CODEN: JASMAN

The Journal of the Acoustical Society of America

ISSN: 0001-4966

Vol. 123, No. 1 January 2008

ACOUSTICAL NEWS-USA 1

USA Meeting Calendar 1

ACOUSTICAL NEWS-INTERNATIONAL 3

Standards Meeting Calendar 3

REVIEWS OF ACOUSTICAL PATENTS 13

LETTERS TO THE EDITOR

On fiber optic probe hydrophone measurements in a cavitating liquid (L) — Aaldert Zijlstra, Claus Dieter Ohl — 29

Power-output regularization in global sound equalization (L) — Nick Stefanakis, John Sarris, George Cambourakis, Finn Jacobsen — 33

Comment on "A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties" [J. Acoust. Soc. Am. 119, 2334–2350 (2006)] (L) — Geoffrey Stewart Morrison — 37

UNDERWATER SOUND [30]

Modal group time spreads in weakly range-dependent deep ocean environments — Ilya A. Udovydchenkov, Michael G. Brown — 41

Parabolic equation solution of seismo-acoustics problems involving variations in bathymetry and sediment thickness — Jon M. Collis, William L. Siegmann, Finn B. Jensen, Mario Zampolli, Elizabeth T. Küsel, Michael D. Collins — 51

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

Prediction of negative dispersion by a nonlocal poroelastic theory — Abir Chakraborty — 56

TRANSDUCTION [38]

Evaluation of the angular spectrum approach for simulations of near-field pressures — Xiaozheng Zeng, Robert J. McGough — 68

Comparative measurements of loudspeakers in a listening situation — Mathieu Lavandier, Philippe Herzog, Sabine Meunier — 77

STRUCTURAL ACOUSTICS AND VIBRATION [40]

On the reflection of coupled Rayleigh-like waves at surface defects in plates — Bernard Masserey, Paul Fromme — 88

Faxén relations in solids—a generalized approach to particle motion in elasticity and viscoelasticity — Andrew N. Norris — 99

(Continued)

CONTENTS—*Continued from preceding page*

Approximations of inverse boundary element methods with partial measurements of the pressure field — Nicolas P. Valdivia, Earl G. Williams, Peter C. Herdic — 109

NOISE: ITS EFFECTS AND CONTROL [50]

A ray model for hard parallel noise barriers in high-rise cities — Kai Ming Li, Man Pun Kwok, Ming Kan Law — 121

The effects of environmental and classroom noise on the academic attainments of primary school children — Bridget M. Shield, Julie E. Dockrell — 133

ARCHITECTURAL ACOUSTICS [55]

On boundary conditions for the diffusion equation in room-acoustic prediction: Theory, simulations, and experiments — Yun Jing, Ning Xiang — 145

Spatial correlation and coherence in reverberant acoustic fields: Extension to microphones with arbitrary first-order directivity — Martin Kuster — 154

Subjective and objective assessment of acoustical and overall environmental quality in secondary school classrooms — Arianna Astolfi, Franco Pellerey — 163

ACOUSTIC SIGNAL PROCESSING [60]

A comparison of filter design structures for multi-channel acoustic communication systems — Pierre M. Dumuid, Ben S. Cazzolato, Anthony C. Zander — 174

Guided wave arrays for high resolution inspection — Alexander Velichko, Paul D. Wilcox — 186

PHYSIOLOGICAL ACOUSTICS [64]

Middle-ear circuit model parameters based on a population of human ears — Kevin N. O'Connor, Sunil Puria — 197

Reconciling the origin of the transient evoked ototacoustic emission in humans — Robert H. Withnell, Chantel Hazlewood, Amber Knowlton — 212

Supporting evidence for reverse cochlear traveling waves — W. Dong, E. S. Olson — 222

PSYCHOLOGICAL ACOUSTICS [66]

Sample discrimination of frequency by hearing-impaired and normal-hearing listeners — Joshua M. Alexander, Robert A. Lutfi — 241

The effect of masker level uncertainty on intensity discrimination — Emily Buss — 254

Across-channel interference in intensity discrimination: The role of practice and listening strategy — Emily Buss — 265

Nonconscious control of fundamental voice frequency — Honorata Zofia Hafke — 273

Listeners' sensitivity to "onset/offset" and "ongoing" interaural delays in high-frequency, sinusoidally amplitude-modulated tones — Thomas N. Buell, Sarah J. Griffin, Leslie R. Bernstein — 279

Influences of auditory object formation on phonemic restoration — Barbara G. Shinn-Cunningham, Dali Wang — 295

The role of spectral modulation cues in virtual sound localization — Jinyu Qian, David A. Eddins — 302

Comparison of adaptive psychometric procedures motivated by the Theory of Optimal Experiments: Simulated and experimental results — Jeremiah J. Remus, Leslie M. Collins — 315

SPEECH PRODUCTION [70]

Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002 — Brad H. Story — 327

(Continued)

CONTENTS—*Continued from preceding page*

Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels: Statistical considerations	Michel T.-T. Jackson, Richard S. McGowan	336
Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography	Jan G. Švec, Johan Sundberg, Stellan Hertegård	347
The interplay between the auditory and visual modality for end-of-utterance detection	Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts	354

SPEECH PERCEPTION [71]

Absorption of reliable spectral characteristics in auditory perception	Michael Kiefte, Keith R. Kluender	366
Spectral structure across the syllable specifies final-stop voicing for adults and children alike	Susan Nittrouer, Joanna H. Lowenstein	377
Spectral tilt change in stop consonant perception	Joshua M. Alexander, Keith R. Kluender	386
Training English listeners to perceive phonemic length contrasts in Japanese	Keiichi Tajima, Hiroaki Kato, Amanda Rothwell, Reiko Akahane-Yamada, Kevin G. Munhall	397
The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception	Martin Cooke, M. L. Garcia Lecumberri, Jon Barker	414
Auditory-visual speech perception in normal-hearing and cochlear-implant listeners	Sheetal Desai, Ginger Stickney, Fan-Gang Zeng	428
Perceptual coherence in listeners having longstanding childhood hearing losses, listeners with adult-onset hearing losses, and listeners with normal hearing	Andrea Pittman	441
Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects	Helen E. Cullington, Fan-Gang Zeng	450
Longitudinal changes in speech recognition in older persons	Judy R. Dubno, Fu-Shing Lee, Lois J. Matthews, Jayne B. Ahlstrom, Amy R. Horwitz, John H. Mills	462
Effect of age, presentation method, and learning on identification of noise-vocoded words	Signy Sheldon, M. Kathleen Pichora-Fuller, Bruce A. Schneider	476
Priming and sentence context support listening to noise-vocoded speech by younger and older adults	Signy Sheldon, M. Kathleen Pichora-Fuller, Bruce A. Schneider	489

MUSIC AND MUSICAL INSTRUMENTS [75]

Spectral envelope sensitivity of musical instrument sounds	David Gunawan, D. Sen	500

BIOACOUSTICS [80]

Measurements and predictions of hooded crow (*Corvus corone cornix*) call propagation over open field habitats	Kenneth Kragh Jensen, Ole Næsbye Larsen, Keith Attenborough	507
Harmonic pulsed excitation and motion detection of a vibrating reflective target	Matthew W. Urban, James F. Greenleaf	519
Hearing sensitivity during target presence and absence while a whale echolocates	Alexander Ya. Supin, Paul E. Nachtigall, Marlee Breese	534
Estimating bottlenose dolphin (*Tursiops truncatus*) hearing thresholds from single and multiple simultaneous auditory evoked potentials	James J. Finneran, Dorian S. Houser, Dave Blasko, Christie Hicks, Jim Hudson, Mike Osborn	542
Evidence for double acoustic windows in the dolphin, *Tursiops truncatus*	Vladimir V. Popov, Alexander Ya. Supin, Vladimir O. Klishin, Mikhail B. Tarakanov, Mikhail G. Pletenko	552

(Continued)

CONTENTS—*Continued from preceding page*

JASA EXPRESS LETTERS

Effects of level and background noise on interaural time difference discrimination for transposed stimuli	Anna A. Dreyer, Andrew J. Oxenham	EL1
Structural health monitoring by extraction of coherent guided waves from diffuse fields	Karim G. Sabra, Ankit Srivastava, Francesco Lanza di Scalea, Ivan Bartoli, Piervincenzo Rizzo, Stephane Conti	EL8
Mechanical response measurements of real and artificial brass players lips	Michael J. Newton, Murray Campbell, Jöel Gilbert	EL14

CUMULATIVE AUTHOR INDEX 569

Effects of level and background noise on interaural time difference discrimination for transposed stimuli

Anna A. Dreyer
Harvard University/MIT Division of Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139
adreyer@alum.mit.edu.

Andrew J. Oxenham
Department of Psychology, University of Minnesota, Elliott Hall, 75 East River Parkway, Minneapolis, Minnesota 55455
oxenham@umn.edu.

Abstract: Just-noticeable interaural time differences were measured for low-frequency pure tones, high-frequency sinusoidally amplitude-modulated (SAM) tones, and high-frequency transposed stimuli, at multiple levels with or without a spectrally notched diotic noise to prevent spread of excitation. Performance with transposed stimuli and pure tones was similar in quiet; however, in noise, performance was poorer for transposed stimuli than for pure tones. Performance with SAM tones was always poorest. In all conditions, performance improved slightly with increasing level. The results suggest that the equivalence postulated between transposed stimuli and pure tones is not valid in the presence of a spectrally notched background noise.

PACS numbers: 43.66.Pn [Q-JF]
Date Received: July 12, 2007 **Date Accepted:** October 23, 2007

1. Introduction

Binaural hearing allows humans to localize sound sources in space and can assist in the perceptual segregation of competing sound sources. A dominant cue for determining the azimuth of a sound source is the interaural time difference (ITD) (e.g., Macpherson and Middlebrooks, 2002). Sensitivity to ITD is found for long-duration pure tones only below about 1500 Hz; however, when complex stimuli with time-varying temporal envelopes are used, some ITD sensitivity is also observed for sounds containing only high frequencies (e.g., Klumpp and Eady, 1956; Henning, 1974; McFadden and Pasanen, 1976). It has been suggested that any remaining differences between low- and high-frequency processing of ITDs might be attributed to stimulus characteristics, rather than any inherent deficit in temporal sensitivity at high frequencies (Colburn and Esquissaud, 1976). Van de Par and Kohlrausch (1997) developed so-called "transposed stimuli" to specifically test this hypothesis. Transposed stimuli are designed to produce temporal discharge patterns in the auditory nerve in response to high-frequency modulated stimuli that resemble the discharge patterns in response to low-frequency pure tones. Transposed stimuli are constructed by modulating a high-frequency pure-tone carrier with a halfwave-rectified low-frequency sinusoid.

Using transposed stimuli, Bernstein (2001) and Bernstein and Trahiotis (2002) showed that ITD-based discrimination and lateralization of transposed stimuli was as good as, or better than, that for pure tones at frequencies (or modulation rates) up to 150 Hz. Furthermore, transposed stimuli produced markedly better performance than sinusoidally amplitude-modulated (SAM) tones. However, as (modulation) frequency was increased, lateralization performance using transposed stimuli fell below that using pure tones, suggesting that additional factors beyond peripheral representation also play a role.

The psychophysical data suggest that transposed stimuli are successful to some extent in enhancing the temporal representation of the modulation frequency. However, the question of how well transposed stimuli actually simulate pure-tone temporal discharge patterns in the auditory nerve was tested only recently. Contrary to the idealized model used for motivating transposed stimuli, Dreyer and Delgutte (2006) found that the phase locking to pure tones and transposed stimuli is only similar near rate threshold in the cat auditory nerve. At levels greater than 10 dB above neural threshold, the synchronization index in response to transposed (and SAM) tones was found to decrease rapidly with increasing stimulus level, whereas the synchronization index for pure tones remained more stable with level. Most psychophysical studies using transposed stimuli have been performed well above absolute threshold, where many neural units should be well above their threshold, suggesting a possible discrepancy between the neural and behavioral measures. In other words, based on the synchrony indices from single units in the auditory nerve of cats, one might not expect psychophysical performance using transposed tones to be as good as that using pure tones, in contrast to the available data (Bernstein and Trahiotis, 2002).

One reason for the apparent discrepancy between the neural and psychophysical data may be that information from off-frequency neurons (i.e., neurons with best frequencies that do not match that of the stimulus) determines performance and that, when the stimuli are presented in quiet, there will always be some neurons responding to the stimulus that are close to their rate threshold (e.g., Siebert, 1968). In this way, "off-frequency listening" may play a role in producing equivalent ITD performance for pure-tone and transposed stimuli, even at levels far above threshold.

We tested this hypothesis by presenting the stimuli in a background noise, which was designed to limit the ability of listeners to use off-frequency information, while minimizing the amount of direct interference (i.e., masking) produced by the background noise. Just-noticeable differences in ITD were measured as a function of level for pure tones, transposed stimuli, and SAM tones (as in Bernstein and Trahiotis, 2002). Thresholds were measured in quiet (unrestricted listening) and in the presence of a background noise, spectrally shaped to restrict the spectral region over which information was available, thereby more accurately reflecting the conditions tested in the physiological experiments.

2. Methods

Detection of ongoing ITDs was measured behaviorally in humans as a function of level in unrestricted and restricted listening conditions using low-frequency pure tones, 100% modulated SAM tones, and transposed stimuli. The frequency of the pure tones and the modulation rate of SAM tones and transposed stimuli was always 125 Hz, which corresponds to a frequency at which ITD performance using pure-tone and transposed stimuli has been found to be equivalent (Bernstein and Trahiotis, 2002; Oxenham *et al.*, 2004). All stimuli were presented at levels between 30 and 70 dB sound pressure level (SPL) in 10 dB steps. The SAM tones were generated by modulating a 4 kHz sinusoid with a 125 Hz sinusoid. The transposed tones were generated by multiplying a 4 kHz sinusoid with a 125 Hz sinusoid that had been half-wave rectified and lowpass filtered (Butterworth fourth order) at 800 Hz (0.2 times the carrier frequency). The carrier frequency of 4 kHz is above the assumed range of optimal phase locking by individual auditory nerve fibers (ANFs). Pure tones, SAM tones and transposed stimuli were all 500 ms in length, including 100 ms onset and offset ramps to prevent spectral splatter and to de-emphasize the salience of the on- and offsets. For the high-frequency stimuli, only the ongoing portion of the temporal envelope was delayed.

The pure tones, SAM tones, and transposed stimuli were presented either with or without a background noise designed to limit off-frequency listening (e.g., O'Loughlin and Moore, 1981). For the SAM and transposed stimuli, the noise contained a stop-band between 0.9 and $1.1f_c$ (3600–4400 Hz), so that the noise energy was always at least 10% of the center frequency away from the stimulus (see Fig. 1(a)). For the pure-tone stimuli, a bandpass noise was used, which extended from 150 to 350 Hz, resulting in the noise being no less than 20% of the pure-tone frequency away from the tone. A larger spectral gap was chosen to compensate for the

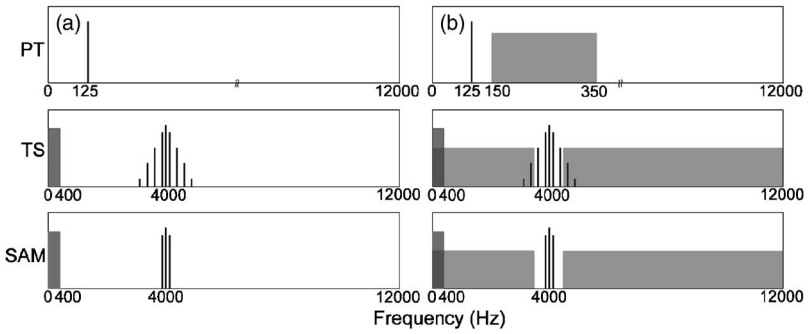


Fig. 1. Schematic illustration of the experimental stimuli. The unrestricted listening condition (a) includes only the low-pass background noise for the high-frequency stimuli (0–400 Hz, shown in dark gray), while the restricted listening condition (b) includes high-pass noise for pure-tone stimuli, and notched noise for SAM tones and transposed stimuli (shown in light gray). The frequency scales for the pure-tone and high-frequency stimuli differ. The frequency axis for the pure-tone stimuli is expanded from 0–400 Hz with a break in the pure-tone frequency axis at 400 Hz, while no such expansion or break is used for representing the high-frequency stimuli.

relatively broader auditory filters at low frequencies (e.g., Glasberg and Moore, 1990), and no noise was added below the pure-tone frequency, because it was deemed unlikely that off-frequency listening would play a role below 100 Hz (e.g., Sek and Moore, 1994). Interaurally correlated noise was used, as informal pilot tests suggested that listeners found it difficult to reliably lateralize transposed stimuli in uncorrelated notched noise within physically realizable ITD limits. Greater interference from uncorrelated noise has also been observed by Ito *et al.* (1982); however, because the task involved distinguishing a right- from a left-leading ITD, correlated noise should not provide additional lateralization cues (e.g., Nuetzel and Hafter, 1976).

For all presentations of SAM tones and transposed stimuli, an additional uncorrelated noise, low-pass filtered with a cutoff frequency of 400 Hz (see Fig. 1(b)), was presented at equal levels to both ears to prevent additional lateralization cues from auditory distortion products. The spectrum level of the low-pass-filtered noise ranged from 29 to 69 dB (re: 20 μPa) for the 30 and 70 dB SPL probe tone, respectively. Adequate masking of distortion products by background noise was confirmed for the 70 dB SPL tone, using the method described by Vliegen and Oxenham (1999).

All noise, when present, was gated on 400 ms prior to the first stimulus interval, and was gated off 200 ms after the second interval, within each two-interval trial, for a total duration of 2.1 s. Both the onset and offset of the noise were shaped with 100 ms raised-cosine ramps.

The stimuli were digitally generated and played through a LynxStudio LynxOne soundcard with 24 bit resolution at a 32 kHz sampling rate. The stimuli were presented to the subjects via Sennheiser HD 580 headphones after being passed though a TDT PA4 programmable attenuator and a TDT HB6 headphone buffer. All testing was conducted in a double-walled, sound-attenuating chamber.

Four normal-hearing subjects (thresholds below 20 dB HL at octave frequencies between 125 and 8000 Hz) served as listeners. Each subject began with a training phase with feedback until they reached asymptotic performance. Training for each subject lasted between 4 and 6 h. Feedback was also provided during the data collection phase. Subjects were paid an hourly wage for their services.

A preliminary masking experiment was performed to determine the levels of the background noise necessary to adequately limit off-frequency listening, while not directly masking the target sounds. Detection thresholds for the stimuli in quiet and in the notched (4 kHz carrier) or bandpass (125 Hz tone) noise were measured using a two-interval forced-choice procedure with a two-down one-up adaptive tracking rule (Levitt, 1971) at spectrum levels in the noise passband of 10, 20, 30, and 40 dB (re: 20 μPa). The masking noise had the same duration spectral characteristics as the noise in the main discrimination study. Subjects' diotic detection

Table 1. Background noise spectrum levels in dB (re: 20 μPa) used for the high-frequency (SAM and transposed) and pure-tone stimuli. Noise levels were the same for all subjects in the presence of the pure tones, but differed for each subject in the presence of the high-frequency stimuli.

		Stimulus level (dB SPL)			
		40	50	60	70
SAM and Transposed	Subject 1	4.2	12.1	19.9	27.8
	Subject 2	2.1	9.5	16.9	24.4
	Subject 3	1.8	10.2	18.5	26.8
	Subject 4	4.1	13.3	22.5	31.6
	Mean (S.D.)	*3.1 (1.3)*	*11.3 (1.7)*	*19.5 (2.3)*	*27.7 (3.0)*
Pure tone	All Subjects	*18.2*	*29.0*	*39.9*	*50.8*

thresholds in quiet and in the two noise conditions were measured for all listeners. Individual signal thresholds were plotted as a function of noise spectrum level and a linear regression was performed to determine the threshold level as a function of the noise level for each subject. The noise level needed to just mask a tone of a particular level was estimated from this regression. The subjects' diotic detection thresholds for pure tones were similar and therefore the background noise levels were chosen by averaging the detection thresholds for all the subjects. Because the detection thresholds differed somewhat between subjects for the high-frequency stimuli, the noise levels were selected individually for each subject. These thresholds were then used to set the band-restricting noise in the ITD discrimination experiment at a level 20 dB below that needed to mask the signal for each subject individually. The noise spectrum levels used in the discrimination experiments are shown in Table 1.

In the lateralization tasks, discrimination thresholds were measured using a two-interval, two-alternative forced-choice method with a two-down, one-up adaptive procedure that tracks the 71% point on the psychometric function (Levitt, 1971). The listeners' task was to decide whether the right ear led in the first or second presentation. Thresholds were obtained in each listening condition in random order and each condition was run three times. The initial ITD was 500 μs (i.e., $\pm$250 μs) and the initial step size was a factor of 2, which was reduced to 1.4 after two reversals and to 1.18 after another two reversals. For each run, testing continued until four reversals at the smallest step size occurred and the threshold was taken as the mean of these four reversals. Final thresholds were geometrically averaged for each listener and stimulus condition over the three adaptive runs.

3. Results

One of the four subjects had some difficulty detecting any ITDs when the tone was at a level of 30 dB SPL and, because of these missing data, only thresholds between 40 and 70 dB SPL are shown. The pattern of results was rather similar across the four subjects, and so only the (geometric) mean results from the ITD discrimination experiment are shown in Fig. 2. Consider first the results without the additional masking noise (Panel A). Thresholds for the pure tones (diamonds) and transposed stimuli (squares) are very similar, in line with previous results showing similar pure-tone and transposed-tone thresholds at a frequency of 125 Hz (e.g., Bernstein and Trahiotis, 2002). Thresholds using SAM tones are elevated relative to those for pure and transposed tones by a factor of between 2 and 3, again broadly consistent with previous studies (Bernstein and Trahiotis, 2002). A repeated-measures analysis of variance (RMANOVA, with Huynh-Feldt correction for sphericity, where appropriate), carried out on the log-transformed data in Panel A confirmed these observations. There was a main effect of condition ($F_{2,6}=20.7, P=0.002$) and a main effect of level ($F_{3,9}=6.54, P=0.043$), with no significant interaction between the factors ($F_{6,18}<1$, n.s.). A contrast analysis revealed a significant linear trend for thresholds to decrease with increasing level ($F_{1,3}=447.0, P<0.001$). Finally, thresholds in

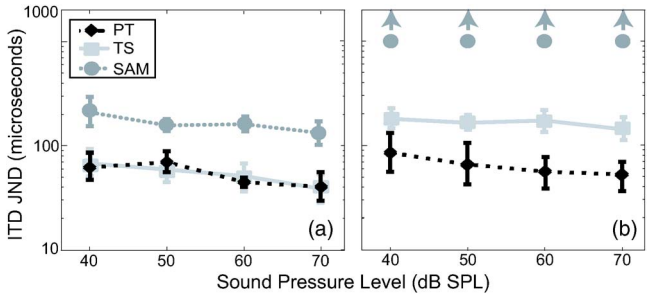


Fig. 2. (Color online) Interaural temporal just noticeable differences measured in an adaptive listening task for pure tones (PT; diamonds), transposed stimuli (TS; squares) and SAM tones (circles) without (a) and with (b) noise added to restrict the listening band. Upward arrows represent SAM tone conditions where the ITD thresholds exceeded 1000 μs. Error bars represent ±1 standard error of the mean.

the SAM condition were shown to be significantly higher than those in the pure-tone and transposed-tone conditions ($P<0.05$ in both cases), which were not significantly different from each other ($F_{1,3}<1$, n.s.).

Consider next the conditions in which the listening band was restricted by noise (Fig. 2(b)). Consistent with results from the unrestricted listening conditions, performance was poorest with the SAM tones. In fact, in this case listeners could not consistently perform the task at all for ITDs of 1000 μs or less, as indicated by the upward-pointing arrows in Fig. 2(b). In contrast to the results without the additional masking noise, performance with the transposed stimuli was consistently poorer than with the pure tones. The difference between pure-tone and transposed thresholds in noise was observed with all four subjects to varying degrees. One subject had thresholds in the transposed conditions that were on average only a factor of 1.2 higher than in the pure-tone conditions, whereas the others showed larger effects, with thresholds being higher by a factor of between 2.7 and 4.5. Averaged across subjects and levels, thresholds in the transposed conditions were about a factor of 3 higher than in the pure-tone conditions. There also remained a trend for improved thresholds with increasing level, which is particularly apparent in the pure-tone condition. A RMANOVA, carried out on just the pure-tone and transposed-tone data with noise, confirmed these observations: significant effects of level ($F_{3,9}=4.4, P=0.036$) and condition ($F_{1,3}=12.7, P=0.038$) were found, again with no interaction ($F_{3,9}<1$, n.s.).

A comparison of performance in the pure-tone and transposed-tone conditions with and without noise showed a significant effect of noise for the transposed-tone condition ($F_{1,3}=49.1, P=0.006$), but not for the pure-tone condition ($F_{1,3}<1$, n.s.). In both cases, there was no interaction between the effects of noise and level ($F_{3,9}<1$, n.s.). Overall, therefore, the additional noise had a substantial effect on thresholds in the transposed and SAM conditions, but not in the pure-tone condition.

4. Discussion

A comparison of ITD discrimination using pure tones and transposed stimuli found similar performance in the absence of background noise, consistent with earlier studies. However, when background noise was added, thresholds in the transposed-tone conditions deteriorated by a factor of about 3 on average, whereas pure-tone thresholds remained very similar. In all conditions, small improvements in performance were found with increasing stimulus level. The results show that the equivalence of transposed and pure tones at low (modulation) frequencies (Bernstein and Trahiotis, 2002) does not hold in the presence of noise. Caution should therefore be exercised in treating the information provided by the two types of stimuli as equivalent.

The question remains as to why the noise affects thresholds for the transposed and SAM stimuli more than for the pure-tone stimuli. The noise in our experiment was designed to restrict off-frequency listening and, as such, the results are consistent with the physiological

findings that the auditory-nerve synchrony to transposed tones is similar to that found for pure tones only when the firing rate of the fibers is close to threshold. As mentioned earlier, in the absence of noise, there will always be some off-frequency fibers whose rates are close to threshold, even at high stimulation levels. Thus, our results are consistent with the idea that the noise restricts the "listening band" to those fibers whose characteristic frequencies are close to that of the stimulus, leading to poor performance in the high-frequency conditions, where off-frequency listening is necessary to maintain performance.

Other interpretations are possible, however. For instance, a number of studies have shown that the presence of one or more simultaneous interfering stimuli can reduce sensitivity to target ITDs (e.g., Buell and Hafter, 1991; Woods and Colburn, 1992; Stellmack and Dye, 1993). However, it has been shown that a continuous diotic background noise in spectrally flanking regions produces little or no interference when it is presented continuously, instead of being gated with the target (Trahiotis and Bernstein, 1990). Furthermore, a recent study using transposed stimuli found no significant effects of interference by low-frequency noise, even when the noise was gated synchronously with the target (Bernstein and Trahiotis, 2004). Thus, while it remains a possibility that the diotic noise produced more spectrally remote interference for the transposed stimuli than for the pure-tone stimuli, this is made less likely by the fact that our noise was presented in a quasi-continuous manner (starting 400 ms before the first target and ending 200 ms after the second target). Another possibility is that the noise produced some direct masking effects. This may also explain our informal finding in pilot runs that listeners found the task much more difficult in a background of interaurally uncorrelated noise. Ito *et al.* (1982) also found that an interaurally uncorrelated broadband background noise with a spectrum level 10 dB below that of the narrowband target noise could increase ITD thresholds substantially. However, in their condition that was most comparable to ours (diotic noise masker), the amount of interference was relatively small for most subjects. Furthermore, our targets were presented 20 dB above masked threshold and were always separated from the noise by a spectral gap.

Finally, it is worth noting that even in the restricted listening cases, performance did not degrade with increases in level between 40 and 70 dB SPL. In fact a small but significant improvement with level was found. These findings do not mirror the physiological results of Dreyer and Delgutte (2006), who found that the synchronization index of the auditory-nerve discharge pattern in cats decreased with increasing level, particularly with SAM and transposed tones. This may be a species difference, or may reflect additional (perhaps efferent) influences, not found in the auditory nerve of an anesthetized cat, which help to maintain good representations of the temporal envelope across a wide dynamic range. Another possibility is that the noise itself shifts the operating point of neurons to higher levels, thereby effectively increasing the dynamic range (Palmer and Evans, 1982). Similar apparent discrepancies between auditory-nerve data and psychophysical performance have also been noted for other tasks, such as intensity discrimination (e.g., Viemeister, 1983).

Acknowledgments

This work was supported by NIH Grant Nos. T32 DC 00038 (AAD) and R01 DC 03909 (AJO). Leslie Bernstein, Christophe Micheyl, Bertrand Delgutte, Steve Colburn, and one anonymous reviewer provided helpful comments on earlier versions of this manuscript.

References and links

Bernstein, L. R. (**2001**). "Auditory processing of interaural timing information: New insights," J. Neurosci. Res. **66**, 1035–1046.

Bernstein, L. R., and Trahiotis, C. (**2002**). "Transposed stimuli improve sensitivity to envelope-based interaural timing information for stimuli having center frequencies of up to 10 kHz," J. Acoust. Soc. Am. **111**, 2466–2466.

Bernstein, L. R., and Trahiotis, C. (**2004**). "The apparent immunity of high-frequency "transposed" stimuli to low-frequency binaural interference," J. Acoust. Soc. Am. **116**, 3062–3069.

Buell, T. N., and Hafter, E. R. (**1991**). "Combination of binaural information across frequency bands," J. Acoust. Soc. Am. **90**, 1894–1900.

Colburn, H. S., and Esquissaud, P. (**1976**). "An auditory-nerve model for interaural time discrimination of high-

frequency complex stimuli," J. Acoust. Soc. Am. **59**, S23.
Dreyer, A. A., and Delgutte, B. (**2006**). "Phase locking of auditory-nerve fibers to the envelopes of high-frequency sounds: Implications for sound localization," J. Neurophysiol. **96**, 2327–2341.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Henning, G. B. (**1974**). "Detectability of interaural delay in high-frequency complex waveforms," J. Acoust. Soc. Am. **55**, 84–90.
Ito, Y., Colburn, H. S., and Thompson, C. L. (**1982**). "Masked discrimination of interaural time delays with narrow-band signal," J. Acoust. Soc. Am. **72**, 1821–1826.
Klumpp, R. G., and Eady, H. R. (**1956**). "Some measurements of interaural time difference thresholds," J. Acoust. Soc. Am. **28**, 859–860.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Macpherson, E. A., and Middlebrooks, J. C. (**2002**). "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," J. Acoust. Soc. Am. **111**, 2219–2236.
McFadden, D., and Pasanen, E. G. (**1976**). "Lateralization at high frequencies based on interaural time differences," J. Acoust. Soc. Am. **59**, 634–639.
Nuetzel, J. M., and Hafter, E. R. (**1976**). "Lateralization of complex waveforms: Effects of fine structure, amplitude, and duration," J. Acoust. Soc. Am. **60**, 1339–1345.
O'Loughlin, B. J., and Moore, B. C. J. (**1981**). "Off-frequency listening: Effects on psychoacoustical tuning curves obtained in simultaneous and forward masking," J. Acoust. Soc. Am. **69**, 1119–1125.
Oxenham, A. J., Bernstein, J. G. W., and Penagos, H. (**2004**). "Correct tonotopic representation is necessary for complex pitch perception," Proc. Natl. Acad. Sci. U.S.A. **101**, 1421–1425.
Palmer, A. R., and Evans, E. F. (**1982**). "Intensity coding in the auditory periphery of the cat: Responses of cochlear nerve and cochlear nucleus neurons to signals in the presence of bandstop masking noise," Hear. Res. **7**, 305–323.
Sek, A., and Moore, B. C. J. (**1994**). "The critical modulation frequency and its relationship to auditory filtering at low frequencies," J. Acoust. Soc. Am. **95**, 2606–2615.
Siebert, W. M. (**1968**). "Stimulus transformations in the peripheral auditory system," in *Recognizing Patterns*, edited by P. A. Kolers, and M. Eden (MIT Press, Cambridge, MA.), pp. 104–133.
Stellmack, M. A., and Dye, R. H. (**1993**). "The combination of interaural information across frequencies: The effects of number and spacing of the components, onset asynchrony, and harmonicity," J. Acoust. Soc. Am. **93**, 2933–2947.
Trahiotis, C., and Bernstein, L. R. (**1990**). "Detectability of interaural delays over select spectral regions: Effects of flanking noise," J. Acoust. Soc. Am. **87**, 810–813.
Van de Par, S., and Kohlrausch, A. (**1997**). "A new approach to comparing binaural masking level differences at low and high frequencies," J. Acoust. Soc. Am. **101**, 1671–1680.
Viemeister, N. F. (**1983**). "Auditory intensity discrimination at high frequencies in the presence of noise," Science **221**, 1206–1208.
Vliegen, J., and Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am. **105**, 339–346.
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.

Structural health monitoring by extraction of coherent guided waves from diffuse fields

Karim G. Sabra
School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0405, USA
karim.sabra@me.gatech.edu.

Ankit Srivastava, Francesco Lanza di Scalea, and Ivan Bartoli
NDE & SHM Laboratory, Department of Structural Engineering,
University of California, San Diego, La Jolla, California 92093-0085, USA
ansrivas@ucsd.edu; flanza@ucsd.edu; ibartoli@ucsd.edu

Piervincenzo Rizzo
Department of Civil and Environmental Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA
prizzo@engr.pitt.edu.

Stephane Conti
Marine Physical Laboratory, Scripps Institution of Oceanography, La Jolla, California 92037, USA
sconti@ucsd.edu.

Abstract: Recent theoretical and experimental studies in a wide range of applications have demonstrated that Green's functions (impulse responses) can be extracted from cross-correlation of diffuse fields using only passive sensors. This letter demonstrates the passive-only reconstruction of coherent Lamb waves (dc–500 kHz) in an aluminum plate of thickness comparable to aircraft fuselage and wing panels. It is further shown that the passively reconstructed waves are sensitive to the presence of damage in the plate as it would be expected in a typical "active" guided wave test. This proof-of-principle study suggests the potential for a structural health monitoring method for aircraft panels based on passive ultrasound imaging reconstructed from diffuse fields.

PACS numbers: 43.40.Le, 43.40.Sk, 43.40.Hb, 43.35.Zc [JGM]
Date Received: September 5, 2007 **Date Accepted:** October 30, 2007

1. Introduction

Effective structural health monitoring (SHM) techniques able to detect, locate and quantify damage are needed to ensure the proper performance and maintenance of current and future structures. SHM methods based on ultrasonic guided stress waves are particularly suitable for probing components with waveguide geometries, such as the skin panels of aircraft fuselage and wings (Rose, 1999; Giurgiutiu *et al.*, 2004; Lanza di Scalea *et al.*, 2007).

Guided-wave SHM is traditionally performed in one of two ways: an "active" approach involving at least a source and a receiver, or a "passive" approach involving only receivers. The two approaches offer complementary performances in terms of damage detectability: while the active version is used to detect existing defects in a postmortem mode, the passive version is a mode of acoustic emission monitoring where defects are detected as they grow in real time.

Diffuse, apparently incoherent fields in structures are generally considered "noise," and would be discarded in conventional SHM. However, from the long-time average of the cross correlation of such noise fields recorded at two sensing points, it is possible to reconstruct the coherent impulse response between the two passive sensors (Sabra *et al.*, 2007). A coherent waveform emerges from noise cross-correlation function (NCF) once the contributions of the diffuse noise sources traveling through both sensors are accumulated over time.

Theoretical and experimental studies have demonstrated the relationship between the Green's function and the NCF for various environments and frequency ranges such as seismology (Shapiro *et al.*, 2005; Sabra *et al.*, 2005c), underwater acoustics (Sabra *et al.*, 2005a, 2005b), civil engineering (Farrar and James 1997; Snieder and Cafak, 2006), low-frequency (<5 kHz) modal properties identification of hydrofoils (Sabra *et al.*, 2007) and ultrasonics (Weaver and Lobkis 2001; Larose *et al.*, 2006; Van Wijk, 2006).

The effects of attenuation and the spatio-temporal statistics of the noise sources often cause the amplitudes of the extracted coherent waveforms of the NCF to differ from those of the theoretical impulse response (Sabra *et al.*, 2005a; Roux *et al.*, 2005; Snieder, 2007). However, Larose *et al.* (2006) were able to image the reflections from a cylindrical hole drilled through an aluminum block using Rayleigh and bulk shear waves which were reconstructed passively. Indeed multiple scattering (e.g., due to a physical boundary or distributed scatterers) typically helps in achieving a fully diffuse wavefield, and thus providing a better estimate of the impulse response (Paul *et al.*, 2005; Larose *et al.*, 2006; Snieder, 2007). A detailed analysis of the role of multiple scattering on the accuracy of the passive imaging technique remains an open question.

This letter shows that passive-only guided-wave SHM can be achieved in a multiply scattering aluminum plate of thickness comparable to an aircraft panel. The passive reconstruction of coherent waves from diffuse fields retains the advantages of the active guided-wave monitoring without a controlled source. The potential could thus exist for passive ultrasound imaging of the aircraft fuselage and wings by exploiting the diffuse vibrations generated by flow unsteadiness during flight (vortices, turbulence) and/or by scattered wavefields (e.g., due to rivets, stiffeners or other structural details).

2. Theoretical background on coherent wave reconstruction from diffuse fields

The expected value of the temporal NCF between two sensors, $C_{12}(t)$, can be computed from the long-time cross correlation of the field $S_1(t)$ measured by sensor 1, and the field $S_2(t)$ measured by sensor 2

$$C_{12}(t) = \int_0^I S_1(\tau) S_2(\tau + t) d\tau, \tag{1}$$

where T is the observation period.

Although defined here in terms of a single temporal integration, the NCF may also be constructed from an ensemble average of shorter duration time averages. The existence of a fully diffuse ambient noise field in the structure is key for the implementation of the passive guided-wave SHM (see, for instance, Weaver and Lobkis, 2004, for further discussions on diffuse field characteristics). This condition ensures the uniform spatial and temporal distribution of the noise sources so that all propagation paths between the two sensors are fully illuminated. In this case, the formal relationship between the Fourier transform of the Green's function between the two sensors, $\widetilde{G}_{12}(\omega)$, and the expected value of the Fourier transform of the NCF, $\widetilde{C}_{12}(\omega)$, can be expressed as

$$\langle \widetilde{C}_{12}(\omega) \rangle = i\beta(\widetilde{G}_{12}(\omega) - \widetilde{G}^*_{21}(\omega)) \tag{2}$$

where $\langle\rangle$ stands for ensemble average and * stands for complex conjugate. The constant β is related to the power spectrum of the noise excitation, the nature of the recorded signals and the propagation medium. Thus, for a fully diffuse noise field, the NCF is a symmetric function of time. The energy equipartition of the diffuse field is a necessary and sufficient condition to extract the full Green's function from the NCF.

3. Experimental results

Extraction of Lamb coherent waves from diffuse fields was performed on a simple aluminum plate, 1.58 mm (1/16 in.) in thickness (Fig. 1(a)). The excitation was provided by multiple

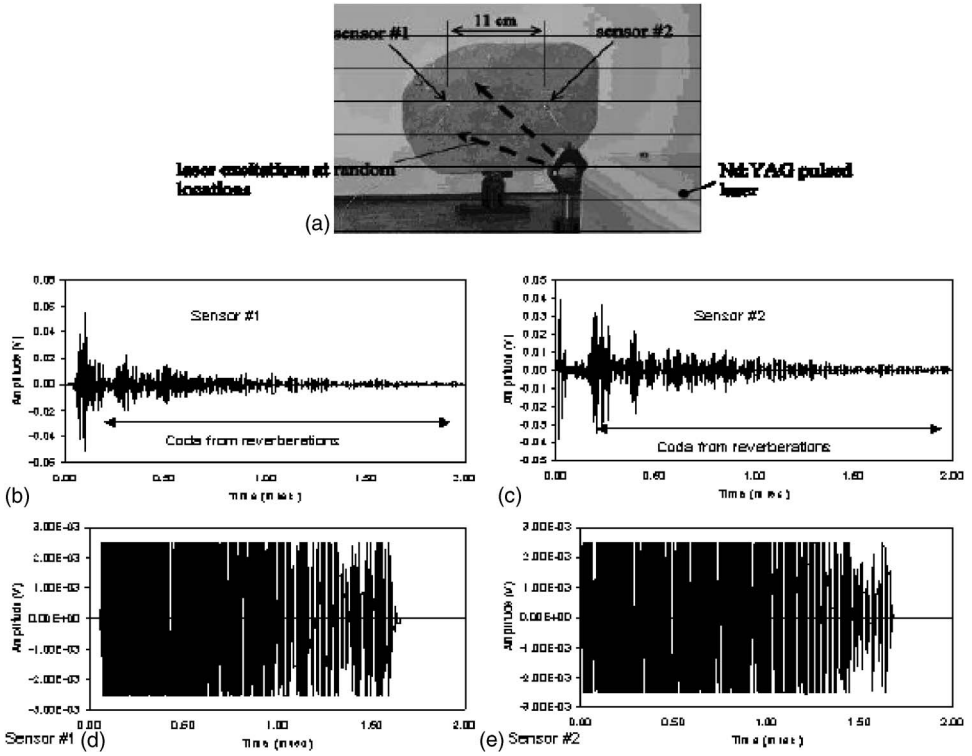


Fig. 1. Experiment on reconstructions of coherent Lamb waves from diffuse fields. (a) A 1.58-mm-thick aluminum plate subjected to random pulsed laser excitation. Plate edges were cut irregularly to randomize the field through multiple reverberations. (b), (c) Raw signals collected by the two sensors following a typical laser excitation showing long coda resulting from reverberations. (d), (e) Same waveforms as (b), (c) but clipped to same level equal to seven times the standard deviation of the electronic noise.

broadband irradiations from an Nd:yttrium—aluminum—garnet pulsed laser (~10 ns pulse duration at 1064 nm). The edges of the plate were cut at irregular geometries with no symmetries to produce multiple reverberations and chaotic trajectories of the pulsed excitation to generate diffuse fields. Two commercial acoustic emission sensors (PICO model, Physical Acoustics Corporation, broadband response between 100 kHz and 1 MHz), sensitive to out-of-plane displacements, were spaced 11 cm apart.

Shown in Figs. 1(b) and 1(c) are typical waveforms collected by the two sensors following one laser excitation. The data clearly show a long "coda" which is caused by the reverberating field. Cross correlation was then performed on the coda signals after signal processing involving an amplitude thresholding to assign uniform weights to the multiple reverberations (Larose *et al.*, 2006; Sabra *et al.*, 2007). Figures 1(d) and 1(e) represent the same waveforms shown in Figs. 1(b) and 1(c) where the amplitude was clipped to a same level threshold approximately equal to seven times the standard deviation of the electronic noise level. The data were further randomized by averaging the cross-correlation results obtained from 50 distinct laser excitations aimed at different locations in the plate. The averaged cross-correlation results are shown in Fig. 2(a) for a low-frequency band (dc–15 kHz) and in Fig. 2(b) for a high-frequency band (330–480 kHz).

The theoretical impulse responses relating out-of-plane excitation at sensor 1 to out-of-plane displacements at sensor 2 are also shown for comparison with the reconstructed waves in Figs. 2(c) and 2(d). The predicted Green's functions were obtained by calculating the forced solutions in the plate using a semianalytical finite element (SAFE) model in the two frequency

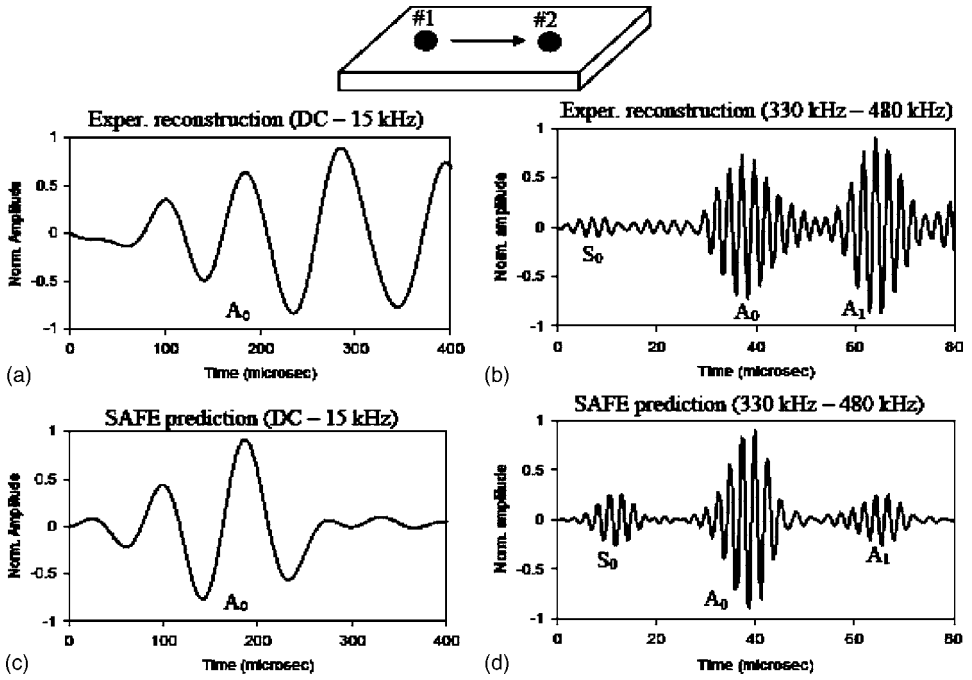


Fig. 2. Reconstruction of coherent guided waves between two sensors in the aluminum plate subjected to random laser excitation. (a) Wave reconstruction in the dc–15 kHz frequency range; (c) corresponding SAFE prediction of out-of-plane Green's function; (b) wave reconstruction in the 330–480 kHz frequency range; (d) corresponding SAFE prediction of out-of-plane Green's function. Notice agreement in arrival times and phase between reconstructed and predicted waves.

bands of interest. These solutions were obtained from Fourier synthesis of the forced harmonic solutions for a point load at the required sensor distance (Hayashi *et al.*, 2003). The computed responses were further convolved with the frequency response of the sensors which were determined experimentally from the laser broadband excitation.

A good agreement is shown in Fig. 2 in both arrival times and phase content between the experimentally reconstructed responses and the SAFE predicted responses. Notice that the later cycles reconstructed in Fig. 2(a) (dc–15 kHz frequency range) are likely due to edge reflections that are absent in the predictions which considered an infinite plate. Amplitude comparisons are more difficult because they are affected by any difference in sensitivity between the two sensors. The reconstruction in the high-frequency band (330–480 kHz) of Fig. 2(b) successfully extracts both A_0 and A_1 modes, again in good agreement with the predictions of Fig. 2(d). The faster S_0 mode is not extracted as efficiently, although it is still discernable. This is associated with the mode shapes of the symmetric mode, which, in the frequency band considered, have large in-plane cross-sectional displacement and small out-of-plane cross-sectional displacement. Overall, Fig. 2 demonstrates that an estimate of the structural impulse response between the two sensors can be obtained passively from the time derivative of the NCF without the use of an active source.

A second set of experiments was conducted to verify that the reconstructed coherent waveforms were suitable for defect detection in the same aluminum plate. A 6.35 mm (1/4 in.)-diam hole was drilled in the plate between sensor 1 and sensor 2, now spaced 18 cm apart (Fig. 3(a)). Figure 3(b) shows the waveforms reconstructed from the random laser excitations in the frequency band 80–160 kHz. The reconstructions are shown for the case of pristine plate, half-thickness hole, and through-thickness hole. As a result of wave scattering, the reductions of amplitude, with respect to the pristine plate case, were, respectively, 22% (half-

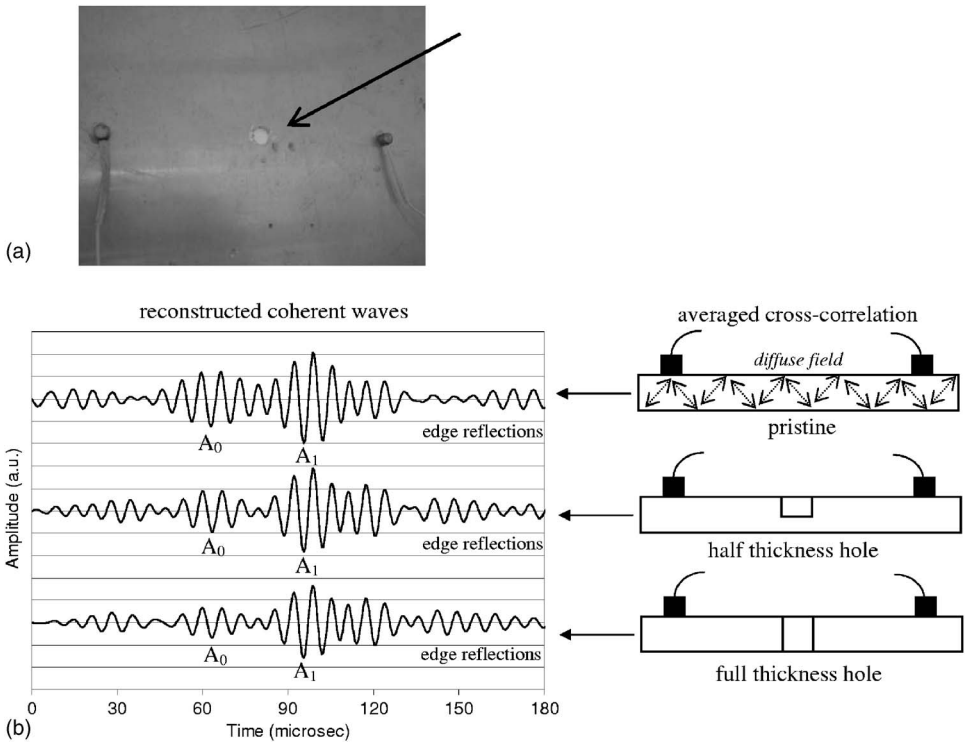


Fig. 3. (a) Picture of hole in test aluminum plate. (b) Coherent waves (80–160 kHz band) reconstructed from random laser excitations of the plate for the case of pristine conditions, 6.35-mm-diam hole extending for half the thickness, and 6.35-mm-diam hole extending through the thickness. Notice the reduction in amplitude of modes S_0 and A_0 caused by wave scattering at the hole. The weak arrival in the $[0-30\ \mu s]$ interval is due to the mode S_0.

thickness hole) and 40% (through-thickness hole) for the mode A_0, and similarly 8% and 18% for mode $A1$. The weak arrival in the $[0-30\ \mu s]$ interval is due to the mode S_0 (see Fig. 2). The root-mean-square value (measured at long time $t>4$ ms) of these coherent signals (i.e., "noise level")—for all three-plate conditions—was equal to 2.5% of mode A_1 amplitude for the pristine case. These values illustrate the fact that not only can the arrival-time structure of the NCF be used for guided-wave SHM but also the NCF's amplitude information.

4. Conclusions

The extraction of Green's functions in a structure by the cross-correlation of random noise provides the possibility of active SHM with none, or a limited number of, ultrasonic secondary sources generating a diffuse field. The method was here shown to have the potential for damage detection in a plate-like structure. The Green's function thus extracted was compared to theoretical predictions, and satisfactory conformance was found in the arrival times and phases of Lamb modes. It was also shown that the reconstructed responses are sensitive to damage in the plate.

Aircraft fuselage and wing structures lend themselves to the formation of diffuse fields because they are geometrically complex (rivets, holes and stiffeners causing scattering) and naturally subjected to random excitations in flight. Furthermore, in order to monitor structural "hot-spots" within the wing structure, it may be more practical to only embed passive sensors during the manufacturing process as opposed to actuators. The diffuse fields in the vicinity of the sensors could also be generated remotely by exciting the outer part of the wing structure during maintenance operations with a few controlled ultrasonic sources strategically located. Hence it is worth investigating the potential for passive-only ultrasonic imaging of these structures during operation.

Acknowledgments

This project was funded in part by Air Force Office of Scientific Research Contract No. FA9550-07-1-0016 (Dr. Victor Giurgiutiu, Program Manager) and by the 2006 Research Fellowship Award from the American Society for Nondestructive Testing.

References and links

Farrar, C., and James, G. (**1997**). "System identification from ambient vibration measurements on a bridge," J. Sound Vib. **205**, 1–18.

Giurgiutiu, V., Zagrai, A., and Bao, J. (**2004**). "Damage identification in aging aircraft structures with piezoelectric waver active sensors," J. Intell. Mater. Syst. Struct. **15**, 673–687.

Hayashi, T., Song, W. J., and Rose, J. L. (**2003**). "Guided wave dispersion curves for a bar with an arbitrary cross section, a rod and rail example," Ultrasonics **41**, 175–183.

Lanza di Scalea, F., Matt, H., Bartoli, I., Coccia, S., Park, G., and Farrar, C. (**2007**). "Health monitoring of UAV wing skin-to-spar joints using guided waves and macro fiber composite transducers," J. Intell. Mater. Syst. Struct. **18**, 373–388.

Larose, E., Lobkis, O. I., and Weaver, R. L. (**2006**). "Passive correlation imaging of a buried scatterer," J. Acoust. Soc. Am. **119**, 3549–3552.

Paul, A., Campillo, M., Margerin, L., Larose, E., and Derode, A. (**2005**). "Empirical synthesis of time-asymmetrical Green functions from the correlation of coda waves, "J. Geophys. Res. **110**, L003521.

Rose, J. L. (**1999**). *Ultrasonic Waves in Solid Media* (Cambridge University Press, Cambridge, U.K).

Roux, P., Sabra, K. G., Kuperman, W., and Roux, A. (**2005**). "Ambient noise cross correlation in free space: Theoretical approach," J. Acoust. Soc. Am. **117**, 79–84.

Sabra, K. G., Roux, P., and Kuperman, W. A. (**2005a**). "Arrival-time structure of the time-averaged ambient noise cross-correlation function in an oceanic waveguide," J. Acoust. Soc. Am. **117**, 164–174.

Sabra, K. G., Roux, P., and Kuperman, W. A. (**2005b**). "Emergence rate of the time domain Green's function from the ambient noise cross correlation," J. Acoust. Soc. Am. **118**, 3524–3531.

Sabra, K. G., Gerstoft, P., Roux, P., Kuperman, W. A., and Fehler, M. C. (**2005c**). "Surface wave tomography from microseisms in Southern California," Geophys. Res. Lett. **32**, L14311.

Sabra, K. G., Winkel, E. S., Bourgoyne, D. A., Elbing, B. R., Ceccio, S. L., Perlin, M., and Dowling, D. R. (**2007**). "On using cross-correlation of turbulent flow-induced ambient vibrations to estimate the structural impulse response. Application to structural health monitoring," J. Acoust. Soc. Am. **121**, 1987–2005.

