up to 50% by using CE versus CP techniques. In addition, for a majority of the phantoms, the increase in eSNR from CE resulted in a modest reduction in estimate variance versus depth of penetration.
 [DOI: 10.1121/1.2908293]

PACS number(s): 43.80.Vj [FD] Pages: 4599–4607

I. INTRODUCTION

The radio frequency (rf) spectrum of ultrasonic backscatter contains information that can be used to noninvasively characterize the structural and mechanical properties of tissue. Imaging techniques based on quantifying the ultrasonic backscatter have been successfully used to diagnose and monitor disease, such as cancer, in clinical settings. Furthermore, these imaging approaches have been used to differentiate different kinds of tissues.[1–5] Parametric images enhanced by scatterer parameters, i.e., the average scatterer size and acoustic concentration, have been constructed for test phantoms[6] and tissues.[7] In clinical settings, these imaging techniques have been successful in diagnosing prostate cancer, ocular tumors, and cardiac and vascular abnormalities.[1,8–10]

Because of the weak scattering condition in tissues, the backscattered signals often have low echo signal-to-noise ratio (eSNR). eSNR specifically refers to the ratio of the echo signals from ultrasonic scatter to background noise (i.e., electronic noise).[11] As with any estimation scheme, low eSNR leads to increased bias and variance in scatterer size estimates.[12] To improve the eSNR, either the amplitude or the pulse duration of the transmitted sound wave can be increased. However, the maximum pressure amplitude (or maximum negative peak pressure level of the propagating sound wave) that can be transmitted into a biological medium is limited because of the possibility of bioeffects.[13] Because of the possibility of bioeffects in diagnostic ultrasound, ultrasonic imaging devices are limited in the United States by the Food and Drug Administration to a spatial peak temporal average intensity of 720 mW/cm^2 and a mechanical index (MI$<$1.9), where the MI is the rarefactional pressure in Megapascals divided by the square root of the frequency in Megahertz.[14] As a result of this pressure amplitude limit, in many cases, the penetration depth can only be increased safely by increasing the pulse duration.[15] However, increasing the duration of the excitation waveform would cause a decrease in axial resolution.

A method to increase the pulse duration while retaining the spatial resolution is through coded excitation and pulse compression. Coded excitation was first used in radar applications to increase the signal energy.[16] Coded signals of long duration were used to excite the source and then the received echoes were filtered or compressed to restore the spatial resolution of the system. Examples of coded excitation schemes that have been used with pulse compression include binary codes, such as Barker codes or Golay codes, or frequency modulated (FM) waveforms, also known as chirps.[17]

In the current study, coded excitation using pulse compression (CEPC) has been used to increase the eSNR of signals backscattered from tissue-mimicking phantoms. Scatterer size estimates were then obtained from the backscattered waveforms by using CEPC and conventional pulsing (CP) techniques and compared. Specifically, linear FM chirps were used as the coded excitation waveform. Section II discusses the methodology of the experimental setup and the implementation of the CEPC routines. Results of the experimental measurements are given in Sec. III. Discussion of the results and some final conclusions are given in Sec. IV.

[a]Also at Department of Engineering and Natural Sciences, University of Applied Sciences, Merseburg 06217, Germany.
[b]Electronic mail: oelze@uiuc.edu

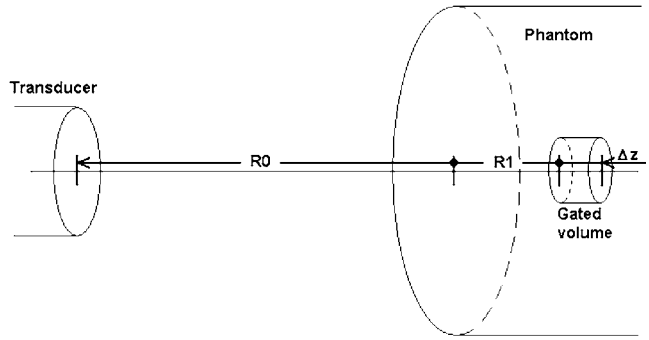


FIG. 1. The distance between the surface of the transducer and the surface of the phantom is R_0, the distance from the the surface of the phantom to the surface of the ROI is R_1, and the length of the gated region is Δz. The focal depth is at R_0+R_1.

II. METHODS AND MATERIALS

A. Experimental setup

Weakly focused ($f/3$ and $f/4$, respectively) single-element transducers were placed in a tank filled with degassed water (21 °C). The first transducer had a nominal center frequency of 5 MHz ($f/3$) and a −6 dB bandwidth of 84% estimated from a pulse-excited signal reflected from a planar surface. The second transducer had a nominal center frequency of 10 MHz ($f/4$) and a 91% −6 dB bandwidth estimated from a pulse-excited signal reflected from a planar surface. The −6 dB depth of field (DOF) and beam width of the transducers were measured by using the wire technique.[18] The −6 dB DOFs were estimated to be 9.1 and 16.9 mm for the 5 and 10 MHz sources, respectively. The −6 dB beam widths were estimated to be 850 and 613 μm for the 5 and 10 MHz transducers, respectively. The transducers were used to measure the backscatter from tissue-mimicking (TM) phantoms. In experiments, a TM phantom was placed in a tank of degassed water. The transducer was positioned so that the beam axis was perpendicular to the face of the phantom and the focus of the transducer was inside the surface of the phantom.

Figure 1 shows the geometry of the transducer and the phantom. For the experiment by using CP [Fig. 2(a)], the transducer was excited by a Panametrics 5800 pulser/receiver (Waltham, MA) and connected through a diplexer (Ritec, Warwick, RI). For the experiment by using CEPC [Fig. 2(b)], the transducer was excited by an arbitrary waveform generator (AWG) (Tabor Electronics WW1281, Tel Hanan, Israel) with a sampling frequency of 100 MHz. From the AWG, the signal was amplified by a 50 dB rf power amplifier (ENI 325LA, Rochester, NY). The amplified signal was connected to the transducer through the diplexer. For both experiments, the backscattered echo signals were received by a Panametrics 5900 pulser/receiver (Waltham, MA) through the diplexer and then acquired by a 12 bit analog-to-digital (A/D) card (Strategic Test UF3025, Akersberga-Stockholm, Sweden) and recorded on the hard drive of a personal computer. The backscattered power spec-