Shapiro, N. M., Campillo, M., Stehly, L., and Ritzwoller, M. (**2005**). "High resolution surface-wave tomography from ambient seismic noise," Science **29**, 1615–1617.

Snieder, R., and Cafak, E. (**2006**). "Extracting the building response using seismic interferometry; theory and application to the Millikan Library in Pasadena, California," Bull. Seismol. Soc. Am. **96**, 586–598.

Snieder, R. (**2007**). "Extracting the Green's function of attenuating heterogeneous media from uncorrelated waves," J. Acoust. Soc. Am. **121**, 2637–2643.

Van Wijk, K. (**2006**). "On estimating the impulse response between receivers in a controlled ultrasonic model," Geophysics **71**, SI79–SI84.

Weaver, R. L., and Lobkis, O. I. (**2001**). "Ultrasonics without a source: Thermal fluctuation correlations at MHz frequencies," Phys. Rev. Lett. **87**, L134301.

Weaver, R. L., and Lobkis, O. I. (**2004**). "Diffuse fields in open systems and the emergence of the Green's function," J. Acoust. Soc. Am. **116**, 2731–2734.

Mechanical response measurements of real and artificial brass players lips

Michael J. Newton and Murray Campbell
School of Physics, University of Edinburgh, Edinburgh EH9 3JZ United Kingdom
m.newton@ed.ac.uk

Jöel Gilbert
Laboratoire d'Acoustique de L'Université du Maine (UMR CNRS 6613),
Avenue Olivier Messiaen, 72085 Le Mans Cedex 9, France

Abstract: Mechanical frequency responses of human and artificial lips in brass instrument playing have been measured using a high-speed digital video technique, in an attempt to classify the true nature of the "lip-reed." Four semiprofessional human players were used, and three notes played on a trombone were studied. All measurements revealed a strong mechanical resonance with "outward striking" behavior; the played note always sounded above this frequency. Several measurements also showed a weaker second resonance, above the played frequency, with "inward striking" behavior. The Q values of the dominant resonances in human lips were lower than those typical of artificial lips.

PACS numbers: 43.75.Fg, 43.75.Yy [TR]
Date Received: July 18, 2007 **Date Accepted:** October 9, 2007

1. Introduction

Brass instruments produce sound as a result of self-sustained oscillations of the player's lips. The lips are destabilized by application of an overpressure from the lungs, causing a pressure difference to be established between the mouth cavity and the mouthpiece. A complex nonlinear coupling between the resulting airflow, the lips themselves and the resonances of the instrument air column allows self-sustained oscillation and the production of a musical note.

The mechanical properties of the lips are important in determining the playing frequency of a note.[1] A player must adjust the tension and mass distribution of the lip tissue in order to tune any mechanical resonances so that they may usefully interact with the instrument (see, for example Refs. 2–6). Up to now a full description of real human lips has proven difficult, and computational models have relied on parameters obtained from experiments with artificial lips, such as the pioneering work of Cullen *et al.*[7] The principal aim of the present study was to measure the mechanical properties of human lips when formed into playable embouchures.

Several computational models of the lip-reed have been suggested. A lumped element model with one degree of freedom presents the simplest description. Each lip is condensed into a single mass attached to a solid boundary by a spring-damper system. The lips move symmetrically, so only one degree of mechanical freedom is required. This model already describes an impressive array of brass instrument behavior. It fails, however, to reproduce an important feature of human players: the ability to "lip" a note above and below the relevant acoustical resonance frequency of the air column of the instrument.

This problem can be resolved by ascribing two degrees of mechanical freedom to the lips,[4,8,9] in a manner similar to that which has been used to model the human vocal folds.[10,11] Experiments using artificial lips[12] in combination with computational modeling have shown that an "outward-inward striking" resonance pair, in the terminology of classical reed physics,[1] can provide the necessary flexibility. The terms outward striking and inward striking here refer to the manner in which the aperture of a musical reed is expected to respond to a steady increase

in supply pressure. The aperture of an outward striking, or "blown open,"[3] reed tends to widen as the supply pressure is increased. Conversely, the inward striking, or "blown closed," reed tends to close. However, no direct measurements have yet provided convincing verification of outward-inward resonance pairs in the lips of human players.

The conventional light transmission method of measuring the mechanical response of artificial lips, described in Sec. 2.1, involves measurement of the modulation of a laser beam passing between the lips. Since this technique could not be used with human players, an alternative technique using a high speed digital video camera was evolved; this is described in Sec. 2.2. The new technique was first validated by comparing mechanical response measurements performed on artificial lips using both the video method and the conventional transmission method. This validation is described in Sec. 3.1. Finally the video method was applied to human subjects in an attempt to classify unambiguously the phase behavior of the human lip-reed. The results of this experiment are presented in Sec. 3.2.

2. Experimental Methods

In the first part of this study a set of artificial lips was used to play a tenor trombone in the first position. The setup was based on one previously documented.[7] The lips were formed from water-filled latex tubes of 20 mm diameter. The principal control parameter was the internal water pressure, which was adjusted by varying the height of a water column. The lips were stretched across a circular aperture on the front face of a box representing the mouth cavity. A transparent mouthpiece[13] was coupled to the lips, which allowed for easy optical access to the lip opening. A compressed air source supplied an overpressure to this cavity, resulting in an air flow which destabilized the lips and generated a musical note.

2.1 Transmission method for measuring mechanical response of artificial lips

The mechanical response of the artificial lips was first measured using a light transmission method, which is a development of the photoelectric method of Backus.[14] This technique has been described previously[7] and has been widely used for experiments with musical valves.[10,15]

An expanded laser beam was directed through the opening between a pair of artificial lips. The beam was then focused on to a photosensitive diode. Oscillation of the lips caused a modulation in the intensity of the beam which was measured by the diode. With a suitable calibration procedure this led directly to a time domain signal representing the oscillating lip opening. The lips were acoustically driven with a sine wave chirp (typically of 10 s duration) from a loudspeaker coupled to the "mouth" cavity behind the lips. The chirp signal amplitude was calibrated so as to maximize the signal-to-noise ratio of the measurements. The acoustic pressure in the cavity was recorded with a Brüel & Kjær type 4192 microphone. It was assumed that the force acting on the lips was proportional to this signal. All signals were generated and recorded using the Brüel & Kjær PULSE system. This allowed a direct calculation of the frequency response function of the lip opening, defined as

$$H(f) = \frac{\overline{G_{hp}(f)}}{\overline{G_{hh}(f)}} \tag{1}$$

and derived from the averaged power spectrum of the lip motion $\overline{G_{hh}(f)}$, and the averaged cross spectrum of the lip motion with the mouthpiece pressure $\overline{G_{hp}(f)}$. The averaging process permitted several repetitions of the excitation sweep in order to further increase the signal-to-noise ratio of the measurement.

2.2 Video method for measuring mechanical response of artificial and human lips

The optical arrangement required for the transmission method made it impossible to use with human players. A new setup was developed whereby the motion of the lips was recorded directly to digital video using a high-speed camera. This type of camera has been extensively applied to the study of lip reeds in self-sustained oscillation.[12,13] The application to the low amplitude oscillations induced by a mechanical response measurement presented a particular

experimental challenge. A primary objective of this study was thus to verify that the new method could consistently reproduce the response curves obtained with the established method. An attempt to measure the mechanical properties of human vocal folds[16] successfully used a similar approach. However, the induced vibrations were driven from a mechanical shaker and not from an acoustical signal. The vocal folds in this study were also relaxed, and not held under tension as for a lip reed embouchure. This was an important issue for the present work: as the lip tension was increased the amplitude response for the same level of loudspeaker driving decreased. Thus it was important to carefully monitor the high speed camera signal to ensure that a sufficiently good signal-to-noise ratio could be obtained.

To implement the video method, the optics (laser, lenses, diode) from the previous setup were replaced with a high speed digital camera (Vision Research Inc., Phantom v4.1) and a high intensity light source (Schott KL1500 LCD). The camera was oriented to capture the same open area region as with the transmission method. The artificial lips were again driven by a calibrated sine sweep. Each frame from the video was analyzed to deduce the instantaneous lip opening. Concatenation of this information provided a time domain signal describing the oscillating lip opening, directly analogous to that produced by the diode in the transmission method.

An important difference between the two methods was the effective sample rate: in the transmission method this was governed by the signal aquisition system (typically 60 kHz), while for the video method it was limited by the frame rate of the camera (typically 2 kHz at a resolution of 128×256 pixels). This reduced the upper frequency range of the measurements, but this was not a significant problem as the relevant frequency range of the important lip resonances was typically 80–200 Hz. The microphone signal was digitally sampled using the Brüel & Kjær PULSE system, together with a trigger signal that allowed synchronization of the camera video with the pressure recording.

Only a single repetition of the excitation signal could be used. The sweep was generally of shorter duration (around 4 s) than with the laser-diode setup, which further limited the signal-to-noise ratio of the measurements. Despite these apparent shortfalls, the new system was able to faithfully reproduce the artificial lip response curves measured with the transmission method, as will be shown in Sec. 3.1.

To adapt the video method for use on human players, a special double-shanked mouthpiece was constructed. In the normal playing configuration the mouthpiece coupled the player to the trombone via the right-hand shank, with the left-hand shank remaining closed. Upon depression of a control valve the right-hand shank was closed off and the left-hand shank opened, coupling the player to a cavity driven by the loudspeaker. This allowed the player to form an embouchure designed to play a specific note, before subjecting the embouchure to forcing by the calibrated sine sweep. The arrangement is shown schematically in Fig. 1.

3. Results

3.1 Comparison of mechanical response curves of artificial lips obtained using the transmission and video methods

Figure 2 shows three typical mechanical response curves of the artificial lips obtained with the transmission method. The plot illustrates the importance of the principal control parameter for the lips: the internal water pressure. As the internal pressure is increased the resonances of the lips are seen to increase in frequency. This can be associated with an increase in the effective stiffness of the lips: the latex tubing is stretched tighter by the increase in water volume.

Clearly evident from Fig. 2 is the presence of a pair of resonances, between which lies the playing frequency. For example, in the low water pressure curve the first resonance lies at 136 Hz, and the second at 184 Hz. The playing frequency was 174 Hz, close to F2. The lower resonance always shows an outward striking ($-90°$) phase behavior, in the terminology of classical reed physics.[1,17] The upper resonance shows an inward striking ($+90°$) behavior. The con-

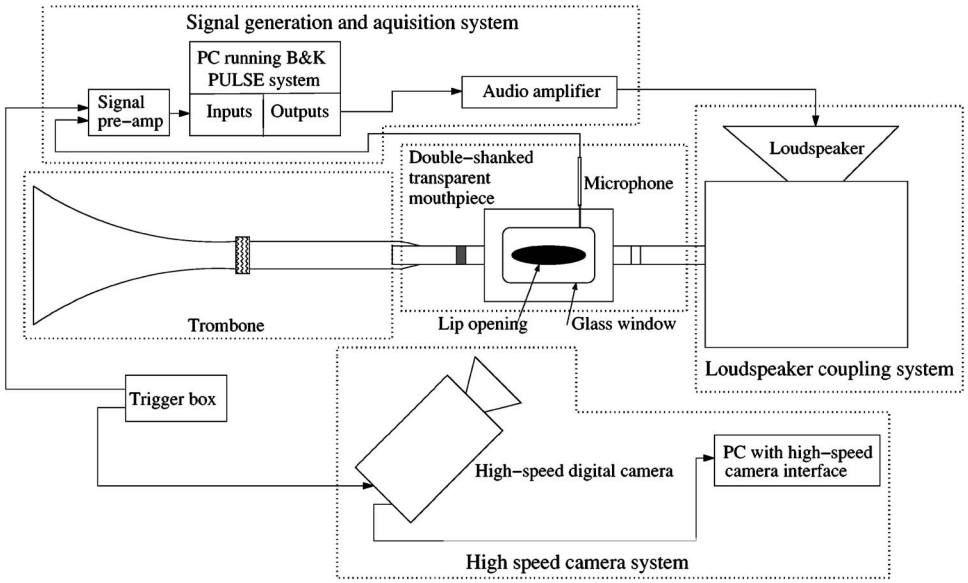


Fig. 1. The setup for mechanical response measurements on human lips using the video method.

sistent appearance of two resonances in such curves has led to the suggestion that a model incorporating at least two degrees of freedom should be required to simulate computationally realistic oscillations of the lip reed.[7,8,12]

Mechanical response measurements using both the transmission method and the video method are shown together in Fig. 3. A satisfactory feature of this plot is the close agreement between the curves in both magnitude and phase. In particular, the resonance frequencies and the phase angles at these frequencies match to within 5%. This is close to the tolerance between consecutive measurements with the transmission method, leading to the conclusion that the video method appears to be reliable.

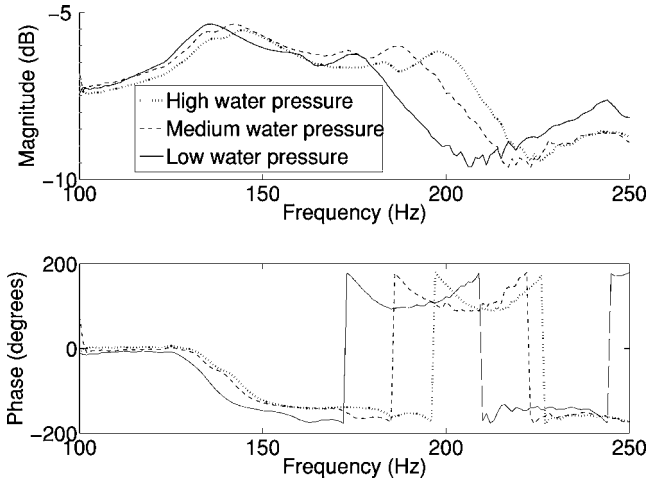


Fig. 2. A plot of three mechanical response curves for different artificial lip embouchures, obtained using the transmission method.

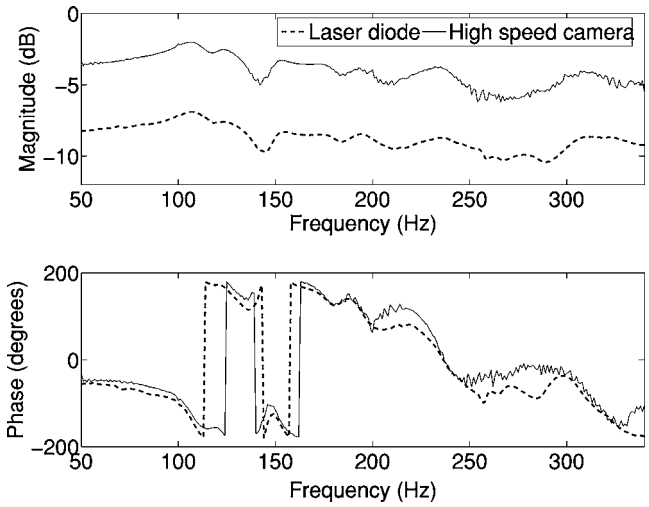


Fig. 3. A comparison between the transmission and video methods for measurement of the mechanical response of the artificial lips. The magnitude curves have been offset by 5 dB for display purposes.

3.2 Mechanical response properties of human lips

A selection of mechanical response curves obtained using human lips is presented in Fig. 4. The embouchures studied corresponded to three played notes: $B_1^\flat$ (the pedal note), $B_2^\flat$ and F_3. These represent the lower range of playable notes on the tenor trombone.

All curves reveal one dominant resonance. This resonance consistently lies below the frequency of the played note, which together with its −90° phase behavior leads to the suggestion that it acts like the outward striking reed of Helmholtz.[17] In the case of the $B_1^\flat$ pedal note shown in the figure this resonance was at 32 Hz, identified from the −90° phase crossing. The played frequency was 58 Hz.

An interesting feature of the curves shown here is the nature of the phase response above the frequency of the played note. After the first resonance it continues down through the inward striking reed angle at +90°. The magnitude curves at this frequency generally show a small peak, though for some measurements the peak disappears below the noise floor.

The progression of the phase response through the inward striking phase angle is generally smooth and consistent. This provides a strong indication that for this frequency, together

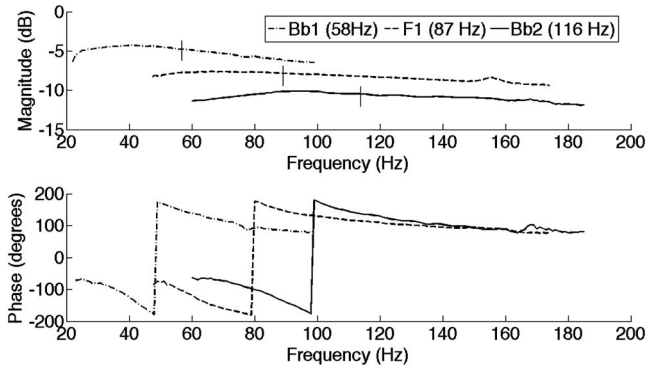


Fig. 4. A plot of three mechanical response curves obtained with human lips, using the video method. The played frequencies for each note are marked as vertical lines. The magnitude curves have been offset by 3 and 6 dB for display purposes.

with a suitable acoustical resonance, the embouchure would be able to accept a positive flow of energy from the flow through the lips and so sustain oscillations. For the $B_1^{\flat}$ pedal note this resonance was at 74 Hz, again identified from the appropriate phase crossing.

However, the presence of a particular mechanical resonance does not immediately reveal the frequency at which an embouchure will play. Rather, for a system containing two resonances the complex interaction with the acoustical modes of the air column means that a coupling between the two mechanical resonances can occur such that the played frequency will fall between them. In this study the playing frequencies of all the notes fell between the two resonances in the same manner as observed with artificial lips.

Previous studies on the transverse motion of the lips during playing have suggested that the lip surfaces follow an eliptical trajectory.[4,8,9] The results here could support such a motion, but could also be described by a simple two mass transverse model such as that frequently used for the human vocal folds.[10,11] The coupled nature of the resonances means that it is difficult to relate particular resonances to specific degrees of freedom.

A second interesting feature of the plots is the shape of the lip resonances. The peaks are of a significantly lower Q value as compared with those of the artificial lips. Typical Q values for the artificial lips[7] range from 8 to 10. In this study the lower outward striking resonances displayed Q values from 1.2 to 1.8. It is possible that the relative broadness of the resonances would make it easier for a human player to "lip" a note far away from the standing wave frequencies of the air column. This result could have implications for parameter choices in the computational modeling of the lip reed.

A recent study measured the mechanical impedance of artificial and human lips.[18] A shaker incorporating an impedance head was placed in contact with a single lip. The Q values of the dominant human lip resonance were lower than those of the artificial lip when filled with either water or glycerine. Interestingly the glycerine-filled artificial lip displayed typical Q values of 4–6 which were a closer match to those of the real lip, typically 1.4–1.6. This result appears to lie in close agreement with the current study.

The general difference in Q values between real and artificial lips is likely to be due to differences in the internal damping of the lips. The artificial lips consist of water surrounded by latex, whereas human lip tissue is constructed from a lattice of skin and muscle cells containing water. One study[19] has deduced some relevant mechanical parameters for human lips and the results here appear to be in broad agreement.

4. Conclusions

This paper has demonstrated the feasibility of measuring the mechanical response properties of human lips formed into playable embouchures. Since the conventional experimental techniques used for artificial lips could not be directly transferred to human players, a video method involving the use of a high speed digital camera was developed and validated.

Application of the video method to human lips has shown that at least two mechanical resonances may be significant in the interaction between the lips and the instrument air column. The magnitude curves suggest that for the range of notes studied the lower resonance is dominant. However, it has been shown elsewhere that a second, apparently weaker resonance can play an important role in the destabilization of an artificial lip embouchure.[12]

The two resonances that are frequently observed appear to have the appropriate phase behavior for a useful reed destabilization, in a manner similar to outward-inward striking resonance pairs frequently observed with artificial lips.

The Q values of the human lip resonances were typically 10–20% of those seen with the artificial lips. This result is in broad agreement with a recent study on the mechanical impedance of human and artificial lips.

Acknowledgments

This work was funded by a Ph.D. grant from the EPSRC, UK. The authors wish to thank A. Downie and D. Low for technical assistance in the construction of the setups, and the four trombone players who kindly volunteered their time and lips.

References and links

[1]J. Elliot and J. M. Bowsher, "Regeneration in brass wind instruments," J. Sound Vib. **83**(2), 181–217 (1982).
[2]J. Saneyoshi, H. Teramura, and S. Yoshikawa, "Feedback oscillations in reed woodwind and brass wind instruments," Acustica **62**, 194–210 (1987).
[3]N. H. Fletcher, "Autonomous vibration of simple pressure-controlled valves," J. Acoust. Soc. Am. **93**(4), 2172–2180, (1993).
[4]S. Yoshikawa, "Acoustical behavior of brass player's lips," J. Acoust. Soc. Am. **97**(3), 1929–1939 (1995).
[5]F.-C. Chen and G. Weinreich, "Nature of the lip reed," J. Acoust. Soc. Am. **99**(2), 1227–1233 (1996).
[6]D. M. Campbell, "Nonlinear dynamics of musical reed and brass wind instruments," Contemp. Phys. **40**(6), 415–431 (1999).
[7]J. S. Cullen, J. Gilbert, and D. M. Campbell, "Brass instruments: Linear stability analysis and experiments with an artificial mouth," **86**, 704–724 (2000).
[8]S. Adachi and M. Sato, "Trumpet sound simulation using a two-dimensional lip vibration model," J. Acoust. Soc. Am. **99**(2), 1200–1209 (1996).
[9]D. C. Copley and W. J. Strong, "A stroboscopic study of lip vibrations in a trombone," J. Acoust. Soc. Am. **99**(2), 1219–1223 (1996).
[10]N. Ruty, X. Pelorson, A. van Hirtum, I. Lopez-Arteaga, and A. Hirschberg, "An in vitro setup to test the relevance and the accuracy of low-order vocal folds models," J. Acoust. Soc. Am. **121**(1), 479–490 (2007).
[11]N. J. C. Lous, G. C. J. Hofmans, N. J. Veldhuis, and A. Hirschberg, "A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design," **84**, 1135–1150 (1998).
[12]O. Richards, "Investigation of the lip reed using computational modeling and experimental studies with an artificial mouth," Ph.D. thesis, University of Edinburgh, UK (2003).
[13]S. Bromage, J. Gilbert, and D. M. Campbell, "Experimental investigation of the open area of the brass player's vibrating lips," *Proc. Forum Acusticum*, Budapest, Hungary (2005).
[14]J. Backus, "Vibrations of the reed and the air column in the clarinet," J. Acoust. Soc. Am. **33**, 800–809 (1961).
[15]C. E. Vilain, X. Pelorson, A. Hirschberg, L. le Marrec, W. O. Root, and J. Willems, "Contribution to the physical modeling of the lips. Influence of the mechanical boundary conditions," **89**, 882–887 (2003).
[16]J. G. Svec, J. Horacek, F. Sram, and J. Vesely, "Resonance properties of the vocal folds: In vivo laryngoscopic investigation of the externally excited laryngeal vibrations," J. Acoust. Soc. Am. **108**(4), 1397–1407 (1988).
[17]H. J. F. Helmholtz, *On the sensations-of-tone (1877)* (Dover, New York, 1954).
[18]B. Gazengel, T. Guimezanes, J.-P. Dalmont, J. B. Doc, S. Fagart, and Y. Leveille, "Experimental investigation of the influence of the mechanical characteristics of the lip on the vibrations of the single reed," *Proc. ISMA*, Barcelona, Spain (2007).
[19]S. Yoshikawa and Y. Muto, "Lip-wave generation in horn players and the estimation of lip-issue elasticity," Acta. Acust. Acust. **89**, 145–162 (2003).

ACOUSTICAL NEWS—USA

Elaine 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

Jeffrey E. Boisvert—For contributions to modeling of elastic shells.

Sarah Hawkins—For contributions to speech perception and phonetics.

Special notice regarding the Distinguished Service Citation and Honorary Fellowship in the Acoustical Society of America

The Distinguished Service Citation is awarded to any present or former Member or Fellow of the Society in recognition of outstanding service to the Society.

An Honorary Fellowship is awarded from time to time to a rare individual for eminence in, or outstanding service to, acoustics; candidates in general should not be members of the Society.

Nominations may be made in writing by any Member or Fellow. They shall be submitted to the Committee on Medals and Awards for review and forwarded to the Executive Council for approval.

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, Florida [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; Web: http://asa.aip.org].

2009

18–22 May — 157th Meeting of the Acoustical Society of America, Portland, Oregon [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; Web: 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.

 0001-4966/2008/123(1)/1/2/$23.00

Classified by subject and indexed by author and inventor. Pp. 395. Out of Print

Volumes 21-30, 1949-1958: JASA, Contemporary Literature and Patents. Classified by subject and indexed by author and inventor. Pp. 952. Price: ASA members $20; Nonmembers $75

Volumes 31-35, 1959-1963: JASA, Contemporary Literature and Patents. Classified by subject and indexed by author and inventor. Pp. 1140. Price: ASA members $20; Nonmembers $90

Volumes 36-44, 1964-1968: JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 485. Out of Print.

Volumes 36-44, 1964-1968: Contemporary Literature. Classified by subject and indexed by author. Pp. 1060. Out of Print

Volumes 45-54, 1969-1973: JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 540. Price: $20 (paperbound); ASA members $25 (clothbound); Nonmembers $60 (clothbound)

Volumes 55-64, 1974-1978: JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 816. Price: $20 (paperbound); ASA members $25 (clothbound); Nonmembers $60 (clothbound)

Volumes 65-74, 1979-1983: JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 624. Price: ASA members $25 (paperbound); Nonmembers $75 (clothbound)

Volumes 75-84, 1984-1988: JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 625. Price: ASA members $30 (paperbound); Nonmembers $80 (clothbound)

Volumes 85-94, 1989-1993: JASA and Patents. Classified by subject and indexed by author and inventor. Pp. 736. Price: ASA members $30 (paperbound); Nonmembers $80 (clothbound)

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

Volumes 105-114, 1999-2003: JASA and Patents. Classified by subject and indexed by author and inventor. Pp.616, Price: ASA members $50; Nonmembers $90 (paperbound)

ACOUSTICAL STANDARDS NEWS

Susan B. Blaeser, Standards Manager
ASA Standards Secretariat, Acoustical Society of America, 35 Pinelawn Rd., Suite 114E, Melville, NY 11747 [Tel.: (631) 390-0125; Fax: (631) 390-0217; e-mail: asastds@aip.org]

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

American National Standards (ANSI Standards) developed by Accredited Standards Committees S1, S2, S3, and S12 in the areas of acoustics, mechanical vibration and shock, bioacoustics, and noise, respectively, are published by the Acoustical Society of America (ASA). In addition to these standards, ASA publishes Catalogs of Acoustical Standards, both National and International. To receive copies of the latest Standards Catalogs, please contact Susan B. Blaeser.
Comments are welcomed on all material in Acoustical Standards News.
This Acoustical Standards News section in JASA, as well as the National and International Catalogs of Acoustical Standards, and other information on the Standards Program of the Acoustical Society of America, are available via the ASA home page: http://asa.aip.org.

Animal Bioacoustics Standards Subcommittee Launched

The Acoustical Society of America is pleased to announce the formation of a new standards subcommittee focused on the subject of Animal Bioacoustics. The formation of this subcommittee was approved by Accredited Standards Committee S3, Bioacoustics, to provide an opportunity for American National Standards to be developed by experts in this specialized subject. The scope of the subcommittee includes: "Standards, specifications, methods of measurement and test, instrumentation and terminology in the field of psychological and physiological acoustics, including aspects of general acoustics, which pertain to biological safety, tolerance and comfort of non-human animals, including both risk to individual animals and to the long-term viability of populations. Animals to be covered may potentially include commercially grown food animals; animals harvested for food in the wild; pets; laboratory animals; exotic species in zoos, oceanaria or aquariums; or free-ranging wild animals."

Membership in the subcommittee is open to companies, government agencies, or professional, scientific or trade associations, with a direct and material interest in the work of the subcommittee. Members of the subcommittee may also elect to become members of Accredited Standards Committee S3 if they wish.

The subcommittee operates according to operating procedures that are accredited by the American National Standards Institute (ANSI) and meets the ANSI requirements for openness, balance, and due process. Organizations wishing to learn more about this subcommittee or the other standards committees and U.S. Technical Advisory Groups administered by the Acoustical Society of America should contact the Standards Secretariat at the telephone number given above or by e-mail at asastds@aip.org.

Standards Meetings Calendar—National

• 10–14 November 2008

Meetings of the National Standards Committees S1-Acoustics, S2-Mechanical Vibration and Shock, S3-Bioacoustics, S3/SC1-Animal Bioacoustics, and S12-Noise, and the ten U.S. TAGs administered by ASA will be held in conjunction with the 156th meeting of the Acoustical Society of America in Miami, Florida.

Standards Meetings Calendar—International

• 26–30 May 2008

Meetings of **ISO/TC 43**, in conjunction with **ISO/TC 43/SC 1** and **ISO/TC 43/SC 2**, will be held in Boras, Sweden.

• 27–31 May 2008

Meetings of **ISO/TC 108/SC 5** will be held in Kyoto, Japan.

• 3–7 November 2008

Meetings of **ISO/TC 108, ISO/TC 108/SC 3** and **ISO/TC 108/SC 6** will be held in St. Louis, Missouri.

Details about these meetings may be obtained from the Secretariat.

ACCREDITED STANDARDS COMMITTEE ON ACOUSTICS, S1

(P. Battenberg, Chair; R.J. Peppin, Vice Chair)

Scope: Standards, specifications, methods of measurement and test, and terminology in the field of physical acoustics including architectural acoustics, electroacoustics, sonics and ultrasonics, and underwater sound, but excluding those aspects which pertain to biological safety, tolerances and comfort.

S1 Working Groups

S1/Advisory-Advisory Planning Committee to S1 (P. Battenberg);

S1/WG1-Standard Microphones and their Calibration (V. Nedzelnitsky);

S1/WG4-Measurement of Sound Pressure Levels in Air (M.A. Nobile, Chair; E. Dunens, Vice Chair);

S1/WG5-Band Filter Sets (A.H. Marsh);

S1/WG17-Sound Level Meters and Integrating Sound Level Meters (B.M. Brooks);

S1/WG19-Insertion Loss of Windscreens (A.J. Campanella);

S1/WG20-Ground Impedance (Measurement of Ground Impedance and Attenuation of Sound Due to the Ground (K. Attenborough, Chair; J.M. Sabatier, Vice Chair);

S1/WG22-Bubble Detection and Cavitation Monitoring (Vacant);

S1/WG25-Specification for Acoustical Calibrators (P. Battenberg);

S1/WG26-High Frequency Calibration of the Pressure Sensitivity of Microphones (A.J. Zuckerwar);

S1/WG27-Acoustical Terminology (J.S. Vipperman);

S1 Inactive Working Groups

S1/WG15-Noise Canceling Microphones (R. McKinley, Chair)

S1/WG16-FFT Acoustical Analyzers (R.J. Peppin, Chair)

S1/WG21-Electromagnetic Susceptibility (EMS) of Acoustical Instruments (J.P. Seiler, Chair)

S1/WG24-Design Response of Weighting Networks for Acoustical Measurements (G.S.K. Wong, Chair)

S1 Standards on Acoustics

ANSI S1.1-1994 (R 2004) American National Standard Acoustical Terminology.

ANSI S1.4-1983 (R 2006) American National Standard Specification for Sound Level Meters. This Standard includes **ANSI S1.4A-1985 (R 2001)** Amendment to ANSI S1.4-1983.

ANSI S1.6-1984 (R 2006) American National Standard Preferred Frequencies, Frequency Levels, and Band Numbers for Acoustical Measurements.

ANSI S1.8-1989 (R 2006) American National Standard Reference Quantities for Acoustical Levels.

ANSI S1.9-1996 (R 2006) American National Standard Instruments for the Measurement of Sound Intensity.

ANSI S1.11-2004 American National Standard Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters.

ANSI S1.13-2005 American National Standard Measurement of Sound Pressure Levels in Air.

ANSI S1.14-1998 (R 2003) American National Standard Recommendations for Specifying and Testing the Susceptibility of Acoustical Instruments to Radiated Radio-Frequency Electromagnetic Fields, 25 MHz to 1 GHz.

ANSI S1.15/Part 1-1997 (R 2006) American National Standard Measurement Microphones, Part 1: Specifications for Laboratory Standard Microphones.

ANSI S1.15/Part 2-2005 American National Standard Measurement Microphones, Part 2: Primary Method for Pressure Calibration of Laboratory Standard Microphones by the Reciprocity Technique.

ANSI S1.16-2000 (R 2005) American National Standard Method for Measuring the Performance of Noise Discriminating and Noise Canceling Microphones.

ANSI S1.17/Part 1-2004 American National Standard Microphone Windscreens—Part 1: Measurements and Specification of Insertion Loss in Still or Slightly Moving Air.

ANSI S1.18-1999 (R 2004) American National Standard Template Method for Ground Impedance.

ANSI S1.20-1988 (R 2003) American National Standard Procedures for Calibration of Underwater Electroacoustic Transducers.

ANSI S1.22-1992 (R 2007) American National Standard Scales and Sizes for Frequency Characteristics and Polar Diagrams in Acoustics.

ANSI S1.24 TR-2002 (R 2007) ANSI Technical Report Bubble Detection and Cavitation Monitoring.

ANSI S1.25-1991 (R 2007) American National Standard Specification for Personal Noise Dosimeters.

ANSI S1.26-1995 (R 2004) American National Standard Method for Calculation of the Absorption of Sound by the Atmosphere.

ANSI S1.40-2006 American National Standard Specifications and Verification Procedures for Sound Calibrators. (*Revision of ANSI S1.40-1984*).

ANSI S1.42-2001 (R 2006) American National Standard Design Response of Weighting Networks for Acoustical Measurements.

ANSI S1.43-1997 (R 2007) American National Standard Specifications for Integrating-Averaging Sound Level Meters.

ACCREDITED STANDARDS COMMITTEE ON MECHANICAL VIBRATION AND SHOCK, S2

(R.L. Eshleman, Chair; A.T. Herfat, Vice Chair)

Scope: Standards, specifications, methods of measurement and test, and terminology in the field of mechanical vibration and shock, and condition monitoring and diagnostics of machines, including the effects of exposure to mechanical vibration and shock on humans, including those aspects which pertain to biological safety, tolerance and comfort.

S2 Working Groups

S2/WG 1-S2 Advisory Planning Committee (R.L. Eshleman, Chair; A.T. Herfat, Vice Chair);

S2/WG2-Terminology and Nomenclature in the Field of Mechanical Vibration and Shock and Condition Monitoring and Diagnostics of Machines (D.J. Evans);

S2/WG3-Signal Processing Methods (T.S. Edwards);

S2/WG4-Characterization of the Dynamic Mechanical Properties of Viscoelastic Polymers (W. Madigosky, Chair; J. Niemiec, Vice Chair);

S2/WG5-Use and Calibration of Vibration and Shock Measuring Instruments (D.J. Evans, Chair; B.E. Douglas, Vice Chair);

S2/WG6—Vibration and Shock Actuators (G.B. Booth);

S2/WG7—Acquisition of Mechanical Vibration and Shock Measurement Data (B.E. Douglas);

S2/WG8—Analysis Methods of Structural Dynamics (B.E. Douglas);

S2/WG9—Training and Accreditation (R.L. Eshleman);

S2/WG10—Measurement and Evaluation of Machinery for Acceptance and Condition (R.L. Eshleman, Chair; H.C. Pusey, Vice Chair);

S2/WG10/Panel 1—Balancing (R.L. Eshleman);

S2/WG10/Panel 2—Operational Monitoring and Condition Evaluation (R. Bankert);

S2/WG10/Panel 3—Machinery Testing (R.L. Eshleman);

S2/WG10/Panel 4—Prognosis (A.J. Hess);

S2/WG10/Panel 5—Data Processing, Communication, and Presentation (K. Bever);

S2/WG11—Measurement and Evaluation of Mechanical Vibration of Vehicles (A.F. Kilcullen);

S2/WG12—Measurement and Evaluation of Structures and Structural Systems for Assessment and Condition Monitoring (B.E. Douglas, Chair);

S2/WG13—Shock Test Requirements for Shelf-Mounted and Other Commercial Electronic Systems (B. Lang, Chair);

S2/WG39—Human Exposure to Mechanical Vibration and Shock—Parallel to ISO/TC 108/SC 4 (D.D. Reynolds, Chair; R. Dong, Vice Chair).

S2 Inactive Working Groups

S2/WG54—Atmospheric Blast Effects (J.W. Reed, Chair; J.H. Keefer, Vice Chair).

S2 Standards on Mechanical Vibration and Shock

ANSI S2.1-2000/ISO 2041:1990 American National Standard Vibration and Shock -Vocabulary. (Nationally Adopted International Standard).

ANSI S2.2-1959 (R 2006) American National Standard Methods for the Calibration of Shock and Vibration Pickups.

ANSI S2.4-1976 (R 2004) American National Standard Method for Specifying the Characteristics of Auxiliary Analog Equipment for Shock and Vibration Measurements.

ANSI S2.7-1982 (R 2004) American National Standard Balancing Terminology.

ANSI S2.8-2007 American National Standard Technical Information Used for Resilient Mounting Applications.

ANSI S2.9-1976 (R 2006) American National Standard Nomenclature for Specifying Damping Properties of Materials.

ANSI S2.16-1997 (R 2006) American National Standard Vibratory Noise Measurements and Acceptance Criteria of Shipboard Equipment.

ANSI S2.17-1980 (R 2004) American National Standard Techniques of Machinery Vibration Measurement.

ANSI S2.19-1999 (R 2004) American National Standard Mechanical Vibration—Balance Quality Requirements of Rigid Rotors, Part 1: Determination of Permissible Residual Unbalance, Including Marine Applications.

ANSI S2.20-1983 (R 2006) American National Standard Estimating Air Blast Characteristics for Single Point Explosions in Air, with a Guide to Evaluation of Atmospheric Propagation and Effects.

ANSI S2.21-1998 (R 2007) American National Standard Method for Preparation of a Standard Material for Dynamic Mechanical Measurements.

ANSI S2.22-1998 (R 2007) American National Standard Resonance Method for Measuring the Dynamic Mechanical Properties of Viscoelastic Materials.

ANSI S2.23-1998 (R 2007) American National Standard Single Cantilever Beam Method for Measuring the Dynamic Mechanical Properties of Viscoelastic Materials.

ANSI S2.24-2001 (R 2006) American National Standard Graphical Presentation of the Complex Modulus of Viscoelastic Materials.

ANSI S2.25-2004 American National Standard Guide for the Measurement, Reporting, and Evaluation of Hull and Superstructure Vibration in Ships.

ANSI S2.26-2001 (R 2006) American National Standard Vibration Testing Requirements and Acceptance Criteria for Shipboard Equipment.

ANSI S2.27-2002 (R 2007) American National Standard Guidelines for the Measurement and Evaluation of Vibration of Ship Propulsion Machinery.

ANSI S2.28-2003 American National Standard Guide for the Measurement and Evaluation of Vibration of Shipboard Machinery.

ANSI S2.29-2003 American National Standard Guide for the Measurement and Evaluation of Vibration of Machine Shafts on Shipboard Machinery.

ANSI S2.31-1979 (R 2004) American National Standard Methods for the Experimental Determination of Mechanical Mobility, Part 1: Basic Definitions and Transducers.

ANSI S2.32-1982 (R 2004) American National Standard Methods for the Experimental Determination of Mechanical Mobility, Part 2: Measurements Using Single-Point Translational Excitation.

ANSI S2.34-1984 (R 2005) American National Standard Guide to the Experimental Determination of Rotational Mobility Properties and the Complete Mobility Matrix.

ANSI S2.42-1982 (R 2004) American National Standard Procedures for Balancing Flexible Rotors.

ANSI S2.43-1984 (R 2005) American National Standard Criteria for Evaluating Flexible Rotor Balance.

ANSI S2.46-1989 (R 2005) American National Standard Characteristics to be Specified for Seismic Transducers.

ANSI S2.48-1993 (R 2006) American National Standard Servo-Hydraulic Test Equipment for Generating Vibration—Methods of Describing Characteristics.

ANSI S2.60-1987 (R 2005) American National Standard Balancing Machines—Enclosures and Other Safety Measures.

ANSI S2.61-1989 (R 2005) American National Standard Guide to the Mechanical Mounting of Accelerometers.

ANSI S2.70-2006 American National Standard Guide for the Measurement and Evaluation of Human Exposure to Vibration Transmitted to the Hand. (*Revision of ANSI S3.34-1986*)

ANSI S2.71-1983 (R 2006) American National Standard Guide to the Evaluation of Human Exposure to Vibration in Buildings (*Reaffirmation and redesignation of ANSI S3.29-1983*)

ANSI S2.72/Part 1-2002 (R 2007)/ISO 2631-1:1997 (*Reaffirmation and redesignation of ANSI S3.18/Part 1-2002/ISO 2631-1:1997*) American National Standard Mechanical vibration and shock—Evaluation of human exposure to whole-body vibration—Part 1: General requirements. (Nationally Adopted International Standard)

ANSI S2.72/Part 4-2003 (R 2007) / ISO 2631-4:2001 (*Reaffirmation and redesignation of ANSI S3.18/Part 4-2003/ISO 2631-4:2001*) American National Standard Mechanical vibration and shock—Evaluation of human exposure to whole-body vibration—Part 4: Guidelines for the evaluation of the effects of vibration and rotational motion on passenger and crew comfort in fixed-guideway transport systems. (Nationally Adopted International Standard)

ANSI S2.73-2002 / ISO 10819:1996 (R 2007) (*Reaffirmation and redesignation of ANSI S3.40-2002/ISO 10819:1996*) American National Standard Mechanical vibration and shock—Hand-arm vibration—Method for the

measurement and evaluation of the vibration transmissibility of gloves at the palm of the hand. (Nationally Adopted International Standard)

ACCREDITED STANDARDS COMMITTEE ON BIOACOUSTICS, S3

(C.A. Champlin, Chair; D.A. Preves, Vice Chair)

Scope: Standards, specifications, methods of measurement and test, and terminology in the fields of psychological and physiological acoustics, including aspects of general acoustics, which pertain to biological safety, tolerance and comfort.

S3 Working Groups

S3/Advisory—Advisory Planning Committee to S3 (C.A. Champlin, Chair; D.A. Preves, Vice Chair);

S3/WG35—Audiometers (R.L. Grason);

S3/WG36—Speech Intelligibility (R.S. Schlauch);

S3/WG37—Coupler Calibration of Earphones (B. Kruger);

S3/WG39—Human Exposure to Mechanical Vibration and Shock (D.D. Reynolds, Chair; R. Dong, Vice Chair);

S3/WG43—Method for Calibration of Bone Conduction Vibrators (J.D. Durrant);

S3/WG48—Hearing Aids (D.A. Preves);

S3/WG51—Auditory Magnitudes (R.P. Hellman);

S3/WG56—Criteria for Background Noise for Audiometric Testing (J. Franks);

S3/WG59—Measurement of Speech Levels (L.A. Wilber and M.C. Killion, Co-Chairs);

S3/WG60—Measurement of Acoustic Impedance and Admittance of the Ear (Vacant);

S3/WG62—Impulse Noise with Respect to Hearing Hazard (J.H. Patterson, Chair; R. Hamernik, Vice Chair);

S3/WG67—Manikins (M.D. Burkhard);

S3/WG72—Measurement of Auditory Evoked Potentials (R.F. Burkard);

S3/WG76—Computerized Audiometry (A.J. Miltich);

S3/WG79—Methods for Calculation of the Speech Intelligibility Index (C.V. Pavlovic);

S3/WG81—Hearing Assistance Technologies (L. Thibodeau and L.A. Wilber, Co-Chairs);

S3/WG82—Basic Vestibular Function Test Battery (C. Wall);

S3/WG83—Sound Field Audiometry (T.R. Letowski);

S3/WG84—Otoacoustic Emissions (G.R. Long);

S3/WG88—Standard Audible Emergency Evacuation and Other Signals (Vacant);

S3/WG89—Spatial Audiometry in Real and Virtual Environments (J. Besing);

S3/WG91—Text-to-Speech Synthesis Systems (A.K. Syrdal and C. Bickley, Co-Chairs).

S3 Liaison Group

S3/L-1 S3 U. S. TAG Liaison to IEC/TC 87 Ultrasonics (W.L. Nyborg).

S3 Inactive Working Groups

S3/WG58 Hearing Conservation Criteria (J.D. Royster and L.H. Royster);

S3/WG71 Artificial Mouths (R. McKinley);

S3/WG80 Probe-tube Measurements of Hearing Aid Performance (W.A. Cole);

S3/WG85 Allocation of Noise-Induced Hearing Loss (R.A. Dobie, Chair).

S3 Standards on Bioacoustics

ANSI S3.1-1999 (R 2003) American National Standard Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms.

ANSI S3.2-1989 (R 1999) American National Standard Method for Measuring the Intelligibility of Speech over Communication Systems.

ANSI S3.4-2007 American National Standard Procedure for the Computation of Loudness of Steady Sounds.

ANSI S3.5-1997 (R 2007) American National Standard Methods for Calculation of the Speech Intelligibility Index.

ANSI S3.6-2004 American National Standard Specification for Audiometers. (*Revision of ANSI S3.6-1996*).

ANSI S3.7-1995 (R 2003) American National Standard Method for Coupler Calibration of Earphones.

ANSI S3.13-1987 (R 2007) American National Standard Mechanical Coupler for Measurement of Bone Vibrators.

ANSI S3.20-1995 (R 2003) American National Standard Bioacoustical Terminology.

ANSI S3.21-2004 American National Standard Methods for Manual Pure-Tone Threshold Audiometry. (*Revision of ANSI S3.21-1978*).

ANSI S3.22-2003 American National Standard Specification of Hearing Aid Characteristics. (*Revision of ANSI S3.22-1996*).

ANSI S3.25-1989 (R 2003) American National Standard for an Occluded Ear Simulator.

ANSI S3.35-2004 American National Standard Method of Measurement of Performance Characteristics of Hearing Aids under Simulated Real-Ear Working Conditions.

ANSI S3.36-1985 (R 2006) American National Standard Specification for a Manikin for Simulated *in situ* Airborne Acoustic Measurements.

ANSI S3.37-1987 (R 2007) American National Standard Preferred Earhook Nozzle Thread for Postauricular Hearing Aids.

ANSI S3.39-1987 (R 2007) American National Standard Specifications for Instruments to Measure Aural Acoustic Impedance and Admittance (Aural Acoustic Immittance).

ANSI S3.41-1990 (R 2001) American National Standard Audible Emergency Evacuation Signal.

ANSI S3.42-1992 (R 2007) American National Standard Testing Hearing Aids with a Broad-Band Noise Signal.

ANSI S3.44-1996 (R 2006) American National Standard Determination of Occupational Noise Exposure and Estimation of Noise-Induced Hearing Impairment.

ANSI S3.45-1999 American National Standard Procedures for Testing Basic Vestibular Function.

ANSI S3.46-1997 (R 2002) American National Standard Methods of Measurement of Real-Ear Performance Characteristics of Hearing Aids.

ACCREDITED STANDARDS SUBCOMMITTEE ON ANIMAL BIOACOUSTICS, S3/SC1

(D.K. Delaney, Chair; VACANT, Vice Chair)

Scope: Standards, specifications, methods of measurement and test, instrumentation and terminology in the field of psychological and physiological acoustics, including aspects of general acoustics, which pertain to biological safety, tolerance and comfort of non-human animals, including both risk to individual animals and to the long-term viability of populations. Animals to be covered may potentially include commercially grown food animals; animals harvested for food in the wild; pets; laboratory animals; exotic species in zoos, oceanaria or aquariums; or free-ranging wild animals.

S3/SC1 Working Groups

S3/SC1/WG1—Animal Bioacoustics Terminology (A.E. Bowles);

S3/SC1/WG2—Effects of Sound on Fish and Turtles (R.R. Fay and A.N. Popper, Co-Chairs);

S3/SC1/WG3—Passive Acoustic Monitoring for Marine Mammal Mitigation for Seismic Surveys (A.M. Thode).

ACCREDITED STANDARDS COMMITTEE ON NOISE, S12

(R.D. Hellweg, Chair; W.J. Murphy, Vice Chair)

Scope: Standards, specifications, and terminology in the field of acoustical noise pertaining to methods of measurement, evaluation, and control, including biological safety, tolerance and comfort and physical acoustics as related to environmental and occupational noise.

S12 Working Groups

S12/Advisory—Advisory Planning Committee to S12 (R.D. Hellweg, Chair; W.J. Murphy, Vice Chair);

S12/WG3—Measurement of Noise from Information Technology and Telecommunications Equipment (K. X. C. Man, Chair);

S12/WG11—Hearing Protector Attenuation and Performance (E.H. Berger, Chair);

S12/WG13—Method for the Selection of Hearing Protectors that Optimize the Ability to Communicate (D. Byrne, Chair);

S12/WG14—Measurement of the Noise Attenuation of Active and /or Passive Level Dependent Hearing Protective Devices (W.J. Murphy, Chair);

S12/WG15—Measurement and Evaluation of Outdoor Community Noise (P.D. Schomer);

S12/WG18—Criteria for Room Noise (R.J. Peppin);

S12/WG23—Determination of Sound Power (B.M. Brooks and J. Schmitt, Co-Chairs);

S12/WG31—Predicting Sound Pressure Levels Outdoors (R.J. Peppin, Chair; L. Pater, Vice Chair);

S12/WG32—Revision of ANSI S12.7-1986 Methods for Measurement of Impulse Noise (A.H. Marsh);

S12/WG36—Development of Methods for Using Sound Quality (P. Davies and G.L. Ebbit, Co-Chairs);

S12/WG38—Noise Labeling in Products (R.D. Hellweg and J. Pope, Co-Chairs);

S12/WG40—Measurement of the Noise Aboard Ships (S. Antonides, Chair; S. Fisher, Vice Chair);

S12/WG41—Model Community Noise Ordinances (L.S. Finegold, Chair; B.M. Brooks, Vice Chair);

S12/WG44—Speech Privacy (G.C. Tocci, Chair; D. Sykes, Vice Chair);

S12/WG45—Measurement of Occupational Noise Exposure from Telephone Equipment (K.A. Woo, Chair; L.A. Wilber, Vice-Chair);

S12/WG46—Acoustical Performance Criteria for Relocatable Classrooms (T. Hardiman and P.D. Schomer, Co-Chairs);

S12/WG47—Underwater Noise Measurements of Ships (M. Bahtiarian, Chair);

S12/WG48—Railroad Horn Sound Emission Testing (J. Erdreich, Chair; J.J. Earshen, Vice Chair);

S12/WG49—Noise from Hand-Operated Power Tools, Excluding Pneumatic Tools (B.M. Brooks, Chair)

S12 Liaison Groups

S12/L-1 IEEE 85 Committee for TAG Liaison—Noise Emitted by Rotating Electrical Machines (Parallel to ISO/TC43/SC 1/WG13) (R.G. Bartheld, Chair);

S12/L-2 Measurement of Noise from Pneumatic Compressors Tools and Machines (Parallel to ISO/TC43/SC 1/WG9) (Vacant);

S12/L-3 SAE Committee for TAG Liaison on Measurement and Evaluation of Motor Vehicle Noise (parallel to ISO/TC 43/SC1/WG8) (R.F. Schumacher, Chair);

S12/L-4 SAE Committee A-21 for TAG Liaison on Measurement and Evaluation of Aircraft Noise (J.D. Brooks, Chair);

S12/L-5 ASTM E-33 on Environmental Acoustics (to include activities of ASTM E33.06 on Building Acoustics, parallel to ISO/TC 43/SC 2 and ASTM E33.09 on Community Noise) (K.P. Roy. Chair);

S12/L-6 SAE Construction-Agricultural Sound Level Committee (I. Douell, Chair);

S12/L-7 SAE Specialized Vehicle and Equipment Sound Level Committee (T.M. Disch, Chair);

S12/L-8 ASTM PTC 36 Measurement of Industrial Sound (R.A. Putnam, Chair; B.M. Brooks, Vice Chair).

S12 Inactive Working Groups

S12/WG9 Annoyance Response to Impulsive Noise (L.C. Sutherland, Chair);

S12/WG19 Measurement of Occupational Noise Exposure (J.P. Barry and R. Goodwin, Co-Chairs);

S12/WG27 Outdoor Measurement of Sound Pressure Level (G.A. Daigle, Chair);

S12/WG29 Field Measurement of the Sound Output of Audible Public-Warning Devices (Sirens) (P. Graham, Chair);

S12/WG37—Measuring Sleep Distrbance Due to Noise (K.S. Pearsons, Chair).

S12 Standards on Noise

ANSI S12.1-1983 (R 2006) American National Standard Guidelines for the Preparation of Standard Procedures to Determine the Noise Emission from Sources.

ANSI S12.2-1995 (R 1999) American National Standard Criteria for Evaluating Room Noise.

ANSI S12.3-1985 (R 2006) American National Standard Statistical Methods for Determining and Verifying Stated Noise Emission Values of Machinery and Equipment.

ANSI S12.5-2006/ISO 6926:1999 American National Standard Acoustics-Requirements for the Performance and Calibration of Reference Sound Sources Used for the Determination of Sound Power Levels. (Nationally Adopted International Standard).

ANSI S12.6-1997 (R 2002) American National Standard Methods for Measuring the Real-Ear Attenuation of Hearing Protectors. (*Revision of ANSI S12.6-1984*).

ANSI S12.7-1986 (R 2006) American National Standard Methods for Measurements of Impulse Noise.

ANSI S12.8-1998 (R 2003) American National Standard Methods for Determining the Insertion Loss of Outdoor Noise Barriers.

ANSI S12.9/Part 1-1988 (R 2003) American National Standard Quantities and Procedures for Description and Measurement of Environmental Sound, Part 1.

ANSI S12.9/Part 2-1992 (R 2003) American National Standard Quantities and Procedures for Description and Measurement of Environmental Sound, Part 2: Measurement of Long-Term, Wide-Area Sound.

ANSI S12.9/Part 3-1993 (R 2003) American National Standard Quantities and Procedures for Description and Measurement of Environmental Sound, Part 3: Short-Term Measurements with an Observer Present.

ANSI S12.9/Part 4-2005 American National Standard Quantities and Procedures for Description and Measurement of Environmental Sound, Part 4: Noise Assessment and Prediction of Long-Term Community Response.

ANSI S12.9/Part 5-1998 (R 2003) American National Standard Quantities and Procedures for Description and Measurement of Environmental Sound, Part 5: Sound Level Descriptors for Determination of Compatible Land Use.