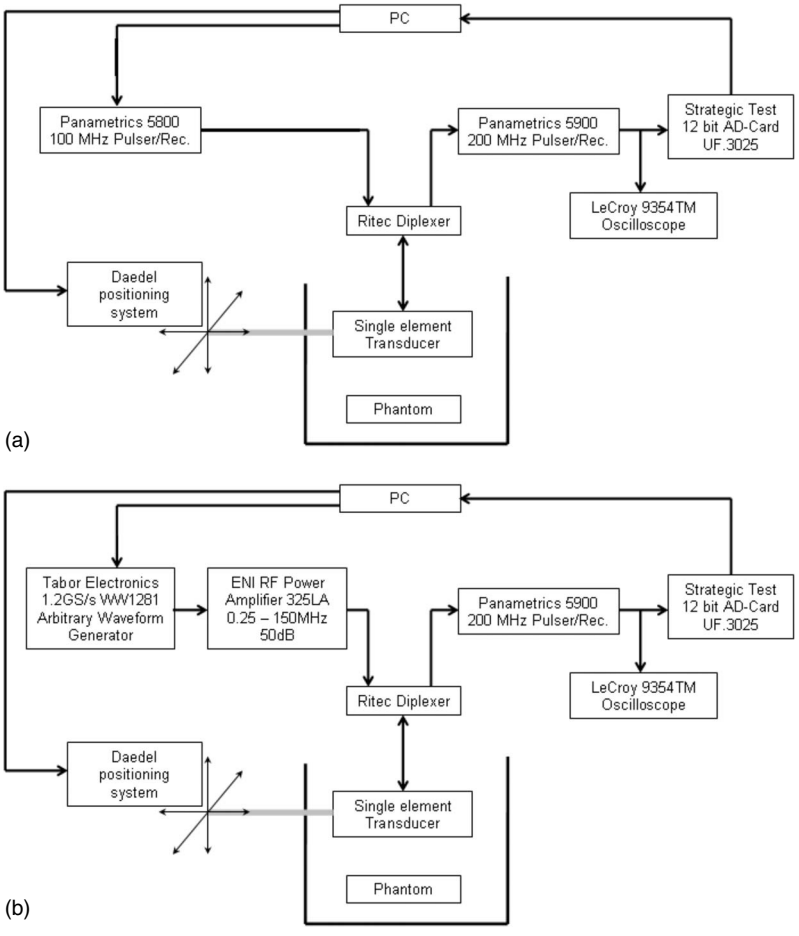


FIG. 2. Block diagram of experimental setup with (a) CP method and (b) CEPC method.

trum was estimated from the backscattered ultrasonic signals and normalized by a reference spectrum.[1] To obtain the reference waveform, a planar surface (Plexiglas®) was placed perpendicular to the beam axis with the surface near the focus of the transducer and the reflected signal recorded.

The A/D card is limited to a voltage range of ±1 V with a dynamic range of 12 bit. However, the amplitude of the reflection from the front surface of the phantom was large compared to the amplitude of the speckle corresponding to inside the phantom. Often, the difference in the signal amplitude from the front surface of the phantom compared to the speckle within the phantom was larger than the dynamic range available to the system. A gain was applied to the backscattered signal through the Panametrics 5900. As a result, the signal corresponding to the reflection off the front surface was clipped to ±1 V; however, the signal corresponding to the speckle more fully spanned the dynamic range of the A/D card. When CP was used, the clipping did not influence the measurements or subsequent scatterer diameter estimates. However, clipping the signal did influence the ability to compress the backscattered signals by using CEPC. During the acquisition process, the A/D card transformed the linear FM chirp (FM sinusoidal waveform), resulting from the reflection off the front surface of the phantom into a pseudochirp (FM rectangular waveform) by clipping the data above 1 V and below −1 V. Because the scattered signal is often −40 dB or more below the signal reflected from the front surface and the dynamic range of the A/D card was 12 bit, the signal reflected from the front surface was allowed to clip rather than attenuate the whole signal. As a result, the dynamic range of the A/D card was used to span the scattered signal. Rectangular waveforms contain high frequency content because of the sharp edges, and this causes side lobes of high amplitudes after compression. To avoid these side lobes, the portion of the echo signal containing the pseudochirp was replaced by artificially created zero-mean white Gaussian noise (WGN) with the same variance as the system noise. The variance of the system noise was estimated from recordings in the region without scatterers (noise window) and was used in subsequent estimates of eSNR. The eSNR value was determined through[11]

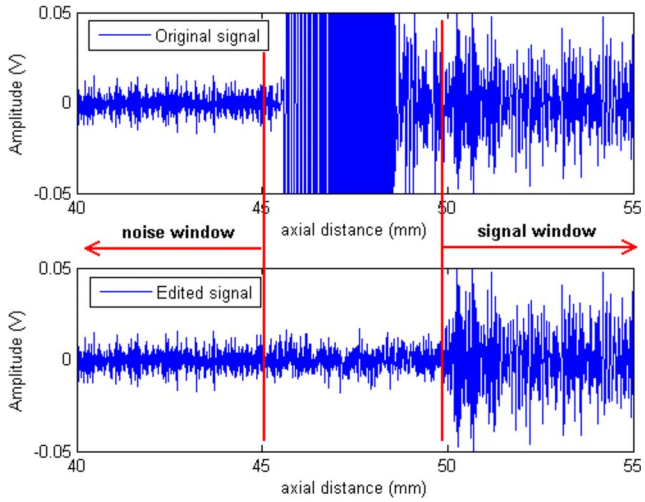


FIG. 3. (Color online) Front surface reflection replaced by zero-mean WGN.

$$\mathrm{eSNR} = 10 \log \frac{\sigma_s^2}{\sigma_n^2}, \tag{1}$$

where σ_s^2 and σ_n^2 are the variances of the signal from the scatterers and from a scatterer free region (noise), respectively. Figure 3 shows an example of the clipped chirp signal from the phantom surface and the clipped signal replaced by zero-mean WGN.

The system noise was much higher from the combined AWG and rf power amplifier system compared to the Panametrics 5800 pulser/receiver. Therefore, the eSNR was much higher for the pulse generation by using the Panametrics 5800 pulser/receiver (CP) than by using the rf power amplifier and AWG (CEPC). In order to quantify the benefits of increasing the eSNR through CEPC versus CP, the eSNR needed to be normalized between the CP signals and the coded excitation signals before pulse compression. To normalize the eSNR, zero-mean WGN was added to the CP signals.