ANSI S12.9/Part 6-2000 (R 2005) American National Standard Quantities and Procedures for Description and Measurement of Environmental Sound, Part 6: Methods for Estimation of Awakenings Associated with Aircraft Noise Events Heard in Homes.

ANSI/ASA S12.10-2002 (R 2007)/ISO 7779:1999 American National Standard Acoustics-Measurement of airborne noise emitted by information technology and telecommunications equipment. (Nationally Adopted International Standard).

ANSI S12.11/Part 1-2003/ISO 10302:1996 (MOD) American National Standard Acoustics-Measurement of noise and vibration of small air-moving devices-Part 1: Airborne noise emission. (Modified Nationally Adopted International Standard).

ANSI S12.11/Part 2-2003 American National Standard Acoustics-Measurement of Noise and Vibration of Small Air-Moving Devices-Part 2: Structure-Borne Vibration.

ANSI/ASA S12.12-1992 (R 2007) American National Standard Engineering Method for the Determination of Sound Power Levels of Noise Sources Using Sound Intensity.

ANSI S12.13 TR-2002 ANSI Technical Report Evaluating the Effectiveness of Hearing Conservation Programs through Audiometric Data Base Analysis.

ANSI/ASA S12.14-1992 (R 2007) American National Standard Methods for the Field Measurement of the Sound Output of Audible Public Warning Devices Installed at Fixed Locations Outdoors.

ANSI/ASA S12.15-1992 (R 2007) American National Standard For Acoustics-Portable Electric Power Tools, Stationary and Fixed Electric Power Tools, and Gardening Appliances-Measurement of Sound Emitted.

ANSI/ASA S12.16-1992 (R 2007) American National Standard Guidelines for the Specification of Noise of New Machinery.

ANSI S12.17-1996 (R 2006) American National Standard Impulse Sound Propagation for Environmental Noise Assessment.

ANSI S12.18-1994 (R 2004) American National Standard Procedures for Outdoor Measurement of Sound Pressure Level.

ANSI S12.19-1996 (R 2006) American National Standard Measurement of Occupational Noise Exposure.

ANSI S12.23-1989 (R 2006) American National Standard Method for the Designation of Sound Power Emitted by Machinery and Equipment.

ANSI S12.42-1995 (R 2004) American National Standard Microphone-in-Real-Ear and Acoustic Test Fixture Methods for the Measurement of Insertion Loss of Circumaural Hearing Protection Devices.

ANSI/ASA S12.43-1997 (R 2007) American National Standard Methods for Measurement of Sound Emitted by Machinery and Equipment at Workstations and Other Specified Positions.

ANSI/ASA S12.44-1997 (R 2007) American National Standard Methods for Calculation of Sound Emitted by Machinery and Equipment at Workstations and Other Specified Positions from Sound Power Level.

ANSI/ASA S12.50-2002 (R 2007)/ISO 3740:2000 American National Standard Acoustics -Determination of sound power levels of noise sources—Guidelines for the use of basic standards. (Nationally Adopted International Standard)

ANSI/ASA S12.51-2002 (R 2007) /ISO3741:1999 American National Standard Acoustics—Determination of sound power levels of noise sources using sound pressure-Precision method for reverberation rooms. This Standard includes Technical Corrigendum 1-2001. (Nationally Adopted International Standard) *This standard replaces ANSI S12.31-1990 and ANSI S12.32-1990.*

ANSI S12.53/Part 1-1999 (R 2004)/ISO 3743-1:1994 American National Standard Acoustics—Determination of sound power levels of noise sources—Engineering methods for small, movable sources in reverberant fields—Part 1: Comparison method for hard-walled test rooms. (Nationally Adopted International Standard) *This standard, along with ANSI S12.53/Part 2-1999, replaces ANSI S12.33-1990.*

ANSI S12.53/Part 2-1999 (R 2004)/ISO 3743-2:1994 American National Standard Acoustics—Determination of sound power levels of noise

sources using sound pressure-Engineering methods for small, movable sources in reverberant fields—Part 2: Methods for special reverberation test rooms. (Nationally Adopted International Standard) *This standard, along with ANSI S12.53/Part 1-1999, replaces ANSI S12.33-1990.*

ANSI S12.54-1999 (R 2004)/ISO 3744:1994 American National Standard Acoustics—Determination of sound power levels of noise sources using sound pressure-Engineering method in an essentially free field over a reflecting plane. (Nationally Adopted International Standard) *This standard replaces ANSI S12.34-1988.*

ANSI S12.55-2006/ISO 3745:2003 American National Standard Acoustics—Determination of sound power levels of noise sources using sound pressure-Precision methods for anechoic and hemi-anechoic rooms. (Nationally Adopted International Standard) *This standard replaces ANSI S12.35-1990.*

ANSI S12.56-1999 (R 2004)/ISO 3746:1995 American National Standard Acoustics—Determination of sound power levels of noise sources using sound pressure-Survey method using an enveloping measurement surface over a reflecting plane. (Nationally Adopted International Standard) *This standard replaces ANSI S12.36-1990.*

ANSI/ASA S12.57-2002 (R 2007)/ISO 3747:2000 American National Standard Acoustics—Determination of sound power levels of noise sources using sound pressure-Comparison method *in situ*. (Nationally Adopted International Standard).

ANSI S12.60-2002 American National Standard Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools.

ANSI S12.65-2006 American National Standard for Rating Noise with Respect to Speech Interference. (*Revision of ANSI S3.14-1977*).

ANSI/ASA S12.68-2007 American National Standard Methods of Estimating Effective A-Weighted Sound Pressure Levels When Hearing Protectors are Worn.

ASA COMMITTEE ON STANDARDS (ASACOS)

ASACOS (P.D. Schomer, Chair and ASA Standards Director)

U.S. TECHNICAL ADVISORY GROUPS (TAGS) FOR INTERNATIONAL STANDARDS COMMITTEES:

ISO/TC 43 Acoustics, **ISO/TC 43/SC 1** Noise (P.D. Schomer, U.S. TAG Chair)

ISO/TC 108 Mechanical vibration, shock and condition monitoring (D.J. Evans, U.S. TAG Chair)

ISO/TC 108/SC2 Measurement and evaluation of mechanical vibration and shock as applied to machines, vehicles and structures (A.F. Kilcullen, and R.F. Taddeo U.S. TAG Co-Chairs)

ISO/TC 108/SC3 Use and calibration of vibration and shock measuring instruments (D.J. Evans, U.S. TAG Chair)

ISO/TC 108/SC4 Human exposure to mechanical vibration and shock (D.D. Reynolds, U.S. TAG Chair)

ISO/TC 108/SC5 Condition monitoring and diagnostics of machines (D.J. Vendittis, U.S. TAG Chair; R. Taddeo, U.S. TAG Vice Chair)

ISO/TC 108/SC6 Vibration and shock generating systems (G. Booth, U.S. TAG Chair)

IEC/TC 29 Electroacoustics (V. Nedzelnitsky, U.S. Technical Advisor)

Standards News from the United States

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

American National Standards Call for Comment on Proposals Listed

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

ASA (ASC S1) (Acoustical Society of America)

New Standards

BSR/ASA S1.44-200x, High-Frequency Calibration of the Pressure Sensitivity of Microphones by Means of Measurements in the Free Field (new standard)

Describes procedures to perform a secondary calibration of the pressure sensitivity of microphones for frequencies above 20 kHz. It utilizes a substitution method, requiring a reference microphone for which the electrostatic actuator frequency response is known. The range of frequencies will be limited to the known frequency response range of the reference microphone. (November 12, 2007)

ASA (ASC S2) (Acoustical Society of America)

New Standards

BSR/ASA S2.62-200x, Shock Test Requirements for Equipment in a Rugged Shock Environment (new standard)

This standard is used for testing equipment that will be subjected to shock. Defines test requirements and severity thresholds for a large range of shock environments, including but not limited to shipping, transport, and rugged operational environments. This standard will allow vendors to better market, and users to more easily identify equipment that will operate or simply survive in rugged shock environments. This standard includes references to various ASTM, IEC, NATO, and US military standards. (December 3, 2007)

ASA (ASC S12) (Acoustical Society of America)

Revisions

BSR/ASA S12.9-Part 5-200x, Quantities and Procedures for Description and Measurement of Environmental Sound—Part 5: Sound Level Descriptors for Determination of Compatible Land Use (revision of ANSI S12.9-Part 5-1998 (R2003))

Provides guidance on the compatibility of various human uses of land with the acoustical environment, using the yearly average total day-night adjusted sound exposure or the yearly average adjusted day-night average sound level to characterize the acoustical environment. (November 12, 2007)

ASTM (ASTM International)

New Standards

BSR/ASTM F1334-200x, Test Method for Determining A-Weighted Sound Power Level of Vacuum Cleaners (new standard) The URL to search for scopes of ASTM standards is: http://www.astm.org/dsearch.htm (October 22, 2007)

IEEE (Institute of Electrical and Electronics Engineers)

Reaffirmations

BSR/IEEE 563-1991 (R2007), Guide on Conductor Self-Damping Measurements (reaffirmation of ANSI/IEEE 563-1991 (R2002))

Presents methods for measuring the inherent vibration damping characteristics of overhead conductors. The intent is to obtain information in a compatible and consistent form that will provide a reliable basis for studying the vibration and damping of conductors in the future, and for comparing data of various investigators. The methods and procedures recommended are not intended for quality-control test purposes. (December 25, 2007)

BSR/IEEE 664-1994 (R200x), Guide for Laboratory Measurement of the Power Dissipation Characteristics of Aeolian Vibration Dampers for Single Conductors (reaffirmation of ANSI/IEEE 664-1994 (R2000))

Describes the current methodologies, including apparatus, procedures, and measurement accuracies, for determining the dynamic characteristics of vibration dampers and damping systems. It provides some basic guidance regarding a given method's strengths and weaknesses. The methodologies and procedures described are applicable to indoor testing only. (December 25, 2007)

American National Standards Final Action

The following American National Standards have received final approval from the ANSI Board of Standards Review.

ASA (ASC S12) (Acoustical Society of America)

New Standards

ANSI/ASA S12.68-2007, Methods of Estimating Effective A-Weighted Sound Pressure Levels when Hearing Protectors Are Worn (new standard)

Reaffirmations

ANSI/ASA S12.10-2002/ISO 7779:1999 (R2007) (incl AMD1), Acoustics—Measurement of airborne noise emitted by information technology and telecommunications equipment (a Nationally Adopted International Standard) (reaffirmation and redesignation of ANSI S12.10-2002/ISO 7779:1999 (incl AMD1)).

ANSI/ASA S12.12-1992 (R2007), Engineering Method for the Determination of Sound Power Levels of Noise Sources Using Sound Intensity (reaffirmation and redesignation of ANSI S12.12-1992 (R2002)).

ANSI/ASA S12.14-1992 (R2007), Methods for the Field Measurement of the Sound Output of Audible Public Warning Devices Installed at Fixed Locations Outdoors (reaffirmation and redesignation of ANSI S12.14-1992 (R2002)).

ANSI/ASA S12.15-1992 (R2007), Acoustics—Portable Electric Power Tools, Stationary and Fixed Electric Power Tools, and Gardening Appliances—Measurement of Sound Emitted (reaffirmation and redesignation of ANSI S12.15-1992 (R2002)).

ANSI/ASA S12.16-1992 (R2007), Guidelines for the Specification of Noise of New Machinery (reaffirmation and redesignation of ANSI S12.16-1992 (R2002)).

ANSI/ASA S12.43-1997 (R2007), Methods for Measurement of Sound Emitted by Machinery and Equipment at Workstations and Other Specified Positions (reaffirmation and redesignation of ANSI S12.43-1997 (R2002)).

ANSI/ASA S12.44-1997 (R2007), Methods for Calculation of Sound Emitted by Machinery and Equipment at Workstations and Other Specified Positions from Sound Power Level (reaffirmation and redesignation of ANSI S12.44-1997 (R2002)).

ANSI/ASA S12.50-2002/ISO 3704-2000 (R2007), Acoustics—Determination of Sound Power Levels of Noise Sources—Guidelines for the Use of Basic Standards (reaffirmation and redesignation of ANSI S12.50-2002/ISO 3704-2000).

ANSI/ASA S12.51-2002/Part 1/ISO 3741:1999 (R2007), Acoustics—Determination of Sound Power Levels of Noise Sources Using Sound Pressure—Precision Method for Reverberation Rooms (reaffirmation and redesignation of ANSI S12.51-2002/Part 1/ISO 3741:1999).

ANSI/ASA S12.57-2002/ISO 3747-2000 (R2007), Acoustics—Determination of Sound Power Levels of Noise Sources Using Sound Pressure—Comparison Method in situ (reaffirmation and redesignation of ANSI S12.57-2002/ISO 3747-2000).

Withdrawals

ANSI S12.30-1990 (R2002), Guidelines for the Use of Sound Power Standards and for the Preparation of Noise Test Codes (withdrawal of ANSI S12.30-1990 (R2002)).

SCTE (Society of Cable Telecommunications Engineers)

Revisions

ANSI/SCTE 62-2007, Measurement Procedure for Noise Figure (revision of ANSI/SCTE 62-2002)

Project Initiation Notification System (PINS)

ANSI Procedures require notification of ANSI by ANSI-accredited standards developers (ASD) of the initiation and scope of activities expected to result in new or revised American National Standards (ANS). Early notification of activity intended to reaffirm or withdraw an ANS and in some instances a PINS related to a national adoption is optional. The mechanism by which such notification is given is referred to as the PINS process. For additional information, see clause 2.4 of the ANSI Essential Requirements: Due Process Requirements for American National Standards.

ASA (ASC S2) (Acoustical Society of America)

BSR/ASA S2.63-200x/ISO 16063-22:2005, Methods for the calibration of vibration and shock transducers—Part 22: Shock calibration by comparison to a reference transducer (identical national adoption of ISO 16063-22:2005)

Specifies the instrumentation and procedures to be used for secondary shock calibration of rectilinear transducers, using a reference acceleration, velocity or force measurement for the time-dependent shock. The methods are applicable in a shock pulse (duration range) of 0,05 ms to 8,0 ms, and a dynamic range (peak value) of 100 m/s2 to 100 km/s2 (time dependent). The methods allow the transducer shock sensitivity to be obtained. Project Need: This standard is aimed at users engaged in shock measurements requiring traceability as stated in ISO 9001 and ISO/IEC 17025. Stakeholders: Engineers, Calibration Laboratories, Industry.

SCTE (Society of Cable Telecommunications Engineers)

BSR/SCTE IPS SP 909-200x, RF-over-Glass Gateway Environmental Requirements (new standard)

Specifies the minimum environmental operating requirements for the RFoG network interface unit. The proposed scope includes but is not limited to: operational and storage temperature and humidity range; RF isolation; electrical surge protection; mechanical shock and vibration; and

regulatory conformance. Project Need: To increase the use of fiber in cable plant. Stakeholders: Cable Telecommunications Industry.

Notice of Withdrawal: ANS at least 10 years past approval date

The following American National Standards have not been revised or reaffirmed within ten years from the date of their approval as American National Standards and accordingly are withdrawn:

ANSI/ASHRAE 68-1997, ANSI/AMCA 330-1997, Laboratory Method of Testing In-Duct Sound Power Measurement Procedure for Fans (also designated ANSI/AMCA 330-86)

Standards News from Abroad

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

International Organization for Standardization (ISO)

Newly Published ISO and IEC Standards

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

ISO Standards

Acoustics (TC 43)

ISO 11689/Cor1:2007, Acoustics—Procedure for the comparison of noise-emission data for machinery and equipment—Corrigendum

MECHANICAL VIBRATION AND SHOCK (TC 108)

ISO 10326-1/Amd1:2007, Mechanical vibration—Laboratory method for evaluating vehicle seat vibration—Part 1: Basic requirements—Amendment 1

IEC Standards

ULTRASONICS (TC 87)

IEC 62127-1 Ed. 1.0 en:2007, Ultrasonics—Hydrophones—Part 1: Measurement and characterization of medical ultrasonic fields up to 40 MHz

IEC 62127-2 Ed. 1.0 en:2007, Ultrasonics—Hydrophones—Part 2: Calibration for ultrasonic fields up to 40 MHz

IEC 62127-3 Ed. 1.0 en:2007, Ultrasonics—Hydrophones—Part 3: Properties of hydrophones for ultrasonic fields up to 40 MHz

ISO Draft Standards

ACOUSTICS (TC 43)

ISO 3822-1/DAmd1.2, Acoustics—Laboratory tests on noise emission from appliances and equipment used in water supply installations—Part 1: Method of measurement—Measurement uncertainty (December 22, 2007)

IEC Draft Standards

104/439/FDIS, IEC 60068-2-6 Ed. 7.0: Environmental testing—Part 2: Tests—Test Fc: Vibration (sinusoidal) (November 23, 2007)

REVIEWS OF ACOUSTICAL PATENTS

Lloyd Rice
11222 Flatiron Drive, Lafayette, Colorado 80026

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

Reviewers for this issue:

GEORGE L. AUGSPURGER, *Perception, Incorporated, Box 39536, Los Angeles, California 90039*
ANGELO CAMPANELLA, *3201 Ridgewood Drive, Hilliard, Ohio 43026–2453*
GEOFFREY EDELMANN, *Naval Research Laboratory. Code 7145, 4555 Overlook Ave. SW, Washington, DC 20375*
JEROME A. HELFFRICH, *Southwest Research Institute, San Antonio, Texas 78228*
DAVID PREVES, *Starkey Laboratories, 6600 Washington Ave. S., Eden Prarie, Minnesota 55344*
NEIL A. SHAW, *Menlo Scientific Acoustics, Inc., Post Office Box 1610, Topanga, California 90290*
ROBERT C. WAAG, *Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York 14627*

7,236,426

43.30.Tg INTEGRATED MAPPING AND AUDIO SYSTEMS

Rex Turner and Ronald G. Weber, assignors to Lowrance Electronics, Incorporated
26 June 2007 (Class 367/88); filed 8 September 2004

The most elementary concept of combining a side-scan profiler and a map (e.g., GPS) is described in this patent. The proposed system also stores and plays aloud measured acoustic echoes—a feat that sonar systems have been doing for a very long time.—GFE

7,239,580

43.30.Vh NOISE ADAPTIVE SONAR SIGNAL PROCESSOR

Nathan Intrator *et al.*, assignors to Brown University
3 July 2007 (Class 367/101); filed 8 June 2004

A trivial sonar system is put forth to deal with low signal-to-noise environments by simply searching for a frequency band with less ambient noise and/or absorption and then pinging within it.—GFE

7,242,638

43.30.Vh METHOD AND SYSTEM FOR SYNTHETIC APERTURE SONAR

Ian B. Kerfoot and James G. Kosalos, assignors to Raytheon Company
10 July 2007 (Class 367/88); filed 24 November 2004

A method is described whereby estimations of heading, pitch, roll, etc. are accomplished by correlating changes in successive measured echoes. However, the technique appears to be the well known method of autofocusing originally developed for synthetic aperture radar. Furthermore, significant hurdles, such as insensitivity to yaw, are not addressed.—GFE

7,245,557

43.30.Vh SONAR

Shinji Ishihara *et al.*, assignors to Furuno Electric Company Limited
17 July 2007 (Class 367/88); filed in Japan 20 December 2004

This document asserts that fish finders are limited in shallow depths due to the ring down time of the transducers. A trivial method is put forth to reduce the tailing signal.—GFE

7,245,559

43.30.Vh ACOUSTIC FENCE

Larry R. McDonald and Gary W. Hicks, assignors to Science Applications Incorporated Corporation
17 July 2007 (Class 367/136); filed 18 October 2004

The basic notion of detecting a diving intruder via sound is presented. The diver detection system simply detects and tracks the diver's reflection (via good old arrival time) with multiple sensors.—GFE

7,258,742

43.35.Pt METHOD OF MANUFACTURING POTASSIUM NIOBATE SINGLE CRYSTAL THIN FILM, SURFACE ACOUSTIC WAVE ELEMENT, FREQUENCY FILTER, FREQUENCY OSCILLATOR, ELECTRONIC CIRCUIT, AND ELECTRONIC APPARATUS

Takamitsu Higuchi *et al.*, assignors to Seiko Epson Corporation
21 August 2007 (Class 117/86); filed in Japan 26 March 2003

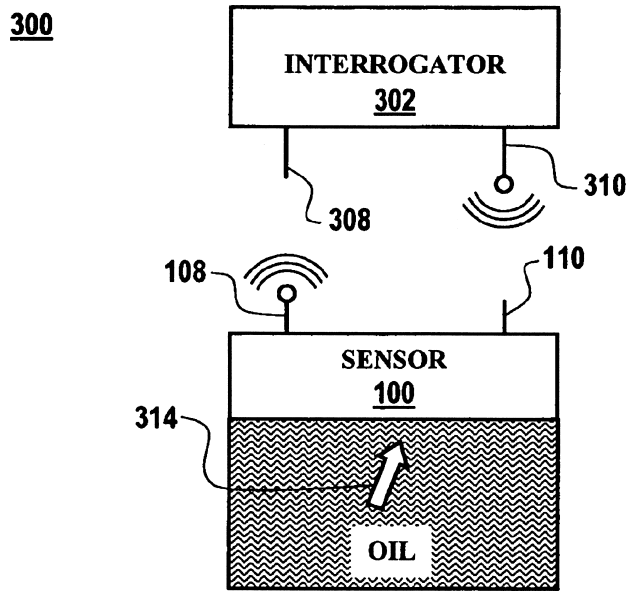
The authors describe their achievements and some of the background work in the development of $KNbO_3$ piezoelectric single crystals for use in surface acoustic wave (SAW) devices. The advantages claimed of $KNbO_3$ over quartz and $LiNbO_3$ are primarily those of high coupling constant k2 and low temperature dependence of the frequency of operation when used in a resonant configuration. The patent is clearly written, including references to their published work and describing the process for depositing the material on quartz and silicon in the preferred crystallographic orientations. Values of the coupling coefficient as high as 0.5 are claimed, making this a very attractive alternative to $LiNbO_3$ for many applications.—JAH

7,219,536

43.35.Zc SYSTEM AND METHOD TO DETERMINE OIL QUALITY UTILIZING A SINGLE MULTI-FUNCTION SURFACE ACOUSTIC WAVE SENSOR

James Z. T. Liu *et al.*, assignors to Honeywell International Incorporated
22 May 2007 (Class 73/54.24); filed 29 April 2005

An acoustic oil quality sensor **300** is claimed that uses a quartz piezoelectric transducer **100** installed inside an engine oil filter and in contact with circulating engine oil **314**. Oil acidity is measured as an upward shift of the fundamental thickness-resonance mode, typically 5 MHz, due to the



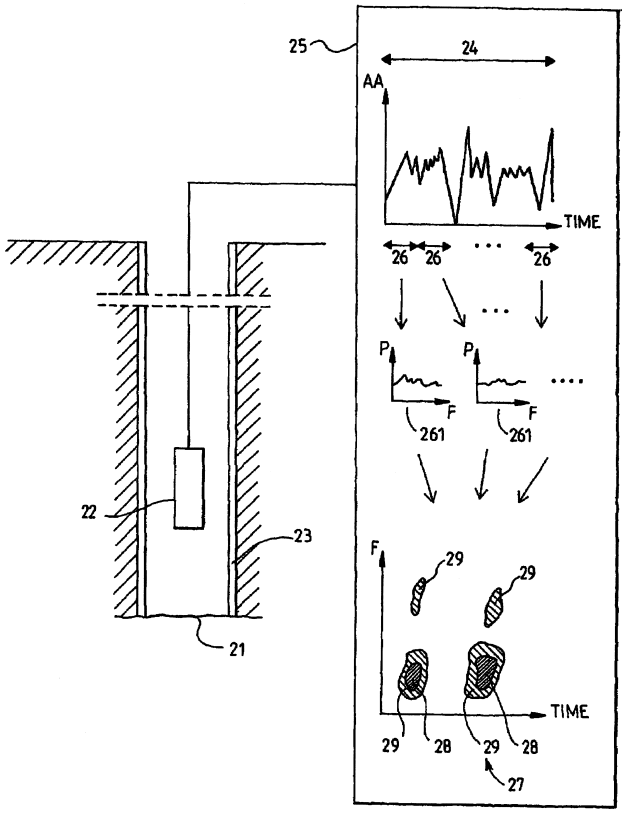
etching away of a metal coating material from the surface of **100**, chosen to simulate critical engine materials that wear. Oil viscosity is measured as the amplitude and phase shift of such vibration vs the driving signal. Rf interrogation **302**, **308**, **310**, **108**, **110** can be remote for processing and readout in the vehicle cabin.—AJC

7,219,762

43.35.Zc METHOD AND APPARATUS FOR ACOUSTIC DETECTION OF A FLUID LEAK BEHIND A CASING OF A BOREHOLE

Simon James and Peter Fitzgerald, assignors to Schlumberger Technology Corporation
22 May 2007 (Class 181/105); filed in the European Patent Office 6 June 2003

A borehole gas leak detector and means to repair that leak are claimed where a hydrophone or geophone is lowered slowly down a well borehole while sensor output is processed **25**. A fast Fourier transform (FFT) **F** of the signal is displayed, typically 0–12 kHz as vs **TIME**, at such a rate that the position of **22** down borehole **23** is evident. Gas leakage appears as certain



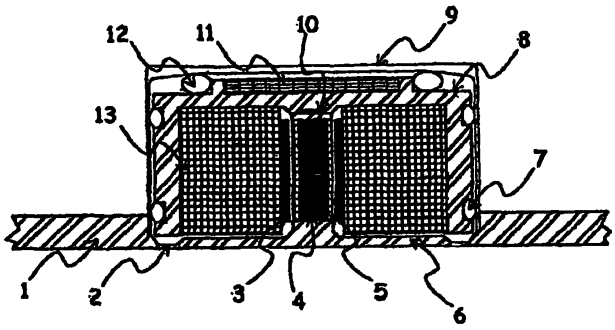
frequency bands **28**, **29** of the local acoustic noise characteristic of a leak. A repair method is also claimed where a following device (not shown) perforates casing **23** at the indicated borehole position and injects a sealing fluid through the casing to seal that hole and its environs to stop the gas leak.—AJC

7,224,814

43.38.Ct TWO-WAY COMMUNICATION DEVICE

Bryan Paul Cross, Doncaster, South Yorkshire and John Anthony Moran, Brigg, North East Lincolnshire, both of United Kingdom
29 May 2007 (Class 381/190); filed in United Kingdom 5 May 2001

Steel front panel face plate **1** is manufactured with flexural supports **2** so that diaphragm **5** can vibrate when actuated by magnetostrictive rod **4** via drive coil **13**. The device works in reverse since sense coil **5** allows the



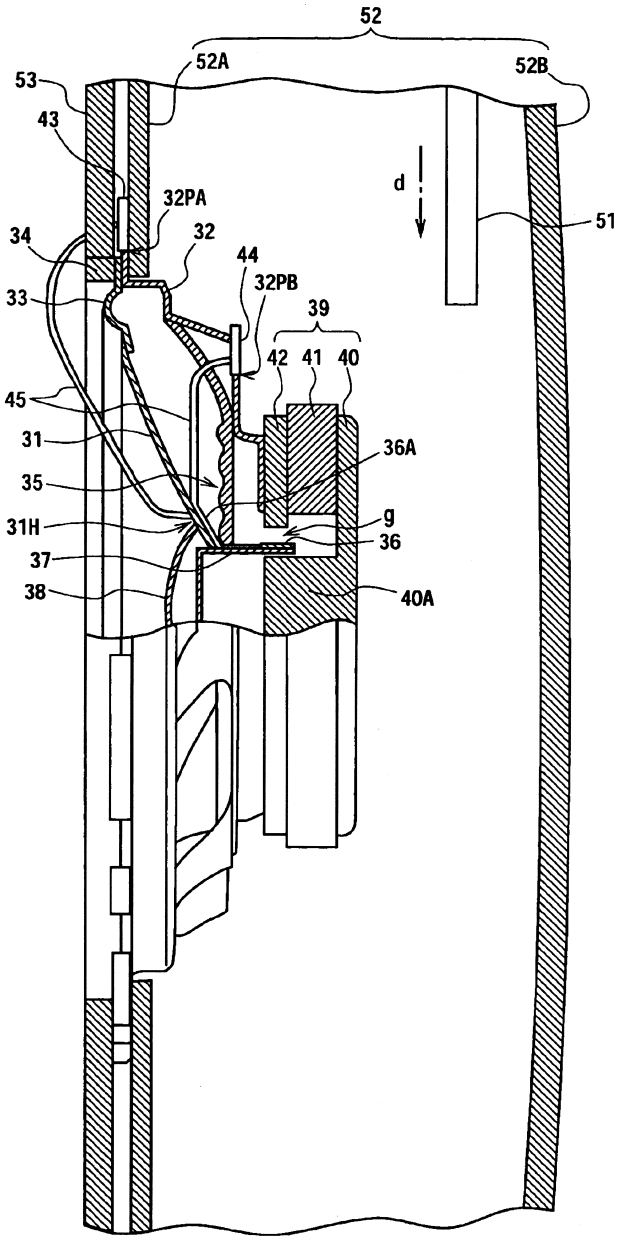
device to act as a microphone. Magnet **3** biases the magnetostrictive rod and prestress spring **12** is said to maximize the strain in rod **4**. The assembly is protected from the elements by outer case **9** and front panel **1**. The device is said to be useful for intercoms.—NAS

7,233,680

43.38.Dv SPEAKER DEVICE

Shinji Kobayashi and Haruto Kusunoki, assignors to Sony Corporation
19 June 2007 (Class 381/396); filed in Japan 17 November 2003

This reviewer thinks that this invention was developed to solve what the patent describes as the problem with using loudspeakers in automobile doors where one side may be exposed to "wet." What exactly is the problem



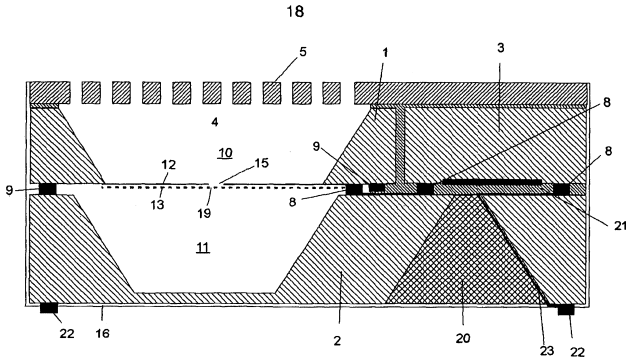
is unclear, but the patent implies that having lead wires **45**, as well as connectors **43** and **44**, on both the interior side and the exterior side of the loudspeaker solves the problem.—NAS

7,221,767

43.38.Fx SURFACE MOUNTABLE TRANSDUCER SYSTEM

Matthias Mullenborn *et al.*, assignors to Sonion Mems A/S
22 May 2007 (Class 381/174); filed 20 December 2002

A fabrication method is claimed for making microphone **18** from a silicon wafer. The microphone comprises acoustic pressure sensing dia



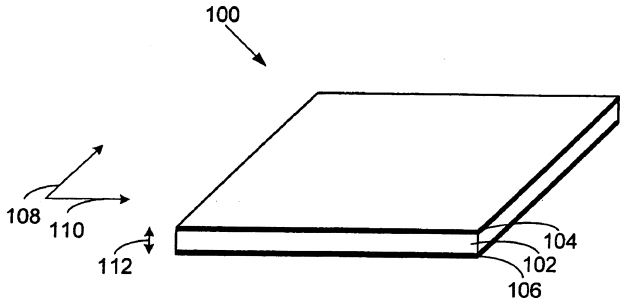
phragm **12**, protective grid **5**, electronics package **3**, and EMI shield **16**. Two or more adjacent microphone elements can be fashioned from a single silicon substrate to achieve sound directivity. Electrical contact for charged backing grid **19** is via insulated solder bumps **8**. Hole **15** is not explained, but may be a front static pressure-relief vent.—AJC

7,259,503

43.38.Fx ELECTROACTIVE POLYMERS

Qibing Pei *et al.*, assignors to SRI International
21 August 2007 (Class 310/363); filed 18 January 2006

This patent discloses new applications of electro-active polymers (EAPs), building on the authors' previous patents in the field. The inventors describe a variety of mechano-electrical transducer configurations that use their EAP material as the working element. These materials are made by casting thin elastomer films and then poling them at high voltage, creating a material with a preferred actuation direction. It is asserted that this material



(which is generally a silicone elastomer) can tolerate and generate strains of 100% in the 3 direction and 500% in the 1 direction. The material responds to voltages by shrinking and expanding along the poling axis, with consequent expansion and shrinkage in the perpendicular directions, resulting in nearly zero volume change. Many applications are outlined in the patent, but little data are given on the performance of these actuators, which has usually not been as good as asserted here.—JAH

7,236,427

43.38.Hz VESSEL HULL TRANSDUCER MODULAR MOUNTING SYSTEM

Terrence K. Schroeder, assignor to SWCE
26 June 2007 (Class 367/188); filed 11 March 2005

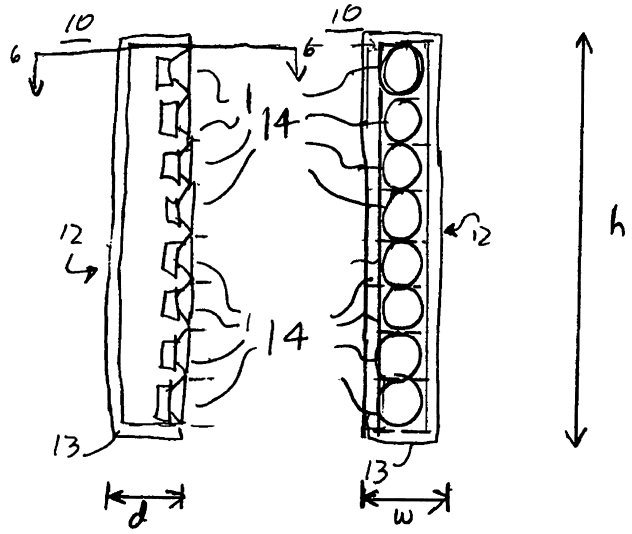
A modular hull mounting system is proposed for acoustic transducers that allows them to be swapped without dry docking.—GFE

7,260,235

43.38.Hz LINE ELECTROACOUSTICAL TRANSDUCING

Clifford A. Henricksen and Kenneth D. Jacob, assignors to Bose Corporation
21 August 2007 (Class 381/335); filed 16 October 2000

The patent teaches that a tall line array of small, closely spaced loudspeakers displays a number of desirable attributes when used for live sound



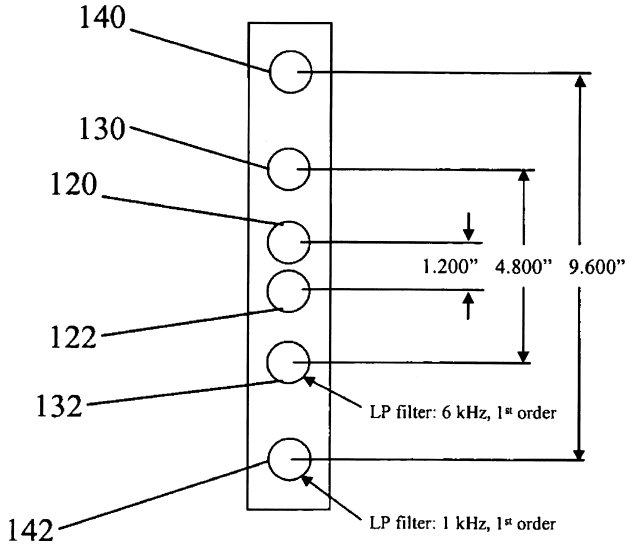
reinforcement. The basic concept is a little tricky to patent since it has been used for more than 40 years. As a result, the patent claims are limited to a specific two-cabinet portable configuration.—GLA

7,260,228

43.38.Hz OPTIMUM DRIVER SPACING FOR A LINE ARRAY WITH A MINIMUM NUMBER OF RADIATING ELEMENTS

Charles Emory Hughes and Kirk Samuel Lombardo, assignors to Altec Lansing, a division of Plantronics, Incorporated
21 August 2007 (Class 381/89); filed 10 March 2004

This patent describes a simple loudspeaker line array that uses a special combination of loudspeaker spacing and lowpass filtering. "The present application utilizes a spacing arrangement of the radiating elements in an array that is neither logarithmic nor equidistantly spaced." The embodiment shown is about 11 in. high and incorporates six 1-in.-diam loudspeakers operating down to 1000 Hz. The vertical pattern is free of significant secondary lobes up to about 3000 Hz.—GLA

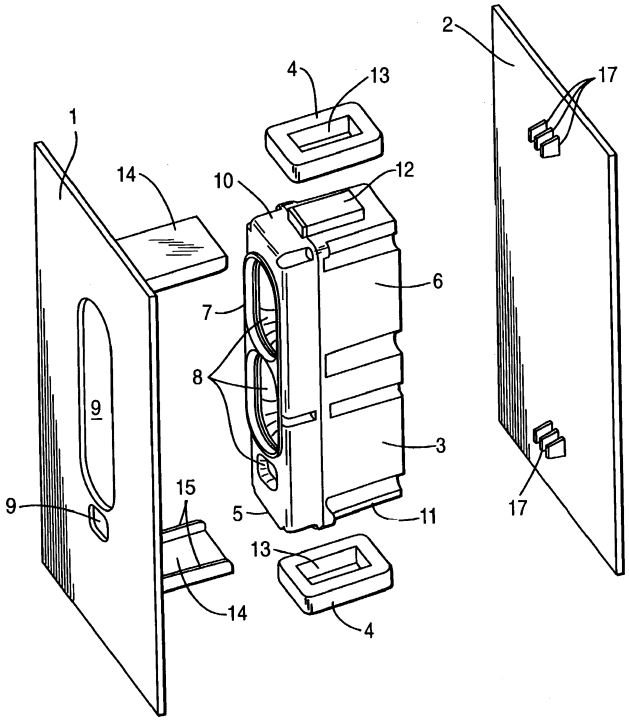


7,221,772

43.38.Ja ELECTRONIC DEVICE COMPRISING A LOUDSPEAKER UNIT

Michel Evenisse *et al.*, assignors to Thomson Licensing
22 May 2007 (Class 381/386); filed in the European Patent Office 19 December 2001

Annular-shaped vibration-absorbing bodies **4** isolate loudspeaker unit



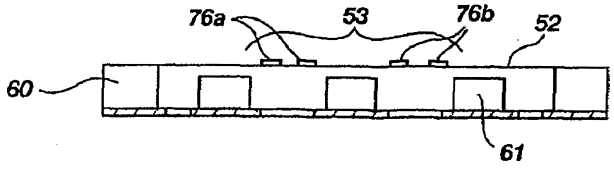
3 from mounting walls **1** and **2**, which may be part of a television cabinet.—NAS

7,251,342

43.38.Ja SINGLE END PLANAR MAGNETIC SPEAKER

David Graebener, assignor to American Technology Corporation
31 July 2007 (Class 381/431); filed 2 March 2001

The illustration is a section through a single-ended planar magnetic loudspeaker. Conductors **76a** and **76b** are affixed to diaphragm **52**. Magnets **61** are mounted to a rigid backplate that is partially perforated. A squished-out magnetic field is formed between each pair of magnets and the lines of flux cross the conductors at angles approaching 90°. Commercial versions



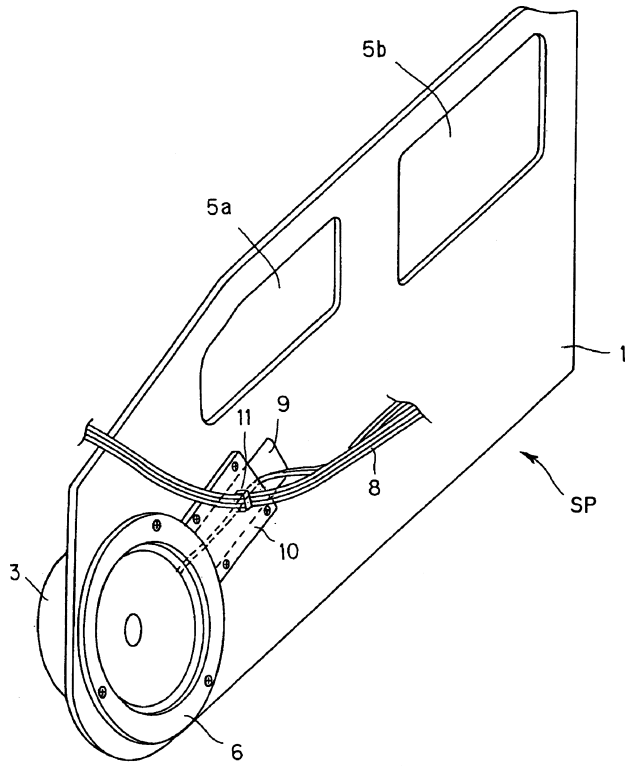
incorporating small ceramic bar magnets have been used as tweeters (enclosed rear cavity) or as wide-range bipolar loudspeakers. According to the patent, replacing the ceramic magnets with newer high-energy magnets increases flux density more than anticipated. The patent claims are composed in painfully knotted English in an effort to extract a novel invention from the preceding observation.—GLA

7,227,969

43.38.Ja INNER PANEL LOUDSPEAKER APPARATUS

Koji Maekawa *et al.*, assignors to Pioneer Corporation
5 June 2007 (Class 381/345); filed in Japan 14 November 2002

Slot **9** and cover **11** form a port to housing unit **3** which is part of car door inner panel **1**. The cables for speaker **6** are run from harness **8** to the



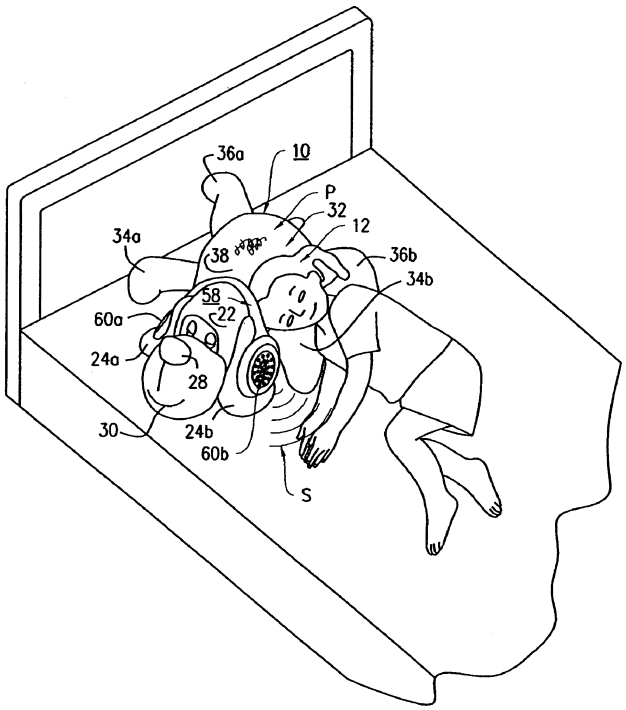
speaker via slot **9**. Cover **10** also secures wire clip **11**. The patent includes formulas for Helmholtz resonators and both open and closed tube modes.—NAS

7,227,965

43.38.Md PILLOW IN THE FORM OF A STUFFED TOY OR 3-D CHARACTER TOY HAVING TWO HEADPHONE SPEAKERS MOUNTED ON THE EARS OF THE TOY

Joseph A. Sutton, assignor to Jay Franco & Sons, Incorporated
5 June 2007 (Class 381/124); filed 24 April 2006

A pillow **10**, which may be in the shape of a stuffed animal or a three-dimensional character toy, is fitted with speakers **60a** and **60b** in the form of a headset **58** over head **22**. An audio player can be placed in a



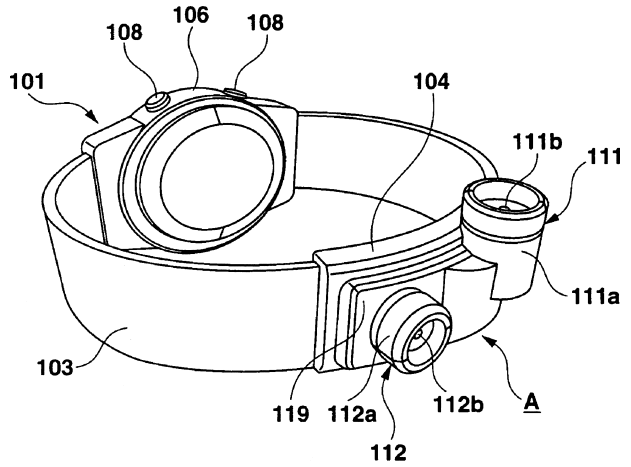
pocket in appendage **34a**. A user can thus listen to the output of the audio player "while laying his/her head **12** on the plush toy pillow **10** while relaxing."—NAS

7,251,197

43.38.Si WRIST-WORN COMMUNICATIONS APPARATUS

Kaoru Yoshida and Yoshiyuki Murata, assignors to Casio Computer Company, Limited
31 July 2007 (Class 368/10); filed in Japan 30 January 2003

The Dick Tracy two-way wrist radio has become a reality. Microphone **112** and tiny loudspeaker **111** are mounted on the palm side of a wrist band while the display and control elements **101** take the place of a conventional wristwatch. When the wearer places his hand to his ear, the microphone faces his mouth while the loudspeaker is positioned to squirt sound upward



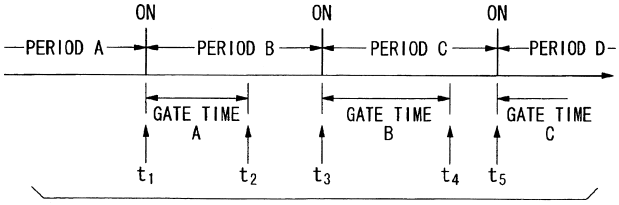
toward his ear. Since the speaker assembly appears to be about 8 mm in diameter, any usable directionality must lie above 10 kHz or so. However, it may be that the surfaces provided by the wrist and the cheek form a crude waveguide.—GLA

7,253,350

43.38.Si VIBRATION SOURCE DRIVING DEVICE

Masao Noro *et al.*, assignors to Yamaha Corporation
7 August 2007 (Class 84/600); filed in Japan 22 October 1999

Hey dude, suppose your cell phone vibrator could be synchronized to



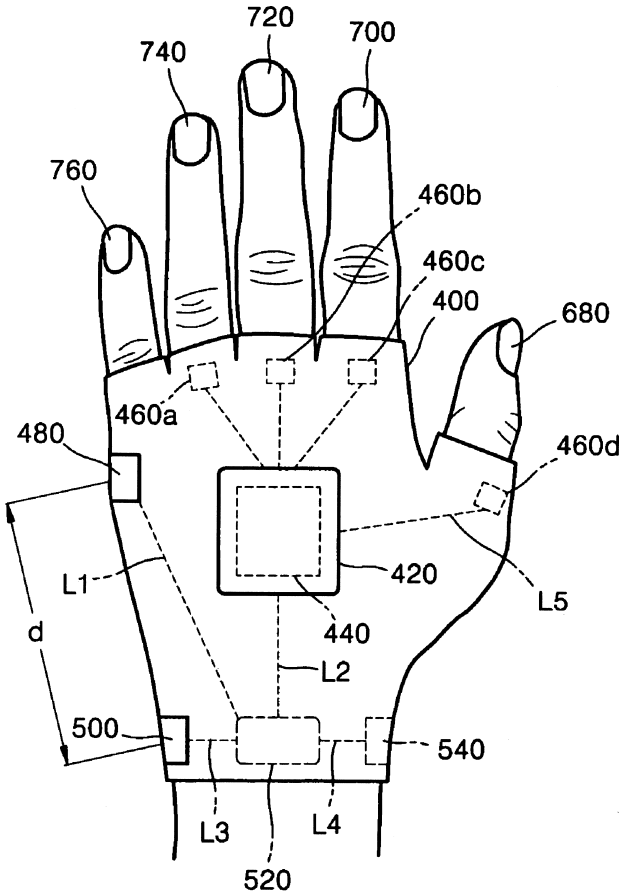
your ring tone melody? Cool, huh.—GLA

7,254,376

43.38.Si WEARABLE PHONE AND METHOD OF USING THE SAME

Tae-suh Park and Sang-goog Lee, assignors to Samsung Electronics Company, Limited
7 August 2007 (Class 455/100); filed in Republic of Korea 27 June 2003

The illustration shows an interesting glove phone that senses finger



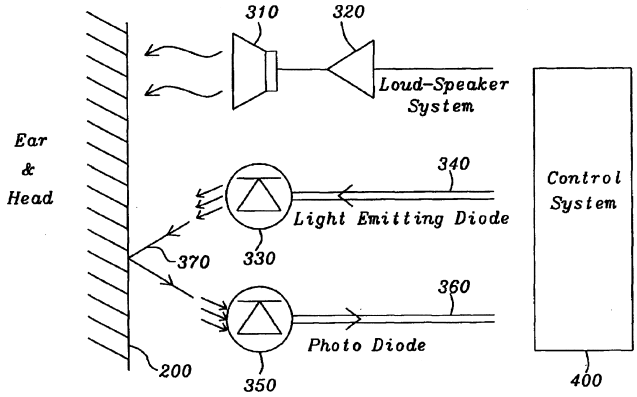
movements to provide a virtual keypad.—GLA

7,260,422

43.38.Si MONOLITHIC OPTICAL AUTOCOMPENSATION READ-OUT CIRCUIT FOR DISTANCE DEPENDENT LOUDNESS CONTROL IN MOBILE PHONES

Horst Knoedgen, assignor to Dialog Semiconductor GmbH
21 August 2007 (Class 455/569.1); filed in the European Patent Office 6 February 2003

This is an unusual patent. The actual operation of the invention is buried in a lengthy discussion of photons, CMOS technology, spectral response, and IR pulses. The goal is to control the loudness of a cellular phone



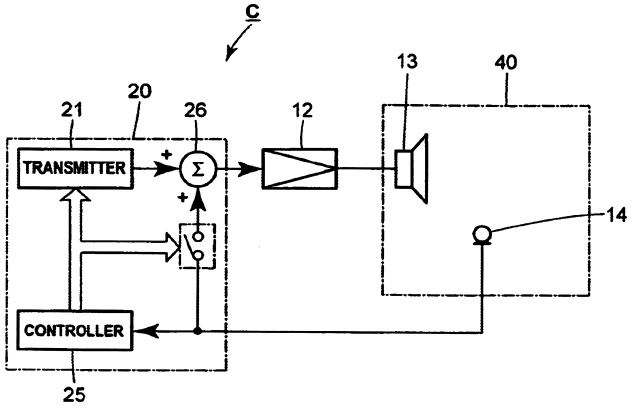
loudspeaker in relation to its distance from the user's ear. Fairly elaborate optical sensing technology is used to create an appropriate measuring system.—GLA

7,249,511

43.38.Tj METHOD OF DETECTING RESONANT FREQUENCY, METHOD OF SELECTING RESONANCE FREQUENCY, AND DEVICE FOR DETECTING RESONANT FREQUENCY SENSOR

Daisuke Higashihara, assignor to TOA Corporation
31 July 2007 (Class 73/579); filed in Japan 9 December 2002

This patent explains that a loudspeaker in a reverberant space may be "difficult to listen to" because of resonant peaks in the response. Acoustic feedback is not mentioned in the disclosure or the claims and we are led to conclude that the patent deals only with sound reproduction rather than sound reinforcement. Instead of relying on the measured response to identify strong peaks, a two-step process is described. First, a conventional response curve is run. A second curve is then run with the microphone live, as in a PA



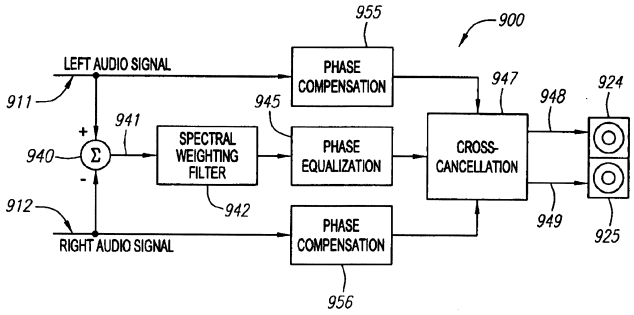
system. The difference between the two curves is used to identify peaks that can be attenuated with notch filters. In fact, this is a variation of a procedure commonly employed to equalize sound reinforcement systems. Unfortunately, the ring frequencies thus identified do not necessarily match the peaks of the playback-only system. To this reviewer, the method appears to be nonfunctional for its intended application.—GLA

7,254,239

43.38.Vk SOUND SYSTEM AND METHOD OF SOUND REPRODUCTION

Lawrence R. Fincham, assignor to THX Limited
7 August 2007 (Class 381/17); filed 11 February 2002

Home theater surround-sound installations are expected to reproduce a variety of formats ranging from two-channel stereo to full 6.1 surround sound, and this can produce unwanted anomalies in certain situations. This patent describes a technique that is especially well suited for driving rear



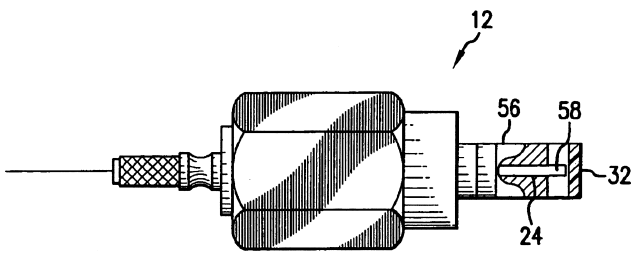
surround speakers in a way that maintains desired localization and "spread" regardless of the playback format. The patent is clearly written and includes much useful information about home surround-sound reproduction.—GLA

7,224,465

43.38.Zp FIBER TIP BASED SENSOR SYSTEM FOR MEASUREMENTS OF PRESSURE GRADIENT, AIR PARTICLE VELOCITY AND ACOUSTIC INTENSITY

Balakumar Balachandran *et al.*, assignors to University of Maryland
29 May 2007 (Class 356/480); filed 21 January 2005

A fiber tip Fabry Perot (FTFB) sound-pressure sensor and signal processor is claimed. Mirror **32** on a mylar film diaphragm that deflects under acoustic pressure changes the phase and intensity of light from and reflected back to optical fiber **58**. The resulting light sum is detected by a photodiode



(not shown). This small device can be arranged in one or more pairs to detect a pressure gradient, hence acoustic velocity, on one or more axes.—AJC

7,240,766

43.40.Tm SOUND DAMPENING PAD

Brandon Rogers and Darren R. McCracken, assignors to Day International, Incorporated
10 July 2007 (Class 181/209); filed 14 January 2004

A sound damping pad is described that attenuates sound from printing presses in order to protect the hearing of their operators.—GFE

7,243,395

43.40.Tm BUMPER DEVICE

Bradley J. Haymond, assignor to Illinois Tool Works Incorporated
17 July 2007 (Class 16/86 R); filed 26 June 2003

Closing cabinets just got quieter thanks to this patent that describes a more quiet bumper.—GFE

7,246,681

43.40.Tm RADIO FREQUENCY SHIELDED AND ACOUSTICALLY INSULATED ENCLOSURE

Walter J. Christen, assignor to Imedco AG
24 July 2007 (Class 181/285); filed 6 August 2003

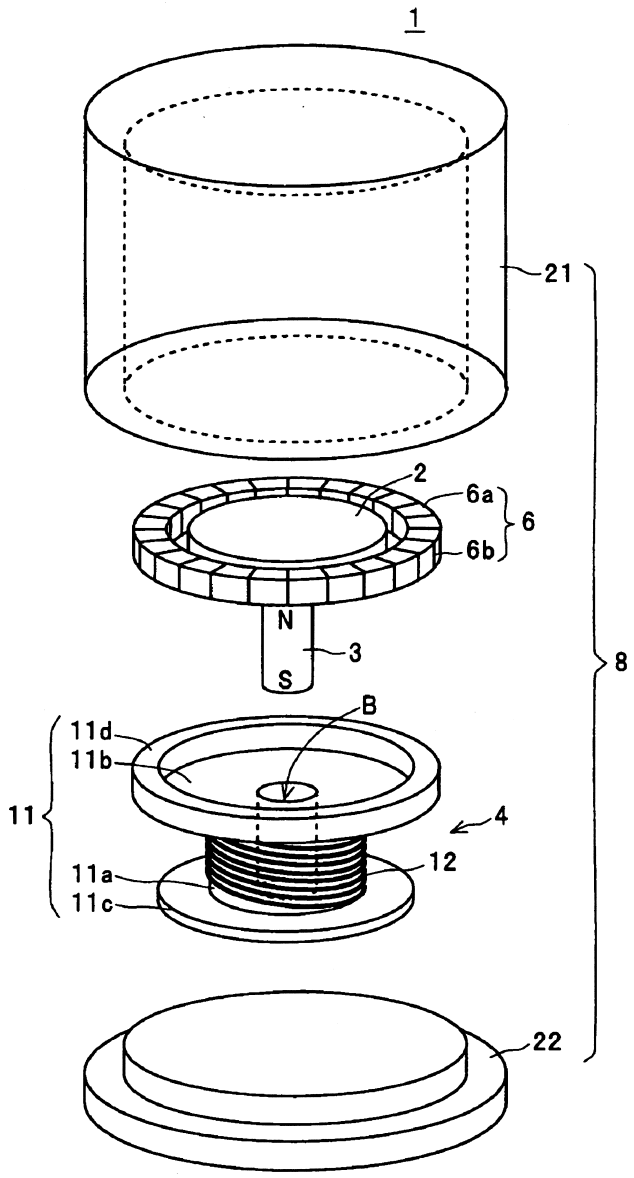
A cost-effective acoustic vibration damping system is described, in conjunction with a radio-frequency absorbing room, for isolating magnetic resonance imaging equipment from ambient noise signals.—GFE

7,219,547

43.40.Yq ANGULAR VELOCITY SENSOR AND ANGULAR VELOCITY DETECTOR

Takahiko Suzuki, assignor to TDK Corporation
22 May 2007 (Class 73/504.04); filed in Japan 16 May 2003

An angular velocity sensor **1** is claimed where alternating current through coil **12** causes an alternating magnetic flux in core **3** and vibrates disk **2**, both of magnetic material. Disk **2** undergoes magnetostrictive radial expansion-contraction adjacent to toroidal core **6a**. Toroidal core **6a** and case **21**–**22**, each also of magnetic material, complete a magnetic circuit back to core **3**. Winding **6b** on toroidal core **6a** picks up no ac signal for a stationary sensor **1** because the alternating magnetic flux is parallel to the



plane of its windings. Rotation of sensor **1** on its axis of symmetry will result in an acceleration coriolis force on **2** such that the magnetostrictive motion is tilted on the circumference from being purely radial. This tilted motion is no longer parallel to the toroidal winding planes, so an ac voltage appears on output winding **6a** that is proportional to the sensor body angular velocity on its axis.—AJC

7,242,952

43.60.Dh PORTABLE TERMINAL DEVICE AND METHOD OF GENERATING CALL SOUND

Katsuya Shirai *et al.*, assignors to Sony Corporation
10 July 2007 (Class 455/467); filed in Japan 4 August 2003

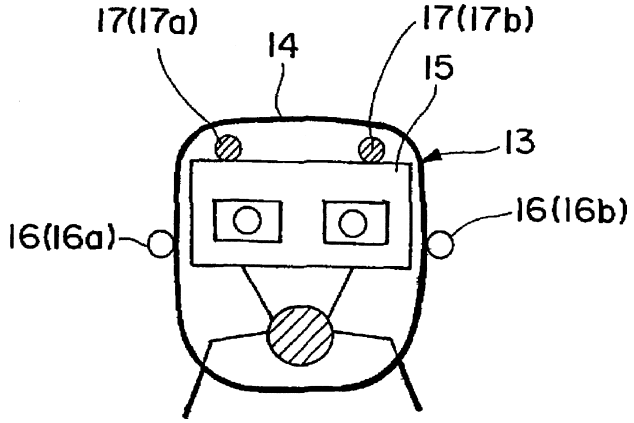
How would you like to have your pulse, temperature, acceleration, electroconductivity, blood flow rate, and more constantly measured just so that your cell phone could change its ring tone based on your mood?—GFE

7,215,786

43.60.Jn ROBOT ACOUSTIC DEVICE AND ROBOT ACOUSTIC SYSTEM

Kazuhiro Nakadai *et al.*, assignors to Japan Science and Technology Agency
8 May 2007 (Class 381/94.1); filed in Japan 9 June 2000

This patent deals with a variation on a robotic auditory perception mechanism previously discussed in United States Patent 7,016,505, reviewed in J. Acoust. Soc. Am. **121**(5), 2401 (2007). The same noise-canceling, two-microphone sound pickup system is here applied to the task of source localization and to the subsequent control of a motor for head



orientation. Just as described in the prior patent, each external microphone is associated with an internal noise-canceling microphone. The source direction analyzer is based on an auditory epipolar geometry together with a harmonic analyzer. A comparison is provided between an auditory system of the type described here and a system based on head-related transfer functions.—DLR

7,248,701

43.55.Ka DYNAMIC ACOUSTIC RENDERING

Alan Gerrard and Nam Do, assignors to Creative Technology, Limited
24 July 2007 (Class 381/17); filed 6 October 2005

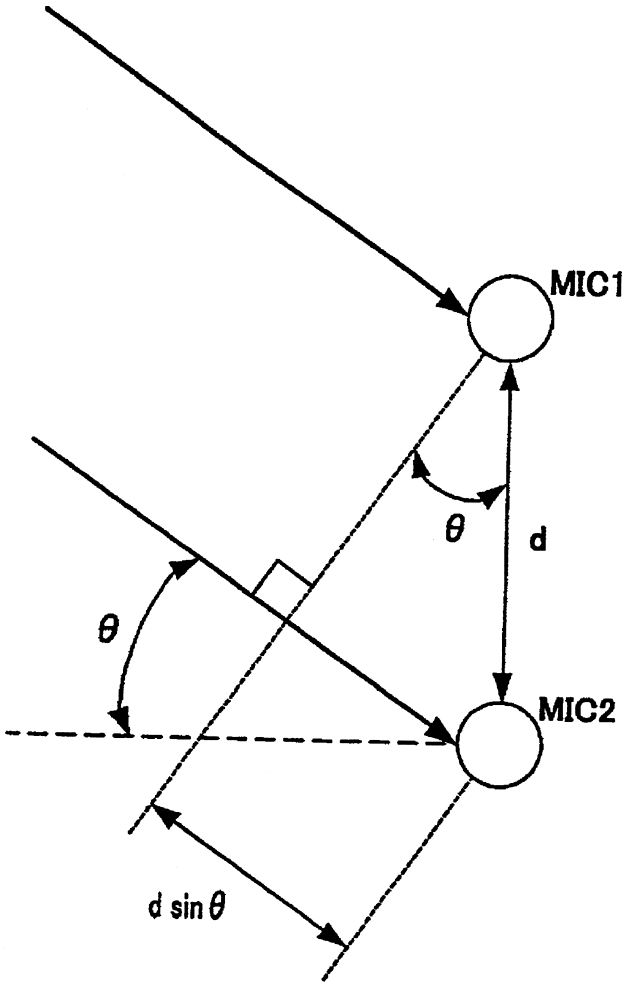
In order to more realistically replicate sound propagation (reflections, occlusions, and absorption) in video games, a method is described that takes advantage of the simulated graphics in 3D environments. Essentially, the graphical polygons serve a dual purpose: visual rendering and acoustic surface interaction.—GFE

7,227,960

43.60.Jn ROBOT AND CONTROLLING METHOD OF THE SAME

Toshihiko Kataoka, assignor to International Business Machines Corporation
5 June 2007 (Class 381/92); filed in Japan 28 May 2001

This patent describes an alternative method of source direction detection, differing from that of United States Patent 7,215,786, reviewed above.



Here, the emphasis is on a delay beamformer type of direction analyzer. A correlation of the two microphone signals provides a delay value which is used to compute the source incidence angle.—DLR

7,251,338

43.66.Ts METHOD FOR HANDLING DATA OF A HEARING DEVICE AND HEARING DEVICE

Ruedi Suter and Andreas Nickisch, assignors to Phonak AG
31 July 2007 (Class 381/314); filed 10 June 2002

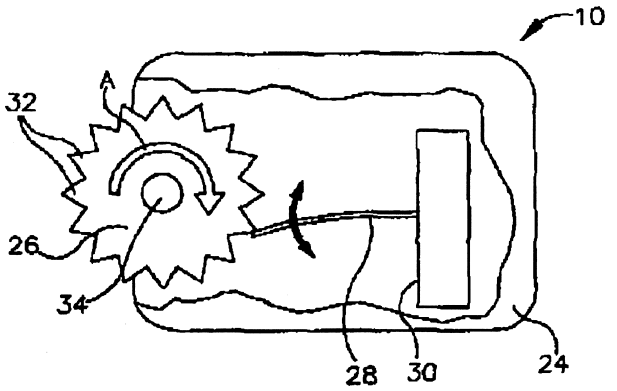
A security feature is implemented via a personal code that allows only specialists who have previously entered data into a hearing aid to enter data again. The security feature is also applied to new or repaired hearing devices that are sent from a manufacturer after being selectively enabled to unconditionally accept data. These features help prevent hearing aid wearers from going to a different specialist who may simply copy data from an existing hearing device into a new hearing device.—DAP

7,251,339

43.66.Ts WIRELESS REMOTE CONTROL FOR A HEARING INSTRUMENT

JAMES GREGORY RYAN, assignor to Gennum Corporation
31 July 2007 (Class 381/315); filed 31 March 2004

A passive remote control utilizes the user's mechanical finger actuation to cause audio pulses to be transmitted to program a hearing aid. The basic embodiment in the remote unit is a cogged wheel rotationally positioned against one or more reeds that are cantilevered to contact each cogged



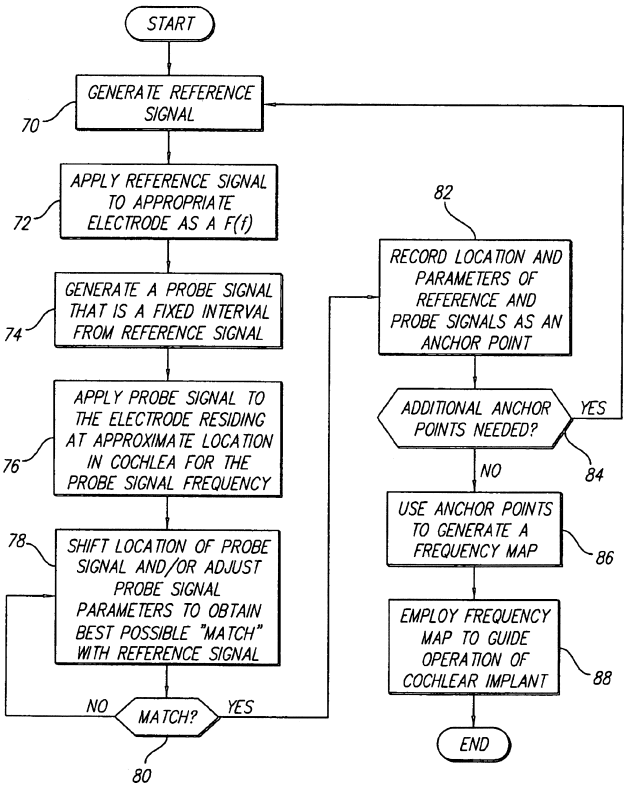
wheel. The resulting audio pulses cause changes in volume level or memory selection. In the hearing aid, circuitry and software include a full-wave rectifier, low-pass filter and pulse time/space detection to decode the audio pulses.—DAP

7,251,530

43.66.Ts OPTIMIZING PITCH AND OTHER SPEECH STIMULI ALLOCATION IN A COCHLEAR IMPLANT

Edward H. Overstreet *et al.*, assignors to Advanced Bionics Corporation
31 July 2007 (Class 607/55); filed 9 December 2003

The temporal structure of the stimulating wave form is controlled by allowing cochlear implant users to make corrections and adjustments in stimulus parameters so that pitch is accurately perceived. A probe signal having a fixed interval relationship to a reference signal, for example one



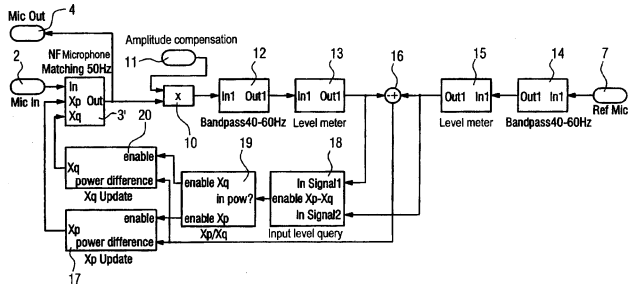
octave higher, and the reference signal are applied to an appropriate electrode of the implanted system. The locations of the probe and reference signals are shifted until the two signals sound matched to the user. These locations may be used to frequency map other signals to correct locations within the cochlea.—DAP

7,254,245

43.66.Ts CIRCUIT AND METHOD FOR ADAPTATION OF HEARING DEVICE MICROPHONES

Georg-Erwin Arndt *et al.*, assignors to Siemens Audiologische Technik GmbH
7 August 2007 (Class 381/312); filed in Germany 11 March 2003

A directional microphone system has a high-pass filter transfer function at low frequencies, thereby causing reduced low-frequency sensitivity relative to that of an omni-directional microphone. As a result, differences in



low-frequency sensitivity and phase may occur between two or more microphones in a directional hearing aid device that may vary over time, depending on environmental effects and ingress of contaminants. These differences may be reduced adaptively by selective filtering in a feedback regulation loop. One of the microphones is designated as a reference microphone to which the other microphones are equalized.—DAP

7,254,246

43.66.Ts METHOD FOR ESTABLISHING A BINAURAL COMMUNICATION LINK AND BINAURAL HEARING DEVICES

Andreas Jakob, assignor to Phonak AG
7 August 2007 (Class 381/315); filed 22 January 2002

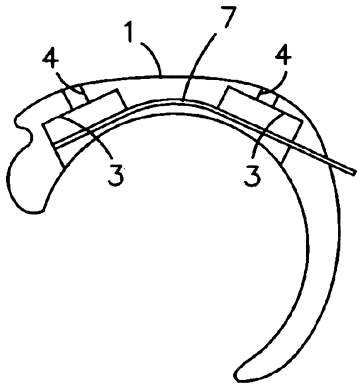
A low-delay, bi-directional communication link between hearing devices in a binaural fitting is established across the head via conduction through the wearer's body and by at least one wire. The wire may be detachably connected to the hearing devices via magnetic attraction and capacitive, ferromagnetic metal, or conductive polymer contacts. Connection to the wearer's body from the hearing devices is made with capacitive or conductive polymer electrodes. Either control signals or audio signals may be transmitted. Another implementation is a wireless transmitter-receiver unit connected to two wires going between the hearing devices.—DAP

7,254,247

43.66.Ts HEARING AID WITH A MICROPHONE IN THE BATTERY COMPARTMENT LID

Lasse Kragelund *et al.*, assignors to Oticon A/S
7 August 2007 (Class 381/322); filed in Denmark 7 December 2001

At least one microphone **4** on a printed circuit board (PCB) **7** is embedded in the battery drawer of a hearing aid. Inlets are provided in the



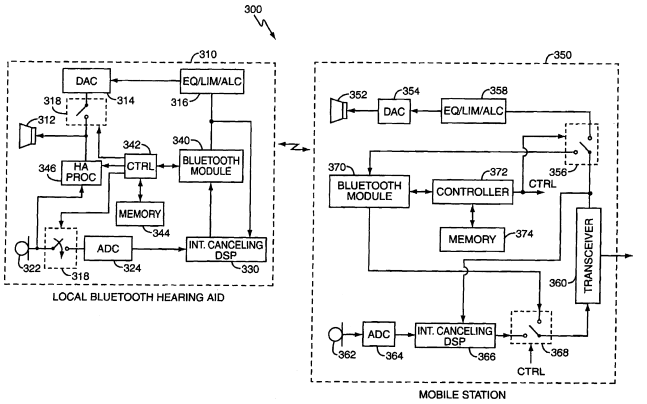
battery compartment to allow sound to reach the microphones, which may be of MEMS construction. The PCB may be made of flexible material.—DAP

7,257,372

43.66.Ts BLUETOOTH ENABLED HEARING AID

Matt Andrew Kaltenbach *et al.*, assignors to Sony Ericsson Mobile Communications AB
14 August 2007 (Class 455/41.2); filed 30 September 2003

This patent focuses on signal processing for hearing impaired persons that is incorporated into wireless headsets or other headworn devices to



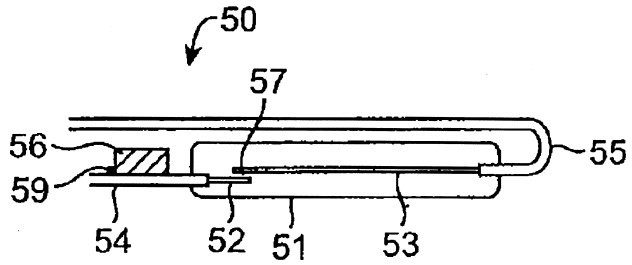
improve suppression of environmental noise, electromagnetic interference, and acoustic echoes while communicating with nearby mobile stations such as cellular telephones.—DAP

7,260,232

43.66.Ts REMOTE MAGNETIC ACTIVATION OF HEARING DEVICES

Adnan Shennib, assignor to InSound Medical, Incorporated
21 August 2007 (Class 381/315); filed 7 February 2006

One or more magnetic switches in an in-the-canal hearing device are activated and deactivated, respectively, by the wearer bringing a hand-held magnet into close or far away proximity. Each switch assembly may include one or more miniature reed switches. A miniature latching magnet maintains



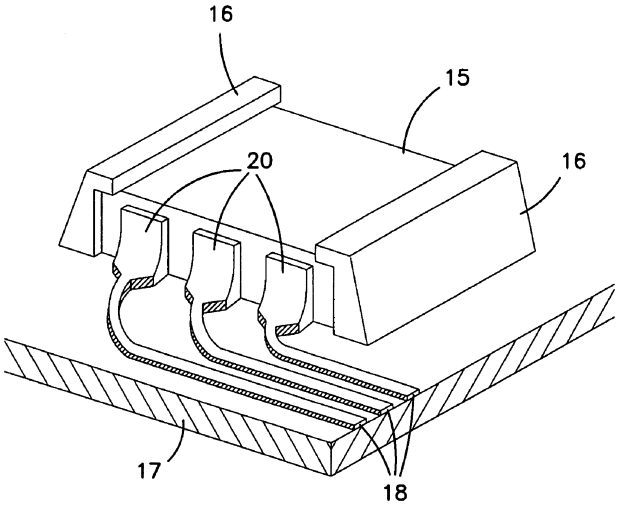
the reed switches in desired positions after removal of the external hand-held magnet. The switches can control power on/off, output sound volume, and frequency response.—DAP

7,260,233

43.66.Ts HEARING AID OR SIMILAR AUDIO DEVICE AND METHOD FOR PRODUCING A HEARING AID

Klaus L. Svendsen and Per Lundberg, assignors to Oticon A/S
21 August 2007 (Class 381/322); filed in Denmark 10 July 2002

Conventional printed circuit boards may be eliminated in hearing aids by forming electrical leads **18**, including those for the battery, directly onto the molded surfaces of the hearing aid housing and reflowing components onto these electrical leads. The advantages are said to be increased use of



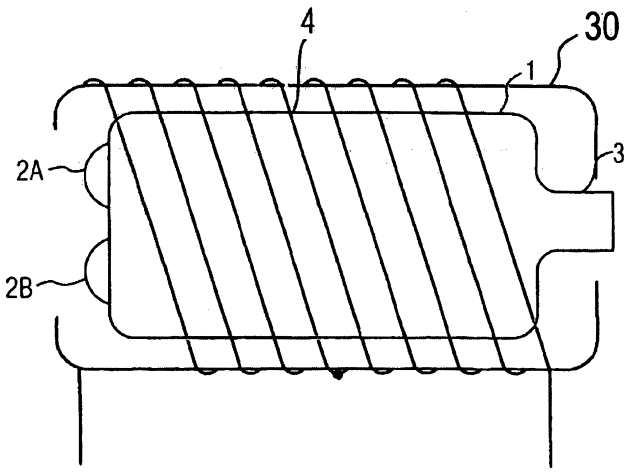
automation, reduced overall size, and a reduced number of electrical wires. The method may be utilized on the inside surfaces of behind-the-ear cases and on inner surfaces of custom hearing aid faceplates.—DAP

7,260,234

43.66.Ts SPACE-SAVING ANTENNA ARRANGEMENT FOR HEARING AID DEVICE

Thomas Kasztelan *et al.*, assignors to Siemens Audiologische Technik GmbH
21 August 2007 (Class 381/324); filed in Germany 12 August 2002

An antenna coil for wireless transmission between a hearing device and another device is wound around the microphone or receiver in the hearing device. An additional compensating coil cancels the magnetic leakage fields emanating from the receiver so they do not disturb the wireless



transmissions. The compensating coil and antenna coil can be implemented as a single coil with a center tap. A shielding plate made of mu-metal, ferrite, or sheet iron may be added around the receiver to further attenuate the magnetic leakage fields.—DAP

7,251,605

43.71.Ky SPEECH TO TOUCH TRANSLATOR ASSEMBLY AND METHOD

Robert V. Belenger and Gennaro R. Lopriore, assignors to The United States of America as represented by the Secretary of the Navy
31 July 2007 (Class 704/271); filed 19 August 2002

The device described in this patent would perform a phonetic analysis of the speech signal and generate controls for a vibrator mechanism which uses the distinctive phonemic features to produce a touch-sensitive version

ENGLISH PHONEMES (WORD SOUNDS) WITH A SET OF PRESSURE FINGER ACTUATION CODES

0=PRESSURE FINGER NOT ACTUATED
1=PRESSURE FINGER ACTUATED

	CONSONANT SOUNDS				PRESSURE FINGER ACTUATION CODES PRESSURE FINGER #					
					1	2	3	4	5	6
1	p	as in	s l p		0	0	0	0	0	1
2	p	as in	p e n		0	0	0	0	1	0
3	b	as in	b i t		0	0	0	0	1	1
4	m	as in	m a p		0	0	0	1	0	0
5	w	as in	w i t		0	0	0	1	0	1
6	ou	as in	o u t		0	0	0	1	1	0
7	f	as in	f a t		0	0	0	1	1	1
8	v	as in	v a t		0	0	1	0	0	0
9	t	as in	t h i n		0	0	1	0	0	1
10	th	as in	t h i s		0	0	1	0	1	0
11	st	as in	s t e p		0	0	1	0	1	1
12	t	as in	t i p		0	0	1	1	0	0
13	d	as in	d i p		0	0	1	1	0	1
14	n	as in	n i p		0	0	1	1	1	0
15	l	as in	l i p		0	0	1	1	1	1
16	tt	as in	u t t e r		0	1	0	0	0	0
17	s	as in	s i p		0	1	0	0	0	0
18	z	as in	z i p		0	1	0	0	1	0
19	r	as in	r e d		0	1	0	0	1	1
20	ss	as in	m i s s i o n		0	1	0	1	0	0
21	s	as in	v i s i o n		0	1	0	1	0	1
22	ck	as in	s l c k		0	1	0	1	1	0
23	k	as in	k i s s		0	1	0	1	1	1
24	g	as in	g i v e		0	1	1	0	0	0
25	ng	as in	k i n g		0	1	1	0	0	1
26	y	as in	y e t		0	1	1	0	1	0
27	i	as in	b i t e		0	1	1	0	1	1
28	h	as in	h i t		0	1	1	1	0	0

	VOWEL SOUNDS				PRESSURE FINGER ACTUATION CODES PRESSURE FINGER #					
					1	2	3	4	5	6
29	ee	as in	b e e t		0	1	1	1	0	1
30	i	as in	b i t		0	1	1	1	1	0
31	i	as in	b i d		0	1	1	1	1	1
32	ai	as in	a i d		1	0	0	0	0	0
33	a	as in	a t		1	0	0	0	0	1
34	ur	as in	h u r t		1	0	0	0	1	0
35	e	as in	b e t		1	0	0	0	1	1
36	a	as in	a b o u t		1	0	0	1	0	0
37	u	as in	p u t t		1	0	0	1	0	1
38	a	as in	f a t h e r		1	0	0	1	1	0
39	oo	as in	f o o d		1	0	0	1	1	1
40	oo	as in	f o o t		1	0	1	0	0	0
41	oe	as in	t o e		1	0	1	0	0	1
42	aw	as in	l a w		1	0	1	0	1	1

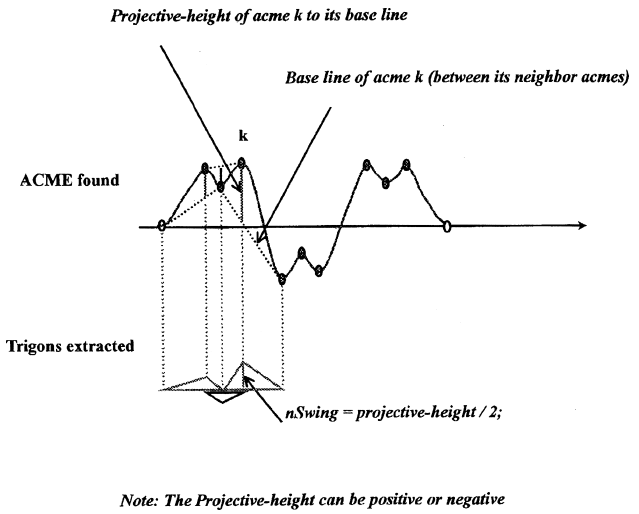
of the phonetic sequences contained in the speech signal. Such a mechanism would allow a person with impaired hearing to understand the spoken item.—DLR

7,251,596

43.72.Ar METHOD AND DEVICE FOR ANALYZING A WAVE SIGNAL AND METHOD AND APPARATUS FOR PITCH DETECTION

Lianshan Zhu and Tao Yu, assignors to Canon Kabushiki Kaisha
31 July 2007 (Class 704/203); filed in China 31 December 2001

The patent describes a speech analysis method based on the detection of sequences of local wave form peaks and valleys, resulting in a series of triangular regions, referred to as "wave trigons." An algorithm is then described for analyzing the patterns of positive and negative trigons to perform



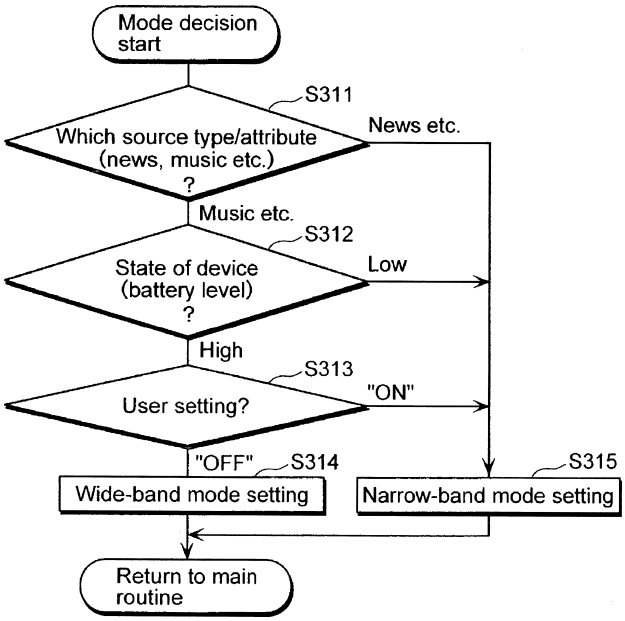
voicing detection, speech pitch analysis, and to detect noise segments and high-frequency regions in the speech signal. The resulting pitch and frequency data are in turn used to detect speech presence, thus marking the beginning and end of sentences in the speech signal.—DLR

7,260,540

43.72.Gy ENCODING DEVICE, DECODING DEVICE, AND SYSTEM THEREOF UTILIZING BAND EXPANSION INFORMATION

Shuji Miyasaka *et al.*, assignors to Matsuhita Electric Industrial Company, Limited
21 August 2007 (Class 704/500); filed in Japan 14 November 2001

In order to extend battery life in portable devices, two encoded bit streams are transmitted and decoded: (1) the encoded sound signal and (2) a bit stream containing information on the amount of band expansion required to achieve pseudo-wide bandwidth via predicting high-frequency content. A selecting unit in the decoder selects either a first or a second sound signal



depending on whether the signal being transmitted is speech or music. If it is music, the decoder facilitates reproduction of the second sound signal that has a wider bandwidth than that of the first sound signal.—DAP

7,177,802

43.72.Gy PITCH CYCLE SEARCH RANGE SETTING APPARATUS AND PITCH CYCLE SEARCH APPARATUS

Kaoru Sato *et al.*, assignors to Matsushita Electric Industrial Company, Limited
13 February 2007 (Class 704/207); filed in Japan 2 August 2001

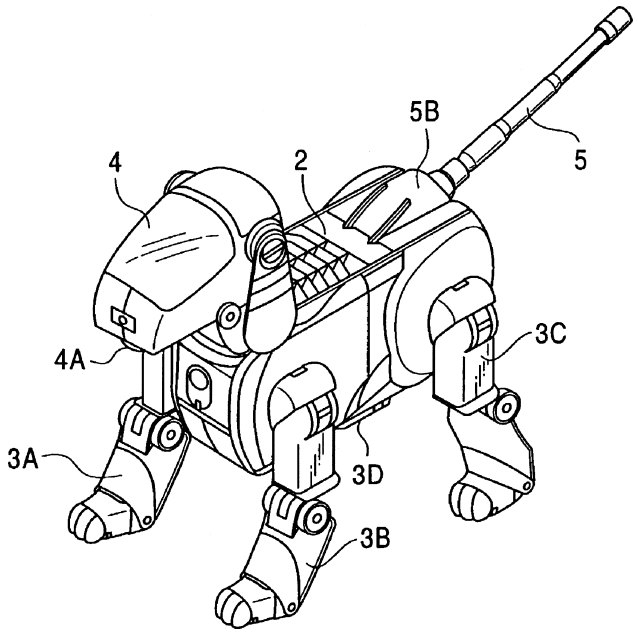
The patent describes a modification to the conventional method of pitch coding as used in any of a variety of speech coding techniques. The question at hand concerns the encoding of the length of a pitch period and the interaction between the pitch value encoded in a particular subframe and the number of full or half periods used to represent the current pitch value. In particular, at low speech pitches, when the pitch cycle is long, conventional methods are said to disallow the use of half periods. A scheme is presented by which an extra bit is obtained to use for coding the nearest half period by juggling the range of the adaptive pitch period codebook, considering the overall speech distortion figure. The patent refers only to a 1985 IEEE publication as reference to the conventional technology and does not mention specific coding standards.—DLR

7,203,642

43.72.Ja ROBOT CONTROL APPARATUS AND METHOD WITH ECHO BACK PROSODY

Kazuo Ishii *et al.*, assignors to Sony Corporation
10 April 2007 (Class 704/231); filed in Japan 11 October 2000

The patent describes a robotic toy in the form of a four-legged animal, such as a dog. The toy includes CCD image sensors as "eyes," microphones as "ears," various body position sensors, surface touch sensors, temperature sensors, body position control motors, and a loudspeaker for generating speech output or various other sounds as might be appropriate. Software described in some detail includes speech recognition as well as touch detection and certain object recognition capabilities. The issue addressed in the



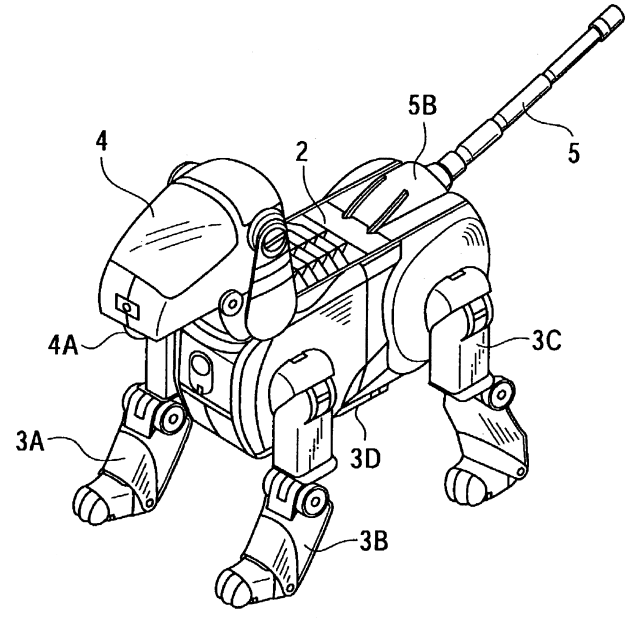
patent is the case where the toy has been spoken to, but no immediate audible or movement response is generated. In this case, the concern is that the user might easily believe the unit is not functioning properly unless there is some immediate response to spoken input. The patented answer to this is that the toy should echo back a modulated tone which echoes the pitch and amplitude patterns of the spoken input. Whenever there is a valid speech or other sound response to the user's speech, that response takes precedence over the modulated tone echo.—DLR

7,222,076

43.72.Ja SPEECH OUTPUT APPARATUS

Erika Kobayashi *et al.*, assignors to Sony Corporation
22 May 2007 (Class 704/275); filed in Japan 22 March 2001

The patent describes a robotic toy in the form of a four-legged animal, such as a dog. Here, the emphasis is on the speech synthesizer controls, specifically regarding a mechanism for interleaving synthesizer outputs originating from competing activities. A case is described in which a prior synthesizer output is under way at the time that the user performs an action which initiates a competing synthesizer output. An example is cited in which



the synthesizer is currently generating the phrase "Where is an exit?" Before that phrase is complete, the user bumps the toy, causing the immediate synthesis of the interjection, "Ouch." The result is an amalgam of the two outputs, something like, "Where is an e—ouch—xit."—DLR

7,236,929

43.72.Kb ECHO SUPPRESSION AND SPEECH DETECTION TECHNIQUES FOR TELEPHONY APPLICATIONS

Richard Hodges, assignor to Plantronics, Incorporated
26 June 2007 (Class 704/233); filed 3 December 2001

Another echo suppression system is described for telephony; this one combining noise estimation, energy detection, hysteresis, trip delay, etc. to suppress that ever annoying self echo.—GFE

7,246,151

43.72.Kb SYSTEM, METHOD AND APPARATUS FOR COMMUNICATING VIA SOUND MESSAGES AND PERSONAL SOUND IDENTIFIERS

Ellen Isaacs and Alan Walendowski, assignors to AT&T Corporation
17 July 2007 (Class 709/206); filed 21 May 2004

Why be limited to visual emoticons when we can communicate via "earcons?" This patent describes a method to send universal sound icons (e.g., yes, no, hungry) to recipient's telephones.—GFE

7,158,934

43.72.Ne SPEECH RECOGNITION WITH FEEDBACK FROM NATURAL LANGUAGE PROCESSING FOR ADAPTATION OF ACOUSTIC MODEL

Hitoshj Honda *et al.*, assignors to Sony Corporation
2 January 2007 (Class 704/244); filed in Japan 30 September 1999

Training a natural-language speech recognizer is not a simple task. In an unsupervised approach, the latest recognition results are used for further training, but the correctness is unknown and results are not good. Supervised methods work better with known correctness, but the user either has to speak a certain amount of material before recognition can proceed or ongoing "trial" recognition results must first be confirmed as correct, then used for training. Either way, an undesirable burden is placed on the user. By including results fed back from an accomplished task, this system minimizes the user burden by making the assumption that if the user chooses to proceed with the task, then the spoken material can be assumed to have been recognized correctly. A system of feedback zones is defined as a part of the system for managing task feedback.—DLR

7,200,559

43.72.Ne SEMANTIC OBJECT SYNCHRONOUS UNDERSTANDING IMPLEMENTED WITH SPEECH APPLICATION LANGUAGE TAGS

Kuansan Wang, assignor to Microsoft Corporation
3 April 2007 (Class 704/257); filed 29 May 2003

This patent describes a speech description language, speech application language tags (SALT), being promoted by the assignee for use in speech recognition and speech understanding systems. SALT is a type of

software application interface with an XML-based structure suitable for understanding as well as recognition. For example, a language model might include context free grammars (CFGs) to describe phrases such as New York, New York City, etc. But functional phrases, such as "Could you show me…" or "Please show me…" do not easily fit within such CFGs. The SALT system would allow *n*-gram descriptions of such functional phrases to be embedded within a CFG-type grammar. Where a speech recognizer might return a word list as a result, a speech understander might return a set of semantic tree elements. Another aspect of the SALT system is a choice of modes of operation which allow more flexibility in setting up the event responses required for use in an event-driven architecture, such as the assignee's Windows™ operating systems.—DLR

7,254,545

43.72.Ne SPEECH CONTROLS FOR USE WITH A SPEECH SYSTEM

Stephen R. Falcon *et al.*, assignors to Microsoft Corporation
7 August 2007 (Class 704/275); filed 2 November 2005

This patent addresses the problem of a user having to know whether to manually or vocally launch a speech-enabled application before speaking desired commands. One or more reusable speech controls are included in a speech-enabled application to provide standardized interactions between the user and the application, thus reducing redundant programming efforts. A question control provides a specific format to request user input. An announcer control is a mechanism for delivering prerecorded verbal feedback to the user. A command control specifies which grammar is to be used for user-initiated speech and communicates back to the application that recognition has occurred. A word trainer control supports recording voice tags.—DAP

7,257,540

43.72.Ne VOICE BROWSER APPARATUS AND VOICE BROWSING METHOD

Fumiaki Ito *et al.*, assignors to Canon Kabushiki Kaisha
14 August 2007 (Class 704/275); filed in Japan 27 April 2000

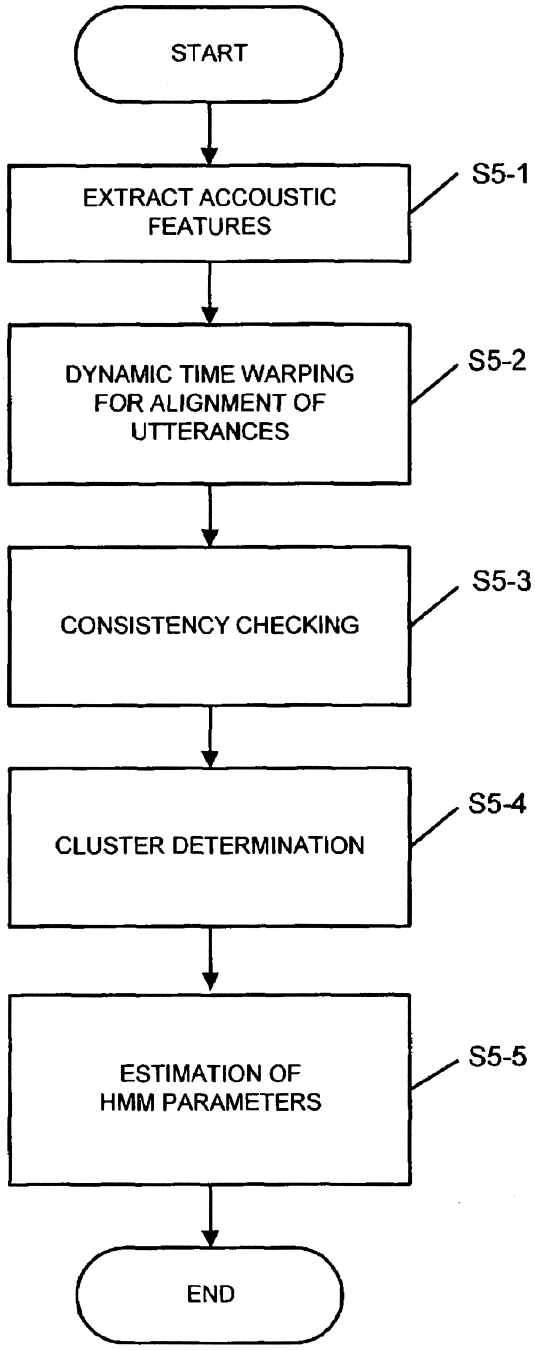
The goal is to provide a voice browser for gaining access by means of voice interaction to Web documents written in a markup language such as HTML. The browser uses several predetermined rules for defining Web output contents and users voice input candidates, and the user or a content creator is allowed to designate which rule is used. Voice output contents and voice input candidates are determined from contents written in HTML according to specific rules. Recognition is performed on user voice input signals and the result is checked for a match against the input candidates.—DAP

7,260,532

43.72.Ne HIDDEN MARKOV MODEL GENERATION APPARATUS AND METHOD WITH SELECTION OF NUMBER OF STATES

David Llewellyn Rees, assignor to Canon Kabushiki Kaisha
21 August 2007 (Class 704/256); filed in United Kingdom 26 February 2002

A speech model generator produces multiple-state hidden Markov models (HMMs) to represent received speech signals. A signal processor determines a sequence of feature vectors indicative of the received speech signal. The feature vectors are variably clustered into several groups and the



goodness of fit between the clusters of vectors and the HMM is determined. A selection unit then determines the optimal number of states in the HMM, each associated with a probability density function.—DAP

7,248,702

43.75.Tv SOUND ENHANCEMENT SYSTEM

Thomas Nelson Packard, Syracuse, New York
24 July 2007 (Class 381/61); filed 6 January 2003

A system to add overtones and transient attack sounds is described that consists of a square root filter. The embodiment is designed to create a more realistic aural experience for music listeners.—GFE

7,251,352

43.80.Vj MARKING 3D LOCATIONS FROM ULTRASOUND IMAGES

Frank Sauer *et al.*, assignors to Siemens Corporate Research, Incorporated
31 July 2007 (Class 382/128); filed 16 August 2002

The position of a target is located using an ultrasound imaging system by tracking the orientation of the transducer with respect to an external three-dimensional coordinate system, obtaining two-dimensional ultrasound images from the transducer, marking a desired target on the ultrasound image, and calculating the three-dimensional position of the marker using the tracking data.—RCW

7,255,678

43.80.Vj HIGH FREQUENCY, HIGH FRAME-RATE ULTRASOUND IMAGING SYSTEM

James I. Mehi *et al.*, assignors to VisualSonics Incorporated
14 August 2007 (Class 600/446); filed 10 October 2003

The ultrasound center frequency of this system is 20 MHz or higher and the transducer is mechanically scanned to obtain images at a rate of 15 frames per second or greater.—RCW

7,258,674

43.80.Vj ULTRASONIC TREATMENT AND IMAGING OF ADIPOSE TISSUE

Robert Cribbs *et al.*, assignors to LipoSonix, Incorporated
21 August 2007 (Class 601/2); filed 20 February 2003

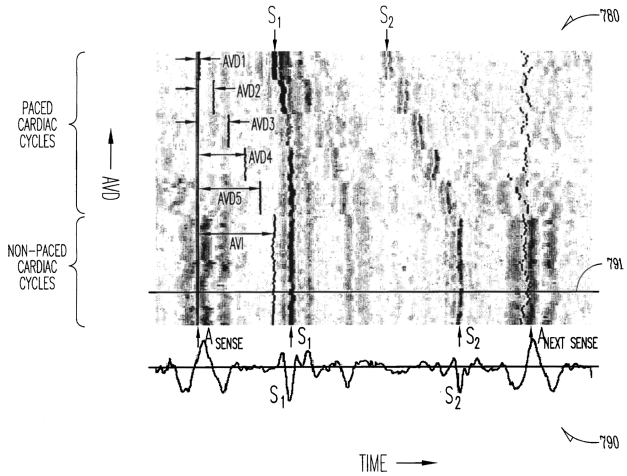
High-intensity focused ultrasound (HIFU) is used to destroy adipose tissue. A means is included to map the three-dimensional position of the adipose tissue. A controller identifies the adipose tissue and establishes a protocol for tissue destruction. A sensor coupled to the controller provides feedback to facilitate safe operation of the HIFU transducer.—RCW

7,260,429

43.80.Vj METHOD AND APPARATUS FOR PHONOCARDIOGRAPHIC IMAGE ACQUISITION AND PRESENTATION

Krzysztof Z. Siejko *et al.*, assignors to Cardiac Pacemakers, Incorporated
21 August 2007 (Class 600/514); filed 2 December 2002

A record of acoustic signals is presented in a two-dimensional format with cardiac cycle as the vertical axis and time as the horizontal axis. Each signal segment is a sound indicative of mechanical and electrical heart



events. The signals are aligned using a selected mechanical or electrical event in the heart.—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.

On fiber optic probe hydrophone measurements in a cavitating liquid (L)

Aaldert Zijlstra
Faculty of Science, Physics of Fluids, University of Twente, 7500 AE Enschede, The Netherlands

Claus Dieter Ohl[a)]
Faculty of Science, Physics of Fluids, University of Twente, 7500 AE Enschede, The Netherlands and Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

(Received 4 July 2007; revised 17 October 2007; accepted 1 November 2007)

The measurement of high-pressure signals is often hampered by cavitation activity. The usage of a fiber optic probe hydrophone possesses advantages over other hydrophones, yet when measuring in a cavitating liquid large variations in the signal amplitude are found; in particular when the pressure signal recovers back to positive values. With shadowgraphy the wave propagation and cavity dynamics are imaged and the important contributions of secondary shock waves emitted from collapsing cavitation bubbles are revealed. Interestingly, just adding a small amount of acidic acid reduces the cavitation activity to a large extent. With this treatment an altered primary pressure profile which does not force the cavitation bubbles close to fiber tip into collapse has been found. Thereby, the shot-to-shot variations are greatly reduced.
 [DOI: 10.1121/1.2816578]

PACS number(s): 43.35.Ei, 43.35.Yb [AJS] Pages: 29–32

I. INTRODUCTION

Accurately registering the wave shapes in medical applications such as in shock wave lithotripsy,[1,2] shock wave therapy,[3] histotripsy,[4] or high intensity focused ultrasound[5] is of prime importance for quality assurance of therapeutical devices. Pressure measurements of high-pressure finite amplitude and shock waves are not only demanding because of the high frequencies involved. The recording of negative pressures is also challenging because the sensor has to withstand cavitation.[6] In general, cavitation can occur when the pressure drops below the vapor pressure while nuclei[7] are present which explode into vaporous cavities. When the pressure recovers again the cavities implode thereby focusing "destructive" energy from the liquid onto very small scales. When the sensor is too close to the collapsing cavity it can easily be damaged. Yet, even if the sensor withstands the cavity collapse, measurements of the pressure are often hampered because of the need to distinguish whether the sensor is entrained within a cavity or accurately registering the liquid.

A device which operates in this demanding environment is the fiber optic probe hydrophone developed by Staudenrauss and Eisenmenger (1992).[8] In this device laser light is coupled into a glass fiber. At the fiber tip the light is reflected; the intensity of the reflected light is a function of the jump in the index of refraction from glass to water. The index of refraction is related through the well known Gladstone–Dale relationship to the pressure.[9] Thus, the pressure can be determined from the intensity reflected back into the fiber[10] and registered with a sensitive photodetector. This type of hydrophone is mentioned in the IEC guidelines for quality assurance of lithotripters to measure at the focus[11] because it has several unique advantages compared to the other types of hydrophones: (i) The strong adhesion of water on the glass reduces nucleation of cavitaties on the sensor, (ii) when the fiber tip is broken the fiber can easily be recleaved and cut, (iii) the entrainment of the fiber into a cavity leads to an instantaneous jump in the signal thus entrainment of the sensor is easily detectable, (iv) and, due to the well documented Gladstone–Dale relationship, calibration is a simple task.

Yet, there exist some difficulties with interpreting the signals. When studying waves which are trailed by a negative pressure phase very strong shot-to-shot variations are observed close to the moment when the negative pressure

[a)]Author to whom correspondence should be addressed. Electronic mail: cdohl@ntu.edu.sg

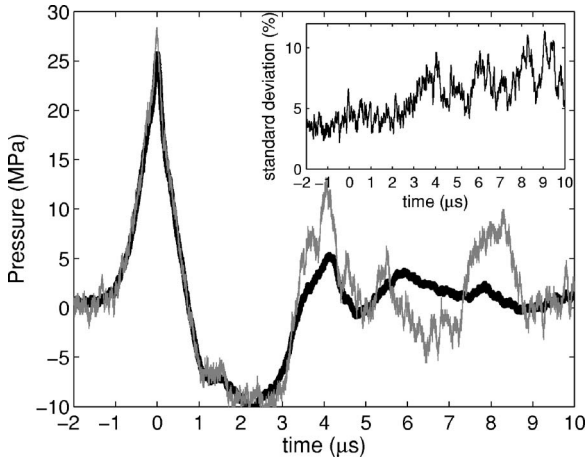


FIG. 1. Pressure signals recorded with a fiber optic probe hydrophone (FOPH) in degassed Millipore water. An averaged signal (thick line) and a randomly selected single recording (thin line) are compared. The inset depicts the standard deviation (SD) of the signal as a function of time. The scale is given in percentage of the average peak pressure. Interestingly the SD increases at time $t=3.2$ μs, that is, when the pressure recovers.

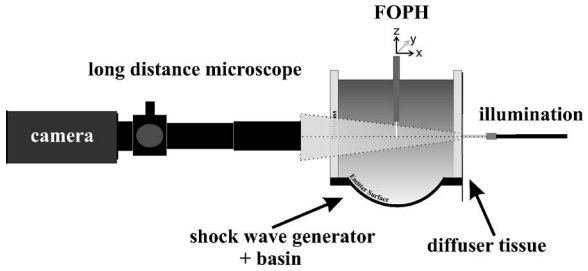


FIG. 2. Experimental setup used both for pressure measurements and for shadowgraphy.

recovers. Averaging the signal over multiple events removes these oscillations but leaves the experimenter with some discomfort on how to interpret the data. An example of the signal which is achieved by this averaging procedure is depicted in Fig. 1 (thick line). The measurements were done in partially degassed Millipore (Milli-Q synthesis A10) water.

The contribution of cavitation on the waveform has been reported by Pishchalnikov *et al.*[12] They find a shortening of the tensile wave and explain it with the loss of energy from the tensile wave due to the growth of cavitation bubbles in the liquid. In recent simulations by Liebler *et al.*[13] the nonlinear wave equation model was coupled with an effective medium which described the gas phase. They revealed that the waveform can be greatly altered: not only the tensile phase is shortened but also a second pressure increase following the shortened tensile phase is found in the simulations and backed convincingly with experiments. The altered waveform, which has also been reported by Arora *et al.*,[14] is explained by Liebler *et al.*[13] by the presence of cavitation bubbles disturbing the focusing of the diffracted waves from the transducer edge.

The averaged waveform over 18 pressure signals presented in Fig. 1 is compared with a randomly chosen recording from this set (thin line in Fig. 1). The variations between the single and the averaged signal remain small during the initial pressure rise to 26 MPa and drop to −10 MPa. This initial variability can be explained with the noise of the laser source and measurement noise of the photodetector. However, at time $t=3$ μs—that is, when the pressure recovers—clear differences appear. They are detailed in the inset of Fig. 1 by plotting the standard deviation of the signal.

What causes the loss of reproducibility in the data? Candidates are cavitation bubbles emitting pressure waves, yet we cannot exclude cavitation occurring on the glass fiber which might affect the light transmission in the glass fiber and has been suggested by Pishchalnikov and colleagues.[12] We are now looking for a way to decrease cavitation activity only: In the work of Liebler *et al.*,[13] it was reported that the pressure signal is affected by adding small amounts of a mild acid (acetic acid). Their explanation was: cavitation is reduced by chemically dissolving calcite particles which serve as cavitation nuclei in the water, a hypothesis put forward by Eisenmenger and Pecha.[15]

High-speed photography of the interaction of the wave with a fiber optic sensor is used to document the effect of cavitation and its reduction on the glass fiber sensor. Next, we describe the experimental setup which allows to conduct this task and visualize the interaction of the pressure wave with the glass fiber tip. The pressure waves are generated with a piezo-electric device used for shock wave therapy (Piezolith 100, Richard Wolf GmbH, Knittlingen, Germany).[16] The pressure measurements are done with the fiber optic hydrophone (FOPH 500, RP Acoustics, Stuttgart, Germany) which is positioned in the focus with a motorized three axis translation stage. The shock wave and the cavitation dynamics are visualized with a shadowgraphy technique using stroboscopic illumination (Fig. 2). The image of the shock wave and the cavitation bubbles are illuminated with a frequency doubled Nd:yttrium–aluminum–garnet laser pulse of 7 ns duration (Solo PIV, New Wave Research, wavelength 532 nm) fed into a fluorescent cell filled with an ethanol-dye mixture (0.417 mg/ml, LDS 698, Exciton Inc., Dayton, U.S.) and then coupled into a glass fiber to the shock wave generator located on a second table. The fluorescent light allows for speckle free illumination. The light escapes the other fiber end and then becomes slightly diffused with a tissue paper. The images are taken with a charge coupled device camera (Imager 3S, LaVision, Goettingen, Germany) which is equipped with a long distance microscope (K2, Infinity, U.S.) and a CF3 objective. The optical resolution is 1.52 μm/pixel. The timing unit (BNC 555, Berkeley Nucleonics, CA), which is triggered through an induction coil connected to the shock wave generator, fires the laser at variable delays with respect to the shock wave. In contrast to prior high-speed framing studies on shock wave inception of cavitation, for example, by Ohl,[7] we emphasized in this work on the quality of the pictures to resolve the shock waves and freeze the bubbles in stop motion. This was achieved with a single frame technique, e.g., the photographs are from different experimental runs with the time of camera exposure shifted. The travel time of the wave from the transducer to the acoustic focus where the fiber optic probe hydrophone is located is approximately 50 μs.