Briefly, the variance of the system noise was estimated at a time window before the reflection of the front surface for the chirp setup. WGN was then created by using the randn( ) function in MATLAB, which creates normally distributed numbers with mean zero, variance 1. This was then multiplied with the square root of the estimated noise variance

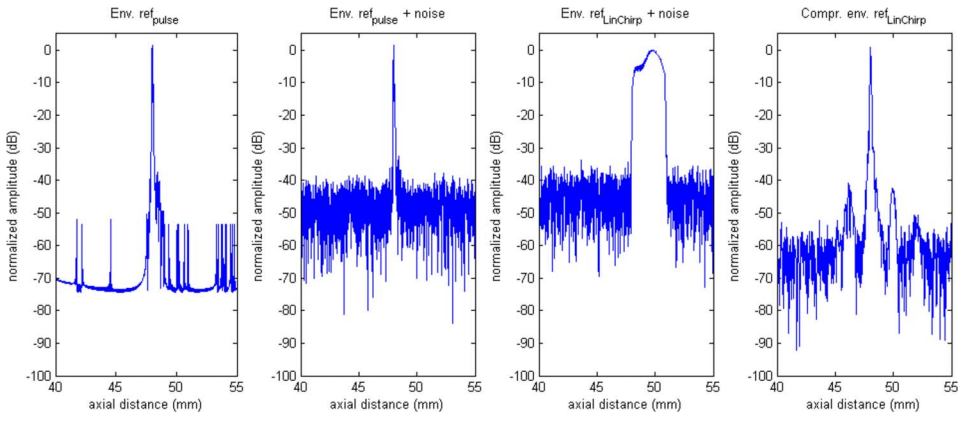


FIG. 4. (Color online) eSNR equalization through noise addition, shown for reference signals with pulsed and chirp signal. Also, the compressed signal is shown.

from the CEPC setup. This WGN was added to the data taken with the pulsed setup. For the signals taken as reflection of a planar surface in pulsed and chirp setup, the noise was created as well and added with the same algorithm. Figure 4 shows the envelopes of recorded signals reflected from a planar surface without and with noise added in comparison, for pulsed setup, and chirp setup before and after compression.

B. Backscatter coefficient estimation

A broadband substitution method valid for weakly focused transducers was used to estimate the backscatter coefficient. The backscatter coefficient σ_b was estimated from the normalized power spectral density (PSD) of the backscattered echo signal by

$$\sigma_b(f) = \frac{0.36(R_0 + R_1)^2}{A_0 \Delta z} W(f), \tag{2}$$

where A_0 is the area of the transducer.[6] $W(f)$ is the average PSD of the backscattered echo signals divided by the average PSD of the reference signal. Corrections for frequency-dependent attenuation and the reference spectrum were incorporated through

$$W(f) = \frac{1}{N_l} \sum_{l=1}^{N_l} \left(\frac{\gamma}{2}\right)^2 \frac{|S_m(f,Z_l)|^2}{|S_0(f,Z_l)|^2} e^{4\alpha_m(f)(R_1+(\Delta z)/2)}, \tag{3}$$

where N_l is the number of gated waveforms of length Z_l that have been obtained from the sample and γ is the amplitude reflection coefficient of the planar surface that was used to obtain the reference waveform. $S_m(f,Z_l)$ is the Fourier transform of the backscattered echo signal and $S_0(f,Z_l)$ is the Fourier transform of the reference signal. $\alpha_m(f)$ is the frequency-dependent attenuation coefficient of the sample medium. The frequency-dependent attenuation was estimated by using an insertion loss method. Briefly, the signal from a transducer was recorded with a hydrophone with and without the phantom material in between. The frequency-dependent attenuation was estimated by dividing the absolute value of the Fourier transforms of the signals with and without the phantom material and by the thickness of the phantom.[19,20] The backscattered echo signal waveforms were gated from regions of interest (ROIs) corresponding to inside the phantom by using a Hann window of width Δz.

C. Scatterer size estimation

The TM phantoms contained randomly located glass spheres of varying diameters, which were modeled by using established theory.[21,22] Specifically, theoretical backscatter coefficients versus frequency were calculated for different glass bead diameters (Fig. 5). The theoretical backscatter coefficients σ_{Faran} calculated for different diameters of glass beads were stored in a look up table (LUT) for faster processing speed. From the LUT, estimates of the glass bead diameter could be obtained by comparing the measured backscatter coefficients from the phantoms with σ_{Faran}. The estimated scatterer diameter was the value that minimized the average squared deviation (MASD) between the measured and theoretical backscatter coefficients given from the LUT.[23] The MASD is mathematically represented by

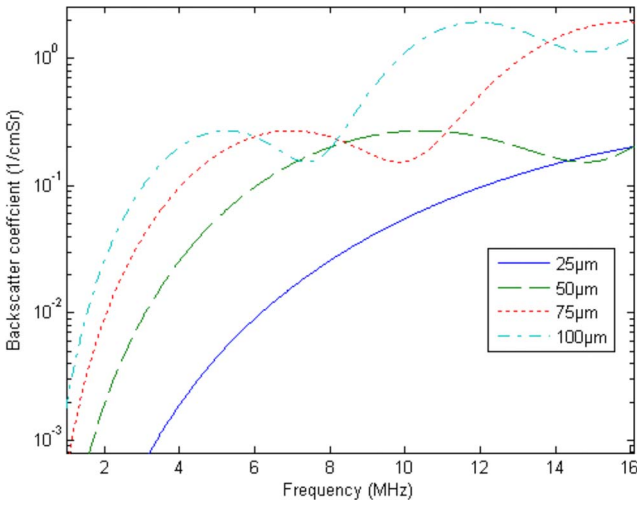


FIG. 5. (Color online) Calculated backscatter coefficient for glass beads with four different diameters vs frequency.

$$\text{MASD} = \min(\varepsilon\{(X(f,D) - \varepsilon\{X(f,D)\})^2\})_D, \tag{4}$$

with

$$X(f,D) = 10 \log_{10}\left(\frac{\sigma_b(f)}{\sigma_{\text{Faran}}(f,D)}\right). \tag{5}$$