A stroboscopic picture sequence from the plain water experiment is shown in Fig. 3 and they are conducted in the

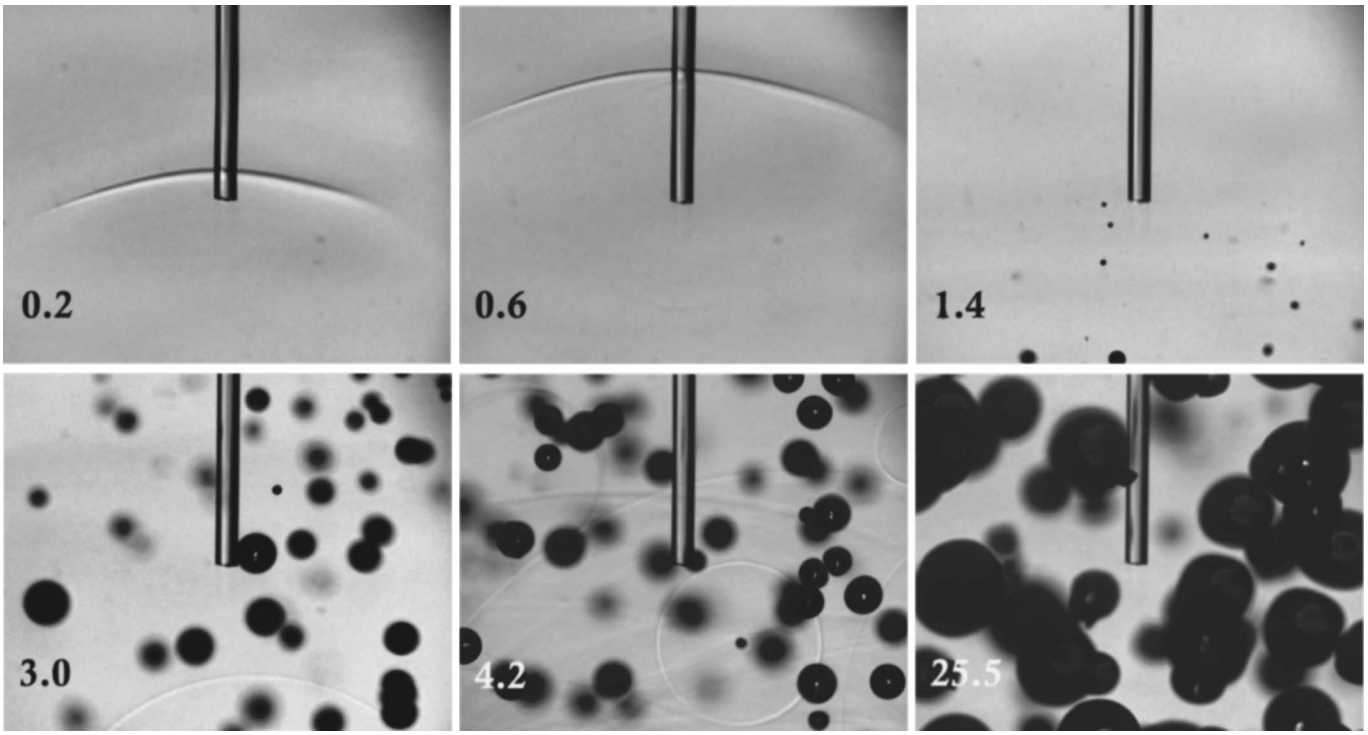


FIG. 3. Flash photography of the shock wave passage and following cavitation activity at the tip of the FOPH. The time in the individual frames is given relative to the shock wave impact on the the glass fiber tip and is given in microseconds. Please note, the spherical shock waves at times 3.0 and 4.2 μs. The image size is 2.0×1.6 mm^2.

same setup as the recording of the pressure signals shown in Fig. 1. The first frame displays the wave shortly after it passed over the glass fiber tip. The time $t=0$ is defined when it just meets with the glass fiber tip. Shadowgraphic imaging technique is sensible to the second derivative of the index of refraction.[17] Thus, the structure depicted in stop motion in the first two frames of Fig. 3 is the image of the pressure maximum of the wave. At $t=1.4$ μs cavitation bubbles nucleate in the liquid. Spherical pressure waves appear shortly after and reach about the position of the fiber tip in the fourth frame, $t=3.0$ μs. This correlates with the start of the signal deviations in the pressure recordings. The spherical waves are created from collapsing cavitation bubbles. Presumably, the second pressure increase following the tensile phase causes the shrinkage of the bubbles. The origin of the pressure increase is likely to be the diffraction effect reported in Liebler *et al.*[13]

Let us now have a look at the cavity dynamics after the acetic acid (4 vol %) has been added to the water. This concentration was chosen because Liebler *et al.*[13] observed excellent agreement of a bubble-free simulation and fiber optic probe hydrophone (FOPH) measurements. It was suggested that cavitation activity is suppressed and the wave is not altered by the interaction with bubbles. Figure 4 depicts the shadowgraphy sequence in the solution with acetic acid. In the first two frames the primary wave and also the reflected wave from the tip of the fiber are visualized. The initial pressure wave form is hardly altered as compared to the water case. The second frame shows on an enlarged scale in detail the shock reflection and the propagation of bending waves along the glass fiber. The first cavities appear at $t=1.4$ μs but only two, a tenth of the number of bubbles from the plain water case. A second difference is that at later times we do not find spherical waves emitted from collapsing cavitation bubbles. This is interesting, because a few bubbles are nucleated in the field of view and thus might qualify for a forced collapse. The addition of acid to the liquid causes a strong reduction on the number of bubbles. But it also causes a second effect, it changes the time averaged pressure recordings. We find in the plain water case a second overpressure peak after the pressure recovered from the negative pressure, $t=4.1$ μs in Fig. 1 which is absent here, very similar to the finding of Liebler *et al.*[13] Presumably, the second overpressure is a nonlinear effect of the diffraction wave propagation in the bubbly/compressible liquid. By adding acid much less

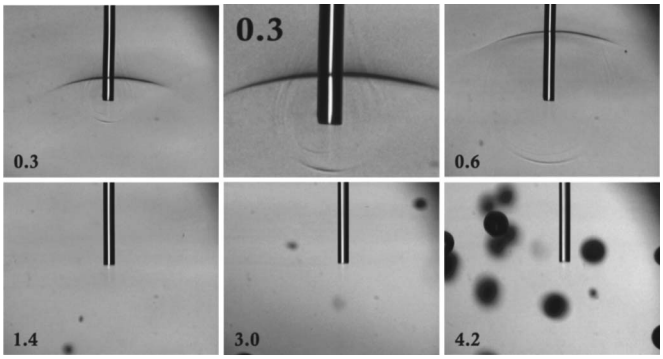


FIG. 4. Flash photography of the shock wave passage and following cavitation activity at the tip of the FOPH in a water-acid solution (4 vol % acetic acid). The time in the individual frames is given relative to the shock wave impact on the the glass fiber tip. The second frame is a magnified view of the shock wave reflection from the tip at $t=0.3$ μs. Please note also the bending waves traveling along the glass fiber. The image size is 2.0×1.6 mm^2 and the enlarged image 0.9×0.7 mm^2.

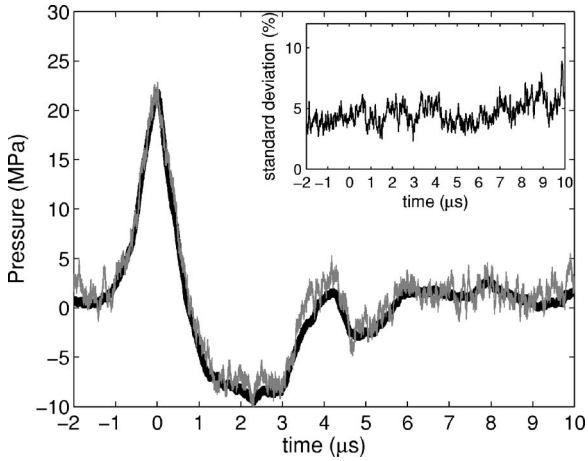


FIG. 5. Pressure signals recorded with a FOPH device in degassed Millipore water with 4 vol % acetic acid. An averaged signal (thick line) and a randomly selected single recording (thin line) are compared. The inset depicts the standard deviation (SD) of the signal as a function of time. Here, the SD stays constant for the measured time. The positive peak pressure recorded with FOPH is slightly less than shown in Fig. 1 which we explain with some spatial shift of the fiber tip of less than 100 μm during the addition of acid and stirring of the liquid.

bubbles are created and the diffraction wave becomes unaffected. Therefore, the remaining cavities are not compressed and do not cause secondary shock waves, which then in return leads to a more reproducible signal. This finding is strengthened by the plot of the standard deviation in Fig. 5. It stays constant over the measured time.

Our finding is that the pressure signal obtained in a cavitation liquid consists of a superposition of the pressure induced by the direct wave and from secondary waves emitted by collapsing bubbles. We extend the results from Liebler *et al.*[13] that not only the waveform is modified by its passage through the bubbly liquid. Additionally, the forced collapse and the thereby emitted secondary pressure and shock waves are picked up by a fiber optic probe hydrophone. The spatial randomness of the cavitation events leads to the noise-like signal registered when the pressure recovers from the tensile phase.

ACKNOWLEDGMENT

This work has been funded through the VIDI Grant from NWO (The Netherlands).

[1]M. R. Bailey, V. A. Khokhlova, O. A. Sapozhnikov, S. G. Kargl, and L. A. Crum, "Physical mechanisms of the therapeutical effect of ultrasound (A review)," Acoust. Phys. **49**, 369–388 (2003).
[2]O. A. Sapoznhnikov, A. D. Maxwell, B. MacConaghy, and M. R. Bailey, "A mechanistic analysis of stone fracture in lithotripsy," J. Acoust. Soc. Am. **121**, 1190–1202 (2007).
[3]S. Warden, "A new direction for ultrasound therapy in sports medicine," Sports Med. **33**, 95107 (2003).
[4]Z. Xu, A. Ludomirsky, L. Y. Eun, T. L. Hall, B. C. Tran, J. B. Fowlkes, and C. A. Cain, "Controlled ultrasound tissue erosion," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 726–736 (2004).
[5]J. E. Kennedy, G. R. ter Haar, and D. Cranston, "High intensity focused ultrasound: Surgery of the future?," Br. J. Radiol. **76**, 590–599 (2003).
[6]C. E. Brennen, *Cavitation and Bubble Dynamics* (Oxford University Press, Oxford, 1995).
[7]C. D. Ohl, "Cavitation inception following shock wave passage," Phys. Fluids **14**, 3512–3521 (2002).
[8]J. Staudenraus and W. Eisenmenger, "Fibre-optic probe hydrophone for ultrasonic and shock-wave measurements in water," Ultrasonics **31**, 267–272 (1993).
[9]J. E. Parsons, C. A. Cain, and J. B. Fowlkes, "Cost-effective assembly of a basic fiber-optic hydrophone for measurement of high-amplitude therapeutic ultrasound fields," J. Acoust. Soc. Am. **119**, 1432–1440 (2006).
[10]J. Krücker, A. Eisenberg, M. Krix, R. Lötsch, M. Pessel, and H.-G. Trier, "Rigid piston approximation for computing the transfer function of an angular response of a fiber-optic hydrophone," J. Acoust. Soc. Am. **107**, 1994–2003 (2000).
[11]International Electrotechnical Committee, "Ultrasonics—pressure pulse lithotripters—characteristics of fields," IEC Standard 61846 (1998).
[12]Y. A. Pishchalnikov, O. A. Sapozhnikov, M. R. Bailey, I. V. Pishchalnikova, J. C. Williams Jr., and J. A. McAteer, "Cavitation selectively reduces the negative pressure phase of lithotripter shock pulses," Acoust. Res. Lett. Online **6**, 280–285 (2005).
[13]M. Liebler, T. Dreyer, and R. E. Riedlinger, "Nonlinear modeling of interactions between ultrasound and cavitation bubbles," Acta Acust. **1**, 165–167 (2006).
[14]M. Arora, C. D. Ohl, and D. Lohse, "Effect of nuclei concentration on cavitation cluster dynamics," J. Acoust. Soc. Am. **121**, 3432–3436 (2007).
[15]W. Eisenmenger and R. Pecha, "Eine neue Art von Kavitationskeimen," engl. "New Species of Cavitation Nuclei," In *Fortschritte der Akustik, DAGA'03, Deutsche Gesellschaft für Akustik e.V.*, 842–843 (2003).
[16]N. Bremond, M. Arora, C. D. Ohl, and D. Lohse, "Controlled multibubble surface cavitation," Phys. Rev. Lett. **96**, 224501 (2006).
[17]G. S. Settles, *Schlieren and Shadowgraph Techniques* (Springer-Verlag, Berlin, 2001).

Power-output regularization in global sound equalization (L)

Nick Stefanakis,[a] John Sarris, and George Cambourakis
School of Electrical and Computer Engineering, National Technical University of Athens, Heroon Polytechniou 9, 157 73, Athens, Greece

Finn Jacobsen
Acoustic Technology, Ørsted DTU, Technical University of Denmark, Ørsted Plads, Building 352, DK-2800 Kgs, Lyngby, Denmark

(Received 24 June 2007; revised 31 October 2007; accepted 1 November 2007)

The purpose of equalization in room acoustics is to compensate for the undesired modification that an enclosure introduces to signals such as audio or speech. In this work, equalization in a large part of the volume of a room is addressed. The multiple point method is employed with an acoustic power-output penalty term instead of the traditional quadratic source effort penalty term. Simulation results demonstrate that this technique gives a smoother decline of the reproduction performance away from the control points. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816580]

PACS number(s): 43.55.Br, 43.38.Md, 43.60.Pt [NX] Pages: 33–36

I. INTRODUCTION

Traditionally, equalization in room acoustics uses digital filters to pre-process the input signal before it is fed to a set of loudspeakers so that the spectral coloration and the reverberation tail associated with the transmission path are removed.[1,2] Recently, techniques that allow the zone of equalization to be extended to a much larger region inside the room have been proposed.[3,4] In these techniques a plane propagating wave was generated in a rectangular enclosure in a region that occupied almost the complete volume of the room.

This paper is also concerned with equalization in a large part of the volume of a rectangular room. The process is studied below the Schroeder frequency where the modal density and the modal overlap is low and the sound field is dominated by discrete modes.[5] The method is based on the conventional multiple point technique which minimizes a cost function that expresses the difference between the desired complex sound pressure and the sound pressure that is actually reproduced at a small number of sampling points in the room.[2,3] Instead of the quadratic effort penalty term used in traditional regularization, a sound power-output penalty term is introduced. The main advantage of this new technique is shown to be a smoother decline of the reproduction performance away from the control points.

II. SOURCE POWER OUTPUT IN THE REPRODUCTION OF SOUND FIELDS IN ROOMS

One possible principle of sound field reproduction is based on the fact that given a spatial volume any sound field can be reproduced perfectly in both space and time, given a complete description of the acoustic pressure and pressure gradient on the hypothetical surface that bounds the spatial volume. This principle is mathematically expressed by the Kirchhoff–Helmholtz integral equation.[6] For sound field reproduction purposes, a continuous layer of dipole and monopole sources should be assigned the values of the sound pressure and the pressure gradient that correspond to the sound field. If the sources that generate the original sound field are outside the volume then as much sound energy flows into the volume as out of it. It follows that reproduction of a sound field generated by sources outside the volume implies that the total sound power output of the source layer should be zero. This suggests that the monopole and dipole strengths should be adjusted so that one part of the source layer absorbs the sound power that is emitted from another part.

A well-known physical implementation of the Kirchhoff–Helmholtz integral equation is wave field synthesis (WFS).[7] Classical WFS assumes that the reproduction environment is anechoic and therefore does not perform well in a real reproduction environment.[8] Nevertheless, the mechanisms of power output minimization and power absorption[9] between the reproduction sources still hold for sound field reproduction in closed spaces, as pointed out by Gauthier *et al.*, who observed that sound absorption becomes important in reverberant enclosures when using optimal control techniques.[8] It was also observed that, at frequencies where the modal density is low, optimal control acts to create a sound energy flow over the sensor array and mainly prevents a pure standing wave pattern. This relation between spatial sound field reproduction and the suppression of the standing wave pattern that is likely to occur is also related to Santillán's work, where the reproduction of a plane wave serves as a solution for global equalization in a rectangular room.[3] These observations suggest that when spatial sound field reproduction is desired, the total sound power emitted from the reproduction sources should avoid the peaks that are likely to occur near the natural frequencies of the room. The addition of a power-output penalty term in the cost function of the multiple point method would therefore seem to be a promising strategy for spatial sound field reproduction.

[a]Author to whom correspondence should be addressed. Electronic mail: nstefan@mobile.ntua.gr

III. CONTROL MODEL

Suppose that it is desired to control the sound field in a spatial region inside an enclosure that is surrounded by L reproduction sources. The pressure in this spatial region is sampled by M monitor sensors placed at $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M\}$, and they provide a measure of the performance of reproduction in the entire listening space. The pressure at the monitoring sensors subject to the L source excitations can be written as[2]

$$\mathbf{p} = \mathbf{Z}^{(m)}\mathbf{q}, \tag{1}$$

where $\mathbf{p}$ is a column vector with the M complex sound pressures at the monitor sensors [Pa], $\mathbf{q}$ is a column vector with the complex strengths of the L sources [m^3/s], and $\mathbf{Z}^{(m)}$ is an M by L matrix with $Z_{ml}^{(m)}$ being the acoustic transfer function from the lth source to the mth field point at $\mathbf{r}_m$. Assume that the reproduction system includes N control sensors placed at discrete points in the room, $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$. It is assumed here that $N < M$, corresponding to a limited number of control sensors, covering only a small part of the listening space. The need for a compact control sensor array is based on the observation that a great number of control sensors that occupy the entire listening area not only would increase the computational cost but also interfere with the listeners inside the room. The system is informed about the performance of the reproduction in the controlled region by the difference between the target pressure and the reproduced pressure field

$$\mathbf{e} = \mathbf{p}_d - \mathbf{Z}^{(c)}\mathbf{q}, \tag{2}$$

where $\mathbf{p}_d$ is the vector with the desired sound pressures at the N control sensors, and $\mathbf{Z}^{(c)}$ is the transfer matrix with the transfer functions from the L sources to the N control sensors.

The proposed control approach suggests the use of a cost function defined as

$$J^{(\lambda)} = \mathbf{e}^H\mathbf{e} + \lambda\mathbf{q}^H\mathbf{W}\mathbf{q}, \tag{3}$$

which should be minimized. Here λ is a real positive scalar that weights the contribution of the penalty term in the cost function, the quantity $\mathbf{q}^H\mathbf{W}\mathbf{q}$ expresses the total sound power emitted by the reproduction sources [W], and $\mathbf{W}$ is a symmetric and positive definite matrix with W_{ij} representing the real part of the transfer function from the ith to the jth source.[9] For distributed sources, each element of the matrix $\mathbf{W}$ is calculated with proper integration of the transfer function on the surface of each source. Equation (3) implies the addition of an λ-weighted sound power-output penalty term instead of the source effort penalty term used in standard regularization,[10]

$$J^{(\mu)} = \mathbf{e}^H\mathbf{e} + \mu\mathbf{q}^H\mathbf{q}. \tag{4}$$

Substitution of Eq. (2) into Eq. (3) yields

$$J^{(\lambda)} = \mathbf{q}^H(\mathbf{Z}^{(c)H}\mathbf{Z}^{(c)} + \lambda\mathbf{W})\mathbf{q} - \mathbf{q}^H\mathbf{Z}^{(c)H}\mathbf{p}_d - \mathbf{p}_d^H\mathbf{Z}^{(c)}\mathbf{q} + \mathbf{p}_d^H\mathbf{p}_d, \tag{5}$$

which is a quadratic function of $\mathbf{q}$. Under the condition that $\lambda\mathbf{W} + \mathbf{Z}^{(c)H}\mathbf{Z}^{(c)}$ is also positive definite, the optimal vector that minimizes $J^{(\lambda)}$ can be found by

$$\mathbf{q}^{(\lambda)} = (\lambda\mathbf{W} + \mathbf{Z}^{(c)H}\mathbf{Z}^{(c)})^{-1}\mathbf{Z}^{(c)H}\mathbf{p}_d. \tag{6}$$

The optimum source strengths derived here should be compared with those obtained by standard regularization,[10]

$$\mathbf{q}^{(\mu)} = (\mu\mathbf{I} + \mathbf{Z}^{(c)H}\mathbf{Z}^{(c)})^{-1}\mathbf{Z}^{(c)H}\mathbf{p}_d. \tag{7}$$

It can be seen that the identity matrix has been replaced by the fully populated matrix $\mathbf{W}$. The achieved quality of the reproduction of each of the two optimum source strengths is measured over the entire listening space with the use of the M monitor sensors. Similar to Eq. (2) the error at the monitor sensors is measured as

$$\mathbf{e}^{(m)} = \mathbf{p}_d^{(m)} - \mathbf{Z}^{(m)}\mathbf{q}, \tag{8}$$

where $\mathbf{p}_d^{(m)}$ is now the vector with the desired complex pressures at the monitor sensors. The quality of the performance is quantified over the entire listening space at the monitor sensors by the global reproduction errors, defined as

$$E_{\mathrm{LS}}^{(\lambda)} = \left(\frac{(\mathbf{p}_d^{(m)} - \mathbf{Z}^{(m)}\mathbf{q}^{(\lambda)})^H(\mathbf{p}_d^{(m)} - \mathbf{Z}^{(m)}\mathbf{q}^{(\lambda)})}{\mathbf{p}_d^{(m)H}\mathbf{p}_d^{(m)}} \right)^{1/2} \tag{9}$$

and

$$E_{\mathrm{LS}}^{(\mu)} = \left(\frac{(\mathbf{p}_d^{(m)} - \mathbf{Z}^{(m)}\mathbf{q}^{(\mu)})^H(\mathbf{p}_d^{(m)} - \mathbf{Z}^{(m)}\mathbf{q}^{(\mu)})}{\mathbf{p}_d^{(m)H}\mathbf{p}_d^{(m)}} \right)^{1/2} \tag{10}$$

for power-output penalty and effort penalty regularization.

IV. SIMULATION RESULTS

In the simulations presented in what follows, the conventional modal sum of the sound field in a lightly damped rectangular enclosure proposed by Morse[11] is used in the form described by Bullmore *et al.*[12] The sources are modeled as square pistons that vibrate with the normal velocity $u_l = q_l/A$, where $A = a^2$ is the area of the piston sources and a is their side length. The piston sources are assumed parallel to the xz plane.

A shallow rectangular room is modeled with dimensions $L_x = 2$ m, $L_y = 3.2$ m, and $L_z = 0.2$ m as shown in Fig. 1. Five square piston sources with side length equal to 0.1 m are placed along each of the walls at $x = 0$ and $x = L_x$, and 16 control sensors centered in the room are used to optimize the source strengths. All the modes up to 1100 Hz are used to model the sound field inside the room, and the damping factor is set equal to 0.03 for all the modes. An array of 405 monitor sensors is spread in the room, covering the dashed rectangle in Fig. 1, with lower left and upper right corner at (0.3, 0.3) m and (1.7, 2.9) m, respectively. The distance between the monitor sensors is 0.1 m in both the x and y direction, while a distance of 0.3 m is preserved between the outer monitor sensors and the closest vertical wall in the room.

Following Santillán's approach, equalization in the entire volume of the room is obtained by the generation of a plane wave traveling in the y direction.[3] This requires that only the desired y-axial modes are significantly excited in the room. The reproduction error for this type of sound field at 300 Hz is illustrated as a function of the penalization param-

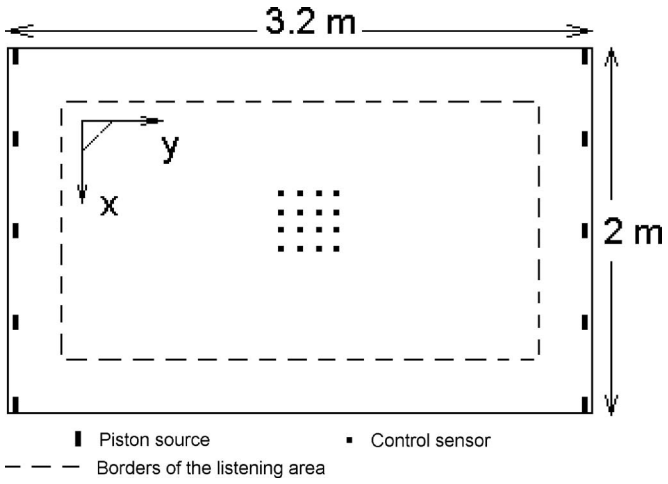


FIG. 1. Schematic diagram of the room for the two-dimensional problem. The sensor plane is at $z=0.1$ m while the x coordinates of the centers of the sources are at 0.05, 0.5, 1, 1.5 and 1.95 m. The y coordinates of their centers are at 0.05 m for the left sources and at 3.15 m for the right sources.

eters λ and μ at the control sensors and at the monitor sensors in Fig. 2. As can be seen in Fig. 2(a), the residual error as calculated at the control sensors is an increasing function of both λ and μ. Nevertheless, as seen in Fig. 2(b), the global reproduction error is minimized with both techniques for nontrivial values of the penalization parameters. It can also be seen that the two techniques give very similar curves except that the effort penalty curve is shifted to the right since μ must be larger in order to achieve a balance between $\mathbf{e}^H\mathbf{e}$ and the effort penalty term similar to the balance between $\mathbf{e}^H\mathbf{e}$ and the power penalty term in the two cost functions. The minimum errors are 0.263 and 0.395 for $E_{\mathrm{LS}}^{(\lambda)}$ and $E_{\mathrm{LS}}^{(\mu)}$, respectively, corresponding to the optimum regularization parameters of $\lambda=16\times0.5$ and $\mu=16\times1000$, as they are chosen from a range of values of the form $N\times[5\times10^{-4},10^{-3},5\times10^{-3},10^{-2},\ldots,5\times10^{4}]$, where N is the

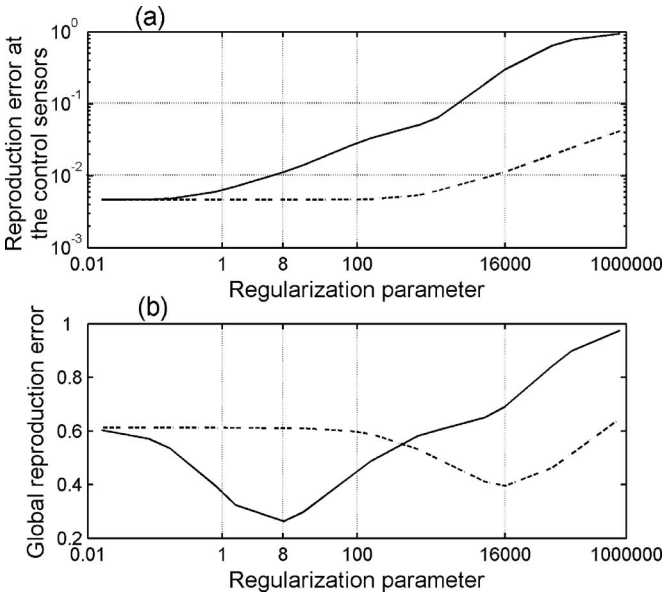


FIG. 2. Variation of the reproduction error as a function of the penalty parameter λ (solid line) and μ (dashed line) calculated (a) at the monitor sensors and (b) at the control sensors.

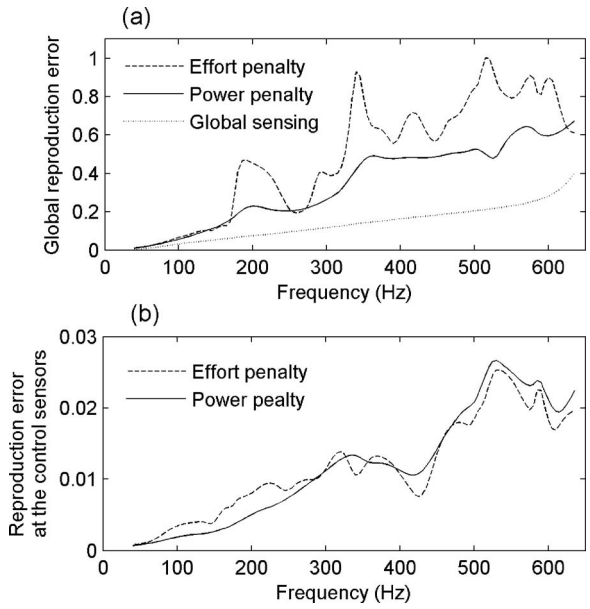


FIG. 3. Reproduction error as a function of the frequency (a) at the monitor sensors and (b) at the control sensors. The penalization parameter is $\lambda=8$ for power penalty regularization and $\mu=16\,000$ for effort penalty regularization. In (a) is also shown the global reproduction error for global sensing.

number of the control sensors. Another interesting point is that even if the power penalty with optimum λ gives a global reproduction error lower than the one obtained with the effort penalty with optimum μ, they both give similar reproduction errors at the control sensors (about 0.01 in Fig. 2(a)).

The global reproduction errors for the two strategies are plotted in the frequency range of concern in Fig. 3(a) where $\lambda=16\times0.5$ and $\mu=16\times1000$. In the same figure the global reproduction error is also shown for the case where all the monitor sensors are used to optimize the source strengths

$$\mathbf{q}_m=(\mathbf{Z}^{(m)H}\mathbf{Z}^{(m)})^{-1}\mathbf{Z}^{(m)H}\mathbf{p}_m^{(d)}. \tag{11}$$

This corresponds to the best that can be done in the rather unrealistic case where a control sensor plane covering the entire listening space is used. Such a great number of control sensors are used in order to provide an independent measure of the optimal performance of the system but would, of course, be impractical in any real problem. The term *global sensing* will be used for this type of optimization. Among the E_{LS} values of the two regularization methods in Fig. 3(a) it can be seen that the proposed strategy clearly outperforms traditional regularization, which is characterized by strong peaks at distinct frequencies. In Fig. 3(b), the reproduction error as a function of the frequency is plotted at the control sensors. It can be seen that the quality of the reproduction result at the control sensor locations is very similar for both techniques. This leads to the conclusion that the improvement in the global performance for the proposed technique is caused by better reproduction results outside the region of control.

The best way of reproducing a progressive wave field in the control volume would be to activate only the y-axial modes by a piston that covers an entire wall and drive the corresponding piston on the opposite, "receiving wall" in such a way that it absorbs the incident wave, that is, in anti-

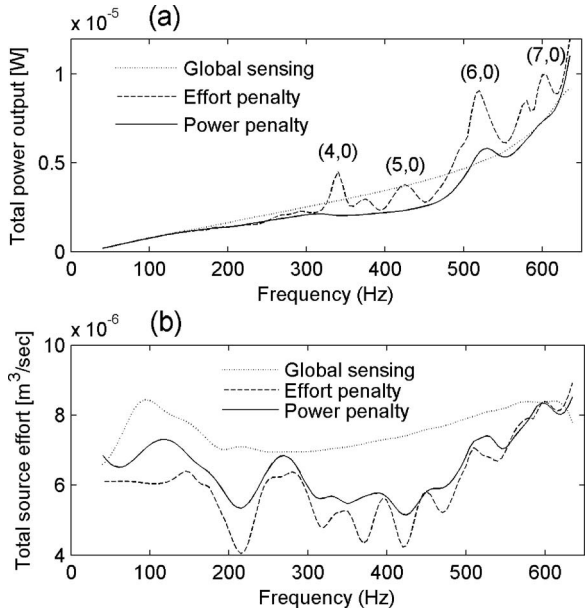


FIG. 4. (a) Total sound power output of the reproduction sources as a function of the frequency and (b) total source effort. The mode notation in (a) is (n_x, n_y).

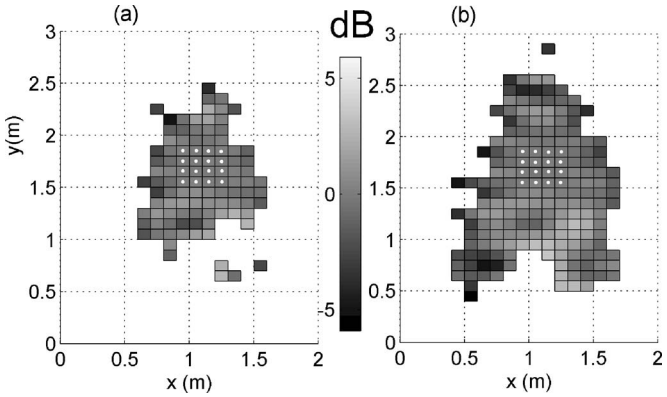


FIG. 5. Distribution of the sound pressure level in dB at 500 Hz using (a) effort regularization and (b) power-output regularization. The sound pressure is shown only at positions where the deviations of the reproduced sound pressure from the desired pressure are within ±6 dB. The positions of the control sensors are marked with white dots.

phase with the pressure on the wall, leading to a plane wave in the entire volume of the room.[3] The effect of using a small control sensor array without regularization is to generate a modal rearrangement that leads to an accurate reproduction of the progressive wave only in the control region and to inevitable deterioration of the sound field outside of it. For this reproduction mechanism, a much greater number of natural modes can contribute to the sound field. With penalization of the effort, the global performance is improved as a result of decreasing the source strengths and therefore the amplitude of the sound field outside the control region. However, this penalty term does not force the control system to avoid the unwanted modes. It forces it to use the modes with natural frequencies close to the excitation frequency, since the stronger the resonant term, the smaller the effort from the sources required to reproduce a given amplitude. Inspection of Figs. 3(a) and 4(a) shows that the error peaks in the global performance of effort penalty regularization are connected to peaks in the power output of the system that occur near the characteristic frequencies of undesired modes. Although this kind of modal rearrangement reproduces the sound field accurately in the control region, it does not avoid degradation outside the region, especially when the excitation frequency is close to a natural frequency of an unwanted mode. On the other hand, it seems that the power penalty makes the system behave more as a global sensing mechanism, with activation of the *y*-axial modes and suppression of the unwanted modes. An interesting difference between the two methods can be seen in Figs. 4(a) and 4(b). The total power output with traditional regularization is always greater than with the proposed technique, but the total source effort with the proposed technique is always greater than with the traditional method.

The reduced global error achieved with the proposed technique is related to a smoother decline of the performance away from the control points. As can be seen in Fig. 5, the percentage of the listening area where the deviations of the reproduced sound pressure from the desired pressure are within ±6 dB at 500 Hz is 42% for traditional regularization and 67% for the proposed technique.

V. CONCLUSIONS

The sound power output of a sound reproduction system in a room is related to standing waves. Therefore a power penalty term in the cost function obtained by a multiple point equalization system can lead to the suppression of undesired modes and to an extension of the region of equalization in the frequency range where the modal density is low. It remains to be seen, how this would be possible in real life. Matters concerning causality as well as the calculation and incorporation of matrix **W** in the solution should be addressed in future work.

[1]J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," J. Sound Vib. **102**, 217–228 (1985).
[2]F. Asano and D. C. Swanson, "Sound equalization in enclosures using modal reconstruction," J. Acoust. Soc. Am. **98**, 2062–2069 (1995).
[3]A. O. Santillán, "Spatially extended sound equalization in rectangular rooms," J. Acoust. Soc. Am. **110**, 1989–1997 (2001).
[4]J. C. Sarris, F. Jacobsen, and G. E. Cambourakis, "Sound equalization in a large region of a rectangular enclosure," J. Acoust. Soc. Am. **116**, 3271–3274 (2004).
[5]M. R. Schroeder, "The statistics of frequency responses in large rooms, Die statistischen Parameter der Frequenzkurven von grossen Räumen (in German)," Acustica **4**, 594–600 (1954).
[6]P. A. Nelson, "Active control of acoustic fields and the reproduction of sound," J. Sound Vib. **177**, 447–477 (1994).
[7]A. K. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," J. Acoust. Soc. Am. **93**, 2764–2778 (1993).
[8]P. A. Gauthier, A. Berry, and W. Woszczyk, "Sound-field reproduction in-room using optimal control techniques," J. Acoust. Soc. Am. **117**, 662–678 (2005).
[9]S. J. Elliott, P. Joseph, P. A. Nelson, and M. E. Johnson, "Power output minimization and power absorption in the active control of sound," J. Acoust. Soc. Am. **90**, 2501–2512 (1991).
[10]P. A. Nelson, "A review of some inverse problems in acoustics," Int. J. Acoust. Vib. **6**, 118–134 (2001).
[11]P. M. Morse, *Vibration and Sound*, 2nd ed. (McGraw–Hill, New York, 1948).
[12]A. J. Bullmore, P. A. Nelson, A. R. D. Curtis, and S. J. Elliott, "The active minimization of harmonic enclosed sound fields, part II: A computer simulation," J. Sound Vib. **117**, 15–33 (1987).

Comment on "A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties" [J. Acoust. Soc. Am. 119, 2334–2350 (2006)] (L)

Geoffrey Stewart Morrison[a)]
Department of Cognitive & Neural Systems, Boston University, 677 Beacon Street, Boston, Massachusetts 02215

(Received 15 March 2007; revised 1 October 2007; accepted 2 October 2007)

In a recent paper by Wassink [J. Acoust Soc. Am. **119**, 2334–2350 (2006)] the spectral overlap assessment metric (SOAM) was proposed for quantifying the degree of acoustic overlap between vowels. The SOAM does not fully take account of probability densities. An alternative metric is proposed which is based on quadratic discriminant analysis and takes account of probability densities in the form of *a posteriori* probabilities. Unlike the SOAM, the *a posteriori* probability-based metric allows for a direct comparison of vowel overlaps calculated using different numbers of dimensions, e.g., three dimensions (Fl, F2, and duration) versus two dimensions (Fl and F2). © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804633]

PACS number(s): 43.70.Jt [AL] Pages: 37–40

I. INTRODUCTION

Wassink (2006) presented a metric for quantifying the degree of overlap between pairs of vowel categories in a two-dimensional (2D) first and second formant (F1–F2) space, and a three-dimensional (3D) Fl–F2 duration space. Wassink's spectral overlap assessment metric (SOAM) is based on the assumption that the acoustic properties of vowel categories can reasonably be represented by multivariate normal distributions of normalized formant and duration values. SOAM calculation consists of the following steps: (1) The mean vectors and covariance matrices of the distributions of the two vowel categories are estimated via least-squares fits to sample data. (2) Ellipsoids extending two standard deviations from the category means are calculated. (3) A regular grid of points is projected into the 2D/3D space. (4) The proportion of the number of points falling within the intersection of the two ellipsoids relative to the number of points falling within a single ellipsoid is calculated. This results in two proportions, one for when the single ellipsoid corresponds to category *A* and one for when it corresponds to category *B*. (5) The larger of the two proportions is used as the overlap metric (Ω_{SOAM}).

Wassink's SOAM has some similarities with quadratic discriminant analysis (Hastie *et al.*, 2001, Sec. 4.3), but the use of a regular grid does not fully exploit information about the probability densities of the two distributions [an earlier vowel-overlap metric based on planimetric convex polygons, Brubaker and Altshuler (1959), also failed to take account of probability densities]. This results in a practical problem with the SOAM: It cannot be used to compare the overlaps of pairs of vowels in different numbers of acoustic dimensions [this problem was acknowledged by Wassink (2006), p. 2340]. If two vowel categories differ in their Fl and F2 distributions, but have identical duration distributions, then duration does not contribute to the separation of the two categories. In this situation, a desirable property for an overlap metric would be that the value calculated on the first two dimensions (Fl and F2) be the same as the value calculated on all three dimensions (Fl, F2, and duration). However, the 2D Ω_{SOAM} value would be greater than the 3D Ω_{SOAM} value. The 2D SOAM procedure is somewhat akin to measuring the overlap of two spheres by assuming that they are instead two cylinders of equal height and thus reducing the problem to the calculation of the overlap of two circles. This will always overestimate the overlap of the two spheres except when overlap is complete or nil. When one considers probability density, the situation is further complicated by the fact that the density of each sphere is not constant and the overlap of more dense regions should count more than the overlap of less dense regions. Hence, except at the two extremes, the 2D Ω_{SOAM} value will always be greater than the 3D Ω_{SOAM} value irrespective of whether the third dimension contributes to the separation of the two vowel categories.

In this letter an alternative vowel-overlap metric is proposed. Using the proposed metric, overlap values can be directly compared irrespective of the number of acoustic dimensions considered. The proposed metric is based on quadratic discriminant analysis and exploits probability density information in the form of *a posteriori* probabilities. Quadratic discriminant analysis uses separate covariance matrices for each category. If homogeneity of covariance matrices were assumed, then standard analytic statistics such as Pillai's Trace or Wilks's Lambda could be used as metrics of vowel overlap, and both have in fact been used in this way (Hay *et al.*, 2006; Morrison, 2004).

[a)]Now at School of Language Studies, Australian National University, Canberra, Australian Capital Territory 0200, Australia. Electronic mail: geoff.morrison@anu.edu.au

II. CALCULATION OF *A POSTERIORI* PROBABILITY-BASED OVERLAP METRIC

The numerical procedure for calculating the *a posteriori* probability-based overlap metric is presented in the following numbered paragraphs. See EPAPS (2008) for a MATLAB function implementing the procedure.

(l) On the basis of experimental sample data, use least-squares fits to estimate (a) the mean vector and covariance matrix for category A and (b) the mean vector and covariance matrix for category B.
(2) On the basis of the estimated mean vectors and covariance matrices, use a multivariate sample generator to generate a large number of fresh points for each category.
(3) Train a quadratic discriminant analysis model on the generated data.
(4) For every generated point from category A, calculate its *a posteriori* probability for membership of category B.
(5) Calculate the mean of the *a posteriori* probabilities from step 4.
(6) Repeat steps 4 and 5 to calculate the mean *a posteriori* probability of generated points from category B as members of category A.
(7) Add the results of steps 5 and 6. Use this as the overlap metric (Ω_{app}).

The procedure can be expanded to any number of dimensions and any number of categories. For calculation of Ω_{app} for more than two categories: At stage 4, calculate *a posteriori* probabilities for all not-A points as members of category A. At stage 6, cycle through all other categories B,C,D, etc. At stage 7, divide the sum of the *a posteriori* probabilities by one less than the number of categories.

III. CALIBRATION OF *A POSTERIORI* PROBABILITY-BASED OVERLAP METRIC AND COMPARISON WITH SOAM

The *a posteriori* probability-based overlap metric is a number between zero and one. An Ω_{app} close to one indicates a high degree of overlap, and an Ω_{app} close to zero indicates very little overlap. To provide a calibration of the *a posteriori* probability-based overlap metric and a comparison with the SOAM, a series of category A and category B distributions were generated using predetermined parameter values, and overlaps were calculated. Distributions A and B were both spherical with variances set to one and covariances to zero. The means of distribution A were fixed at the origin, and the means of distribution B were roved. Five-hundred-thousand sample points were generated for each category in the *a posteriori* probability-based procedure, and a total of one-million grid points were used in the SOAM procedure. The 2D and 3D Ω_{app} values were calculated using the same set of generated samples: The 2D Ω_{app} values were calculated using a quadratic discriminant model trained only on the first two dimensions $[x,y]$ from the generated samples, and the 3D Ω_{app} values were calculated using all three dimensions $[x,y,z]$. Table I presents the results, which are discussed in the following paragraphs.

TABLE I. SOAM values (Ω_{SOAM}) and *a posteriori* probability-based overlap metric values (Ω_{app}) calculated on two $[x\ y]$ and three $[x\ y\ z]$ dimensions. Distributions A and B were spherical with equal variances. The mean vector of distribution A was fixed at [0 0 0], and the mean vector of B was roved.

		Ω_{SOAM}		Ω_{app}	
B relative to A	B means $[x\ y\ z]$	2D	3D	2D	3D
$B_1=A$	[0 0 0]	1.00	1.00	1.00	1.00
B_2 1σ shift on x axis	[1 0 0]	0.69	0.63	0.80	0.80
B_3 2σ shift on x axis	[4 0 0]	0.00	0.00	0.07	0.07
B_4 3σ shift on x axis	[9 0 0]	0.00	0.00	0.00	0.00
B_5 $\sqrt{2}\sigma$ shift on x axis	[2 0 0]	0.39	0.31	0.45	0.45
B_6 $\sqrt{2}\sigma$ shift on xy plane	$[\sqrt{2}\ \sqrt{2}\ 0]$	0.39	0.31	0.45	0.45
B_7 $\sqrt{2}\sigma$ shift in xyz space	$[\sqrt{\frac{4}{3}}\ \sqrt{\frac{4}{3}}\ \sqrt{\frac{4}{3}}]$	0.50	0.31	0.57	0.45
B_8 $\sqrt{2}\sigma$ shift on xy plane $+\sqrt{2}\sigma$ shift on z axis	$[\sqrt{2}\ \sqrt{2}\ 2]$	0.39	0.11	0.45	0.23
B_9 1σ shift on xy plane $+2\sigma$ shift on z axis	$[\sqrt{\frac{1}{2}}\ \sqrt{\frac{1}{2}}\ 4]$	0.69	0.00	0.80	0.06

Distribution B_1 is identical to A, so Ω_{app} is very close to one and Ω_{SOAM} is exactly one.

As distribution B moves away from distribution A along the x axis (B_2,B_3,B_4) both metrics decrease toward zero. By design, Ω_{SOAM} is exactly zero for separations of two standard deviations or greater. In contrast, Ω_{app} does not suffer from this arbitrary floor effect.

Distributions B_5 and B_6 illustrate that an equal-magnitude shift away from A in one, or two dimensions results in the same value for 2D Ω_{app} and 3D Ω_{app}. The third dimension does not contribute to the separation of the two categories, and this fact is reflected in the identical values for the 2D Ω_{app} and 3D Ω_{app}. In contrast, the 3D Ω_{SOAM} value is smaller than the 2D Ω_{SOAM} value.

Distributions B_5, B_6, and B_7 illustrate that an equal-magnitude shift away from A in one, two, or three dimensions results in the same 3D Ω_{app} values. The 3D Ω_{SOAM} values are also the same for the three distributions. Distribution B_7 differs from distributions B_5 and B_6 in that the magnitude of the shift in two dimensions is less than the magnitude of the shift in three dimensions. This difference is reflected in larger values for both 2D Ω_{app} and 2D Ω_{SOAM} for distribution B_7 compared to distributions B_5 and B_6.

Distributions B_5, B_6, and B_8 illustrate that an equal-magnitude shift away from A in one or two dimensions results in the same 2D Ω_{app} values. The 2D Ω_{SOAM} values are also the same for the three distributions. Distribution B_8 differs from distributions B_5 and B_6 in that the magnitude of the shift in three dimensions is greater than the magnitude of the shift in two dimensions. This is reflected in the fact that the

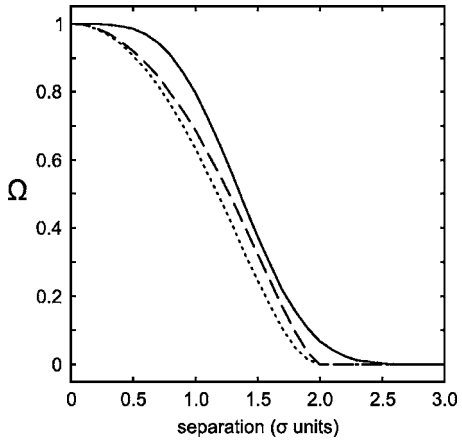


FIG. 1. Relationship between the *a posteriori* probability-based overlap metric (solid line), 2D SOAM (dashed line), 3D SOAM (dotted line), and a unidimensional separation of a pair of equal-variance spherical distributions (separation measured in standard deviation or d' units).

3D Ω_{app} value is smaller for distribution B_8 compared to distributions B_5 and B_6. This is also true for the 3D Ω_{SOAM} value.

Finally, distributions B_7, B_8, and B_9 illustrate that a shift including a shift in the third dimension can result in 3D Ω_{app} values which are substantially smaller than the corresponding 2D Ω_{app} values. The 3D Ω_{SOAM} values are also smaller than the 2D Ω_{SOAM} values; however, this is also true for distributions B_5 and B_6, which do not differ from category *A* on the third dimension. Thus, the difference between the 2D and 3D Ω_{app} values, but not the difference between the 2D and 3D Ω_{SOAM} values, can be taken as an indication of the magnitude of the contribution of the third dimension to the separation between the two distributions.

The relationship between the SOAM and the *a posteriori* probability-based overlap metric is further illustrated in Fig. 1 which graphs the relationship between the overlap metrics and a shift of distribution *B* along the *x* axis. The Ω_{app} (solid line) is identical whether calculated on two or three dimensions. The 3D Ω_{SOAM} (dotted line) is smaller than the 2D Ω_{SOAM} (dashed line), even though the third dimension does not contribute to the separation of the two distributions. Due to not fully taking into account probability densities and using an arbitrary floor, both 2D and 3D Ω_{SOAM} underestimate the degree of overlap and have discontinuities at an *x*-axis separation of two standard deviations.

IV. EXAMPLE OF USE OF *A POSTERIORI* PROBABILITY-BASED OVERLAP METRIC WITH NATURAL DATA

As a demonstration of the *a posteriori* a probability-based procedure applied to natural data, tests were conducted to determine the degree of overlap between /æ/ and /ɛ/ in Midwestern US adult male and in females speakers' productions using data from Hillenbrand *et al.* (1995). Overlap values were calculated using a one-dimensional space (duration), a two-dimensional space (steady-state Fl and F2), a three-dimensional space (steady-state Fl and F2, plus duration), a four-dimensional space (Fl and F2 at 20% of the duration of the vowel, plus the change in Fl and F2 from 20% to 80% of the duration of the vowel), and a five-dimensional space (the four-dimensional space plus duration). First and second formant values for /æ/ and /ɛ/ were log-interval normalized (Nearey and Assmann, 2007), and vowel durations were independently log-interval normalized. Data from 44 men and 47 women were included. Plots of the two- and three-dimensional distributions are provided in Fig. 2, and Table II provides the calculated overlap values for the one- through five-dimensional spaces. Table II also provides Ω_{SOAM} values; however, one should keep in mind that Ω_{SOAM} values cannot be directly compared across different numbers of dimensions. Also, note that the SOAM floor effect results in several Ω_{SOAM} values of zero where Ω_{app} values range from 0.02 to 0.07. The significance levels reported in Table II are based on randomization tests on the Ω_{app} values. Monte Carlo simulations were conducted on Ω_{app} values to determine whether there were significant reductions in vowel overlap when additional dimensions were considered. For both men and women, there was a significant reduction in overlap when dynamic spectral properties were compared to steady-state spectral properties, and when duration properties were considered in addition to spectral properties.

In terms of Fl and F2 steady states (two dimensions), the men produced a substantially and significantly greater overlap between /æ/ and /ɛ/ than did the women. When dynamic spectral information was considered (four dimensions), both men and women had a substantial drop in vowel overlap compared to when only steady-state values were considered. Although it just failed to meet the customary significance level of 0.05, the gender difference in overlap values calculated on dynamic formant data was still relatively large.

When vowel duration was considered, there was a substantial decrease in vowel-category overlap for both men and women. The magnitude of the reduction was, however, greater for men than for women. When duration was included in the calculation of vowel-category overlap, the gender difference was small and not significant. Thus although the men produced a greater spectral overlap between /æ/ and /ɛ/ than did the women, they produced a smaller duration

TABLE II. SOAM values (Ω_{SOAM}) and *a posteriori* probability-based overlap metric values (Ω_{app}) for Midwestern US English /æ/ and /ɛ/ productions in the Hillenbrand *et al.* (1995) data. Overlaps calculated separately for men and women. The *p* values indicate the statistical significance of the unsigned difference between the men's and women's Ω_{app} values calculated via randomization tests.

	Ω_{SOAM}		Ω_{app}		
Dimensions	Men	Women	Men	Women	*p*
1. Duration	0.00	0.15	0.07	0.12	0.278
2. Steady state F1 and F2	0.88	0.56	0.87	0.63	0.041
3. Steady state F1 and F2 + duration	0.00	0.01	0.06	0.09	0.518
4. Initial F1 and F2 +ΔF1 and ΔF2	0.23	0.05	0.32	0.17	0.080
5. Initial Fl and F2 +ΔF1 and ΔF2 + duration	0.00	0.00	0.04	0.02	0.565

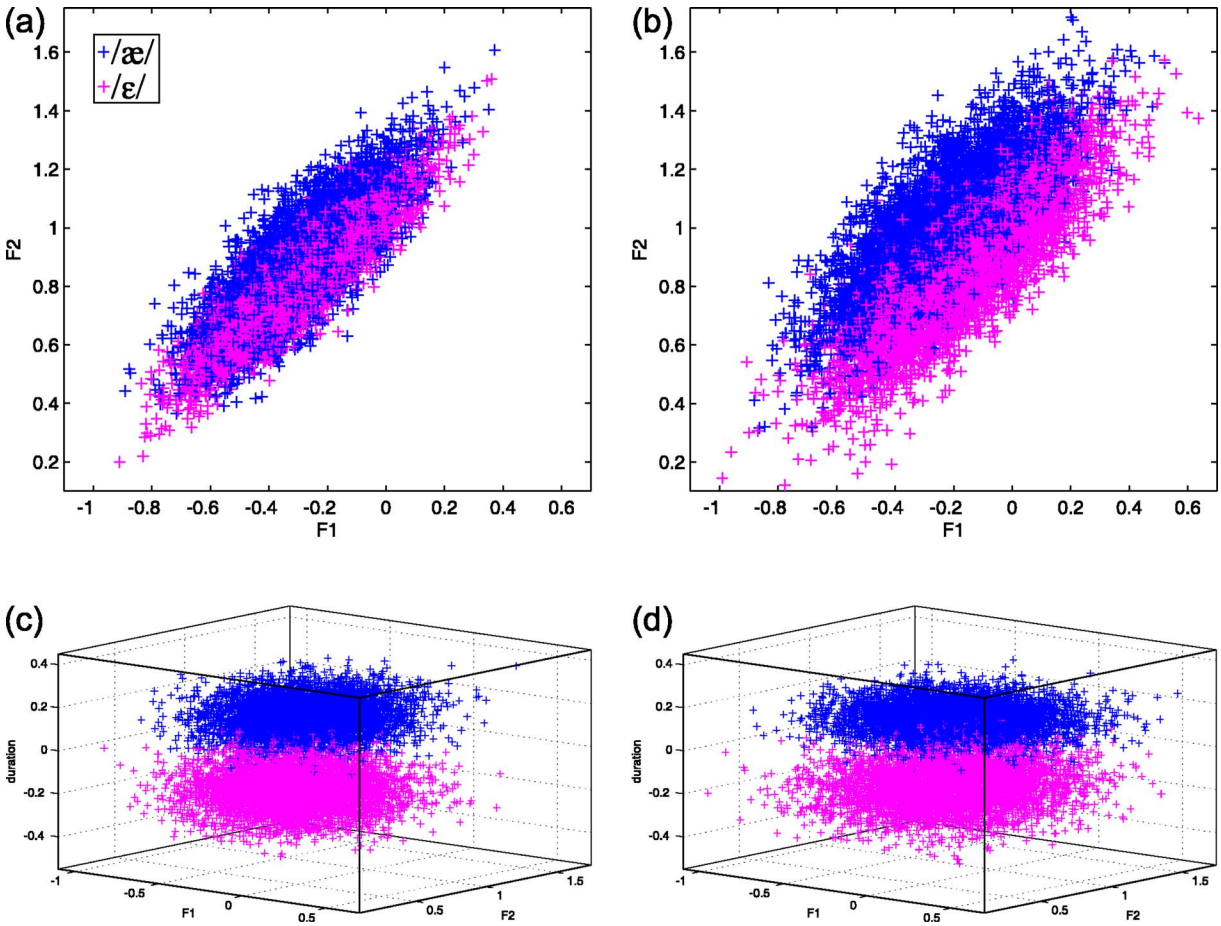


FIG. 2. (Color online) Plots of the two-dimensional distributions (normalized steady state Fl and F2) and three-dimensional distributions (normalized steady state Fl and F2 + normalized duration) of /æ/ and /ɛ/. Plots contain 5000 tokens of each vowel generated on the basis of the mean and covariance matrices calculated from the Hillenbrand *et al.* (1995) data. (a) 2D men's data. (b) 2D women's data. (c) 3D men's data. (d) 3D women's data.

overlap, and hence men and women had a similar degree of overlap between the two vowels when both spectral and duration properties were considered.