$\varepsilon\{\cdots\}$ represents the mean and $X(f,D)$ is the ratio (decibels) of σ_b (in the −6 dB bandwidth of the transducer) to σ_{Faran} for the diameter D. By subtracting the mean $\varepsilon\{X(f,D)\}$ from $X(f,D)$, the result is independent from the magnitude of measured and calculated backscatter coefficients and only depends on the shape, i.e., the frequency dependence, of σ_b.[23]

At depths where the noise was much greater than the echo from the scatterers, the glass bead diameter estimates diverged from the median value. A rf spectrum dominated by white noise would have a slope of zero. After correcting for the frequency-dependent attenuation, the slope would take on the characteristics of the frequency-dependent attenuation correction. Likewise, for a band-limited signal, higher frequencies are attenuated more rapidly than lower frequencies, leading to lower SNR in the higher frequency channels. Compensating for frequency-dependent losses will result in an amplification of noise preferentially at higher frequencies in the analysis bandwidth. The amplification of noise will result in a larger slope for the estimated backscatter coefficient, which leads to an increasing bias in scatterer size estimates. The depth of penetration for obtaining reliable scatterer size estimates was defined to be the depth where the estimate bias (absolute difference between the median glass bead size and the estimated size) was less than 25%.

D. Coded excitation and pulse compression

For coded excitation, a linear FM chirp v was used as the excitation waveform. The chirp waveform was constructed using MATLAB (MathWorks, Natick, MA) and then uploaded to the AWG. The chirp had a center frequency

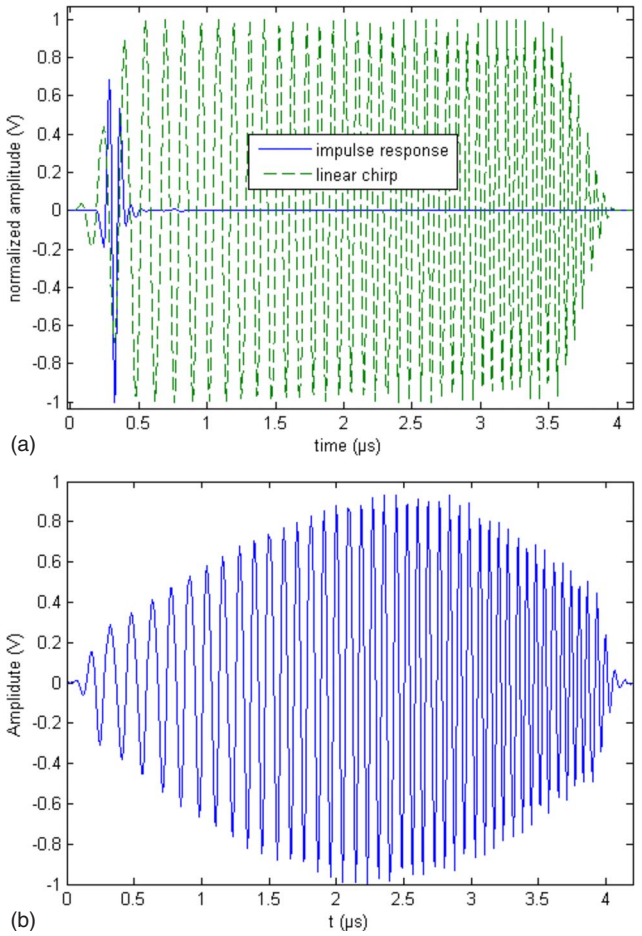


FIG. 6. (Color online) (a) Impulse response (solid line) of the 10 MHz transducer and the linear FM chirp (dashed line) used to excite the transducer. (b) Impulse response of the 10 MHz transducer excited with the linear FM chirp.

equal to the center frequency of the transducer and a time-bandwidth product (TBP) of 40. A TBP of 40 would lead to a gain in eSNR of 16 dB [10 log(TBP)] compared to conventional pulsing.[24] The bandwidth was chosen to be 1.14 times the −6 dB bandwidth of the transducer, which was determined to be the optimum bandwidth for a linear FM chirp.[25] Finally, the chirp was tapered using a 10% Tukey window to suppress sidelobe levels.

Figure 6 shows the impulse response of the 10 MHz transducer, the linear FM chirp used to excite the transducer (a), and the FM chirped signal measured from a planar reflector (b). The resulting sound waveform h_{out} was the convolution of the linear chirp with the impulse response of the system including the transducer. This waveform was propagated through the water and backscattered by the phantom. The received backscattered echo signals were compressed in MATLAB by convolution (or multiplication in frequency domain) with a filter function,

$$G_{pc} = G\beta, \tag{6}$$

where G is the Fourier transform of the received echo signal and β is the filter function. A Wiener filter was chosen because it allows the trade-offs between sidelobes and noise to be controlled. The Wiener filter is given by

TABLE I. Parameters of all four phantoms.

Parameter	Phantom A	Phantom B	Phantom C	Phantom D
Glass diameters (μm)	75–90	9–45	45–53	45–53
Speed of sound (m s^{-1})	1540	1540	1540	1540
Attenuation (dB MHz^{-1} cm^{-1})	0.5	0.7	0.5	0.7

$$\beta = \frac{V^*}{|V|^2 + \gamma_1 \overline{\text{eSNR}}^{-1}}, \tag{7}$$

where V is the Fourier transform of the excitation waveform and V^* designates its complex conjugate. $\overline{\text{eSNR}}$ is the average echo signal-to-noise ratio per frequency channel and is defined as

$$\overline{\text{eSNR}(u|x)} = \frac{|H(u|x)|^2 \varepsilon\{|F(u)|^2\}}{\varepsilon\{|E(u)|^2\}}, \tag{8}$$

where u is the discrete frequency sampling variable and $H(u|x)$, $F(u)$, and $E(u)$ are the Fourier transforms of $h(nT,x)$, $f(x)$, and $e[n]$, respectively.[26] $\varepsilon\{\cdots\}$ represents the expectation value, $e[n]$ is the noise present in the system, $h(nT,x)$ is the spatially varying impulse response of the system, and $f(x)$ is a function representing the scattering object. Because $f(x)$ is unknown for the sample, it is replaced by the average PSD of the gated waveforms divided by the PSD of h_{out},

$$|F(u)|^2 = \frac{\overline{|S_m(f,Z_l)|^2}}{|H_{\text{out}}(u|x)|^2}. \tag{9}$$

E. Test samples

In this study, four different test samples were used. All samples were fluidlike water-based agar materials that contained randomly positioned glass spheres (Potters Industries, Valley Forge, PA). Table I lists the relevant parameters associated with each of the TM phantoms used in this study.[19,20] The TM phantoms contained different concentrations of randomly located glass spheres. The sizes of glass spheres included in this study corresponded to scatterer sizes previously encountered in tissue experiments.[27] All phantoms where circular cylindrical samples with a diameter of 7.6 cm and a length of either 4 cm (phantoms A and B) or 5 cm (phantoms C and D). The phantoms were bounded on the curved surface with a plastic wall and on the flat surface with a 25-μm-thick Saran–Wrap plastic foil (Dow Chemical, Midland, MI).

The Saran–Wrap served as a window for transmitting ultrasound between the surrounding medium (water) and the TM material inside the phantom. To correct for the influence of the Saran–Wrap to the backscattered spectrum, the following equation was used:

$$T(k) = \frac{2Z_i}{2Z_i \cos(k_s * d)} + j\frac{1.69c_s + Z_i^2}{1.69c_s} \sin(k_s * d), \tag{10}$$

where k is the wave number represented by $2\pi f/c$, Z_i is the acoustic impedance (Z_i=14.9 kg m^{-2} s^{-1}), d is the thickness

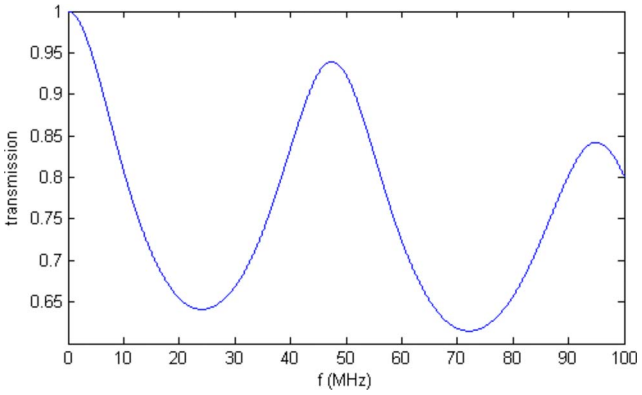


FIG. 7. (Color online) Transmission through two layers of 25-μm-thick Saran–Wrap.

of the Saran–Wrap layer (d=25.1 μm), c_s is the speed of sound inside the Saran–Wrap (c_s=2400 m s^{-1}), and k_s is the wave number inside the wrapping material,[19]

$$k_s = \frac{2\pi f}{c_s} - j(0.05f^{1.5}). \tag{11}$$

Figure 7 shows the transmission coefficient through two 25-μm-thick layers of Saran–Wrap. The backscatter coefficient was corrected for the frequency-dependent transmission through the Saran–Wrap layer by dividing $W(f)$ by the square of the power transmission coefficient (accounting for transmission in and out of the phantom).

III. EXPERIMENTAL RESULTS

The first set of data was obtained by using a weakly focused (f/3) 5 MHz single-element transducer (Panametrics, Waltham, MA) with a −6 dB bandwidth of 4.2 MHz. The linear FM chirp had a length of 8.9 μs and a bandwidth of 4.5 MHz, which led to a TBP of 40. The second set of data was obtained by using a weakly focused (f/4), single-element transducer (Panametrics, Waltham, MA) with a center frequency of 10 MHz and a −6 dB bandwidth of 9.1 MHz. The linear FM chirp had a length of 4 μs and a bandwidth of 10 MHz, which led to a TBP of 40 (an estimated gain in eSNR of 16 dB). A compression of the chirp waveforms from the planar reflector at the focus produced gains in eSNR of approximately 16 dB matching the expected increase. The eSNR gain can be observed in Fig. 4. All four phantoms were scanned 30 mm laterally with a step size of 0.5 mm. In addition, lateral scans were performed at three different distances. R_1 was set to 5, 10, and 15 mm, respectively. R_0 was decreased by the same amount that R_1 was increased, such that the sum of R_0 and R_1 was constant. This allowed for estimates to be obtained from deeper in the phantom while still within the depth of focus of the transducer. The axial length of the ROI Δz was set to 4 mm. The first 9 and 4 mm for the 5 and 10 MHz transducers, respectively, behind the front surface of the phantom was dominated by the chirp reflected from that surface. Therefore, the data sets used for analysis started at 10 or 5 mm behind the front surface.

TABLE II. *ka* ranges for three different scatterer diameters for 5 and 10 MHz measurements.

Average scattering diameter (μm)	*ka* range (5 MHz)	*ka* range (10 MHz)
25	0.12–0.34	0.28–0.75
49	0.24–0.66	0.54–1.45
82	0.40–1.10	0.91–2.45

The results at 5 MHz did not suggest an improvement in scatterer sizes estimates with depth. The reason for this lack of improvement is due to the lower attenuation of ultrasound at 5 MHz. While the gain in eSNR using CEPC over CP within the phantom was typically 10 dB or more, the eSNR for both CP and CEPC was always above 10 dB throughout the depth examined in each phantom. Therefore, at lower ultrasonic frequencies where attenuation is lower, the use of CEPC may not yield significant improvements in scatterer size imaging throughout the depth of field of the imaging source.

For the 10 MHz data, the maximum depth from which data was used was at 15 mm because signals from locations greater than 15 mm deep were already attenuated below the level of noise in the system. Figure 4 shows an example of the signal reflected from a planar surface with WGN added before and after compression. In the compressed image, it can be observed that the spatial resolution has been improved and the eSNR has been increased. The values of *ka* for the analysis bandwidth of the 5 and 10 MHz sources based on the estimated average scatterer radius are listed in Table II. Estimates of scatterer diameters were found to have the best performance in terms of bias and variance when the *ka* range went above 0.5. Figure 8 compares glass bead diameter estimates from ultrasound backscatter by using CP and CEPC from the phantom with the median value of glass bead diameter of 82 μm for CP and CEPC. A small absolute estimate bias is observed between the median value and the estimates by using CEPC at all depths, i.e., better than 25%. For CP, the estimates were close (less than 25% divergence) to the median value at depths smaller than 1.7 cm. The eSNR ranged from 0.1 to 1.6 dB for CP and 1.7 to 16 dB for