Similar results could have been obtained via a direct application of quadratic discriminant analysis. The advantage of using an overlap metric is that it provides an easily interpretable quantitative measure of the degree of overlap between vowels, which appears to have been an important objective in Wassink (2006).

V. CONCLUSION

In conclusion, the *a posteriori* probability-based overlap metric is superior to Wassink's (2006) spectral overlap assessment metric, because it more fully takes account of probability densities and can therefore be used to compare the degree of overlap between vowels when different numbers of acoustic dimensions are considered. Unlike Ω_{SOAM} values, Ω_{app} values can therefore be used to determine whether speakers produce a greater separation between a vowel pair if one considers vowel duration in addition to Fl and F2 values.

ACKNOWLEDGMENTS

This research was supported by the Social Sciences and Humanities Research Council of Canada. Thanks to James M. Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler for making their data available. Thanks to the Associate Editor and anonymous reviewers for comments on earlier versions of this letter.

Brubaker, R. S., and Altshuler, M. W. (**1959**). "Vowel overlap as a function of fundamental frequency and dialect," J. Acoust. Soc. Am. **31**, 1362–1365.

EPAPS Document No. E-JASMAN-123-001801 MATLAB function which calculates the a-posteriori-probability-based overlap metric. Metric designed to measure degree of overlap in vowel distributions in multiple dimensions such as normalized F1, F2, and duration. The document may also be reached via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html) or from ftp.aip.org in the directory /epaps/. See the EPAPS homepage for more information.

Hastie, T., Tibshirani, R., and Friedman, J. (**2001**). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).

Hay, J., Warren, P., and Drager, K. (**2006**). "Factors influencing speech perception in the context of a merger-in-progress," J. Phonetics **34**, 458–484.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Morrison, G. S. (**2004**). "An acoustic and statistical analysis of Spanish mid-vowel allophones," Estudios de Fonética Experimental **13**, 11–37.

Nearey, T. M., and Assmann, P. F. (**2007**). "Probabilistic 'sliding-template' models for indirect vowel normalization," in *Experimental Approaches to Phonology*, edited by M. J. Solé, P. S. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.

Wassink, A. B. (**2006**). "A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties," J. Acoust. Soc. Am. **119**, 2334–2350.

Modal group time spreads in weakly range-dependent deep ocean environments

Ilya A. Udovydchenkov[a] and Michael G. Brown
RSMAS/AMP, University of Miami, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA

(Received 18 June 2007; revised 18 September 2007; accepted 2 October 2007)

The temporal spread of modal group arrivals in weakly range-dependent deep ocean environments is considered. It is assumed that the range dependence is sufficiently weak that mode coupling is predominantly local in mode number. The phrase "modal group arrival" is taken here to mean the contribution to a transient wave field corresponding to a fixed mode number. There are three contributions to modal group time spreads which combine approximately in quadrature. These are the reciprocal bandwidth (the minimal pulse width), a deterministic dispersive contribution that is proportional to bandwidth and grows like range r, and a scattering-induced contribution that grows approximately like $r^{3/2}$. The latter two contributions are shown to be proportional to the waveguide invariant β, a property of the background sound speed profile. The results presented, based mostly on asymptotic theory, are shown to agree well with full-wave numerical wave field simulations and available exact mode theoretical results. Simulations are shown that correspond approximately to conditions during the LOAPEX acoustic propagation experiment. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804634]

PACS number(s): 43.30.Bp, 43.30.Ft, 43.30.Dr [RAS] Pages: 41–50

I. INTRODUCTION

The work reported here was largely motivated by a desire to interpret and understand measurements made during the 2004 long-range acoustic propagation experiment (LOAPEX) that was conducted in the eastern North Pacific ocean.[1] Essential elements of the experiment are that broadband signals in the 50–100 Hz band from a submerged compact source were transmitted to a receiving array at ranges of approximately 50, 250, 500, 1000, 1600, 2300, and 3200 km. The propagation path had weak background range dependence on which smaller scale structure, due mostly to internal waves, was superimposed. The receiving array had several deficiencies (which will not be discussed here) but this array still allows some mode filtering to be performed. These comments suggest that a natural way to describe the wave field and interpret the measurements is to employ a modal description that accounts for the broadband nature of the field and its range evolution in the presence of weak scattering (mode coupling). In this paper we provide such a theoretical framework. Simulated wave fields that correspond approximately to conditions during the LOAPEX experiment are used to illustrate and test the theoretical results presented. Analysis of the LOAPEX measurements will be discussed separately.

Previously, simulation and data-based estimates of modal group time spreads for multimegameter transmissions in the eastern Pacific Ocean have been reported in Refs. 2 and 3. In addition, Colosi and Flatté[4] have presented an extensive set of numerical simulations that were designed to investigate the influence of internal-wave-induced mode coupling on modal group time spreads. The questions that motivated their work included understanding the limitations imposed by mode coupling on acoustic tomography and matched field processing, and understanding the limitations of the adiabatic mode approximation. These issues continue to be of interest. Thus, in addition to providing a basis for interpreting the LOAPEX measurements, earlier measurements, and simulations, the results presented here have implications for aspects of broadband signal coherence and stability.

The remainder of the paper is organized as follows. In Sec. II basic results relating to modal group time spreads are presented. Full-wave numerical simulations in a range-independent environment are presented in Sec. III and compared to a simple asymptotic form of the theory. Motivated by this comparison, a correction to the asymptotic results is derived and shown to be in excellent agreement with numerical simulations. The influence of weak mode coupling on modal group time spreads is considered in Sec. IV. In Sec. V full-wave numerical simulations in a range-dependent environment similar to the LOAPEX environment are presented and used to illustrate and test the theoretical predictions. The results presented are summarized and discussed in the final section.

II. DETERMINISTIC MODAL GROUP TIME SPREADS

In this section basic results relating to modal group time spreads are presented. The phrase "modal group arrival" is taken here to mean the contribution to a transient wave field corresponding to a fixed mode number. The measurement of modal group arrivals requires that the wave field be measured on a vertical array that is sufficiently long and dense to allow the processing steps described in the following to be performed. Receiving array deficiencies will not be discussed here. In this section and the following section it is

[a]Electronic mail: iudovydchenkov@rsmas.miami.edu

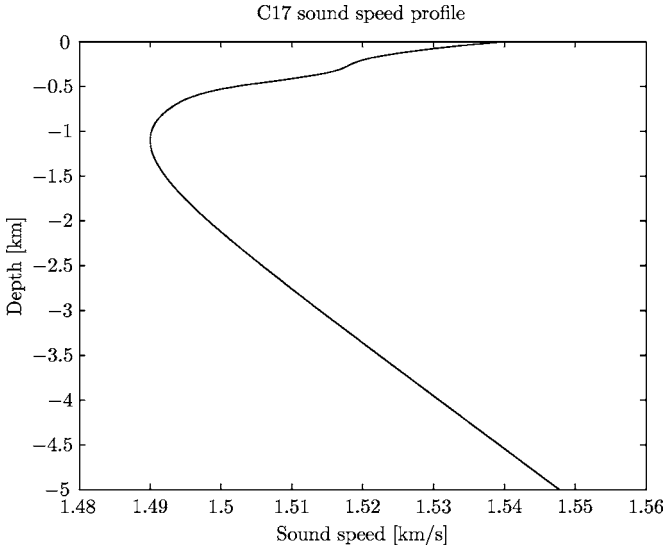


FIG. 1. Sound speed profile used in the simulations shown in Figs. 2–4.

assumed that the sound speed is a function of depth only, $c=c(z)$. Some of the material presented in this section is contained in Refs. 5–7.

It is assumed that the transient acoustic field, where $u(z,r,t)$ denotes acoustic pressure, is generated by a transient point source, with time history $s(t)$, located at $r=0$, $z=z_0$. Thus $u(z,r,t)$ satisfies

$$\nabla^2 u(z,r,t) - c^{-2}(z)\frac{\partial^2 u(z,r,t)}{\partial t^2} = -\delta(z-z_0)\frac{\delta(r)}{2\pi r}s(t). \quad (1)$$

Let $\bar{s}(\sigma)$ denote the Fourier transform of $s(t)$, and similarly for $\bar{u}(z,r,\sigma)$, where $\sigma=2\pi f$ is radian frequency. The solution to Eq. (1) can be written as a Fourier integral:

$$u(z,r,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \bar{s}(\sigma)\bar{u}(z,r,\sigma)e^{-i\sigma t}d\sigma, \quad (2)$$

where

$$\bar{u}(z,r,\sigma) = \frac{i}{4}\sum_{m=0}^{\infty} \phi_m(z_0,\sigma)\phi_m(z,\sigma) \times H_0^{(1)}(\sigma p_m r) \Big/ \int \phi_m^2(z,\sigma)dz \quad (3)$$

is a sum of normal modes, and $H_0^{(1)}$ is the zeroth-order Hankel function of the first kind. The normal modes $\phi_m(z,\sigma)$ satisfy

$$\frac{d^2\phi_m}{dz^2} + \sigma^2(c^{-2}(z) - p_m^2)\phi_m = 0, \quad (4)$$

together with a pair of boundary conditions, where $p_m=k_m/\sigma$ is a discrete value of the horizontal component of the slowness vector $p_r=k_r/\sigma$. (The variables k_m and p_m are discrete samples of the continuous variables k_r and p_r.) Imposition of these conditions leads to a quantization condition

$$p_m = p_r(m,\sigma), \quad m=0,1,2,\ldots, \quad (5)$$

which determines the allowed values of $p_m(\sigma)$, $m=0,1,2,\ldots$. The contribution to the wave field from the mode with frequency σ and mode number m can be written

$$\bar{u}_m(r,\sigma) = \int \bar{u}(z,r,\sigma)\phi_m(z,\sigma)dz \Big/ \int \phi_m^2(z,\sigma)dz. \quad (6)$$

The inverse Fourier transform of $\bar{u}_m(r,\sigma)$ is $u_m(r,t)$ which has the asymptotic form

$$u_m(r,t) \approx \frac{1}{2\pi}\int \frac{\bar{a}_m(\sigma)}{\sqrt{r}}e^{i(k_m(\sigma)r-\sigma t)}d\sigma, \quad (7)$$

where

$$\bar{a}_m(\sigma) = \frac{i}{4}\left(\frac{2}{\pi k_m}\right)^{1/2} e^{-i\pi/4}\phi_m(z_0,\sigma)\bar{s}(\sigma) \Big/ \int \phi_m^2(z,\sigma)dz \quad (8)$$

(here the far-field approximation $\sigma p_m r \gg 1$ is assumed). Under the assumption of a narrow-band signal (i.e., that the spectral content of the source is sharply peaked around the center frequency $\sigma_0=2\pi f_0$) a Taylor series expansion of $k_m(\sigma)$ about σ_0 gives

$$k_m(\sigma) \approx k_m(\sigma_0) + S_g(m,\sigma_0)(\sigma-\sigma_0) + \frac{1}{2}\frac{dS_g}{d\sigma}(m,\sigma_0)(\sigma-\sigma_0)^2, \quad (9)$$

where $S_g(m,\sigma)=dk_m/d\sigma$ is the group slowness. Consistent with the narrow-band assumption and Eq. (9) we shall assume that $\bar{a}_m(\sigma)$ can be approximated by a Gaussian $\bar{a}_m(\sigma)=Ae^{(-\pi(f-f_0)^2/(\Delta f)^2)}$ that is peaked at the center frequency σ_0. This assumption allows the Fourier integral (7) to be evaluated analytically. The result is

$$u_m(r,t) = |A|\sqrt{\frac{\Delta f}{r\Delta t_m^0(r)}}\exp\left(-\frac{\pi(t-S_g(m,\sigma_0)r)^2}{(\Delta t_m^0(r))^2}\right) \times \exp(i(k_m(\sigma_0)r - \sigma_0 t + \gamma)). \quad (10)$$

This represents a slowly varying dispersive wave train whose envelope moves at the group slowness $S_g(m,\sigma_0)$, under which surfaces of constant phase move at the phase slowness p_m. Note that this statement remains valid even if higher order terms are retained in the Taylor series expansion (9). The point to emphasize here is that energy propagates at the group slowness, so the time of arrival of modal energy is $t=S_g r$. The exact form of the phase term γ in Eq. (10) is not important for our purposes. The temporal width of the envelope, i.e., the modal group time spread, is

$$\Delta t_m^0(r) = \sqrt{\Delta t_{bw}^2 + \Delta t_d^2}. \quad (11)$$

Here

$$\Delta t_{bw} = (\Delta f)^{-1} \quad (12)$$

and

$$\Delta t_d = -2\pi r \Delta f \beta(m,\sigma_0)\frac{\partial p_r}{\partial \sigma}(\sigma_0), \quad (13)$$

where

$$\beta(m,\sigma) = -\frac{\partial S_g}{\partial p_r} \tag{14}$$

is the waveguide invariant.[6,8,9] The product $-\beta(m,\sigma)\,\partial p_r(\sigma)/\partial\sigma$ is the derivative $\partial S_g(m,\sigma)/\partial\sigma$. Partial derivatives are used here to emphasize that m is held constant. Simplification of Eqs. (13) and (14) is generally possible if the quantization condition (5) is known. This will be discussed in the following section. It will be shown that these expressions have a particularly simple form when a simple asymptotic form of the quantization condition is used. The superscript 0 is used in Δt_m^0 to distinguish this quantity from Δt_m, defined below, which includes a scattering contribution.

It should be emphasized that the validity of Eqs. (11)–(14) is not limited to asymptotic analysis. These results do, however, require that $\beta(m,\sigma_0)\neq 0$ [so the quadratic term in the Taylor series expansion (9) does not vanish] and that the bandwidth is sufficiently narrow that across this band $\beta(m,\sigma)$ is approximately constant. The geometric interpretation of these constraints is that the slope of the dispersion curve—discussed in Sec. III—for mode number m is finite and nearly constant across the relevant frequency band Δf.

The widths Δf and Δt are defined as the half-widths of the Gaussian distributions at the point where the amplitude of the distribution (in f or in t) is reduced by a factor of $e^{-\pi}$ relative to the peaks. Equivalently, Δf and Δt are the full widths of the Gaussian at the points where the amplitude is reduced by a factor of $e^{-\pi/4}\approx 0.456\approx 0.5$. Thus, for distributions that can be approximated as Gaussians, to a good approximation Δf and Δt can be defined as the full width of the distributions at the half-amplitude points.

Equation (13) can be written as $\Delta t_d = r(\partial S_g(\sigma_0)/\partial\sigma)\Delta\sigma$, where $\Delta\sigma = 2\pi\Delta f$. Note that the validity of the truncated Taylor series expansion (9) rests on the assumption that the curvature of $S_g(\sigma)$ in the $\Delta\sigma$ band centered at $\sigma=\sigma_0$ is small. If this condition is satisfied, then an equivalent expression is $\Delta t_d = r|S_g(m,\sigma_0+\Delta\sigma/2) - S_g(m,\sigma_0-\Delta\sigma/2)|$. The correctness of the latter expression is an obvious consequence of the negligible curvature assumption together with the observation that the time of arrival of modal energy is $t=S_g(m,\sigma)r$. An immediate consequence of the latter observation is the general result

$$\Delta t_d = r\left[\max_{|\sigma-\sigma_0|\le\Delta\sigma/2} S_g(m,\sigma) - \min_{|\sigma-\sigma_0|\le\Delta\sigma/2} S_g(m,\sigma)\right], \tag{15}$$

which holds even when $S_g(m,\sigma)$ is not a monotonic function of σ for a fixed m. Note that according to Eq. (15) $d\Delta t_d/d\Delta\sigma \ge 0$, but that this derivative is generally not constant, as assumed in Eq. (13), Equation (15) is a generalization of Eq. (13). Note that we have shown that Eqs. (11), (12), and (15) are consistent with each other only in the limit of small bandwidth, where Eq. (15) reduces to Eq. (13). But Eqs. (11), (12), and (15) predict the correct behavior in the limits of small and large r, so it is natural to assume that Eq. (11) remains approximately valid when Eq. (15) replaces Eq. (13). Finally, we emphasize that the validity of the results presented so far is not limited to asymptotic validity.

III. ASYMPTOTIC APPROXIMATIONS

In this section full wave numerical simulations in a range-independent environment are presented and compared to the wave field structure predicted by Eqs. (10)–(14). In particular, we focus on the modal group time spread (11). Asymptotic analysis is exploited, particularly in the evaluation of $\beta(m,\sigma)$ (14).

The exact expression for the group slowness is[10,11]

$$S_g(m,\sigma) = \frac{\int \phi_m^2(z,\sigma)c^{-2}(z)dz}{p_m(\sigma)\int \phi_m^2(z,\sigma)dz}. \tag{16}$$

Both this expression and the derivative $dS_g(m,\sigma)/d\sigma$ can be evaluated numerically. Then, using Eqs. (12) and (13), the modal group time spread (11) can be computed. While this process is not difficult to carry out, it offers no computational advantage over numerically evaluating Eqs. (6)–(8) directly. In the following we show that the approximate theoretical results (10)–(14) provide an accurate description of modal energy distributions provided $\beta(m,\sigma_0)\neq 0$ and that under most circumstances simple asymptotic expressions for $\beta(m,\sigma_0)$ and $\partial p_r/\partial\sigma$ provide very good approximations.

In an environment in which $c(z)$ has a single minimum, it is well known[12,13] that modes with two internal (nonreflecting) turning points asymptotically satisfy the quantization condition

$$\sigma I(p_r) = m + \frac{1}{2}, \quad m = 0,1,2,\ldots, \tag{17}$$

where the classical action

$$I(p_r) = \frac{1}{\pi}\int_{\check z(p_r)}^{\hat z(p_r)} (c^{-2}(z) - p_r^2)^{1/2}dz \tag{18}$$

with $c(\check z(p_r)) = c(\hat z(p_r)) = 1/p_r$, where $\check z$ and $\hat z$ are lower and upper turning points, respectively. Note that these equations define $p_m = p_r(m,\sigma)$. For the class of problems for which Eq. (17) is valid

$$S_g(p_r) = \frac{T(p_r)}{R(p_r)}. \tag{19}$$

Here $R(p_r) = -2\pi dI/dp_r$ and $T(p_r) = 2\pi I(p_r) + p_r R(p_r)$ are the range and travel time, respectively, of a ray double loop. Use of Eqs. (17)–(19) leads to a simple explicit expression for Δt_d (13),

$$\Delta t_d = -2\pi r\Delta f\frac{I}{R(I)f_0}\beta(I), \tag{20}$$

where I and $R(I)$ are those values corresponding to the relevant (m,σ_0) pair and $\beta(I) = (2\pi I/R^2)(\partial R/\partial p_r)$.

Using the sound speed profile shown in Fig. 1 the simulations shown in Figs. 2–4 were performed to illustrate some important features of the above-described results, and to test the validity of some of the approximations made. The sound speed profile shown in Fig. 1 consists of a perturbed canonical profile,[14] $c(z) = c_M(z) + dc\,\exp\left(-\frac{1}{2}((z-z_c)/z_w)^2\right)$, $c_M = c_a(1 + \epsilon(e^\eta - \eta - 1))$ with $\eta = 2(z-z_a)/b$. Here z increases upwards,

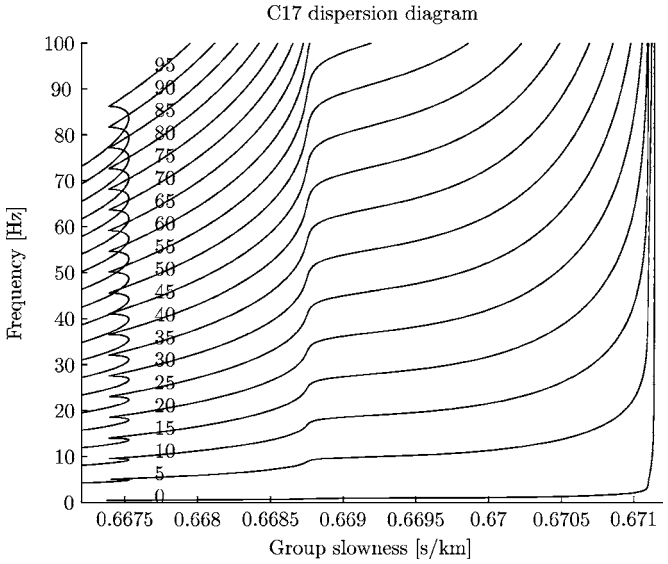


FIG. 2. Dispersion diagram for mode numbers $m=0,5,10,\ldots,95$ in the environment shown in Fig. 1, computed using the simple asymptotic results, Eqs. (17)–(19).

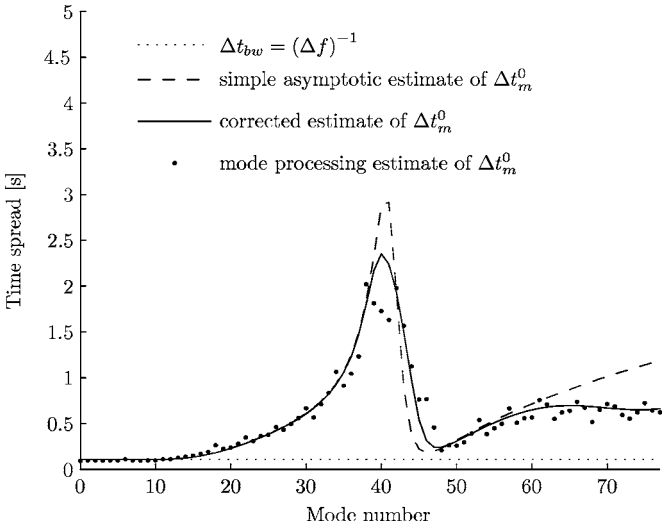


FIG. 4. Theoretical and simulation-based estimates of modal group time spreads Δt_m^0 for the wave field shown in Fig. 3, corresponding to the dispersion diagram shown in Fig. 2. Theoretical estimates make use of Eqs. (11) and (12). In the simple asymptotic estimate of Δt_m^0, Eq. (20) was used to compute Δt_d. Equations (15), (24), and (25) were used to compute Δt_d in the corrected estimate of Δt_m^0.

$z_a=-1.1$ km is the sound channel axis depth, $c_a=1.49$ km/s is the sound speed on the channel axis, $b=1.0$ km is the thermocline depth scale, $\epsilon=0.0057$ is a dimensionless constant, $z_w=0.1$ km is the width of the Gaussian sound speed perturbation that is centered at depth $z_c=-0.35$ km, and $dc=0.008$ km/s is the amplitude of the Gaussian perturbation.

Under commonly encountered experimental conditions $\Delta t_d \gg \Delta t_{bw}$. With this assumption dispersion diagrams provide a complete picture of the modal group arrival structure. These correspond to a family of curves, each corresponding to a fixed m, that are most naturally plotted in (f,S_g) space. For the sound speed profile shown in Fig. 1, the dispersion diagram is shown in Fig. 2. Figure 2 was constructed using the asymptotic results (17)–(19); those equations parametrically define a family of curves $S_g(m;\sigma)$. Because Eqs. (17) and (19) are approximate the dispersion diagram shown in Fig. 2 contains some errors which will be discussed in the following. An exact dispersion diagram can be constructed from the exact quantization condition (5) (which in general must be found numerically) and Eq. (16). Because energy associated with mode number m at frequency σ arrives at time $t=S_g(m,\sigma)r$, the dispersion diagram shows the temporal structure of the wave field for all mode numbers in the frequency band of interest.

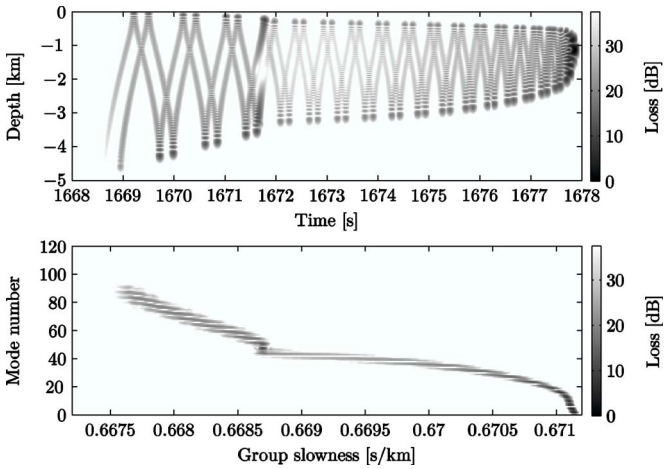


FIG. 3. (Color online) Upper panel: Wave field intensity, shown on a logarithmic scale, computed in the environment shown in Fig. 1 using an axial source with $f_0=75$ Hz and a full computational bandwidth of 18.75 Hz at a range of 2500 km. Lower panel: The corresponding mode-processed wave field shown using the same time axis $t=S_g r$.

The upper panel of Fig. 3 shows full wave numerical simulations at $r=2500$ km with $|f-f_0| \le \Delta f_c/2$, $f_0=75$ Hz, $\Delta f_c=18.75$ Hz, that correspond to conditions shown in Fig. 2. Here Δf_c is the full computational bandwidth. The source spectrum $\bar{s}(\sigma)$ had the shape of a Hanning window, for which $\Delta f \approx \Delta f_c/2$. The simulations shown in Fig. 3 were performed by solving the standard parabolic wave equation in a transformed environment $\bar{c}(\bar{z})$ as described in Refs. 15 and 16. Loosely speaking, the transformation from $c(z)$ to $\bar{c}(\bar{z})$ is constructed in such a way that the solution to the standard parabolic wave equation in $\bar{c}(\bar{z})$ is the same as the solution to the Helmholtz equation in $c(z)$. In particular, we note that $I(p_r), R(p_r)$, and $T(p_r)$ that appear in Eqs. (19) and (20) are preserved under the transformation. The reason for making use of this transformation is that it provides a relatively simple means to construct accurate approximate solutions to the Helmholtz equation in a range-dependent environment; simulations in range-dependent environments will be presented in Sec. V. The lower panel of Fig. 3 shows the mode-processed wave field $|u_m(t)|^2$ corresponding to the wave field $|u(z,t)|^2$ shown in the upper panel. Note that the time axis is the same in both plots, $t=S_g r$ with $r=2500$ km.

An estimate of Δt_m^0, derived from the wave field shown in Fig. 3, is plotted as a function of mode number m in Fig. 4. Also in Fig. 4, Δt_{bw} (12) and Δt_m^0, constructed using Eqs. (11), (12), and (20), are plotted as a function of m. The latter curve is labeled simple asymptotic estimate. Note that according to Eq. (11) a lower bound on Δt_m^0 is predicted to be

$(\Delta f)^{-1}$. Overall, the agreement between simple theory [Eqs. (11), (12), and (20)] and estimates of Δt_m^0 derived from simulations is seen to be good, but a systematic deviation between simulation-based estimates of Δt_m^0 and the simple theoretical prediction of Δt_m^0 is seen in two bands of mode numbers. First, consider $39 \lesssim m \lesssim 46$. For this band of m the slopes of the dispersion curves in the 65–85 Hz frequency band seen in Fig. 2 are seen to be nonconstant. As a result, Eqs. (13) and (20), which assume a constant slope, give poor approximation to Δt_d. To correct this problem one need only replace Eq. (20) by Eq. (15), which gives good agreement with simulation-based estimates of Δt_m^0 even when the simple asymptotic expression (19) is used to compute S_g. The second band of mode numbers where a systematic deviation between theory and simulation-based estimates of Δt_m^0 is seen in Fig. 4 is $60 \lesssim m \lesssim 75$. This range of m corresponds to modes with upper turning depths near the surface. The cause of the misfit between simulations and the simple theory for these mode numbers is the use of Eq. (20) which, in turn, depends on Eqs. (17)–(19). Those equations do not correctly describe near-surface-reflecting modes. It is possible, however, to derive corrections within an asymptotic framework to those equations. This issue will now be addressed.

Corrections to Eqs. (17)–(19) for near-grazing (surface and/or bottom) modes have previously been described by many authors. A straightforward and quite general approach to addressing problems of this type is to note that Eq. (17) is a special case of a general expression (see, e.g., Ref. 17)

$$\sigma J(p_m) = m - \frac{\varphi_u + \varphi_l}{2\pi}. \tag{21}$$

Here $J(p_m)$ has the same form as $I(p_m)$ in Eq. (18) except that the lower and upper bounds on the integral coincide with the modal turning depths, which may coincide with one or both of the boundaries, while φ_u and φ_l are the phases of the upper and lower reflection coefficients, R_u and R_l, respectively. For modes with upper and lower turning depths far from the boundaries $R_u = R_l = -i$, and Eq. (21) reduces to Eq. (17).

In the following we focus on near-surface-grazing modes in an environment with a deep sound speed excess, so $R_l = -i$ and R_u transitions smoothly from $-i$ [$c(\hat{z}) \ll c(0)$, non-surface-reflecting modes] to -1 [$c(\hat{z}) \gg c(0)$, surface-reflecting modes]. For the latter class of modes (21) reduces to $\sigma J(p_m) = m + 3/4$. The desired reflection coefficient, with the properties just described, was derived originally by Murphy and Davis.[18] For $c(\hat{z}) < c(0)$ (RR modes with upper turning depths close to the surface),

$$R_u = \exp\left[i\left(-\frac{\pi}{2} - 2\tan^{-1}\frac{Ai(S)}{Bi(S)} \right) \right], \tag{22}$$

where

$$S = \left[\frac{3}{2}\sigma \int_{\hat{z}}^{0} \sqrt{p_r^2 - c^{-2}(z)}\, dz \right]^{2/3}. \tag{23}$$

[These equations require some modification for surface-reflecting modes, $c(\hat{z}) > c(0)$, but this will not be discussed here.] With $R_l = -i$ and R_u given by Eq. (22), Eq. (21) becomes

$$\sigma I(p_m) = m + \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left[\frac{Ai(S)}{Bi(S)} \right]. \tag{24}$$

Note that Eq. (24) reduces to Eq. (17) for large S (upper turning depth many wavelengths away from the surface) and to $\sigma I = m + 2/3$ for a surface grazing mode ($S = 0$). We note also that Eq. (24) is a special case of Eq. (5.21) in Ref. 12 with the assumption that the ocean bottom is many wavelengths away from the lower modal turning depth. Noting that $k_m = \sigma p_m$ and $S_g = \mathrm{d}k_m/d\sigma$, differentiation of Eq. (24) with respect to σ (with m fixed) gives

$$S_g(p_m, \sigma) = \frac{T(p_m) + 2\left(\frac{3}{2}\right)^{-1/3}\kappa\sigma^{-1/3}\left(\int_{\hat{z}}^{0} \sqrt{p_m^2 - c^{-2}(z)}\, dz \right)^{-1/3} \int_{\hat{z}}^{0} \frac{c^{-2}(z)\, dz}{\sqrt{p_m^2 - c^{-2}(z)}}}{R(p_m) + 2p_m\left(\frac{3}{2}\right)^{-1/3}\kappa\sigma^{-1/3}\left(\int_{\hat{z}}^{0} \sqrt{p_m^2 - c^{-2}(z)}\, dz \right)^{-1/3} \int_{\hat{z}}^{0} \frac{dz}{\sqrt{p_m^2 - c^{-2}(z)}}}, \tag{25}$$

where

$$\kappa = -\frac{1}{\pi(Ai^2(S) + Bi^2(S))}. \tag{26}$$

Elimination of p_m from Eqs. (24) and (25) allows the construction of a dispersion diagram.

An estimate of Δt_m^0, constructed using Eqs. (24) and (25), is plotted as a function of mode number m in Fig. 4 with a solid line. The agreement between the corrected theory and estimates of Δt_m^0 derived from simulations is seen to be excellent even for modes with turning depths close to the pressure release surface. Equations (23)–(26) generalize Eqs. (17)–(19) for near-surface-grazing modes with lower turning depths far from the water/bottom interface. Modes of this type are observed in most single-minimum deep ocean sound channels with a deep sound speed excess, including the LOAPEX environment. Although the validity of Eqs. (23)–(26) is restricted to this class of modes, the type of analysis that we have presented can easily be generalized to treat other classes of modes for which Eqs. (17)–(19) are invalid.

IV. THE SCATTERING-INDUCED CONTRIBUTION TO MODAL GROUP TIME SPREADS

The results presented in Sec. III need to be amended to account for the mode coupling that occurs in realistic range-dependent ocean environments. We focus here on environments consisting of a range-independent background on which a highly structured range-dependent perturbation, due for example to internal waves, is superimposed. Mode coupling is then predominantly local in mode number. The predominance of nearest neighbor mode coupling is discussed in Ref. 17. In Ref. 19 the same authors show that to an excellent approximation ray scattering can be described by a diffusion process in action I. But the quantization condition (17) [or some generalization thereof, such as Eq. (24)] shows that as acoustic energy diffuses in action in the ray description it diffuses (taking discrete steps) in mode number in the mode description. This is an aspect of ray-mode duality that, like other aspects, is particularly transparent when asymptotic mode results, such as Eq. (17), are used. With the above comments as background, we will treat m and I as interchangeable labels in the following discussion, keeping in mind that a quantization condition connects m and I. In the presence of mode coupling the total group delay of modal energy that has been scattered among mode numbers m_i, $i=1,\ldots,n$ can be written

$$t_g = \sum_{i=1}^{n} S_g(m_i,\sigma_0)\Delta r_i, \tag{27}$$

where $\Sigma_{i=1}^{n}\Delta r_i = r$, the total range. Expanding S_g in a Taylor series, making use of Eq. (14) and $R=-2\pi dI/dp_r$, gives

$$t_g \approx S_g(m_0,\sigma_0)r + 2\pi\frac{\beta(I_0)}{R(I_0)}\sum_{i=1}^{n}(I_i - I_0)\Delta r_i. \tag{28}$$

Convenient choices for I_0 are the action at $r=0$ or the action at the final range, and $I_0=(m_0+1/2)/\sigma_0$ (or some generalization thereof). If Δr_i is taken to be constant (so $n\Delta r=r$), then

$$\delta t_s = 2\pi\frac{\beta(I_0)}{R(I_0)}\Delta r\sum_{i=1}^{n}(I_i - I_0) \tag{29}$$

is the scattering-induced arrival time perturbation for acoustic energy whose forward or reversed action history is $I_0, I_1, \ldots, I_n$. In Eq. (29) I_i is the action after the ith scattering event

$$I_i = I_0 + \sum_{j=1}^{i}\delta I_j. \tag{30}$$

Assume that δI_j is a delta-correlated zero-mean random variable,

$$\langle \delta I_j\rangle = 0, \langle \delta I_j \delta I_k\rangle = \langle(\delta I)^2\rangle\delta_{jk}. \tag{31}$$

Let $\Delta I_n = I_n - I_0$. Then $\langle(\Delta I_n)^2\rangle = n\langle(\delta I)^2\rangle$ and, for large n, $\langle(\Sigma_{i=1}^{n}\Delta I_n)^2\rangle \approx n^3\langle(\delta I)^2\rangle/3$. Thus, it follows from Eq. (29) with $\Delta t_s = 2\sqrt{\pi}\langle(\delta t_s)^2\rangle^{1/2}$ that

$$\Delta t_s = 4\pi^{3/2}\frac{|\beta(I_0)|}{R(I_0)}\left(\frac{B}{3}\right)^{1/2} r^{3/2}, \tag{32}$$

where $B=\langle(\delta I)^2\rangle/\Delta r$. Note also that, in terms of B, $\langle(\Delta I(r))^2\rangle = Br$, which can be taken as the definition of B in the continuum limit. It is important to keep in mind that in a scattering environment of the type considered, any wave field contains scattered energy corresponding to many mode number or action histories. Assuming each such mode number or action history is independent, $\langle(\delta t_s)^2\rangle^{1/2}$ is the rms value of the corresponding distribution of travel time perturbations at range r. The quantity Δt_s is taken to be $2\sqrt{\pi}$ times the standard deviation of this distribution so that this quantity, like Δf and Δt_m^0, is defined as the full width of the δt_s distribution at the half-amplitude points. (Recall that 0.5 is an approximation to $e^{-\pi/4}\approx 0.456$ and note that because wave field intensity in the ray description is proportional to ray density it is assumed that wave field intensity is proportional to the action distribution.) We shall refer to Eq. (32) as the scattering-induced contribution to a modal group time spread. It should be noted that a formula similar to Eq. (32) was derived in a different way by Virovlyansky[7] and by Virovlyansky *et al.*[20]

The assumption of delta-correlated action scattering events has been shown[19,21] to be an excellent approximation at long range in deep ocean environments. Using results from the study of stochastic differential equations, scattering in the continuum limit was treated directly in those studies. An advantage of that approach is that it allowed the authors to correctly treat near-axial (where I is close to 0) scattering. The authors solved this problem by pointing out that the relevant Fokker–Planck equation for $\Delta I(r)$ admits an exact solution for a reflecting boundary at $I=0$. [Note that our Eqs. (31) and (32) fail near the sound channel axis as they fail to enforce the condition $I\geq 0$. This issue will not be further discussed. It is important to emphasize, however, that the above-presented results—Eqs. (31) and (32), in particular—are not valid for I close to zero (near-axial modes).]

A final remark concerning Eq. (32) is that this expression is expected to be a good approximation only when $\beta(I_0)$ is representative of β values for all scattered energy that at range r has $I=I_0$. This condition will be satisfied if $\beta(I)$ [i.e., $\beta(m)$ in the relevant frequency band] is a slowly varying function. When this assumption is not satisfied, an improved—relative to Eq. (32)—estimate of δt_s should result from replacing Eq. (29) by

$$\delta t_s = \frac{2\pi}{R(I_0)}\Delta r\sum_{i=1}^{n}\beta(I_i')(I_i - I_0). \tag{33}$$

This comment follows from the observations that: (1) the Taylor series expansion (28) is exact—by the mean value theorem—if, for each term in the sum, $\beta(I_0)/R(I_0)$ is replaced by $\beta(I_i')/R(I_i')$ for some I_i' between I_0 and I_i; and (2) variations in $R(I)$ are negligible compared to variations in $\beta(I)$. The value of I_i' is not known *a priori*, but it might be possible to parametrize $\beta(I_i')$ in such a way that Eq. (33) can be simplified in cases where Eq. (29) is a poor approximation.

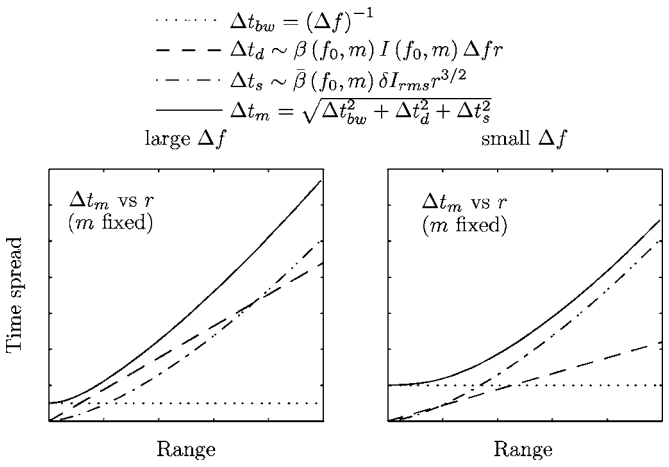


FIG. 5. Plots of Δt_{bw}, Δt_d, Δt_s, and Δt_m vs r under conditions for which there is (left panel) and is not (right panel) a range of r values over which Δt_d is the dominant term. The following parameter values, which are typical of deep ocean conditions, were used to construct these plots: f_0=75 Hz, m =56, β=−0.117, I=0.12 s, R=53.37 km, B=3.3×10^{-7} s^2/km, large Δf =20 Hz (left panel), and small Δf=10 Hz (right panel). The maximum range and time spread are 1000 km and 0.7 s, respectively.

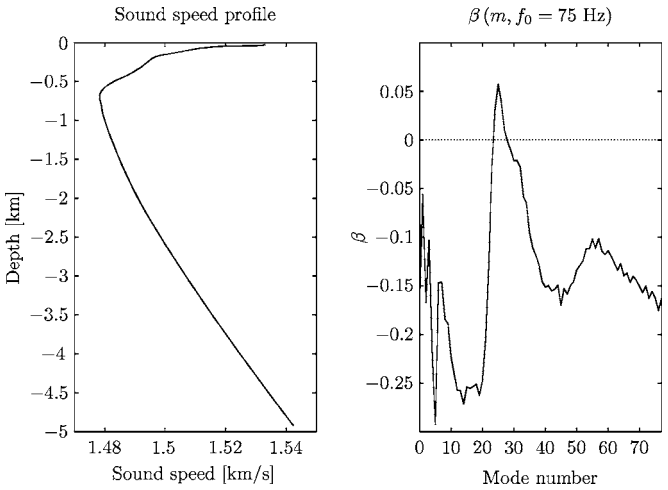


FIG. 6. Left panel: Background sound speed profile used in the wave field simulations shown in Figs. 7–9. Right panel: Corresponding plot of $\beta(m;\sigma_0)$ with $\sigma_0/2\pi = f_0$=75 Hz.

Because Δt_s, Eq. (32), is independent of the deterministic contributions, Δt_{bw} and Δt_d, it is natural to assume that the deterministic and stochastic contributions combine in quadrature, so with the aid of Eq. (11) the total modal group time spread is

$$\Delta t_m(r) = \sqrt{\Delta t_{bw}^2 + \Delta t_d^2 + \Delta t_s^2}. \tag{34}$$

The assumption that the three contributions to Δt_m combine in quadrature will be revisited in the following.

V. SIMULATIONS OF MODAL GROUP TIME SPREADS

In this section full-wave numerical simulations in an environment similar to the LOAPEX environment are presented. The purpose of presenting these simulations is both to test the approximate theoretical predictions that we have presented and to provide a foundation for the interpretation of the LOAPEX measurements (which will be presented elsewhere). The LOAPEX measurements were made at ranges between approximately 50 and 3200 km. For this reason we emphasize in this section the evolution in range of modal group time spreads.

Before presenting wave field simulations we note that a simple consequence of Eq. (34) and the observation that Δt_{bw}, Δt_d, and Δt_s grow like r^0 [Eq. (12)], r [Eqs. (13), (20), or (15)] and $r^{3/2}$ [Eq. (32)], respectively, is that plots of $\Delta t_m(r)$ are of one of two types, as illustrated in Fig. 5.

At short ranges Δt_{bw} is always the dominant contributor to Δt_m and at large ranges Δt_s is always the dominant contributor. There may, but need not, be an intermediate range of r values where Δt_d is the dominant contributor. Whether this regime is present is controlled largely, but not exclusively, by Δf: Large Δf favors small Δt_{bw} and large Δt_d. Thus large Δf favors the presence of Δt_d-dominant regime. A signal processing option is to reduce Δf, but this is generally not desirable, as reducing Δf reduces the relative importance of the deterministic dispersive contribution Δt_d to the total time spread. The deterministic dispersive and the scattering-induced contributions depend on the environment and are mode number dependent. Thus, it is possible that in the same wave field both types of behavior seen in Fig. 5 are simultaneously present for different groups of mode numbers.

To test our theoretical prediction, and the validity of the associated approximations, wave field simulations have been performed at many ranges in an environment similar to the LOAPEX environment. The source center frequency was f_0 =75 Hz and the computational bandwidth was Δf_c =37.5 Hz, corresponding to an effective bandwidth of Δf =18.75 Hz. The background sound speed profile $c(z)$ and corresponding plot of $\beta(m, f_0)$ are shown in Fig. 6. A fairly realistic internal-wave-induced sound speed perturbation $\delta c(z,r)$[22] was superimposed on the background $c(z)$ in the simulations. The parameters of the internal wave (IW) model are the following: maximum IW mode number $j_{\max}$=30, maximum horizontal wave number $k_{\max} = 2\pi/1.0$ km^{-1} $\approx$6.28 km^{-1}, and the minimum horizontal wave number $k_{\min} = 2\pi/3276.8$ km$^{-1} \approx 2\times10^{-3}$ km^{-1}. The sound speed perturbation has a sharp peak at 50 m depth where δc $\approx$2.3 m/s, and then decays rapidly to $\delta c \approx 0.5$ m/s at 200 m depth and to $\delta c \approx 0.1$ m/s at 1 km depth. The assumed strength of the internal wave field used in the simulations shown was the nominal Garrett–Munk value, $E = E_{\mathrm{GM}}$. Figure 7 shows sample wave fields in (z,t) at three ranges, together with the corresponding mode-processed fields in (m,t) where $t = S_g r$. By performing similar processing on wave fields at many ranges, $\Delta t_m(r)$ for all mode numbers can be estimated. Three such plots for m=17, 25, and 56 are shown in Fig. 8. For the simulations shown in Figs. 7 and 8 the source was located on the sound channel axis.

The construction of Fig. 8 requires some explanation. Consistent with the full width at half-amplitude criterion used earlier to define the widths of approximately Gaussian distributions, Δt_m was estimated from wave fields of the type shown in Fig. 7 as the half-difference in time between the latest and earliest points whose intensity is 27 dB below the peak intensity. As noted earlier, Δf was estimated in the

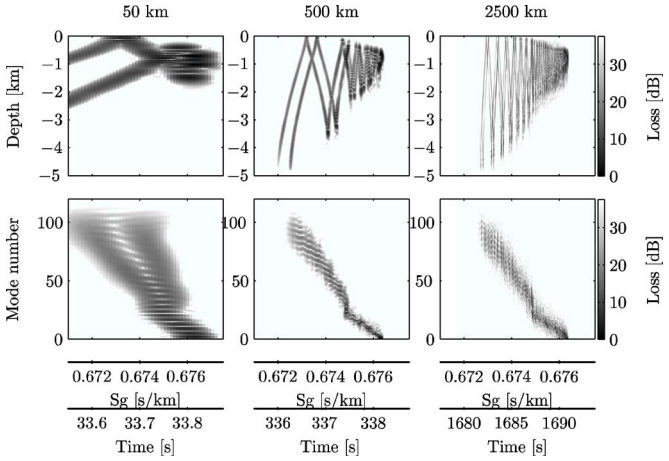


FIG. 7. (Color online) Simulated wave fields with $f_0=75$ Hz, $\Delta f_c=37.5$ Hz at three ranges in the environment shown in Fig. 6, with an internal-wave-induced sound speed perturbation superimposed. Upper panels: Wave field intensity in (z,t). Lower panels: Corresponding mode-processed fields in (m,t) where $t=S_g r$.

same way giving, to a good approximation, $\Delta f=\Delta f_c/2$. The dispersive contribution was computed using Eq. (15). [For mode numbers $20\lesssim m\lesssim 30$ Eq. (15) is a much better estimate of Δt_d than Eq. (20). Outside of this band Eq. (20) gives essentially the same estimate. Only for mode numbers greater than approximately 70 is the near-surface-grazing correction discussed in Sec. III important.] The scattering-induced contribution Δt_s was computed using Eq. (32). The parameter B in Eq. (32) was estimated numerically using cw parabolic equation simulations using a single mode starting field. Computed wave fields were projected onto the normal modes of the background environment. The spreading of energy in mode number [and thus also in I, making use of Eq. (17)] was found to be well approximated by a Gaussian whose variance grew like a constant—our estimate of B—times range. These simulations gave the estimates $B=6.7\times10^{-8}$ s^2/km for $m=17$, $B=1.3\times10^{-7}$ s^2/km for $m=25$, and $B=2.7\times10^{-7}$ s^2/km and $m=56$. These values are in approximate agreement with the estimates of B reported in Refs. 19 and 23.

The agreement between simulations and theory-based estimates of Δt_m in Fig. 8 is seen to be good for $m=17$ and $m=56$. Agreement is not good for $m=25$, however, where the

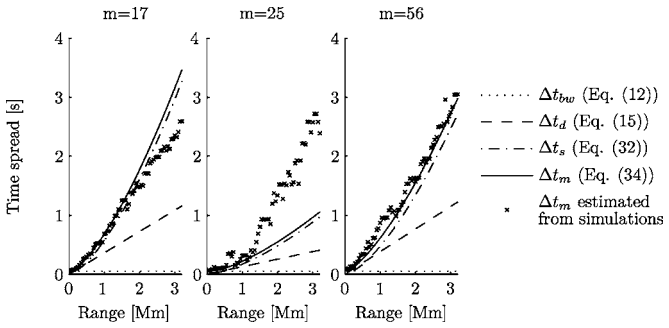


FIG. 8. Predicted and simulated estimates of modal group time spreads Δt_m vs range for three values of m. The simulated estimates of Δt_m were extracted from wave field simulations of the type shown in Fig. 7. 1 Mm =1000 km.

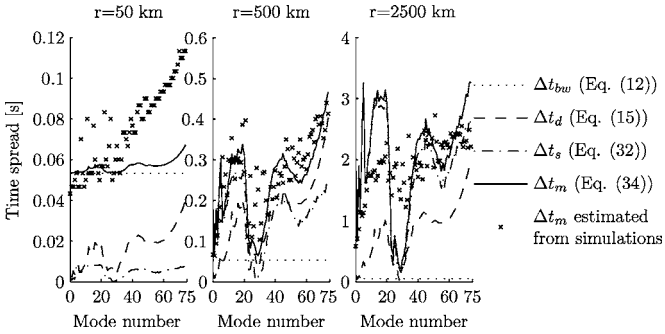


FIG. 9. Predicted and simulated estimates of modal group time spreads Δt_m vs mode number at $r=50$ km, $r=500$ km, and $r=2500$ km. The simulated estimates were extracted from the mode-processed wave fields shown in Fig. 7. Note that the time axes are different in the three subplots.

predicted estimate of Δt_s—and hence also Δt_m—is too small. The reason for this discrepancy is that, as discussed earlier, for m values near $m=25$, $\beta(I)$ is not a slowly varying function. This is clear from the $\beta(m,f_0=75$ Hz) structure shown in Fig. 6. In order to account for this nonuniform β-weighting, Eq. (32) must be replaced by an estimate based on Eq. (33) rather than Eq. (29), as discussed earlier. Nonuniform β-weighting of scattered energy also explains why agreement between theory and simulations in Fig. 8 is better at short range than at long range for $m=17$ and $m=25$. For $m=17$ scattered energy at long range has $|\beta|<|\beta(m=17)|$ (see Fig. 6), so Eq. (32) predicts a value of Δt_s that is too large. Similarly, for $m=25$ scattered energy at long range has $|\beta|>|\beta(m=25)|$ (see Fig. 6), so Eq. (32) predicts a value of Δt_s that is too small.

Figure 9 shows a comparison of predicted and simulation-based estimates of Δt_m as a function of mode number at three ranges: 50, 500, and 2500 km. The corresponding wave fields are shown in Fig. 7. The theoretical prediction of Δt_m is based on Eqs. (34), (32), (15), and (12). The value of B used in Eq. (32) was an empirical m-dependent fit to simulations of the type mentioned earlier: $B(m)=0.44\times(0.5+m/10)\times10^{-7}$ s^2/km. Agreement between the simple theoretical estimates and simulations is generally good. Note, however, that our estimate of Δt_s, based on Eq. (32), is clearly too low for $20\leqslant m\leqslant 30$, especially at $r=500$ km and $r=2500$ km. This behavior is consistent with our explanation of the error associated with Δt_s for $m=25$ in Fig. 8. Also, note that at $r=50$ km theoretical estimates of Δt_m are clearly too small for $m\gtrsim 40$. The cause of this discrepancy is that Eqs. (11), (12), and (15) underpredict the nonscattered energy; this is linked to the bimodal (and highly non-Gaussian) distribution of energy in (m,t) (see Fig. 7 at $r=50$ km).

VI. DISCUSSION AND SUMMARY

The purpose of the work described here is to provide a theoretical framework for a mode-based interpretation of low-frequency broadband measurements on a vertical array at multiple ranges in the deep ocean. Measurements of this type were made during the recent LOAPEX experiment;[1] those data will be analyzed using the results presented here in a separate publication. We have assumed that the environ-

ment consists of a range-independent background on which a small-scale perturbation is superimposed. The extension to a slowly varying (adiabatic) background is straightforward. The perturbation, due for example to internal waves, was assumed to be sufficiently weak that mode coupling is predominantly local in mode number and can be approximately modeled as a diffusive process. There are three contributions to modal group time spreads: The reciprocal bandwidth $\Delta t_{bw}=(\Delta f)^{-1}$, a deterministic dispersive contribution $\Delta t_d \sim I(f_0,m)\beta(m,\sigma_0)r\Delta f$, and a scattering-induced contribution $\Delta t_s \sim \beta(m,\sigma_0)r^{3/2}$. We have argued, but not shown rigorously, that these three contributions combine in quadrature; this dependence was shown to be in generally good agreement with simulations. Under most experimental circumstances the term Δt_{bw} is negligible. Because both Δt_d and Δt_s are proportional to $\beta(m,\sigma_0)$ it may be difficult to experimentally distinguish these contributions from each other. Note, however, that Δt_s is expected to dominate at long range, and that at any range Δt_d can be reduced by reducing Δf (bandpass filtering recorded pressure time histories). The results of implementing this type of processing of the LOAPEX measurements will be described elsewhere.