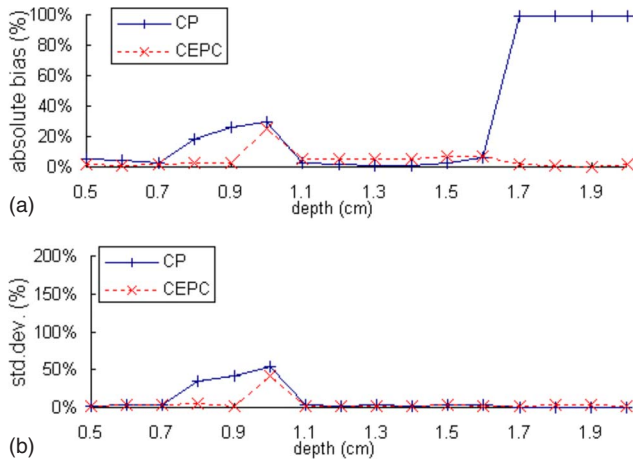


FIG. 8. (Color online) Phantom A: Estimated average scatterer sizes (plot a) and corresponding standard deviations (plot b) of 10 MHz measurements.

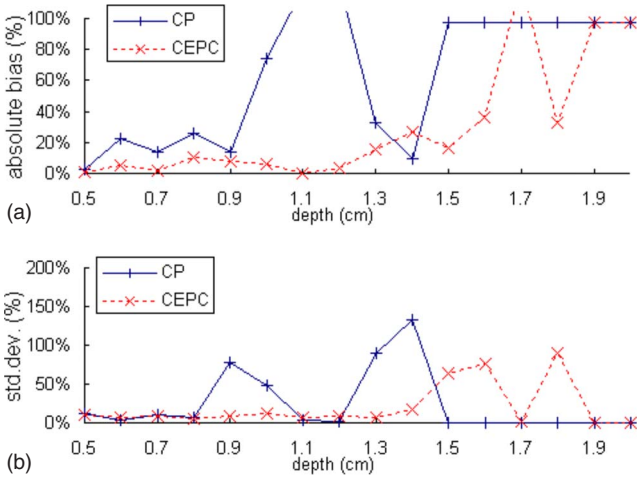


FIG. 9. (Color online) Phantom B: Estimated average scatterer sizes (plot a) and corresponding standard deviations (plot b) of 10 MHz measurements.

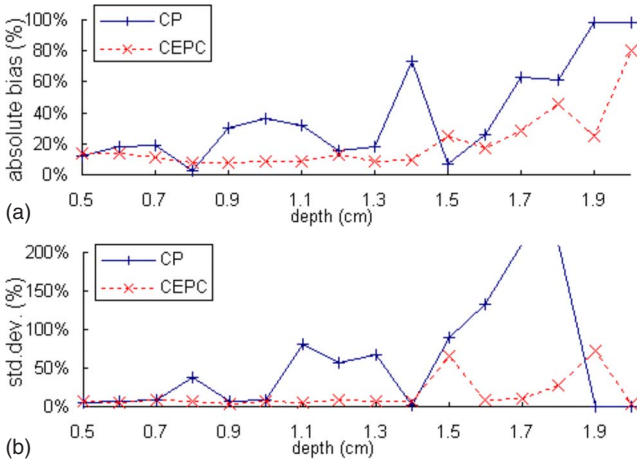


FIG. 11. (Color online) Phantom D: Estimated average scatterer sizes (plot a) and corresponding standard deviations (plot b) of 10 MHz measurements.

CEPC. At 1.7 cm, the eSNR was 0.2 dB (CP). For phantom B (Fig. 9), a bias to the calculated median glass bead diameter (25 μm) was observed. Estimates occurring closer to the surface, i.e., less than 1 cm, have an average estimated scatterer diameter of 40 μm. For a distribution of sizes in the phantom material, scatterer size estimates will be larger than the median value because the backscatter coefficient is proportional to the size of the scatterer to the sixth power. Because the range of glass bead diameters in phantom B is large, i.e., 9–45 μm, an average estimate of 40 μm is reasonable. By using 40 μm as reference value, good estimates were obtained for depths smaller than 1.0 and 1.5 cm for CP and CEPC, respectively. At these depths, estimates started to diverge from the reference value by more than 25% and the standard deviations increased above 70%. eSNR values ranged from 0 to 1.6 dB for CP and from 0 to 16 dB for CEPC. Moreover, at the diverging point, the eSNR was 0.5 dB for CP and 2.4 dB for CEPC. Diverging point is defined as the point in depth, where the deviations from the reference value or the error in the estimates where too high (>25% error). Therefore, the diverging point for CP is considered to occur at 1.0 cm even though the estimate bias appears to decrease between the depths of 1.2–1.4 cm. However, at a depth of 1.2 cm, the estimate bias is well above 100%, and at the depths of 1.3 and 1.4 cm, the standard deviations are above 100%. A similar behavior was observed from phantoms C and D. From phantom C (Fig. 10), CEPC produced accurate estimates until 1.6 cm depth was reached; the eSNR was 5.7 dB at this point. CP only produced accurate estimates in depths smaller than 1.1 cm, where eSNR was larger than 0.5 dB. eSNR values ranged from 0 to 1.8 dB for CP and from 1.2 to 16 dB for CEPC. Similar responses were observed from phantom D (Fig. 11), however, the diverging points occurred at different positions. The estimates were close to the median value (better than 25%) in depths smaller then 1 cm when CP was used and in depths smaller than 1.8 cm when CEPC was used. Again, it was observed that close to the diverging point, the standard deviations significantly increased. eSNR values were 0.6 dB for CP and 1.7 dB for CEPC at the diverging point. Table III summarizes the value of depth at which the bias begins to diverge and the eSNR value where the divergence occurs.

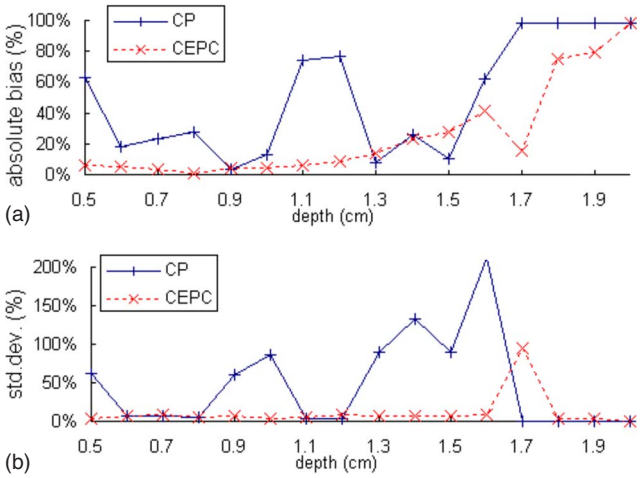


FIG. 10. (Color online) Phantom C: Estimated average scatterer sizes (plot a) and corresponding standard deviations (plot b) of 10 MHz measurements.