With the above comments as background it is insightful to comment on the earlier work of Colosi and Flatté,[4] who investigated using numerical simulations modal group time spreads in deep ocean environments in the presence of internal-wave-induced sound speed fluctuations. They found empirically that Δt_m grows like $r^{3/2}$. This is consistent with our results, but some caveats should be noted. We predict $\Delta t_m \sim r^{3/2}$ when the dominant contribution to Δt_m is Δt_s and when Eq. (32) is valid. The first condition is not expected to hold at shorter ranges where Δt_{bw} and Δt_d are comparable to or larger than Δt_s. The second condition is not expected to hold for small mode numbers or when local variations in $\beta(m,\sigma_0)$ cannot be neglected. Colosi and Flatté[4] also observed a scattering-induced modal group arrival time bias that scales like r^2. We have not addressed the issue of a scattering-induced modal group arrival time bias but we note that such a bias is associated with the breakdown of Eq. (32). When Eq. (32) is invalid (for one of the two reasons just noted) the scattered mode arrival time pdf will in general be non-Gaussian and have a centroid that is displaced relative to the time of arrival of nonscattered energy. Associated with such a pdf is a scattering-induced arrival time bias. It is not clear why such a bias should grow like r^2. But the sign of the bias for low mode numbers is readily explained: Under typical deep ocean conditions S_g is greatest for smallest m values, so scattered low mode number energy will be biased toward larger m and smaller S_g and thus earlier arriving energy. Note, however, that the observation by Colosi and Flatté that the scattering-induced arrival time bias increased with increasing m suggests that in their simulations another effect—possibly nonuniform $\beta(m,\sigma_0)$ (or, equivalently, significant curvature of the group slowness curves in the relevant frequency band)—is more important than the near-axial correction to Eq. (32).

Several extensions to the results presented here will be explored in conjunction with analysis of the LOAPEX measurements. First, Virovlyansky's[19,21] scattered action probability density function $P_I(I;I_0,r)$ correctly treats near-axial scattering and can, in a straightforward manner, be incorporated into the framework that we have described, yielding an improved estimate of Δt_s for small mode numbers. Second, the dependence of the effective action diffusivity B on mode number (or action) is worthy of a more thorough investigation than has been provided here. We have focused on internal-wave-induced sound speed perturbations here, but the result that energy diffuses in action is expected to be a good approximation for any weak small-scale perturbation $\delta c(z,r)$. This expectation needs to be tested. Virovlyansky[7] has derived an approximate expression for the action diffusivity for a general perturbation. Third, if the β-weighted sum (33) could be simplified, this would lead to an improved estimate of δt_s, and in turn Δt_s. In the presence of strong local variations in $\beta(I)$, this task may prove to be very difficult. And finally, we note that because we have not addressed the phase of scattered energy, the results that we have presented cannot be used to address cross-mode coherences or full field statistics. That extension requires a fundamentally different approach.

The observation that both Δt_d and Δt_s are proportional to $\beta(m,\sigma_0)$ is noteworthy inasmuch as these quantities constitute two among many properties of both deterministic and scattering-induced contributions to wave fields that are controlled by $\beta(m,\sigma)$ or its ray counterpart $\alpha(I)$.[6] This list includes travel time dispersion,[6] ray amplitudes at long range,[6] scattered ray amplitudes/ray stability,[24] both constrained and unconstrained scattered ray travel time spreads,[25,26] both spatial and temporal spreads of narrow beams with and without scattering (Beron-Vera and Brown, private communication), and both diffractive (Fresnel zone width) and scattering-induced contributions to the effective width of a ray.[23] The combination of all of these results provides strong support for the idea that in nearly stratified environments wave field structure and stability is largely controlled by $\beta(m,\sigma)$ [or its ray counterpart $\alpha(I)$].

ACKNOWLEDGMENTS

We thank Javier Beron-Vera and Irina Rypina for the benefit of discussions related to this paper. This work was supported by the Office of Naval Research, Code 321, and the National Science Foundation, Agent No. CMG0417425.

[1] J. A. Mercer, R. K. Andrew, B. M. Howe, and J. A. Colosi, "Cruise report: Long-range ocean acoustic propagation experiment (LOAPEX)," Technical report, Applied Physics Laboratory, University of Washington, 2005.
[2] K. E. Wage, A. B. Baggeroer, and J. C. Preisig, "Modal analysis of broadband acoustic receptions at 3515-km range in the North Pacific using short-time Fourier techniques," J. Acoust. Soc. Am. **113**, 801–817 (2003).
[3] K. E. Wage, M. A. Dzieciuch, P. F. Worcester, B. M. Howe, and J. A. Mercer, "Mode coherence at megameter ranges in the North Pacific Ocean," J. Acoust. Soc. Am. **117**, 1565–1581 (2005).
[4] J. A. Colosi and S. M. Flatté, "Mode coupling by internal waves for multimegameter acoustic propagation in the ocean," J. Acoust. Soc. Am. **100**, 3607–3620 (1996).
[5] M. G. Brown, J. Viechnicki, and F. D. Tappert, "On the measurement of modal group time delays in the deep ocean," J. Acoust. Soc. Am. **100**, 2093–2102 (1996).
[6] M. G. Brown, F. J. Beron-Vera, I. Rypina, and I. A. Udovydchenkov, "Rays, modes, wave field structure, and wave field stability," J. Acoust. Soc. Am. **117**, 1607–1610 (2005).

[7]A. L. Virovlyansky, *Ray Theory of Long-Range Sound Propagation in the Ocean* (Institute of Applied Physics, Nizhny Novgorod, 2006) (in Russian).
[8]S. D. Chuprov, "Interference structure of a sound field in a layered ocean," *Ocean Acoustics, Current State* (Nauka, Moscow, 1982), pp. 71–91.
[9]G. A. Grachev, "Theory of acoustic field invariants in layered waveguides," Acoust. Phys. **39**, 33–35 (1993).
[10]R. A. Koch, C. Penland, P. J. Vidmar, and K. E. Hawker, "On the calculation of normal mode group velocity and attenuation," J. Acoust. Soc. Am. **73**, 820–825 (1983).
[11]D. M. F. Chapman and D. D. Ellis, "The group velocity of normal modes," J. Acoust. Soc. Am. **74**, 973–979 (1983).
[12]D. S. Ahluwalia and J. B. Keller, "Exact and asymptotic representations of the sound field in a stratified ocean," in *Wave Propagation and Underwater Acoustics* (Springer, New York, 1977).
[13]W. H. Munk and C. Wunsch, "Ocean acoustic tomography: Rays and modes," Rev. Geophys. Space Phys. **21**, 777–793 (1983).
[14]W. H. Munk, "Sound channel in an exponentially stratified ocean with application to SOFAR," J. Acoust. Soc. Am. **55**, 220–226 (1974).
[15]I. I. Rypina, I. A. Udovydchenkov, and M. G. Brown, "A transformation of the environment eliminates parabolic equation phase errors," J. Acoust. Soc. Am. **120**, 1295–1304 (2006).
[16]H. K. Brock, R. N. Buchal, and C. W. Spofford, "Modifying the sound-speed profile to improve the accuracy of the parabolic-equation technique," J. Acoust. Soc. Am. **82**, 543–552 (1977).
[17]A. L. Virovlyansky, A. Yu. Kazarova, and L. Ya. Lyubavin, "Ray-based description of normal mode amplitudes in a range-dependent waveguide," Wave Motion **42**, 317–334 (2005).
[18]E. L. Murphy and J. A. Davis, "Modified ray theory for bounded media," J. Acoust. Soc. Am. **56**, 1747–1760 (1974).
[19]A. L. Virovlyansky, A. Yu. Kazarova, and L. Ya. Lyubavin, "Statistical description of chaotic rays in a deep water acoustic waveguide," J. Acoust. Soc. Am. **121**, 2542–2552 (2007).
[20]A. L. Virovlyansky, A. Yu. Kazarova, and L. Ya. Lyubavin, "Modal structure of the field under conditions of wave chaos," in *Ocean Acoustics, Proceedings of the 11th L. M. Brekhovskikh's Conference*, Moscow, CEOS, 2006, pp. 40–43 (in Russian).
[21]A. L. Virovlyansky, "Statistical description of ray chaos in an underwater acoustic waveguide," Acoust. Phys. **51**, 71–80 (2005).
[22]J. A. Colosi and M. G. Brown, "Efficient numerical simulation of stochastic internal-wave-induced sound speed perturbation fields," J. Acoust. Soc. Am. **103**, 2232–2235 (1998).
[23]I. I. Rypina and M. G. Brown, "On the width of a ray," J. Acoust. Soc. Am. **122**, 1440–1448 (2007).
[24]F. J. Beron-Vera and M. G. Brown, "Ray stability in weakly range-dependent sound channels," J. Acoust. Soc. Am. **114**, 123–130 (2003).
[25]A. L. Virovlyansky, "Ray travel times at long range in acoustic waveguides," J. Acoust. Soc. Am. **113**, 2523–2532 (2003).
[26]F. J. Beron-Vera and M. G. Brown, "Travel time stability in weakly range-dependent sound channels," J. Acoust. Soc. Am. **115**, 1068–1077 (2004).

Parabolic equation solution of seismo-acoustics problems involving variations in bathymetry and sediment thickness

Jon M. Collis[a] and William L. Siegmann
Rensselaer Polytechnic Institute110 8th Street, Troy, New York 12180

Finn B. Jensen and Mario Zampolli
NATO Undersea Research Center, 19126 La Spezia, Italy

Elizabeth T. Küsel
Northeastern University360 Huntington Avenue, Boston, Massachusetts 02115

Michael D. Collins
Naval Research Laboratory, Washington, D.C. 20375

(Received 21 February 2007; revised 21 September 2007; accepted 27 September 2007)

Recent improvements in the parabolic equation method are combined to extend this approach to a larger class of seismo-acoustics problems. The variable rotated parabolic equation [J. Acoust. Soc. Am. **120**, 3534–3538 (2006)] handles a sloping fluid-solid interface at the ocean bottom. The single-scattering solution [J. Acoust. Soc. Am. **121**, 808–813 (2007)] handles range dependence within elastic sediment layers. When these methods are implemented together, the parabolic equation method can be applied to problems involving variations in bathymetry and the thickness of sediment layers. The accuracy of the approach is demonstrated by comparing with finite-element solutions. The approach is applied to a complex scenario in a realistic environment. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2799932]

PACS number(s): 43.30.Ma, 43.30.Dr, 43.30.Gv [RCG] Pages: 51–55

I. INTRODUCTION

The parabolic equation method is a powerful approach for solving range-dependent propagation problems (i.e., problems in laterally varying media) in ocean acoustics.[1] Since many sediments support shear waves, the development of the elastic parabolic equation[2–5] has been a topic of great interest. In recent years, there has been a focus on improving accuracy for range-dependent seismo-acoustics problems. The introduction of the (u_r,w) formulation,[6] where u_r is the range derivative of the horizontal displacement and w is the vertical displacement, has led to progress in this area. An improved single-scattering solution in the (u_r,w) formulation accurately handles range dependence within purely elastic media.[7] An improved rotated parabolic equation solution accurately handles fluid-solid interfaces of variable slope.[8] In this paper, we combine these approaches to obtain parabolic equation solutions of problems involving sloping ocean bottoms and varying sediment thickness. We demonstrate the accuracy of the approach by making comparisons with finite-element solutions. We also apply the approach to a model problem based on a complex environment off the New Jersey coast.[9] The approach is discussed in Sec. II and applied to examples in Sec. III.

II. PARABOLIC EQUATION SOLUTION

We describe the parabolic equation solution in cylindrical coordinates, where the range r is the horizontal distance from a source and z is the depth below the ocean surface. We work in the far field and remove the cylindrical spreading factor $r^{-1/2}$ from the solution. The parabolic equation method is based on the factorization of the operator in the elliptic wave equation into a product of operators that correspond to outgoing and incoming energy. These factors give rise to the parabolic wave equations[7]

$$\frac{\partial}{\partial r}\begin{pmatrix} u_r \\ w \end{pmatrix} = \pm i(L^{-1}M)^{1/2}\begin{pmatrix} u_r \\ w \end{pmatrix}, \tag{1}$$

where L and M are matrices containing depth operators, the plus sign corresponds to outgoing waves, and the negative sign corresponds to incoming waves. The outgoing wave equation is the basis for most applications of the parabolic equation method, but the incoming wave equation is also required in the single-scattering solution. The parabolic equation method can be applied to seismo-acoustics problems, which involve both fluid and solid layers, by using the vector operators in Eq. (1) in solid layers, using a scalar operator in water layers, and enforcing interface conditions at fluid-solid interfaces.[5]

TABLE I. Elastic properties of the layers in example A, where c_p and c_s are compressional and shear wave speeds, ρ is density, and α_p and α_s are compressional and shear attenuations. The layers are numbered from the sea floor.

Example	Elastic layers	c_p (m/s)	c_s (m/s)	ρ (g/cm^3)	α_p (dB/λ)	α_s (dB/λ)
A	1	2400	1200	1.5	0.1	0.2
	2	4000	2500	3.0	1.0	2.0
	3	4000	2500	3.0	10.0	20.0

[a]Author to whom correspondence should be addressed. Present address: Boston University, Boston, Massachusetts. Electronic mail: jcollis@bu.edu

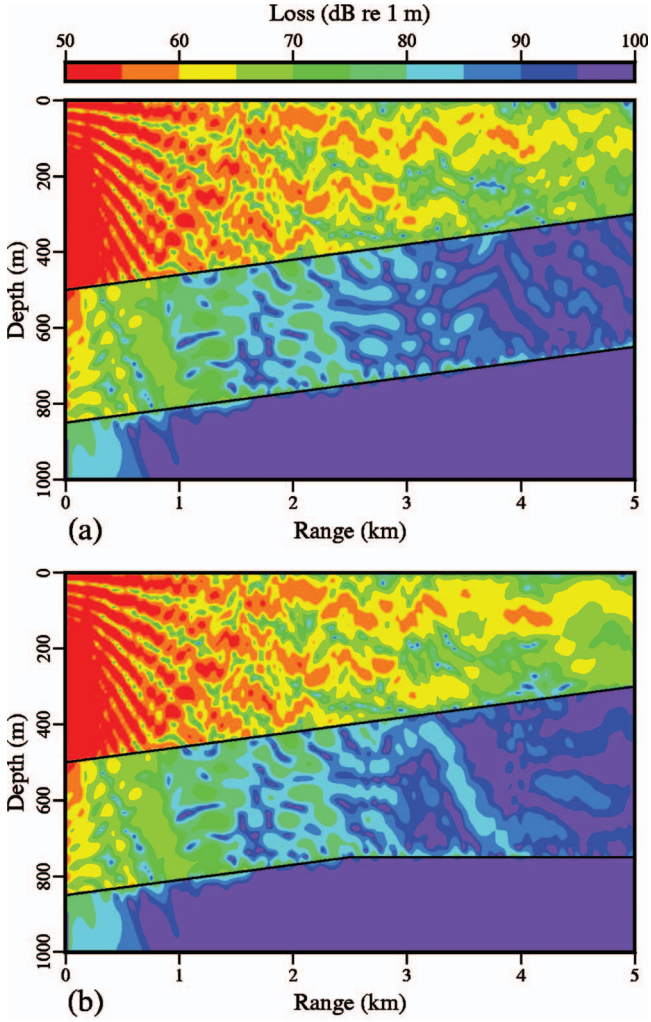


FIG. 1. Compressional transmission loss for Example A. For both cases, the sediment thickness is constant for $r>2.5$ km. (a) The case in which sediment thicknesses remains constant for $r<2.5$ km. (b) The case in which sediment thickness increases for $r>2.5$ km.

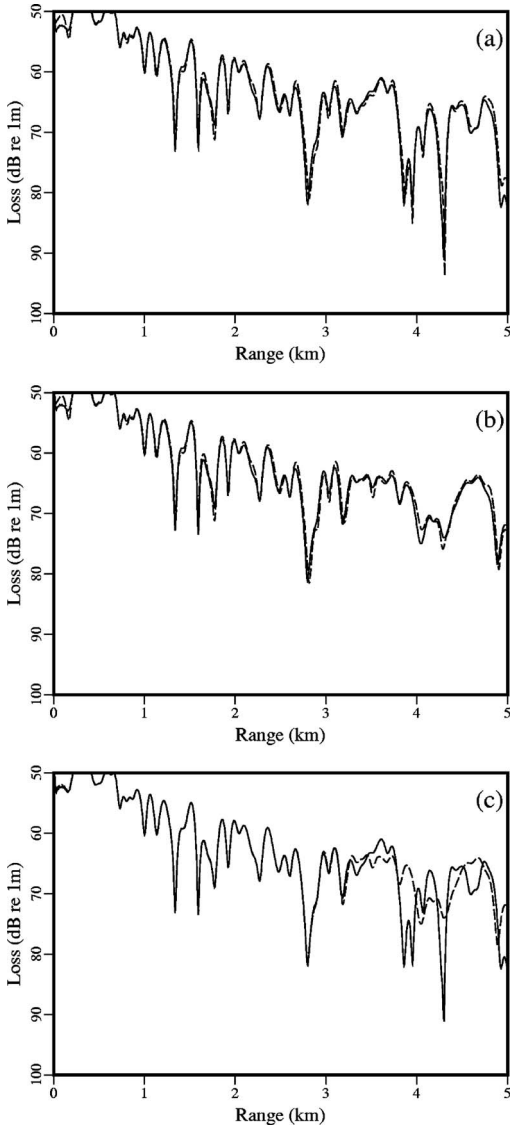


FIG. 2. Compressional transmission loss at $z=25$ m for example A. The dashed curves correspond to reference solutions that were generated using a finite-element model. The solid curves are the parabolic equation solutions for the cases of (a) constant layer thickness and (b) variable layer thickness. (c) Comparison between the parabolic equation solutions of the constant and variable layer thickness cases.

The factorization that gives rise to the parabolic wave equations in Eq. (1) is based on the assumption that the medium is range independent. Range-dependent problems can be solved by approximating the medium in terms of a series

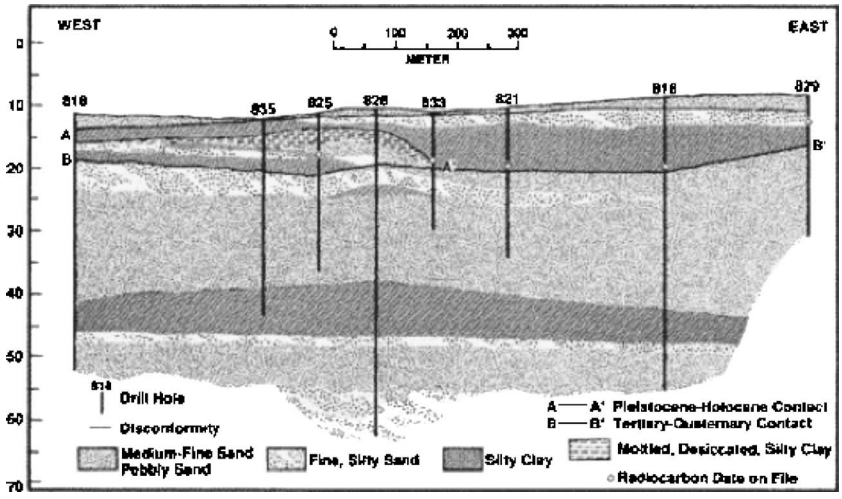


FIG. 3. The Atlantic Generating Site off the coast of New Jersey. Black vertical lines show boreholes where data were acquired and used to estimate the geoacoustic properties of layers. Image reused with permission from Badiey *et al.* (Ref. 11) and the Acoustical Society of America.

TABLE II. Elastic properties of the layers in example B, where c_p and c_s are compressional and shear wave speeds, ρ is density, and α_p and α_s are compressional and shear attenuations. The layers are numbered from the sea floor.

Example	Elastic layers	c_p (m/s)	c_s (m/s)	ρ (g/cm^3)	α_p (dB/λ)	α_s (dB/λ)
B(1)	1	1800	872	2.08	0.24	1.0
B(2)	1	1733	817	2.0	0.37	1.37
	2	1866	927	2.15	0.11	0.63
B(4)	1	1733	817	2.0	0.37	1.37
	2	1997	1117	2.25	0.08	0.55
	3	1700	800	1.95	0.15	0.75
	4	1900	1000	2.15	0.1	0.6
B(8)	1	1900	850	2.3	0.05	0.55
	2	1600	700	1.7	1.0	3.0
	3	1700	900	2.0	0.05	0.55
	4	2100	1300	2.25	0.1	0.55
	5	1790	750	2.25	0.05	0.55
	6	2100	1300	2.25	0.1	0.55
	7	1700	800	1.95	0.15	0.75
	8	1900	1000	2.15	0.1	0.6

of range-independent regions and applying Eq. (1) to propagate the field through each region. For a purely solid medium, the vertical interfaces between range-independent regions can be handled by conserving the displacements and the stresses using the single-scattering solution, which is based on the equations

$$\begin{pmatrix} \sigma_{rr} \\ w \end{pmatrix} = R \begin{pmatrix} u_r \\ w \end{pmatrix}, \tag{2}$$

$$\frac{\partial}{\partial r} \begin{pmatrix} u \\ -\sigma_{rz} - \lambda_0 \dfrac{\partial u}{\partial z} \end{pmatrix} = S \begin{pmatrix} u_r \\ w \end{pmatrix}, \tag{3}$$

$$R = \begin{pmatrix} \lambda + 2\mu & \lambda \dfrac{\partial}{\partial z} \\ 0 & 1 \end{pmatrix}, \tag{4}$$

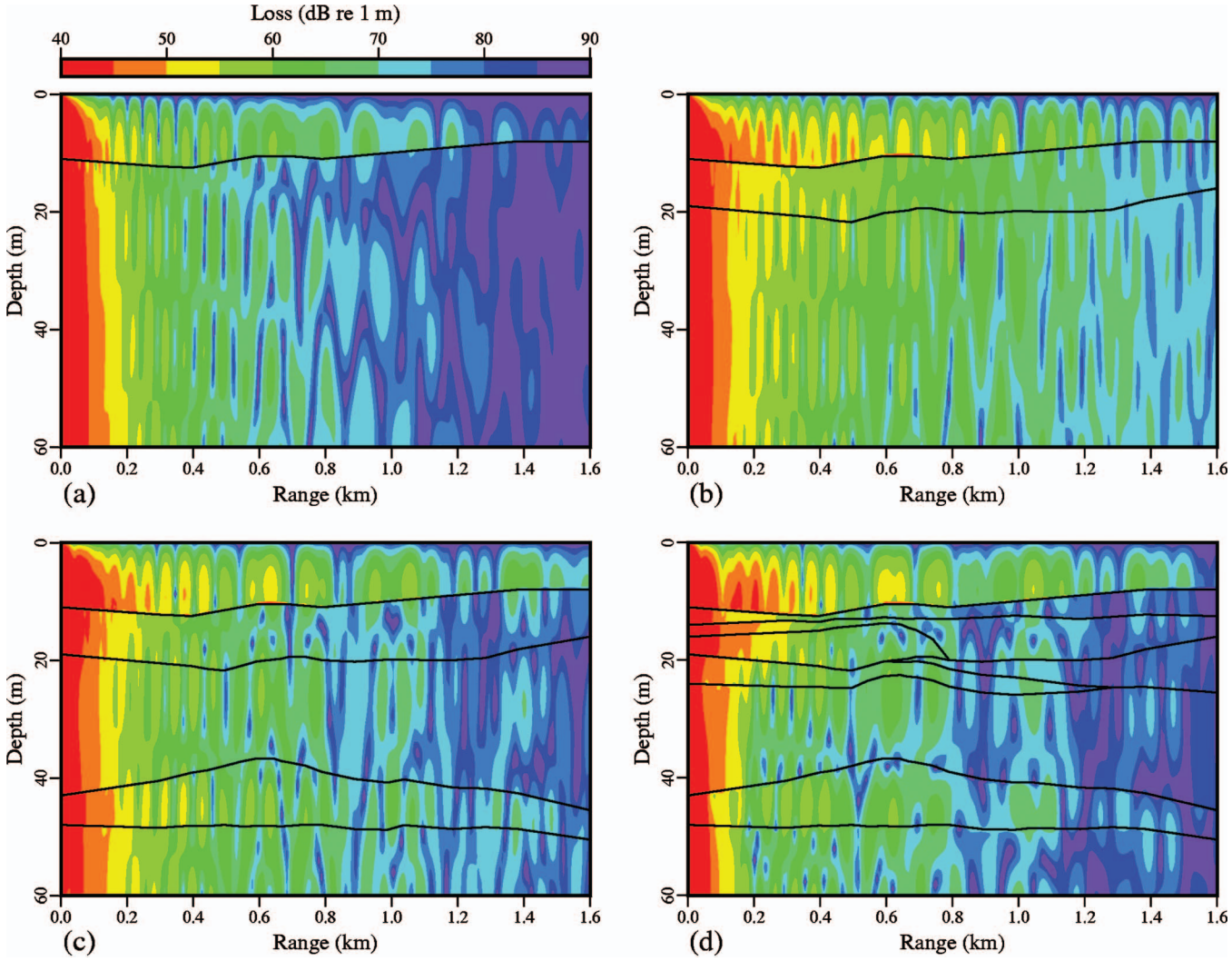


FIG. 4. Compressional transmission loss for example B. Approximations of the sediment structure based on (a) one, (b) two, (c) four, and (d) eight layers.

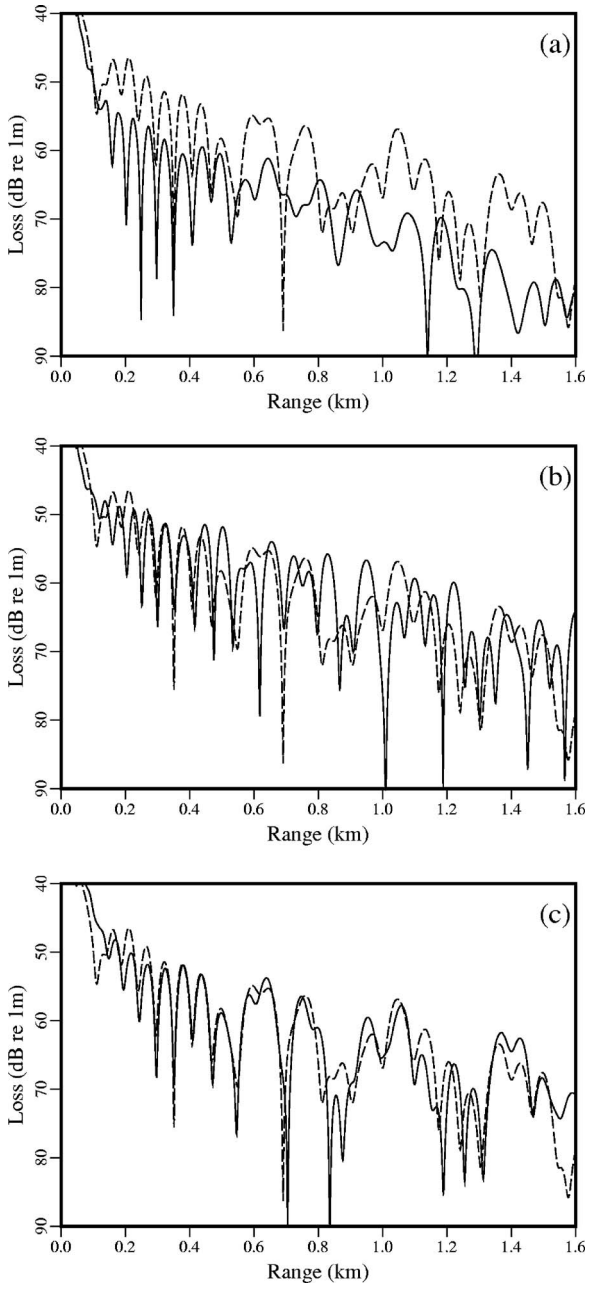


FIG. 5. Compressional transmission loss at $z=5$ m for example B. The dashed curve corresponds to the eight-layer solution. The solid curves correspond to the (a) one, (b) two, and (c) four layer approximations.

$$S = \begin{pmatrix} 1 & 0 \\ (\lambda - \lambda_0)\dfrac{\partial}{\partial z} + \dfrac{\partial \lambda}{\partial z} & \rho\omega^2 + \dfrac{\partial}{\partial z}(\lambda + 2\mu)\dfrac{\partial}{\partial z} \end{pmatrix}, \qquad (5)$$

where λ and μ are the Lamé parameters, the constant λ_0 is a representative value of λ, and σ_{rr} and σ_{rz} are the normal and tangential stresses. The operators R and S relate the quantities that must be conserved across a vertical interface with the dependent variables.

Since σ_{rz} is conserved across both vertical and horizontal interfaces, an auxiliary condition has been added in Eq. (3) in order to avoid singular systems of equations. This condition is obtained by taking the depth derivative of the condition for conservation of u across the vertical interface, which is valid since u is conserved at all depths along the vertical interface. The single-scattering solution is obtained by using the depth operators in Eq. (1) to eliminate the range derivative in Eq. (3) and obtaining an iteration formula for the reflected field in terms of the incident field. We apply a special case of the iteration formula that provides accurate solutions with only one iteration. After obtaining the reflected field, the transmitted field is then used as an initial condition in the next region.

The single-scattering solution is readily generalized to problems involving horizontal interfaces between fluid and solid layers. The condition for conservation of the acoustic pressure p is implemented as part of Eq. (2). The condition for conservation of the particle velocity $\partial p/\partial r$ is implemented as part of Eq. (3). A sloping fluid-solid interface at the ocean bottom is handled by rotating coordinates so that the range direction is tangent to the interface. The ocean bottom is approximated in terms of a series of constant slope regions. When a change in slope is encountered, the solution is propagated slightly beyond the change in slope in order to obtain an initial condition in the next region. The dependent variables are interpolated, extrapolated, and rotated as described in Ref. 8.

III. EXAMPLES

In this section, we apply the variable-rotated single-scattering solution to problems involving variations in bathymetry and layer thicknesses within an elastic bottom. We generate reference solutions using a finite-element model.[10] For each of the examples, the sound speed is 1500 m/s in the water column, and an artificial absorbing layer is used to prevent reflections from the bottom of the computational grid.

For example A, the ocean depth is 500 m at the range of the source and slopes upward at approximately 2.3°. The sediment contains a layer over a half space and we consider two cases. In one case, the sediment thickness is 350 m for all r. In the other case, the sediment thickness is 350 m for $r<2.5$ km and linearly increases to 450 m for 2.5 km$<r<5$ km. Geoacoustic properties of the layers are given in Table I. A 25 Hz source is located at $z=390$ m. In the transmission loss plots appearing in Fig. 1, the significant differences in the water column at long ranges illustrate the influence of sediment thickness on the field. In the plots of transmission loss curves in Fig. 2, we observe that the parabolic equation solutions are in good agreement with the finite-element solution and once again note the differences in the solutions for the two cases.

For example B, we apply the parabolic equation model to a realistic environment involving multiple layers and complex stratigraphy that is based on borehole data from the Atlantic Generating Site (see Fig. 3) on the New Jersey Shelf.[9] This shallow water site has been the location of broadband acoustic experiments over short range tracks.[11–13] Geoacoustic properties of the layers are given in Table II. A 50 Hz source is located at $z=5$ m. For this problem, we consider a series of approximations of the sediment in terms of different numbers of layers. When thicker layers were

formed from multiple thinner layers, the properties were taken to be the average of the constituent layers. In order to test the importance of shear effects in such an environment, the shear speed values were chosen to be greater than those measured at the site. Compressional transmission loss is plotted for cases involving one, two, four, and eight layers in Fig. 4. The major differences between the one and two layer cases are mainly attributable to the fact that there is no interface within the sediment to reflect energy back into the water column in the one layer case, which results in a greater decay rate of amplitude with range. The two and four layer cases have similar amplitudes and are qualitatively similar. The four and eight layer cases are in fair quantitative agreement. The transmission loss curves appearing in Fig. 5 illustrate the differences between the cases more quantitatively.

IV. CONCLUSIONS

The variable rotated parabolic equation and the single-scattering method were combined to obtain an accurate and efficient approach for solving seismo-acoustics problems that involve variations in bathymetry and the thicknesses of elastic sediment layers. The accuracy of the model was demonstrated by making benchmark comparisons with solutions generated using a finite-element model. To illustrate that this parabolic equation solution is applicable to a large class of problems, we presented an example involving complex stratigraphy from the Atlantic Generating Site.

ACKNOWLEDGMENT

This work was supported by the Office of Naval Research, including an ONR Ocean Acoustics Graduate Traineeship Grant to the first author.

[1]F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics* (American Institute of Physics, New York, 1994), pp. 343–412.
[2]R. R. Greene, "A high-angle one-way wave equation for seismic wave propagation along rough and sloping interfaces," J. Acoust. Soc. Am. **77**, 1991–1998 (1985).
[3]M. D. Collins, "A higher-order parabolic equation for wave propagation in an ocean overlying an elastic bottom," J. Acoust. Soc. Am. **86**, 1459–1464 (1989).
[4]B. T. R. Wetton and G. H. Brooke, "One-way wave equations for seismoacoustic propagation in elastic waveguides," J. Acoust. Soc. Am. **87**, 624–632 (1990).
[5]M. D. Collins, "Higher-order Padé approximations for accurate and stable elastic parabolic equations with application to interface wave propagation," J. Acoust. Soc. Am. **89**, 1050–1057 (1991).
[6]W. Jerzak, W. L. Siegmann, and M. D. Collins, "Modeling Rayleigh and Stoneley waves and other interface and boundary effects with the parabolic equation," J. Acoust. Soc. Am. **117**, 3497–3503 (2005).
[7]E. T. Küsel, W. L. Siegmann, and M. D. Collins, "A single-scattering correction for large contrasts in elastic layers," J. Acoust. Soc. Am. **121**, 808–813 (2007).
[8]D. A. Outing, W. L. Siegmann, M. D. Collins, and E. K. Westwood, "Generalization of the rotated parabolic equation to variable slopes," J. Acoust. Soc. Am. **120**, 3534–3538 (2006).
[9]K. P. Bongiovanni, M. Badiey, and W. L. Siegmann, "Shallow-water sediment layer structure and composition effects on range-dependent acoustic propagation at the Atlantic Generating Station (AGS) site," J. Acoust. Soc. Am. **98**, 2249–2261 (1995).
[10]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).
[11]M. Badiey, I. Jaya, and A.H.-D. Cheng, "Shallow water acoustic/geoacoustic experiments near the New Jersey Atlantic Generating Station site," J. Acoust. Soc. Am. **96**, 3593–3604 (1994).
[12]M. Badiey, K. P. Bongiovanni, and W. L. Siegmann, "Interpretation of frequency-dependent transmission-loss interference patterns," IEEE J. Ocean. Eng. **22**, 219–225 (1997).
[13]M. Jaye, M. Badiey, and W. L. Siegmann, "Geoacoustic profile estimation using empirical orthogonal functions for propagation applications," IEEE J. Ocean. Eng. **26**, 795–808 (2001).

Prediction of negative dispersion by a nonlocal poroelastic theory

Abir Chakraborty
India Science Lab, General Motors R & D, Whitefield Road, Bangalore, Karnataka 560066, India

(Received 4 May 2007; revised 27 September 2007; accepted 31 October 2007)

The objective of this work is to show that the negative dispersion of ultrasonic waves propagating in cancellous bone can be explained by a nonlocal version of Biot's theory of poroelasticity. The nonlocal poroelastic formulation is presented in this work and the exact solutions for one- and two-dimensional systems are obtained by the method of Fourier transform. The nonlocal phase speeds for solid- and fluid-borne waves show the desired negative dispersion where the magnitude of dispersion is strongly dependent on the nonlocal parameters and porosity. Dependence of the phase speed and attenuation is studied for both porosity and frequency variation. It is shown that the nonlocal parameter can be easily estimated by comparing the theoretical dispersion rate with experimental observations. It is also shown that the modes of Lamb waves show similar negative dispersion when predicted by the nonlocal poroelastic theory.
 [DOI: 10.1121/1.2816576]

PACS number(s): 43.35.Cg, 43.20.Hq, 43.80.Qf, 43.80.Ev [JJM] Pages: 56–67

I. INTRODUCTION

Quantitative ultrasound has been validated as a useful modality in screening large populations for bone health in general, and osteoporosis in particular. The method relies on two parameters, broadband ultrasonic attenuation (BUA) and/or speed of sound (SOS) in order to assess bone status.[1] There are several studies that document the utility of SOS for this purpose (a comprehensive list is provided by Wear[2]). There are three different measures of wave speed, namely, phase speed (speed of a single sinusoid), group speed (speed of a collection of sinusoids with different frequency content), and signal speed (speed of the front of the wave packet).[3] The SOS measured by ultrasonic technique is closely related to all three of these speed definitions.

The current unsolved problem in the experimentally observed variation of SOS with frequency (called the dispersion relation) is the reduction of the speed with frequency (called negative dispersion). Negative dispersion has been observed in human calcaneus bone for both in vitro[4–8] and in vivo.[2,9] This phenomenon seems to contradict the natural relation of frequency dependent phase velocity and attenuation, called the Kramers–Kronig (KK) relations. For media with an attenuation coefficient increasing linearly with frequency such as bone, the nearly local approximations to the KK relations with one subtraction[10] predicts increase of speed with frequency. On the other hand, the "restricted bandwidth form" of the KK relations has been shown to accurately predict the negative dispersion in bovine cancellous bone.[11] Another explanation of negative dispersion is provided by Marutyan *et al.*[12,13] where it is shown that a combination of fast and slow wave component with positive dispersion can generate a resultant wave of negative dispersion.

There are two theoretical models used for cancellous bone that predict negative dispersion. The first one is the so called "stratified model," which deals with a periodic arrangement of alternate solid and fluid layer. The analysis of wave propagation in periodic layered media is carried out by many researchers.[14–21] Plona *et al.*[22] used the stratified model theory to predict negative dispersion in aluminum/water and plexiglass/water layers and obtained very good agreement with experimental data. As a theory of periodic structures, the stratified model also predicts stop bands in frequency (i.e., regions where no wave propagates) and in turn, the negative dispersion.

The second theory, developed by Lopatnikov and Cheng,[23,24] is based on variational formulation for fluid infiltrated porous bodies. The energy expressions are obtained using macroscopic state parameters obtained from microscopic averaging. This theory treats porosity as an independent variable and predicts a third dilatational wave coming from the evolution of porosity. The theory also predicts decreasing phase speed (in fact, negative phase speed) with frequency and stop band frequencies.

Cancellous (porous) bones are also modeled by Biot's theory of poroelasticity,[25,26] which considers both solid and fluid displacement as unknown variables. The coupling between these two displacement components results in fluid motion in and out of phase with the solid motion. For a one-dimensional (1D) system, the theory predicts two dilatational waves, one is solid borne (fast) wave and the other is fluid borne (slow). The fast wave is the poroelastic counterpart of the P-wave of homogeneous material. The slow wave, also predicted by the stratified theory, has been observed by numerous experiments, thus validating the applicability of Biot's theory. Biot's theory, however, does not predict negative dispersion, stop band frequencies, or negative phase speed as suggested by the two other theories mentioned before.

The hypothesis of the present work is that a nonlocal extension of Biot's theory can accomodate these desired characteristics. The nonlocal theory of elasticity[27] takes into consideration the intrinsic length scale of a material gener-

ated by a repeated atomic or molecular structure. By virtue of this consideration, several discrepancies of classical (local) theory, e.g., infinite stress field near the crack tip, the nondispersive nature of Rayleigh wave, etc., can be explained by the nonlocal theory. The nonlocal theory is particularly useful in predicting the realistic dispersion relation, e.g., the nonlocal dispersion relation $\omega/c_1k=(1+\epsilon^2k^2)^{-1/2}$ (ϵ = nonlocality parameter, k= wave number, ω= frequency, and c_1= characteristic wave speed) closely matches the Born–Karman model dispersion $wa/c_1=2\sin(ka/2)$ for $\epsilon=0.39a$, where a is the lattice spacing. Thus, by suitably choosing the nonlocality parameter, dispersion relations of periodic structures (with band gap frequencies) can be captured within the realm of continuum mechanics.

The nonlocality parameter is dependent on the intrinsic length scale of the structure. For porous materials like cancellous bone, there are several length scales involved in terms of mean trabecular spacing, length, or thickness. A nonlocal theory of poroelasticity for trabecular bone, once developed, can be correlated to any of these length scales for future applications in the diagnosis of osteoporosis. The nonlocality parameter can also be related to the length scale introduced in the theory of Lopatnikov and Cheng.[24]

In this work, a nonlocal extension of Biot's theory of poroelasticity is presented. The nonlocal model has only one parameter ϵ, which can be estimated by comparing the theoretical dispersion rate with experimental observation. It is shown that the nonlocal poroelastic theory has a close resemblance to the averaging based theory of Lopatnikov and Cheng,[24] where a reduction of the effective elastic modulus with frequency is obtained. This reduction is responsible for negative dispersion. As an extension of Biot's theory, the nonlocal poroelastic theory enjoys the dependence of its elastic parameters on porosity. Thus, another important outcome of this analysis is the variation of negative dispersion with porosity. The governing equations of the nonlocal poroelasticity are solved by the method of integral transform where Fourier transform for the time variable and Fourier series for the spatial variable are employed. Thus, the exact solution for one- and two-dimensional (2D) layered media is obtained, which can be used to capture the effect of nonlocality on time domain signals. The formulation for 2D layered media also paves the way for analysis of Lamb wave modes, and their dependence on nonlocality is investigated.

II. MATHEMATICAL FORMULATION

According to Biot's theory of poroelasticity, the solid stresses σ_{ij} and fluid stress $s=-\beta p$, (where p is the fluid pressure and β is the porosity) are related to the solid strains ϵ_{ij} and fluid dilatation γ by the relation

$$\sigma_{ij}=2N\epsilon_{ij}+(A\epsilon_{kk}+Q\gamma)\delta_{ij}, \tag{1}$$

$$s=Q\epsilon_{kk}+R\gamma, \tag{2}$$

where ϵ_{kk} denotes the trace of the solid strain tensor (dilatational part) and A, N, Q, and R are the four porosity (β) dependent material parameters. Among these parameters, A and N are the familiar Lame's parameters and the coefficient Q represents the coupling between the solid and fluid. R is a measure of the pressure required on the fluid to force a certain volume of the fluid into the aggregate while the total volume remains constant.[28] The solid and fluid strains are defined in terms of the solid displacement u_i and fluid displacement U_i as:

$$\epsilon_{ij}=l/2(u_{i,j}+u_{j,i}),\quad \gamma=U_{i,i}, \tag{3}$$

where a comma in the subscript denotes differentiation with respect to the following spatial variable. The governing equations of poroelasticity are

$$\sigma_{ij,j}=\rho_{11}\ddot{u}_i+\rho_{12}\ddot{U}_i+b(\dot{u}_i-\dot{U}_i),$$

$$s_{,i}=\rho_{12}\ddot{u}_i+\rho_{22}\ddot{U}_i-b(\dot{u}_i-\dot{U}_i), \tag{4}$$

where a dot over a quantity represents differentiation with respect to time, t. The coefficient b is related to the Darcy's coefficient of permeability κ, fluid viscosity μ, and porosity β by the relation $b=\mu\beta^2/\kappa$. The mass coefficients ρ_{11}, ρ_{12}, and ρ_{22} are the densities which take into account the fact that the relative fluid flow through the pores is not uniform. The relative importance and characteristics of these parameters are discussed in detail by Biot.[25]

In the theory of nonlocal elasticity, the stresses at a point in the continuum are not only dependent on the strains at that point but also on the strain distribution in a neighborhood Ω about that point, i.e.,

$$\begin{Bmatrix}\Sigma_{ij}\\ S\end{Bmatrix}(\mathbf{x})=\int_\Omega C(|\mathbf{x}-\xi|)\begin{Bmatrix}\epsilon_{ij}\\ \gamma\end{Bmatrix}(\xi)d\Omega(\xi), \tag{5}$$

where $C(|\mathbf{x}-\xi|)$ is a spatially dependent constitutive relation of poroelasticity defined in terms of the local part and functional term as

$$C(|\mathbf{x}-\xi|)=C_0\alpha(|\mathbf{x}-\xi|). \tag{6}$$

The kernel function α is form invariant under arbitrary spatial translations and rotations, i.e., α relates $\mathbf{x}$ with ξ only through the distance $|\mathbf{x}-\xi|$. The kernel function satisfies the following properties:[27]

(i) $\alpha(r)$ is a continuous function of r, with a bounded support Ω, where $\alpha>0$ inside the boundary $\partial\Omega$ and $\alpha=0$ outside;
(ii) $\alpha(r)$ satisfies the normality condition:

$$\int_\Omega \alpha(r)dr=1, \tag{7}$$

(iii) Additionally, $\alpha(r)$ is a Green function of a linear differential operator $\mathcal{L}=1-\epsilon^2\nabla^2$ (∇^2 denotes the Laplacian), i.e., $\mathcal{L}\alpha(|\mathbf{r}-\mathbf{r}'|)=\delta(|\mathbf{r}-\mathbf{r}'|)$, where δ is the Dirac delta function. For 2D solids of infinite extent $\alpha(|\mathbf{r}-\mathbf{r}'|)=(2\pi\epsilon^2)^{-1}K_0(|\mathbf{r}-\mathbf{r}'|/\epsilon)$, where K_0 is the modified Bessel function of the second kind.

If the last condition is substituted in Eq. (5), we get the relation between the nonlocal stress $\{\Sigma_{ij},S\}$ and the classical form of the stress $\{\sigma_{ij},s\}$ as

$$\mathcal{L}\begin{Bmatrix}\Sigma_{ij}\\ S\end{Bmatrix}(\mathbf{x}) = C_0\begin{Bmatrix}\epsilon_{ij}\\ \gamma\end{Bmatrix}(\mathbf{x}) = \begin{Bmatrix}\sigma_{ij}\\ s\end{Bmatrix}(\mathbf{x}), \tag{8}$$

and the governing equation can be modified to

$$\sigma_{ij,j} = \mathcal{L}[\rho_{11}\ddot{u}_i + \rho_{12}\ddot{U}_i + b(\dot{u}_i - \dot{U}_i)],$$

$$s_{,i} = \mathcal{L}[\rho_{12}\ddot{u}_i + \rho_{22}\ddot{U}_i - b(\dot{u}_i - \dot{U}_i)]. \tag{9}$$

At this point, we consider the motion in the x_1-x_3 plane, i.e., motions are invariant of the x_2 direction, (i.e., $u_2=0=U_2$) and apply Helmholtz decomposition to the displacement field

$$u_1 = \phi_{1,1} - \psi_{1,3}, \quad U_1 = \phi_{2,1} - \psi_{2,3},$$

$$u_3 = \phi_{1,3} + \psi_{1,1}, \quad U_3 = \phi_{2,3} - \psi_{2,1}, \tag{10}$$

where ϕ_1, ϕ_2 are the dilatational potentials and ψ_1, ψ_2 are the rotational potentials. Assuming time harmonic solution for the potentials, i.e.,

$$\{\phi,\psi\}_{1,2}(x_1,x_3,t) = \{\hat{\phi},\hat{\psi}\}_{1,2}(x_1,x_3,\omega)e^{I\omega t}, I^2 = -1. \tag{11}$$

The stresses in terms of the potentials take the form

$$\sigma_{11} = P(\phi_{1,11} - \psi_{1,13}) + A(\phi_{1,33} + \psi_{1,13}) + Q\nabla^2\phi_2,$$

$$\sigma_{33} = P\phi_{1,33} + A\phi_{1,11} + 2N(\psi_{1,13}) + Q\nabla^2\phi_2,$$

$$\sigma_{13} = N(2\phi_{1,13} + \psi_{1,11} - \psi_{1,33}),$$

$$s = Q\nabla^2\phi_1 + R\nabla^2\phi_2, \tag{12}$$

where $P=A+2N$. Similarly, the governing equations in terms of the potentials become

$$\bar{P}\nabla^2\hat{\phi}_1 + \bar{Q}\nabla^2\hat{\phi}_2 + \omega^2(M_{11}\hat{\phi}_1 + M_{12}\hat{\phi}_2) = 0, \tag{13}$$

$$\bar{Q}\nabla^2\hat{\phi}_1 + \bar{R}\nabla^2\hat{\phi}_2 + \omega^2(M_{12}\hat{\phi}_1 + M_{22}\hat{\phi}_2) = 0, \tag{14}$$

$$\bar{N}\nabla^2\hat{\psi}_1 + \omega^2(M_{11}\hat{\psi}_1 + M_{12}\hat{\psi}_2) = 0, \tag{15}$$

$$\omega^2(M_{12}\hat{\psi}_1 + M_{22}\hat{\psi}_2) = 0, \tag{16}$$

where $M_{ij}=\rho_{ij}-(-1)^{i+j}Ib/\omega$, $\bar{P}=P-\epsilon^2\omega^2M_{11}$, $\bar{Q}=Q-\epsilon^2\omega^2M_{12}$, and $\bar{R}=R-\epsilon^2\omega^2M_{22}$ are the reduced moduli. These equations are of the same form as the original Biot's equations with the exception in the definitions of the elastic moduli. Here, the effect of nonlocality is manifested in the reduction of the moduli, which is also observed by Lopatnikov and Cheng[24] as an effect of the porosity dynamics.

Equation (16) used to express $\hat{\psi}_2$ in terms of $\hat{\psi}_1$ as $\hat{\psi}_2=(-M_{12}/M_{22})\hat{\psi}_1=r\hat{\psi}_1$, which on substitution to Eq. (15), yields

$$\bar{N}\nabla^2\hat{\psi}_1 + \omega^2(M_{11} + rM_{12})\hat{\psi}_1 = 0. \tag{17}$$

Denoting the phase speed of the shear wave by V_3, the Helmholtz equation for $\hat{\psi}_1$ can be written as

$$V_3^2\nabla^2\hat{\psi}_1 + \omega^2\hat{\psi}_1 = 0, \quad V_3^2 = \bar{N}/(M_{11} + rM_{12}). \tag{18}$$

Assuming harmonic solution in the x_1 direction, say, $\sin(\eta x_1)$ (where η is the wave number in the x_1 direction) it is evident that the x_3 dependency can be written as an exponential function, $\exp(Iqz)$ so that $q^2=\omega^2/V_3^2-\eta^2$. The complete solution for ψ_1 becomes

$$\psi(x_1,x_3,t) = \sum_{n=0}^{N-1}\left[\sum_{m=1}^{M}\{C_{1mn}e^{-Iq_{mn}x_3} + C_{2mn}e^{+Iq_{mn}x_3}\}\sin(\eta_m x_1)\right]e^{I\omega_n t}, \tag{19}$$

where it is assumed that to completely resolve the time and space dependency of ψ_1, N frequency steps and M wave number steps are required. Assuming a similar form for $\hat{\phi}_1$ and $\hat{\phi}_2$

$$\{\hat{\phi}_1,\hat{\phi}_2\}(x_1,x_3,\omega) = \{\tilde{\phi}_1,\tilde{\phi}_2\}e^{-Ikx_3}\cos(\eta x_1), \tag{20}$$

and substituting into Eqs. (13) and (14), the algebraic eigenvalue problem for each ω_n and η_m becomes $[\mathbf{W}]_{nm}\{\tilde{\phi}_1,\tilde{\phi}_2\}^T=0$ where

$$\mathbf{W}_{nm} = \begin{bmatrix}\omega_n^2M_{11} - (k^2+\eta_m^2)\bar{P}, & \omega_n^2M_{12} - (k^2+\eta_m^2)\bar{Q}\\ \omega_n^2M_{12} - (k^2+\eta_m^2)\bar{Q}, & \omega_n^2M_{22} - (k^2+\eta_m^2)\bar{R}\end{bmatrix} \tag{21}$$

For nontrivial solution of $\tilde{\phi}_1$ and $\tilde{\phi}_2$, the determinant of $\mathbf{W}_{nm}$ should be zero, which generates the required characteristic equation to solve for wave number k

$$(\bar{P}\bar{R} - \bar{Q}^2)\xi^4 + \omega^2(2\bar{Q}M_{12} - \bar{R}M_{11} - \bar{P}M_{22})\xi^2 + \omega^4(M_{11}M_{22} - M_{22}^2) = 0, \tag{22}$$

where $\xi^2=k^2+\eta^2$. The solution of Eq. (22) generates the four roots of k, which occur in pairs and are written as $\pm k_1$ and $\pm k_2$. As a result, the solutions for $\tilde{\phi}_1$ and $\tilde{\phi}_2$ become

$$\tilde{\phi}_1(x_3,\eta_m,\omega_n) = \sum_{\alpha=1}^{4}R_{1\alpha}A_{\alpha mn}e^{-1k\alpha mnx_3}, \tag{23}$$

$$\tilde{\phi}_2(x_3,\eta_m,\omega_n) = \sum_{\alpha=1}^{4}R_{2\alpha}A_{\alpha mn}e^{-1k\alpha mnx_3}. \tag{24}$$

It should be noted that the unknown coefficients are kept the same for both ϕ_1 and ϕ_2, i.e., $A_{\alpha mn}$, where the elements of the matrix $\mathbf{R}$ control the linear dependency of the two solutions. The columns of this matrix are obtained (for each wave number k_j) from the relation

$$\begin{bmatrix}W_{11}(k_j), & W_{12}(k_j)\\ W_{12}(k_j), & W_{22}(k_j)\end{bmatrix}\begin{Bmatrix}R_{1j}\\ R_{2j}\end{Bmatrix} = \begin{Bmatrix}0\\ 0\end{Bmatrix}. \tag{25}$$

The complete solution of the potentials are

$$\phi_1(x_1,x_3,t) = \sum_{n=0}^{N-1}\sum_{m=1}^{M}\tilde{\phi}_1(x_3,\eta_m,\omega_n)\cos(\eta_m x_1)e^{-I\omega_n t}, \tag{26}$$

$$\phi_2(x_1,x_3,t) = \sum_{n=0}^{N-1}\sum_{m=1}^{M} \tilde{\phi}_2(x_3,\eta_m,\omega_n)\cos(\eta_m x_1)e^{-I\omega_n t}. \quad (27)$$

These solutions with the relations given in Eq. (10) generate the expressions of the solid and fluid displacements, which is necessary to impose the Dirichlet boundary conditions for a layer media.