IV. DISCUSSION AND CONCLUSIONS

For the 5 MHz measurements, no significant differences in the estimated scatterer diameters were observed between the CP method and the CEPC method. This was due to larger eSNR available throughout the depth of phantom when using 5 MHz. This was not the case for the 10 MHz data, where the attenuation quickly attenuated the signal, resulting in decrease in the ability to accurately estimate scatterer size. The

TABLE III. Summary of estimate results at the 25% bias divergence by using 10 MHz.

	Bias depth (cm)		eSNR (dB)	
Phantom	CP	CEPC	CP	CEPC
A	1.7	2.0	0.19	1.74
B	1.0	1.6	0.35	1.48
C	1.1	1.6	0.42	4.68
D	1.0	1.8	0.52	1.64

use of CEPC at 10 MHz resulted in a significant increase in penetration depth yielding accurate estimates and provided the ability to estimate the correct diameter of the scattering objects in regions where the CP method failed. When comparing the results of phantom A, a gain in penetration depth of at least 0.4 cm was obtained when CEPC was used instead of CP. If a greater depth was examined with phantom A at 10 MHz, the increase in depth of penetration may have been even larger. When comparing the results of phantoms B, C, and D, the gain in penetration depth was about 50% (7, 6, and 8 mm, respectively) when CEPC was used instead of CP.

Any improvement in estimate variance when using CEPC was found to be modest at best; however, it can be concluded that the use of CEPC did not cause the estimate variance to become worse. Variance in scatterer size estimates is driven by electronic noise at low eSNR and by spatial variation noise due to the random location of scatterers in the phantoms.[28] While coded excitation can increase the eSNR related to electronic noise, it cannot reduce the spatial variation noise (the scatterers are still located spatially at random).

The comparison of CP and CEPC suggests that for the 10 MHz measurements, where the eSNR was crucial for the success of the estimations, using CEPC yielded significant benefits. This is highlighted by the fact that CEPC was more successful than CP in deeper regions of the phantoms, where the attenuation was high and the amplitude of the noise was much higher than the amplitude of the signal, even after averaging. The expected gain in penetration depth when CEPC was used led to a higher eSNR, which improved the bias and variance of estimates.

When analyzing the tissue, the attenuation of the ultrasound decreases eSNR with the depth of the ROI and with an increase in the center frequency. In order to keep good measurement conditions in terms of scatterer diameter and frequency ($ka > 0.5$), it is often necessary to use higher frequencies when the scatterers are relatively small. The use of higher frequencies leads to higher attenuation and the need for increasing the eSNR. When coded excitation is used in clinical ultrasound, the eSNR can be safely increased without exceeding the regulatory limits and with pulse compression the spatial resolution can be preserved.

The current study suggests that it is possible to use CEPC when estimating the average diameter of scatterers by measuring backscatter coefficients inside TM phantoms. One advantage of using CEPC over CP techniques was the improvement in accuracy of scatterer diameter estimates in measurement scenarios where the level of attenuation was high enough such that the eSNR was below 1 dB. More important, accurate scatterer diameter estimates were obtained for low eSNR signals, indicating that the ability to estimate scatterer properties is robust. A second advantage was the improvement in variance for estimates by using CEPC over CP. While the improvement in estimate variance was small, it was still quantifiable and will improve the ability to distinguish tissue types using scatterer size imaging.

Although the penetration depth for scatterer size imaging was increased in the phantom studies, the gain was less than predicted. For example, the gain in eSNR by using CEPC was 14–15 dB. If the attenuation was 0.7 dB MHz^{-1} cm^{-1}, then at 10 MHz the loss per centimeter of penetration is 7 dB. Therefore, for CEPC, the expected gain in penetration depth should be a full 2 cm. The largest gain in the experimental results was 0.8 cm, less than half the gain that would be predicted by the increase in eSNR. The reason for this discrepancy may be due to the analysis bandwidth being used. While 10 MHz is the center frequency of the source, the actual bandwidth used includes both lower and higher frequencies. Therefore, the loss of the higher frequencies, which attenuate more rapidly, may cause the bias to the estimate to increase earlier in the phantom. The *ka* range due to the decreasing analysis bandwidth will become smaller and may reduce below 0.5. Low *ka* values correspond to a decreased ability to accurately estimate the scatterer size because the frequency dependence of scattering is not greatly affected by scatterer size at low *ka*.[23]

In future studies, other structural properties of tissue derived from the backscatter coefficients, i.e., scatterer concentration, could be examined by using CEPC. CEPC will be an important tool to analyze tissue by measuring the backscatter coefficient and estimating structural properties, especially when the scatterers are relatively small and higher frequencies have to be used to have a *ka* above 0.5. This could have a significant impact in tissue characterization of superficial cancer sites, e.g., the thyroid, breast, testes, or prostate, where higher frequencies may reveal more detail and important microstructure related to scattering from cells.[27] Future studies will also examine the effects of side lobes on scatterer property estimates, the effects of ROI size on estimate variance and bias by using CEPC, and application of CEPC and quantitative ultrasound to animal models in cancer detection.

Finally, the frequency modulated coded excitation was evaluated for improving the estimation of scatterer size. However, the use of phase based codes, e.g., Golay or Barker codes, could also be used to improve scatterer size estimates with similar expected benefits. Future work will examine the use of alternate coding schemes to improve the penetration depth of scatterer size imaging.

ACKNOWLEDGMENTS

The authors would like to acknowledge the helpful discussions and technical assistance of J. Sanchez, R. Lavarello, A. Haak, and A. Thakkar. This work was supported by a grant from the NIH (Grant No. R21 EB006741).

[1] J. G. Miller, J. E. Perez, J. G. Mottley, E. I. Madaras, P. H. Johnston, E. D. Blodgett, L. J. Thomas, III, and B. E. Sobel, "Myocardial tissue characterization: An approach based on quantitative backscatter and attenuation," Ultrason. Symp. Proc. **2**, 782–793 (1983).
[2] F. L. Lizzi, M. Greenebaum, E. J. Feleppa, M. Elbaum, and D. J. Coleman, "Theoretical framework for spectrum analysis in ultrasonic tissue characterization," J. Acoust. Soc. Am. **73**, 1366–1373 (1983).
[3] F. L. Lizzi, M. Ostromogilsky, E. J. Feleppa, M. C. Rorke, and M. M. Yaremko, "Relationship of ultrasonic spectral parameters to features of tissue microstructure," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **34**, 319–329 (1987).
[4] M. F. Insana, T. J. Hall, and J. L. Fishback, "Identifying acoustic scattering sources in normal renal parenchyma from the anisotropy in acoustic properties," Ultrasound Med. Biol. **17**, 613–626 (1991).
[5] M. L. Oelze, J. F. Zachary, and W. D. O'Brien, Jr., "Characterization of

tissue microstructure using ultrasonic backscatter: Theory and technique optimization using a Gaussian form factor," J. Acoust. Soc. Am. **112**, 1202–1211 (2002).
[6]M. F. Insana and T. J. Hall, "Parametric ultrasound imaging from backscatter coefficient measurements: Image formation and interpretation," Ultrason. Imaging **12**, 245–267 (1990).
[7]M. F. Insana and D. G. Brown, "Acoustic scattering theory applied to soft biological tissues," in *Ultrasonic Scattering in Biological Tissues*, edited by K. K. Shung and G. A. Thieme (CRC, Boca Raton, FL, 1993), pp. 75–124.
[8]E. J. Feleppa, T. Liu, A. Kalisz, M. C. Shao, N. Fleshner, V. Reuter, and W. R. Fair, "Ultrasonic spectral-parameter imaging of the prostate," Int. J. Imaging Syst. Technol. **8**, 11–25 (1997).
[9]F. L. Lizzi, M. Astor, T. Liu, C. Deng, D. J. Coleman, and R. H. Silverman, "Ultrasonic spectrum analysis for tissue assays and therapy evaluation," Int. J. Imaging Syst. Technol. **8**, 3–10 (1997).
[10]A. Nair, B. D. Kuban, N. Obuchowski, and D. G. Vince, "Assessing spectral algorithms to predict atherosclerotic plaque composition with normalized and raw intravascular ultrasound data," Ultrasound Med. Biol. **27**, 1319–1331 (2001).
[11]J. Liu and M. F. Insana, "Coded pulse excitation for ultrasonic strain imaging," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**, 231–240 (2005).
[12]A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).
[13]J. F. Zachary, J. M. Sempsrott, L. A. Frizzell, D. G. Simpson, and W. D. OBrien, Jr., "Superthreshold behavior and threshold estimation of ultrasound-induced lung hemorrhage in adult mice and rats," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **48**, 581592 (2001).
[14]M. L. Oelze and W. D. O'Brien, Jr., "Application of three scattering models to the characterization of solid tumors in mice," Ultrason. Imaging **28**, 83–96 (2006).
[15]N. A. H. K. Rao, "Investigation of a pulse compression technique for medical ultrasound: A simulation study," Med. Biol. Eng. Comput. **32**, 181–188 (1994).
[16]C. E. Cook and W. M. Siebert, "The early history of pulse compression radar," IEEE Trans. Aerosp. Electron. Syst. **24**, 825–833 (1988).
[17]B. Haider, P. A. Lewis, and K. E. Thomenius, "Pulse elongation and deconvolution filtering for medical ultrasound imaging," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **45**, 98–113 (1998).
[18]K. Raum and W. D. O'Brien, Jr., "Pulse-echo field distribution measurement technique of high-frequency ultrasound sources," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 810–815 (1997).
[19]K. A. Wear, T. A. Stiles, G. R. Frank, E. L. Madsen, F. Cheng, E. J. Feleppa, C. S. Hall, B. S. Kim, P. Lee, W. D. O'Brien, Jr., M. L. Oelze, B. I. Raju, K. K. Shung, T. A. Wilson, and J. R. Yuan, "Interlaboratory comparison of ultrasonic backscatter coefficient measurements from 2 to 9 MHz," J. Ultrasound Med. **24**, 1235–1250 (2005).
[20]E. Madsen, F. Dong, G. R. Frank, B. S. Garra, K. A. Wear, T. Wilson, J. A. Zagzebski, H. L. Miller, K. K. Shung, S. H. Wang, E. J. Feleppa, T. Liu, W. D. O'Brien, Jr., K. A. Topp, N. T. Sanghvi, A. V. Zaitsev, T. J. Hall, J. B. Fowlkes, O. D. Kripfgans, and J. G. Miller, "Interlaboratory comparison of ultrasonic backscatter: Attenuation, and speed measurements," J. Ultrasound Med. **18**, 615–631 (1999).
[21]J. J. Faran, "Sound scattering by solid cylinders and spheres," J. Acoust. Soc. Am. **23**, 405–418 (1951).
[22]R. Hickling, "Analysis of echoes from a solid elastic sphere in water," J. Acoust. Soc. Am. **34**, 1582–1592 (1962).
[23]M. F. Insana, R. F. Wagner, D. G. Brown, and T. J. Hall, "Describing small-scale structure in random media using pulse-echo ultrasound," J. Acoust. Soc. Am. **78**, 179–192 (1990).
[24]M. L. Oelze, "Bandwidth and resolution enhancement through pulse compression," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **54**, 768–781 (2007).
[25]M. Pollakowski, H. Ermert, L. von Bernus, and T. Schmeidl, "The optimum bandwidth of chirp signals in ultrasonic applications," Ultrasonics **31**, 417–420 (1993).
[26]J. K. Tsou, J. Liu, and M. F. Insana, "Modeling and phantom studies of ultrasonic wall shear rate measurements using coded pulse excitation," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 724–734 (2006).
[27]M. L. Oelze and J. F. Zachary, "Examination of cancer in mouse models using high-frequency quantitative ultrasound," Ultrasound Med. Biol. **32**, 1639–1648 (2006).
[28]M. L. Oelze and W. D. O'Brien, Jr., "Defining optimal axial and lateral resolution for estimating scatterer properties from volumes using ultrasound backscatter," J. Acoust. Soc. Am. **115**, 3226–3234 (2004).