The relation between the local and global forms of the stresses is[27]

$$\{\Sigma_{ij},S\} = (1+\epsilon^2\nabla^2)\{\sigma_{ij},s\}. \quad (28)$$

Based on the assumption of the displacement (potential) field the above relation becomes

$$\{\Sigma_{ij},S\} = [1-\epsilon^2(\eta_m^2+k_{mn}^2)]\{\sigma_{ij},s\} = \frac{\{\sigma_{ij},s\}}{1+\epsilon^2(\eta_m^2+k_{mn}^2)}, \quad (29)$$

where k_{mn} is any of the four wave numbers at frequency ω_n and wave number η_m and the last relation is possible because $\epsilon \ll 1$. It should be noted that Eq. (29) can also be obtained by integrating the kernel function[29] when α is assumed as

$$\alpha(|\mathbf{x}-\xi|) = \frac{1}{2\pi\epsilon^2}K_0(\epsilon^{-1}\sqrt{(x_1-\xi_1)^2+(x_3-\xi_3)^2}). \quad (30)$$

The integral to be evaluated is

$$\frac{1}{2\pi\epsilon^2}\int_\Omega \alpha(|\mathbf{x}-\xi|)f(\mathbf{x},\xi)d\xi = \frac{f(\mathbf{x})}{1+\epsilon^2(\eta_m^2+k_{mn}^2)}, \quad (31)$$

where

$$f(\mathbf{x}) = \left\{\begin{matrix}\sin(\eta_m x_1)\\ \cos(\eta_m x_1)\end{matrix}\right\} \times \exp(-jk_{mn}x_3)$$

and the expression of $\{\Sigma_{ij},S\}$ obtained is the same as that of Eq. (29).

The nonlocal stress boundary conditions are expressed in terms of the potentials by the relations given in Eq. (12). The harmonic dependency in the x_1 direction implicitly assumes periodicity of the solution in this direction. To avoid any undesired effect of this assumption, the x_1 dimension of the model must be to be taken to be quite large compared to the other (x_3) dimension. Hence, the solutions obtained so far [Eqs. (19), (26), and (27)] are essentially that of a layered media having two lateral boundaries, say at $x_3=0$ and $x_3=L$. The Neumann boundary conditions at these edges are the specifications of the tractions and fluid pressure, which act normal to the surface. The tractions in the x_1 and x_3 directions are related to the stresses by Cauchy's principle

$$t_1 = \Sigma_{11}n_1 + \Sigma_{13}n_3, t_3 = \Sigma_{13}n_1 + \Sigma_{33}n_3, \quad (32)$$

where n_1 and n_3 are the components of the surface normal. In the present case, where the boundaries are parallel to the x_1 axis, $n_1=0$ and $n_3 = = \mp 1$. Thus, the tractions are specified only in terms of the shear stress Σ_{13} and normal stress Σ_{33}. All together, there are three boundary conditions, $\mp\Sigma_{13}$, $\mp\Sigma_{33}$, and $\mp S$ that can be prescribed at each edge. However, when it comes to displacement boundary conditions, there are four components: two solid displacements and two fluid displacements. This ambiguity could be avoided if the boundary conditions are specified in terms of the potentials (since there are only three independent potentials). Thus, it is necessary to drop one displacement component. In the present layer model, the U_1 component of the fluid is not considered while imposing the boundary conditions.

The other three displacement components u_1, u_3, and U_3 can be evaluated at the two edges of the layer to relate to the unknown constants $\{A_1,A_2,A_3,A_4\}$ and $\{C_1,C_2\}$. As a matter of convenience, we consider the boundaries at $x_3=0$ and $x_3=L$ (defining L as the thickness of the layer). Then the displacements are related to the unknown constants by the relation

$$u_1(x_3=0)=v_1, \quad u_3(x_3=0)=w_1,$$

$$u_1(x_3=L)=v_2, \quad u_3(x_3=L)=w_2,$$

$$U_3(x_3=0)=W_1, \quad U_3(x_3=L)=W_2, \quad (33)$$

which can be written in concise form as

$$\{\hat{u}\} = \{v_1,w_1,W_1,v_2,w_2,W_2\}^T = [\mathbf{T}_1]\{a\}^T, \quad (34)$$

where $\{a\}=\{A_1,A_2,A_3,A_4,C_1,C_2\}$. Similarly, the stress boundary conditions can be expressed in terms of the constants as

$$-\Sigma_{13}(x_0=0)=t_{11}, \quad -\Sigma_{33}(x_3=0)=t_{31},$$

$$+\Sigma_{13}(x_3=L)=t_{12}, \quad +\Sigma_{33}(x_3=L)=t_{32},$$

$$-S(x_3=0)=s_1, \quad +S(x_3=L)=s_2,$$

which again can be written as

$$\{\hat{t}\} = \{t_{11},t_{31},s_1,t_{12},t_{32},s_2\}^T = [\mathbf{T}_2]\{a\}^T. \quad (35)$$

Combining Eqs. (34) and (35), the stresses are related to the edge displacements by the frequency-wave number domain based "stiffness" matrix $\mathbf{K}$

$$\{\hat{t}\} = [\mathbf{T}_2][\mathbf{T}_1]^{-1}\{\hat{u}\} = [\mathbf{K}]\{\hat{u}\}. \quad (36)$$

Equation (36) represents the algebraic form of the governing partial differential equations described by Eq. (4). It is worth noting that this form is quite generalized and is not particular to any set of boundary conditions. Thus, different solutions for different boundary conditions can be obtained from this single equation. Also, Eq. (36) represents the system at frequency ω_n and wave number η_m. To obtain the complete solutions, this equation needs to be solved $M \times N$ times. The discrete values of the horizontal wave number, η_m is related to the x_1 window length X_L and M by

$$\eta_m = 2\pi(m-1)/X_L = 2\pi(m-1)/M\Delta x_1. \quad (37)$$

The window length is dictated by the geometry of the structure to be analyzed and M is dictated by the spatial variation of the applied stress in the x_1 direction.

A. Reduction to one-dimensional system

Further simplification to the governing system of equations is possible if we are primarily interested in the bulk

waves propagating through ID porous media. The unknown variables are the solid and fluid displacements in the x_1 direction, u_1 and U_1. The relevant stresses are

$$\sigma_{11} = Pu_{1,1} + QU_{1,1}, \quad s = Qu_{1,1} + RU_{1,1}. \tag{38}$$

The governing equations are simplified to

$$Pu_{1,11} + QU_{1,11} = \rho_{11}(\ddot{u}_1 - \epsilon^2\ddot{u}_{1,11}) + \rho_{12}(\ddot{U}_1 - \epsilon^2\ddot{U}_{1,11}) + b(\dot{u}_1 - \epsilon^2\dot{u}_{1,11}) - b(\dot{U}_1 - \epsilon^2\dot{U}_{1,11}), \tag{39}$$

$$Qu_{1,11} + RU_{1,11} = \rho_{12}(\ddot{u}_1 - \epsilon^2\ddot{u}_{1,11}) + \rho_{22}(\ddot{U}_1 - \epsilon^2\ddot{U}_{1,11}) - b(\dot{u}_1 - \epsilon^2\dot{u}_{1,11}) + b(\dot{U}_1 - \epsilon^2\dot{U}_{1,1}). \tag{40}$$

Assuming the general form of the solutions

$$u_1(x_1,t) = u_0 e^{I(\omega t - kx_1)}, \quad U_1(x_1,t) = U_0 e^{I(\omega t - kx_1)}, \tag{41}$$

and substituting in the governing equations, the algebraic eigenvalue problem becomes

$$\begin{bmatrix} \omega^2 M_{11} - k^2\bar{P}, & \omega^2 M_{12} - k^2\bar{Q} \\ \omega^2 M_{12} - k^2\bar{Q}, & \omega^2 M_{22} - k^2\bar{R} \end{bmatrix} \begin{Bmatrix} u_0 \\ U_0 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}, \tag{42}$$

where M_{ij}, $\bar{P}$, $\bar{Q}$, and $\bar{R}$ are as defined before. Following the same argument of nontrivial solution for u_0 and U_0, the characteristic equation for k is obtained, which has the same form as given in Eq. (22). The only difference in this case is that the other wave number η is zero. Rewriting the characteristic equation as

$$a_0 k^4 + b_0 k^2 \omega^2 + c_0 \omega^4 = 0, \tag{43}$$

the roots become

$$k^2 = \omega^2 \left(\frac{-b_0 \pm \sqrt{b_0^2 - 4a_0c_0}}{2a_0} \right). \tag{44}$$

As the phase speed is defined as $V = \omega/\mathrm{Re}(k)$, the expressions of the fast and slow phase speeds are readily obtained as

$$V_{1,2} = \mathrm{Re}\left[\left(\frac{2a_0}{-b_0 \pm \sqrt{b_0^2 - 4a_0c_0}} \right)^{1/2} \right], \tag{45}$$

where Re( ) denotes the real part of a complex number. Similarly, the imaginary part of the roots of Eq. (44) provides the attenuation, which can be written explicitly for the fast and slow waves as

$$\mathcal{A}_{1,2} = \mathrm{Im}\left[\left(\frac{-b_0 \pm \sqrt{b_0^2 - 4a_0c_0}}{2a_0} \right)^{1/2} \right] \omega, \tag{46}$$

where Im( ) indicates the imaginary part of a complex number. It should be noted that the wave numbers become complex quantities due to the introduction of dynamic tortuosity.[30]

The complete solution of the displacement field is written as

TABLE I. Material properties of trabecular bone considered in this study.

Parameter	Value	Parameter	Value
E_s	20.0 GPa	ν_s	0.32
ρ_s	1960 kg/m^3	ν_b	0.32
ρ_f	930 kg/m^3	n	1.32
K_f	2.2 GPa	τ	0.25
Λ	5.0×10^{-6} m	μ	1.0×10^{-3} N s/m^2

$$u_1(x_1,t) = \sum_{n=0}^{N-1} \left(\sum_{\alpha=1}^{4} R_{1\alpha} A_{\alpha n} e^{-Ik_{\alpha n}x_1} \right) e^{I\omega_n t}, \tag{47}$$

$$U_1(x_1,t) = \sum_{n=0}^{N-1} \left(\sum_{\alpha=1}^{4} R_{2\alpha} A_{\alpha n} e^{-Ik_{\alpha n}x_1} \right) e^{I\omega_n t}, \tag{48}$$

where the elements of $\mathbf{R}$ satisfy a similar relation to Eq. (25). Using the stress-strain relations of Eq. (38), the stresses are related to the unknown constants $\{A_1, A_2, A_3, A_4\}$. Following the same procedure as outlined earlier, the displacements and stresses are evaluated at $x=0$ or $x=L$ to obtain the matrices $\mathbf{T}_1$ and $\mathbf{T}_2$, and in turn the stiffness matrix.

III. VARIATION OF PHASE SPEED

First we evaluate the effect of nonlocality on the dispersion of slow and fast wave components. The material properties of trabecular bone are taken from the work of Williams[28] and listed in Table I. The elastic parameters of Biot, P, Q, R, and N are expressed in terms of Young's modulus (E_s) and Poisson's ratio (ν_s) of the solid phase, bulk modulus (K_f) of the fluid phase, structural parameter τ, Poisson's ratio of the skeletal frame (ν_b), porosity (β), and an exponent n. Similarly, the inertial parameters ρ_{11}, ρ_{12}, and ρ_{22} are related to the density of solid (ρ_s) and fluid (ρ_f), pore size parameter(Λ), fluid viscosity (μ), structural parameter, and porosity. The expressions of these parameters can be found in the literature[28,30] and are not reproduced here.

These parameters along with Eq. (45) are utilized to obtain the variations of the phase speeds with frequency. The nonlocal theory based phase speed variation is compared with the stratified model based predictions. The dispersion relation for the stratified model can be found in Wear.[2] For this model the data presented in Table III of Wear[2] is considered with marrow as fluid and $BV/TV=0.083$. Figure 1 shows the dispersion of the slow wave for different values of the nonlocal parameter ϵ (dashed lines) along with the stratified model prediction (solid line). For $\epsilon=0$, in absence of any dissipation ($b=0$), the classical theory of Biot does not show any reduction of phase speed. However, for nonzero ϵ, the nonlocal theory predicts a negative dispersion. The dispersion rate is defined as $10(V_{600}-V_{300})/3$, where V_{300} and V_{500} are the phase speeds (in m/s) at 300 and 600 kHz, respectively. For the four different values of $\epsilon=5$, 5.5, 6, and 7 $\times10^5$ m, the dispersion rates obtained are −30.5, −37.0, −44.1, and −60.2 m/s/MHz, respectively. The stratified model predicts a dispersion rate of −33.0 m/s/MHz. For the same frequency bandwidth (300–600 kHz), Wear[2] obtained

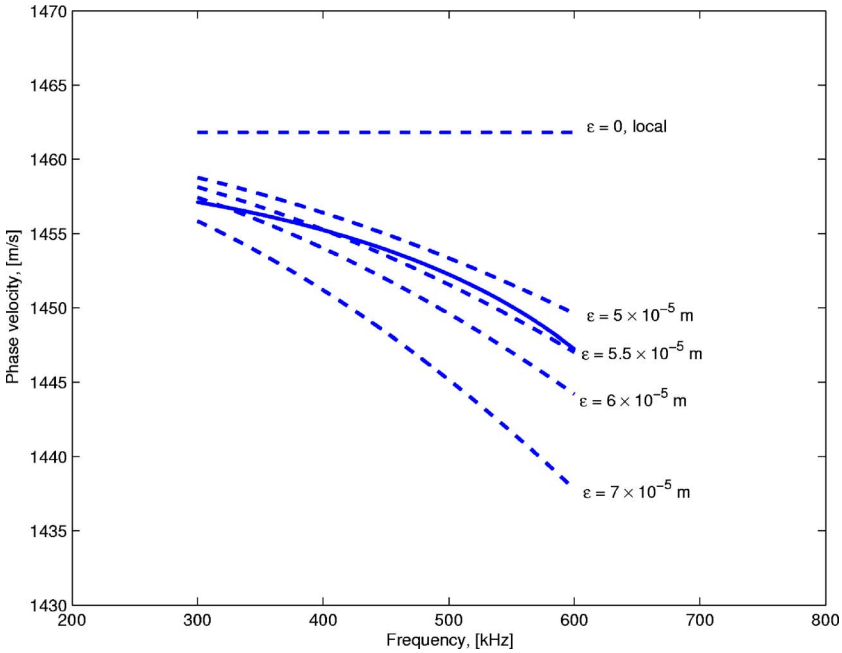


FIG. 1. (Color online) Nonlocal (dashed) and stratified (solid) theory based predictions of phase speed for trabecular bone. For $\epsilon=5.5\times10^{-5}$ m, the nonlocal dispersion matches closely with the stratified model prediction in the frequency range of 400–600 kHz.

a dispersion rate of -59 ± 52 m/s/MHz. Thus, for $\epsilon=7\times10^{-5}$ m the dispersion rate is closest to that observed experimentally.

On the other hand, the speed variation predicted by the nonlocal theory for $\epsilon=5.5\times10^{-5}$ m is the closest to the stratified model[2] between 400 and 600 kHz. However, for frequencies close to 300 kHz, the stratified model is close to the $\epsilon=6\times10^{-5}$ m curve. Thus, the general trend of phase speed variations predicted by these theories is different. This can be attributed to the differences in the formulation of these theories. The stratified theory is much more simplified in its construction than the nonlocal poroelastic theory and requires fewer parameters. The nonlocal theory, on the other hand, utilizes all the Biot's parameters along with the nonlocal parameter (ϵ), which imposes the periodicity of the stratified model. It has been observed earlier that the experimental data of the phase speed vary approximately linearly with frequency between 300 and 700 kHz, whereas the stratified model predicts somewhat nonlinear behavior over this range.[31] The same phenomenon can also be observed here for the stratified model where the nonlocal theory shows almost linear dispersion.

To have comparable phase speed magnitudes with that obtained by Wear,[2] the porosity considered in the previous case is 99%. Next, we study the variation of the dispersion rate with porosity as well as the nonlocal parameter. For ϵ varying between 3 and 7×10^{-5} m and porosity varying between 1% and 99%, the dispersion rate is plotted in Figs. 2 and 3 for the fast and slow wave component, respectively. Comparing these figures, we can readily see that the disper-

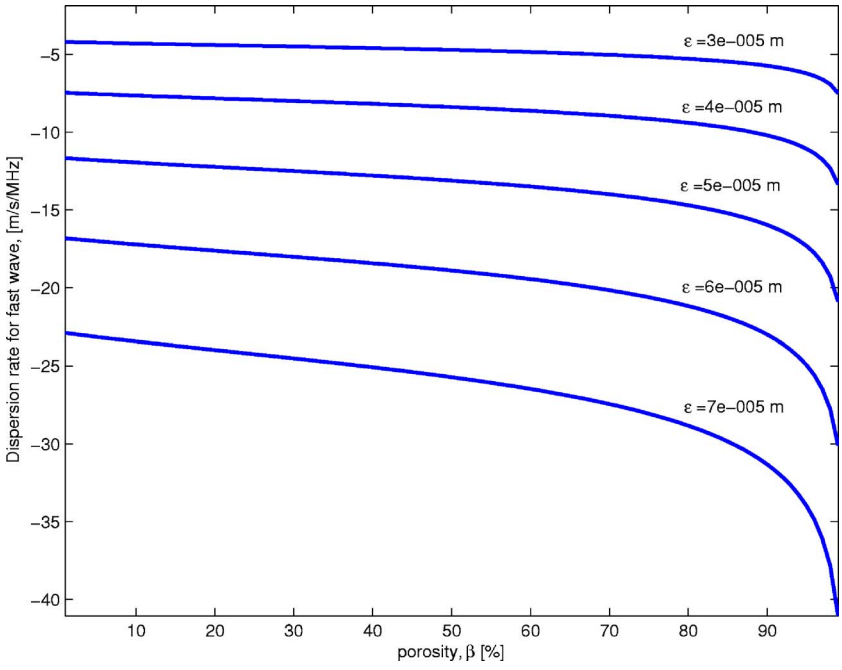


FIG. 2. (Color online) Nonlocal theory prediction of dispersion rate for fast wave for different values of the nonlocal parameter ϵ.

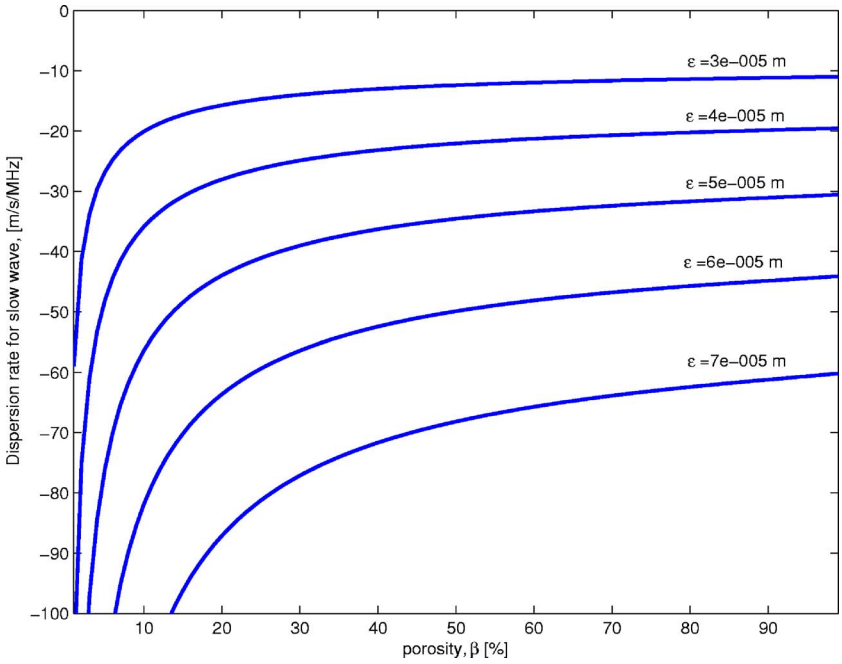


FIG. 3. (Color online) Nonlocal theory prediction of dispersion rate for slow wave for different values of the nonlocal parameter ϵ.

sion rate is much higher for the slow wave compared to the fast wave component. Moreover, at higher porosity, the fast wave shows greater dispersion, whereas for the slow wave the same is true at lower porosity.

Next, the variations of the phase speeds and attenuation with frequency and porosity are studied. The nonlocality parameter is fixed at 6×10^{-5} m. Figure 4 shows the variation of the fast and slow wave speeds with frequency up to 3 MHz. The speed plots are normalized with their maximum values to stress the relative dispersion of the slow and fast waves. The figure reiterates the fact that negative dispersion is greater for slow waves and with increasing porosity fast wave speed decreases and slow wave speed increases.

To study the variation of attenuation [Eq. (46)] with frequency, two different porosity ranges are considered (ϵ fixed at 6×10^{-5} m). For a frequency range of 200 kHz–3 MHz, Fig. 5 shows the variation of attenuation for porosity varying between 5% and 15% and Fig. 6 shows the variation for porosity varying from 75% to 92%. Also shown in the same figures are the predictions by the classical Biot theory ($\epsilon=0$). For all porosity, the classical theory predicts a linear rise of attenuation with frequency for both the fast and slow waves. The nonlocal theory based predictions of attenuation for slow waves do not deviate much from the classical theory predictions, especially at lower porosity. However, the attenuation variations for the fast waves show considerable departure once the frequency exceeds 0.5 MHz. The rapid (nonlinear) increase of the fast wave attenuation, as predicted by the nonlocal theory, is also observed experimentally by

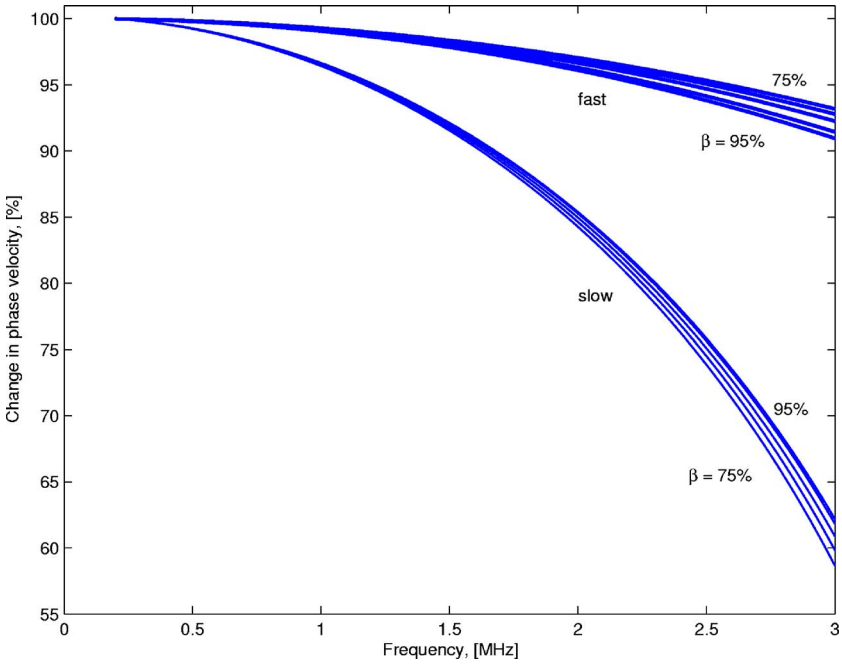


FIG. 4. (Color online) Nonlocal theory prediction of dispersion for slow and fast waves with porosity varying between 75% and 95% ($\epsilon=6.0\times10^{-5}$ m).

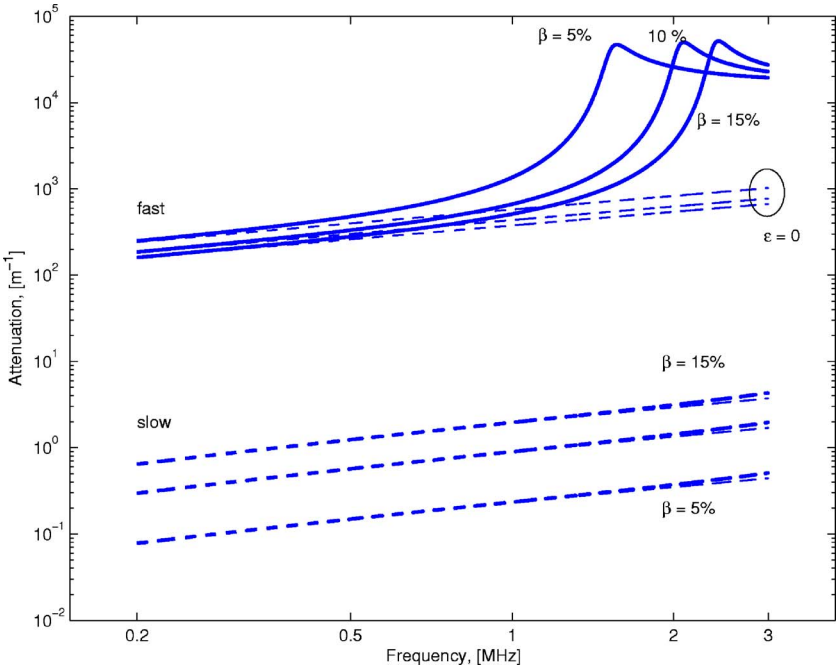


FIG. 5. (Color online) Prediction of attenuation by the classical Biot's theory (thin dashed lines) and the nonlocal poroelastic theory (thick solid and dashed lines, $\epsilon = 6.0 \times 10^{-5}$ m) for porosity varying between 5% and 15%.

Kaczmarek *et al.*[8] and Hosokawa and Otani.[32] Further, it can be seen from Figs. 5 and 6 that with increasing porosity, the attenuation increases considerably for the slow waves, whereas, relatively less variation is observed for the fast waves.

IV. VARIATION OF LAMB WAVE MODES

One important outcome of the previously outlined formulation is the ability of capturing the modes of the Lamb wave, which are important for ultrasonic characterization of materials. By definition, Lamb waves are generated in a doubly bounded media when the edges are traction free. This traction-free criterion is equivalent to solving Eq. (36) for the edge displacement field when the edge tractions are zero. For the nontrivial solution of the displacement field, the determinant of the matrix $\mathbf{K}$, or equivalently, determinant of $\mathbf{T}_2$ should be made equal to zero. In the general method of solution, for a given frequency ω_n and wave number η_m, the x_3 wave numbers are computed from Eq. (22). Next, the elements of the $\mathbf{R}$ matrix are computed and the $\mathbf{T}_2$ matrix is formed. However, in the case of the Lamb wave, the wave number η_m cannot vary independently and must be solved from the equation

$$\det[\mathbf{T}_2(\mathbf{R}(\eta,\omega,k),k(\eta,\omega))] = 0 \tag{49}$$

Once η is obtained for a given frequency, the Lamb wave speed can be obtained from the relation $C_p = \omega/\eta$.

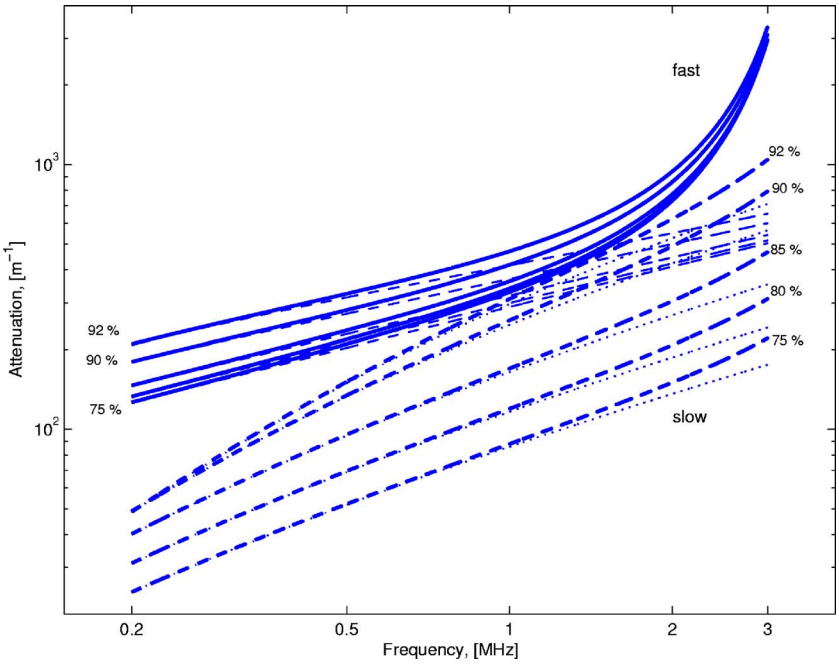


FIG. 6. (Color online) Prediction of attenuation by the classical Biot's theory (thin dashed and dotted lines) and the nonlocal poroelastic theory (thick solid and dashed lines, $\epsilon = 6.0 \times 10^{-5}$ m) for porosity varying from 75% to 92%.

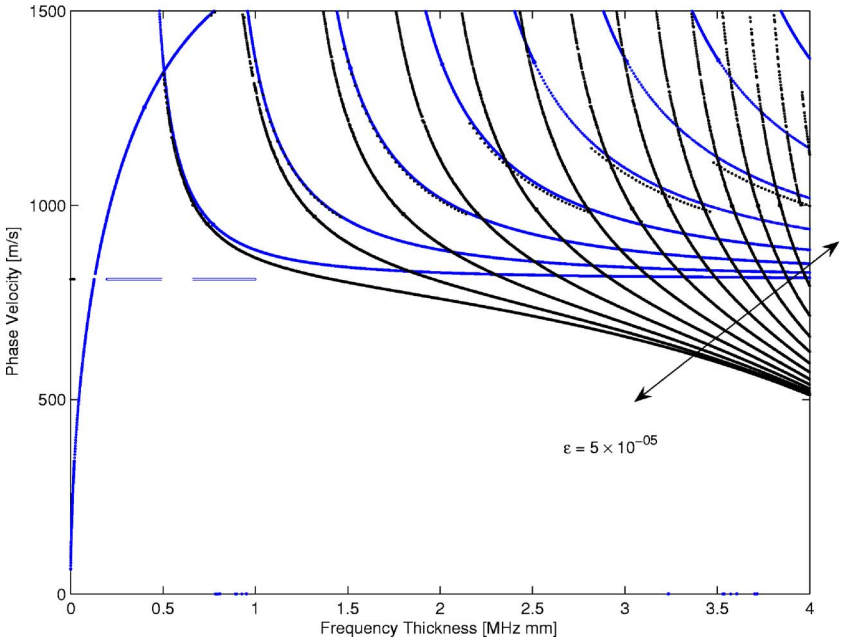


FIG. 7. (Color online) Effect of nonlocality on the Lamb wave modes, $\beta = 10\%$, $\epsilon = 5.0 \times 10^{-5}$ m.

The effect of porosity and nonlocality on the Lamb wave modes is discussed in this section. The Lamb wave modes are plotted in terms of the horizontal (x_1) phase velocity, ω/η. The solution of Eq. (49) is multivalued, unbounded, and complex (although the real part is of interest). One way to solve these equations is to appeal to the strategies of nonlinear optimization, which are based on nonlinear least square methods. There are several choices of algorithms, like the trust-region dogleg method, the Gauss–Newton method with a line search, or the Levenberg–Marquardt method with line search. Here, MATLAB function *fsolve* is used and the default option for medium scale optimization, i.e., the trust-region dogleg method is adopted, which is a variant of the Powell's dogleg method.

Apart from the choice of algorithm there are other subtle issues in root capturing for the solution of wave numbers as the solutions are complicated in nature. Moreover, except for the first one or two modes, all the other roots escape to infinity at low frequency. For isotropic materials, these cutoff frequencies are known a priori. However, no expressions can be found for anisotropic materials and in most of the cases, the modes (solutions) should be tracked backward, i.e., from the high-frequency to the low-frequency region. In general two strategies are essential to capture all the modes within a given frequency band. Initially, the whole region should be scanned for different values of the initial guess, where the initial guess should remain constant for the whole range of frequency. This sweeping opens up all the modes in that region, although they are not completely traced. Subsequently, each individual mode should be followed to the end of the domain or to a preset high value of the solution. For this case, the initial guess should be changed for each frequency to the solution of the previous frequency step. Also, sometimes it is necessary to reduce the frequency step in the vicinity of the high gradient of the modes. Once the Lamb modes are generated they are fed back into the frequency loop to produce the frequency domain solution of the Lamb wave propagation, which through inverse Fourier transform produces the time domain signal.

For the Lamb wave mode computation a cancellous bone sample of 2 mm thickness is considered with three different porosities: 10%, 50%, and 90%. It can be expected that with increasing porosity the phase speeds will reduce. Figure 8 shows the modes for 10% porosity. At frequencies below 0.25 MHz (0.5 MHz mm) only two propagating modes exist. The one that starts from zero is the antisymmetric mode (a_0) which converges at the Rayleigh wave speed (1800 m/s). The other mode starts at around 3200 m/s, and is the first symmetric mode (s_0). It is not easy to distinguish other symmetric and antisymmetric modes, especially if a common strategy is adopted for computation. It is easy to see that for a typical ultrasound evaluation of bone samples (say 2 mm thick and 1 MHz excitation frequency) at least eight modes will be excited and contribute in the overall response.

The effect of nonlocality on the Lamb wave modes was studied earlier for anisotropic materials.[33] In the present study, ε is fixed at 5×10^{-5} m and the resulting Lamb wave modes for 10% porosity are plotted in Fig. 7. We see that the Lamb wave speeds are steadily decreasing with increasing frequency, and instead of converging to the Rayleigh wave speed (as happens in the classical case, Fig. 8), the modes are monotonically going to zero. A similar negative dispersion has also been observed earlier for the anisotropic case.[33]

The Lamb wave modes for 50% and 90% porosity are shown in Figs. 9 and 10, respectively. Comparing the modes for different porosity it can be said that with increasing porosity the phase velocities decrease and the cutoff frequencies (the frequency before which a mode ceases to exist) increase. The effect of nonlocality on these modes is not shown here, although the influence is similar to what is seen for 10% porosity.

It has been observed experimentally that the cutoff frequencies of the Lamb modes can be related to the bulk slow

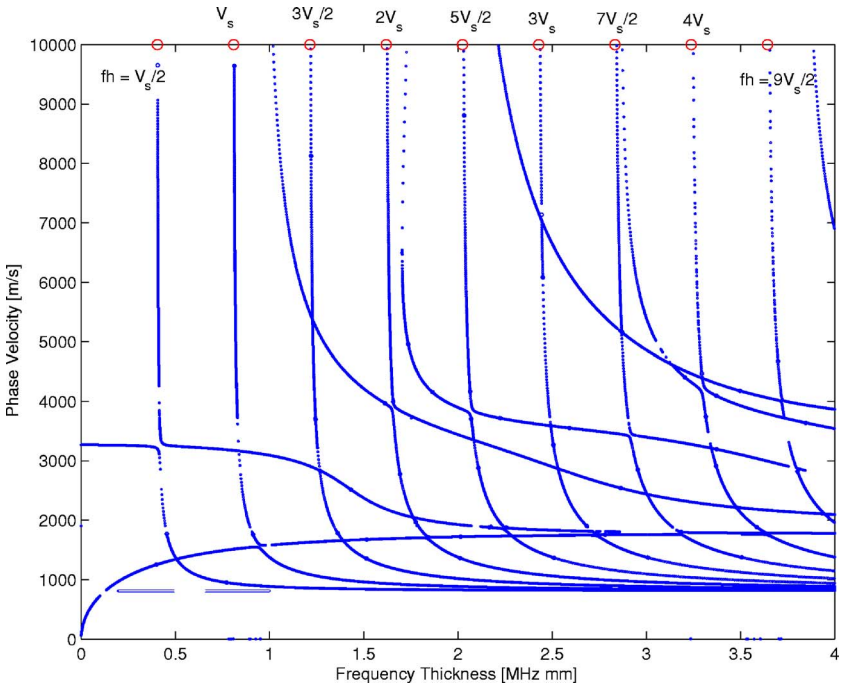


FIG. 8. (Color online) Lamb wave modes for 10% porosity, ϵ=0.

wave speed. This property can be utilized to identify the extra modes appearing due to the slow wave component. For 10% porosity the bulk slow wave speed is 0.809/km/s. It should be noted that the abscissa and ordinate of Figs. 8–10 are the same if the velocity is expressed in km/s rather than m/s. Then the slow wave speed V_s can be related to the frequency multiplied by the thickness value. Now, an interesting observation is that the cutoff frequencies appear as an integer multiple of $V_s/2$. These multiples of $V_s/2$, i.e., $V_s/2, V_s, 3V_s/2, 2V_s, \ldots$ are marked by circles in Fig. 8. As the figure suggests this rule applies very well to the appearance of the modes. Thus, in this figure we have nine slow wave modes. The rest of the seven modes are the regular homogeneous Lamb wave modes. Similarly, in Figs. 9 and 10, there are, respectively, six and five slow wave modes. Since with increasing porosity the slow wave speed increases (Fig. 2), the number of slow wave modes decreases within a given frequency-thickness window. On the other hand, with increasing porosity, both the fast and shear wave speed decrease, and as a result the participation (number) of these modes increases.

V. CONCLUSION

This study brings out the effect of nonlocal elasticity on the dispersion variation of trabecular bone. Negative dispersion, reported in the majority of work[13] related to the dispersion of trabecular bone, is predicted unambiguously by the

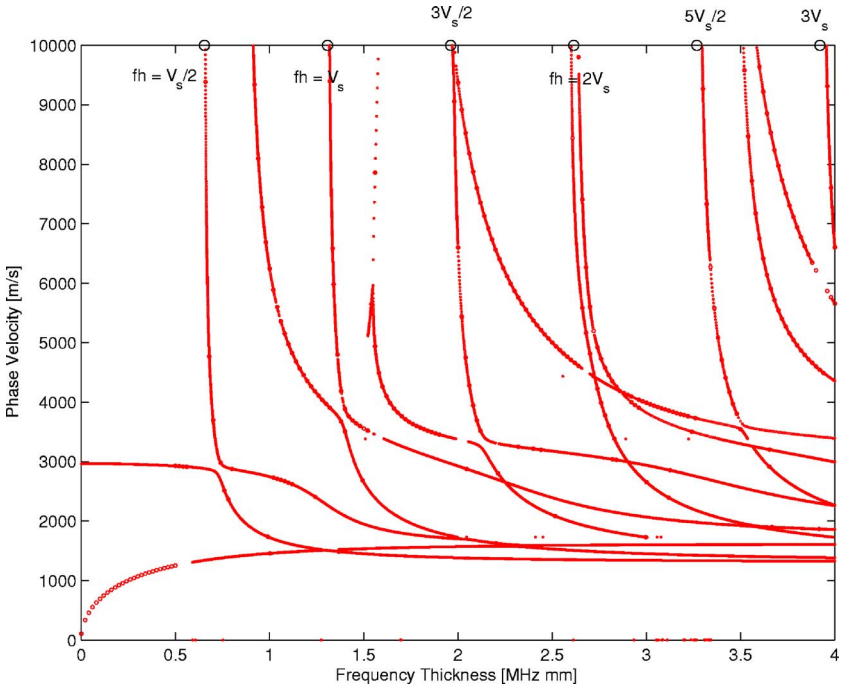


FIG. 9. (Color online) Lamb wave modes for 50% porosity, ϵ=0.

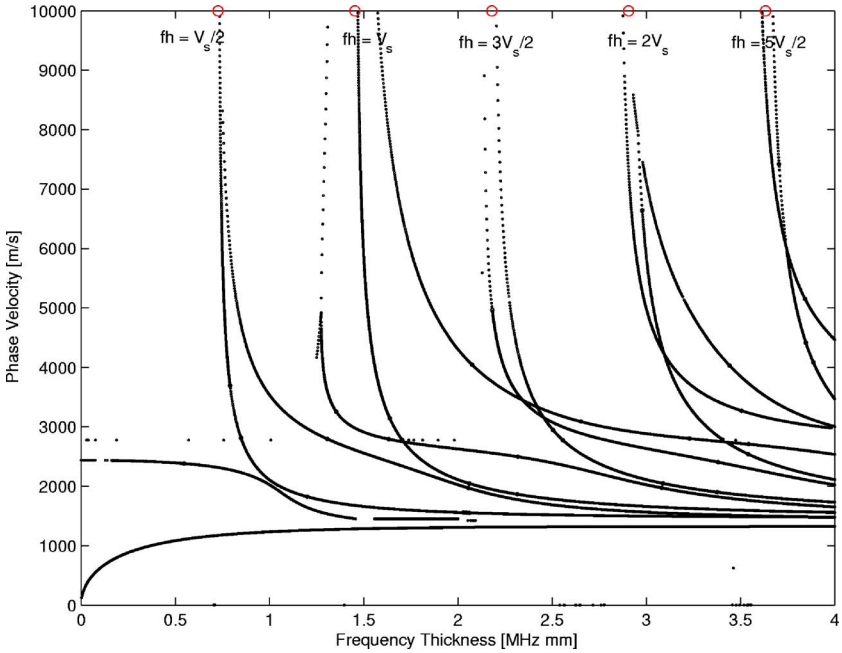


FIG. 10. (Color online) Lamb wave modes for 90% porosity, ϵ=0.

nonlocal poroelasticity. The variation of the dispersion rate on the nonlocal parameter is studied. This study will be useful in estimating the nonlocal parameter for a given experimental observation. The theory also predicts the experimentally observed trend of the variation of attenuation for fast and slow waves. The governing equations of nonlocal poroelasticity are solved exactly in the frequency wave number domain. The stiffness matrix based representation of the layer dynamics is particularly helpful in capturing the Lamb wave modes, which are essential features of ultrasonic studies. This work presents the effect of porosity and nonlocality on the Lamb wave modes. Overall, this work confirms the ultrasound technique based observations of phase speed and attenuation variation in bone and provides a way of predicting the same while working within the limitation of Biot's theory of poroelasticity.

[1]P. Laugier, "An overview of bone sonometry," International Congress Series **1274**, 23–32 (2004).
[2]K. A. Wear, "Group velocity, phase velocity, and dispersion in human calcaneus in vivo," J. Acoust. Soc. Am. **121**, 2431–2437 (2007).
[3]P. Morse and K. Ingard, *Theoretical Acoustics* (Princeton University Press, Princeton, NJ, 1986), Chap. 9.
[4]R. E. Strelitzki, J. A. Evans, and A. Clarke, "The influence of porosity and pore size on the ultrasonic properties of bone investigated using a phantom material," Osteoporosis Int. **7**, 370–375 (1996).
[5]P. Nicholson, G. Lowet, C. Langton, J. Dequeker, and G. Van der Perre, "A comparison of time-domain and frequency-domain approaches to ultrasonic velocity measurement in trabecular bone," Phys. Med. Biol. **41**, 2421–2435 (1996).
[6]P. Droin, G. Berger, and P. Laugier, "Velocity dispersion of acoustic waves in cancellous bone," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **45**, 581–592 (1998).
[7]K. Wear, "Measurement of phase velocity and group velocity in human calcaneus," Ultrasound Med. Biol. **26**, 641–646 (2000).
[8]M. Kaczmarek, J. Kubik, and M. Pakula, "Short ultrasonic waves in cancellous bone," Ultrasonics **40**, 95–100 (2002).
[9]P. Chen and T. Chen, "Measurements of acoustic dispersion on calcaneus using split spectrum processing technique," Med. Eng. Phys. **28**, 187–193 (2006).
[10]K. Waters, M. Hughes, J. Mobley, G. Brandenburger, and J. Miller, "On the applicability of Kramers–Kronig relations for ultrasonic attenuation obeying a frequency power law," J. Acoust. Soc. Am. **108**, 556–563 (2000).
[11]K. Waters and B. Hoffmeister, "Kramers–Kronig analysis of attenuation and dispersion in trabecular bone," J. Acoust. Soc. Am. **118**, 3912–3920 (2005).
[12]K. Marutyan, G. Bretthorst, and J. Miller, "Baysian estimation of the underlying bone properties from mixed fast and slow mode ultrasonic signals," J. Acoust. Soc. Am. **121**, EL8–EL15 (2006).
[13]K. Marutyan, M. Holland, and J. Miller, "Anomalous negative dispersion in bone can result from the interference of fast and slow waves," J. Acoust. Soc. Am. **120**, EL55–EL61 (2006).
[14]D. Bruggeman, "Calculation of physical constants from heterogeneous substances," Ann. Phys. **24**, 636–664 (1935).
[15]A. Tarkov, "The problem of the anisotropy of elastic properties in rocks," Mater. Vses. N.-I. Geol. In-ta Obsch. Seriya. Sb. **5**, 209–213 (1940).
[16]Y. Riznichenko, "Propagation of seismic waves in discrete and heterogenous medium," Izv. Akad. Nauk SSSR, Ser. Geogr. Geofiz. **13**, 115–128 (1949).
[17]G. Postma, "Wave propagation in a stratified medium," Geophysics **20**, 780–806 (1955).
[18]S. Rytov, "Acoustical properties of a finely layered medium," Sov. Phys. Acoust. **2**, 67–80 (1956).
[19]L. Brekhovskikh, *Waves in Layered Media*, (Academic, New York, 1980), p. 81.
[20]B. Gurevich and S. L. Lopatnikov, "Velocity and attenuation of elastic waves in finely layered porous rocks," Geophys. J. Int. **121**, 933–947 (1995).
[21]B. Gurevich, "Effect of fluid viscosity on elastic wave attenuation in porous rocks," Geophysics **67**, 264–270 (2002).
[22]T. Plona, K. Winkler, and M. Schoenberg, "Acoustic waves in alternating fluid/solid layers," J. Acoust. Soc. Am. **81**, 1227–1234 (1987).
[23]S. Lopatnikov and A.-D. Cheng, "Variational formulation of fluid infiltrated porous material in thermal and mechanical equilibrium," Mech. Mater. **34**, 685–704 (2002).
[24]S. Lopatnikov and A.-D. Cheng, "Macroscopic Lagrangian formulation of poroelasticity with porosity dynamics," J. Mech. Phys. Solids **52**, 2801–2839 (2004).
[25]M. Biot, "Theory of propagation of elastic waves in a fluid-saturated porous solid. I. low-frequency range," J. Acoust. Soc. Am. **28**, 168–178 (1955).
[26]M. Biot, "Theory of propagation of elastic waves in a fluid-saturated porous solid. II. Higher-frequency range," J. Acoust. Soc. Am. **28**, 179–191 (1955).
[27]A. Eringen, "Theory of nonlocal elasticity and some applications," Res. Mech. **21**, 313–342 (1987).
[28]J. Williams, "Ultrasonic wave propagation in cancellous and cortical bone: Prediction of some experimental results by Biot's theory," J. Acoust. Soc. Am. **91**, 1106–1112 (1992).

[29]R. Artan and B. Altan, "Propagation of sv waves in a periodically layered media in nonlocal elasticity," Int. J. Heat Mass Transfer **39**, 5927–5944 (2002).
[30]D. Johnson, J. Koplik, and R. Dashen, "Theory of dynamic permeability and tortuosity in fluid-saturated porous media," J. Fluid Mech. **176**, 379–402 (1987).
[31]K. A. Wear, "A stratified model to predict dispersion in trabecular bone," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **48**, 1079–1083 (2001).
[32]A. Hosokawa and T. Otani, "Ultrasonic wave propagation in bovine cancellous bone," J. Acoust. Soc. Am. **101**, 558–562 (1997).
[33]A. Chakraborty, "Wave propagation in anisotropic media with nonlocal elasticity," Int. J. Solids Struct. **44**, 5723–5741 (2007).

Evaluation of the angular spectrum approach for simulations of near-field pressures

Xiaozheng Zeng[a] and Robert J. McGough[b]
Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, USA

(Received 3 May 2007; revised 15 October 2007; accepted 18 October 2007)

The implementation of the angular spectrum approach based on the two-dimensional fast Fourier transform is evaluated for near-field pressure simulations of square ultrasound transducers, where the three-dimensional pressure field is calculated from the normal velocity distribution on the transducer surface. The pressure field is propagated in the spatial frequency domain with the spatial propagator or the spectral propagator. The spatial propagator yields accurate results in the central portion of the computational grid while significant errors are produced near the edge due to the finite extent of the window applied to the spatial propagator. Likewise, the spectral propagator is inherently undersampled in the spatial frequency domain, and this causes high frequency errors in the computed pressure field. This aliasing problem is alleviated with angular restriction. The results show that, in nonattenuating media, the spatial propagator achieves smaller errors than the spectral propagator after the region of interest is truncated to exclude the windowing error. For pressure calculations in attenuating media or with apodized pistons as sources, the spatial and spectral propagator achieve similar accuracies. In all simulations, the angular spectrum calculations with the spatial propagator take more time than calculations with the spectral propagator.
 [DOI: 10.1121/1.2812579]

PACS number(s): 43.38.Hz, 43.20.El, 43.20.Rz, 43.40.Rj [TDM] Pages: 68–76

I. INTRODUCTION

The angular spectrum approach describes the diffraction of acoustic waves from finite apertures by superposing plane waves traveling in different directions[1] and propagating these components in the spatial frequency domain. As opposed to integral approaches that calculate the field at each observation point, the angular spectrum approach computes the pressure field in successive planes with a two-dimensional (2D) fast Fourier transform (FFT), which speeds up these calculations significantly. The angular spectrum approach uses either the normal particle velocity or the pressure as the source, and then the spectral propagator function or the 2D Fourier transform of the spatial propagator is multiplied by the source in the spatial frequency domain to simulate the propagation of acoustic waves. The spectral propagator is described as an analytical function,[2–9] whereas the spatial propagator is calculated analytically and then transformed into the spatial frequency domain with a 2D FFT.[10–12]

The spectral propagator is frequently applied to the angular spectrum approach. Williams[13,14] analyzes the effect of the windowing function, the aliasing errors due to the lack of spectral samples, and the singularity of the spectral propagator. Orofino[3,4] discusses the spatial sampling rate and the angular resolution. Christopher and Parker[10] and Wu *et al.*[6–8] both investigate an angular restriction technique to reduce the aliasing error due to undersampling of the spectral propagator, and Wu and co-workers derive the optimal bandwidth of a low-pass spatial filter that truncates the undersampled spatial frequencies. The spatial propagator is less frequently employed due to the additional 2D FFT calculation required in each plane.

A better understanding of the trade-offs between the spectral propagator and the spatial propagator in terms of numerical accuracy and computational time is clearly needed. Christopher and Parker[10] claim that the spatial propagator produces more accurate results in the context of on-axis simulations for axisymmetric radiators using the fast Hankel transform. However, comparisons relative to a standard reference are not evaluated for the off-axis case. Zemp[12] argues that the spectral propagator is superior to the spatial propagator when the spatial sampling is coarser than $\lambda/2$, which is applicable only in the far field. In many applications, the region of interest is the near field, and comparisons with higher sampling rates are desired.

This paper thoroughly compares the spatial propagator and the spectral propagator for angular spectrum calculations in a homogeneous three-dimensional (3D) domain. The pressure field from a square piston with uniform particle velocity distribution is computed with the fast near-field method[15] as the reference, and for an apodized particle velocity distribution, the Rayleigh–Sommerfeld integral is the reference. The errors in the computed pressure field for the angular spectrum approach using the spatial propagator and the spectral propagator (with and without angular restriction) are then compared. First, the role of angular restriction in angular spectrum calculations with the spectral propagator is established. Second, the errors generated in the edge of the computational grid specific to the spatial propagator are identified and explained. Third, the frequency filtering effect of attenuating media is shown to eliminate the need for angular

[a]Electronic mail: zengxiao@msu.edu
[b]Electronic mail: mcgough@egr.msu.edu

restriction. Fourth, a similar effect is demonstrated for velocity source apodization. The results show that for the uniformly excited source in nonattenuating media, the spatial propagator yields smaller errors in the central portion of the grid, whereas the spectral propagator outperforms the spatial propagator in attenuating media or for an apodized source by achieving similar accuracy in a larger region in less time.

II. THEORY

A. The Rayleigh–Sommerfeld integral

For a planar radiator mounted on an infinite rigid baffle, the radiated time-harmonic pressure field is represented by the Rayleigh–Sommerfeld diffraction integral,[1]

$$p(\mathbf{r},t) = j\rho ck e^{j\omega t} \int_{\mathbf{S}'} u(\mathbf{r}') \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{2\pi|\mathbf{r}-\mathbf{r}'|} d\mathbf{S}', \tag{1}$$

where ρ and c represent the medium density and the speed of sound, respectively, k is the acoustic wavenumber, ω is the driving frequency, u is the distribution of the normal velocity on the radiator with surface area $\mathbf{S}'$, j is $\sqrt{-1}$, and $|\mathbf{r}-\mathbf{r}'|$ is the distance between the source and the observation coordinates. Numerical implementations of Eq. (1) often divide the radiator aperture into point sources and superpose the results.[16] This approach is especially time-consuming in the near-field region due to the large number of abscissas required for the convergence of the 2D integral and the numerical singularity on the piston surface. Equation (1) calculates the near-field reference pressure for radiators with nonuniform velocity distributions, whereas significantly faster approaches such as the fast near-field method[15] are available for pressure calculations with uniform velocity distributions.

B. The fast near-field method

The fast near-field method[15] is a rapidly converging one-dimensional (1D) integral approach that is analytically equivalent to Eq. (1) for a uniform surface velocity u_0. The fast near-field method formula for the near-field pressure is

$$\begin{aligned} p(x,y,z,t) = j\rho c u_0 e^{j\omega t} \frac{1}{2\pi} \Bigg(& s_1 \int_{-l_1}^{l_2} \frac{e^{-jk\sqrt{z^2+\sigma^2+s_1^2}} - e^{-jkz}}{\sigma^2+s_1^2} d\sigma \\ & + l_1 \int_{-s_1}^{s_2} \frac{e^{-jk\sqrt{z^2+\sigma^2+l_1^2}} - e^{-jkz}}{\sigma^2+l_1^2} d\sigma \\ & + s_2 \int_{-l_1}^{l_2} \frac{e^{-jk\sqrt{z^2+\sigma^2+s_2^2}} - e^{-jkz}}{\sigma^2+s_2^2} d\sigma \\ & + l_2 \int_{-s_1}^{s_2} \frac{e^{-jk\sqrt{z^2+\sigma^2+l_2^2}} - e^{-jkz}}{\sigma^2+l_2^2} d\sigma \Bigg), \end{aligned} \tag{2}$$

where $s_1=a-x$, $l_1=b-y$, $s_2=a+x$, $l_2=b+y$, σ is the 1D variable of integration, and a and b present the half-width and the half-height of the rectangular piston, respectively. This approach eliminates the singularities that are encountered in numerical evaluations of the Rayleigh–Sommerfeld integral and the impulse response.[17]

C. The angular spectrum approach

Equation (1) is a 2D convolution, where the source is located in the $z=0$ plane such that

$$p(x,y,z,t) = j\rho ck e^{j\omega t} u(x,y) \otimes h_u(x,y,z), \tag{3}$$

and the *spatial propagator* is defined as

$$h_u(x,y,z) = \frac{e^{-jkr}}{2\pi r}, \tag{4}$$

where $r=\sqrt{x^2+y^2+z^2}$ is the distance from the origin to an arbitrary field point. Applying a 2D Fourier transform to both sides of Eq. (3) transforms the formula into the spatial frequency domain,

$$P(k_x,k_y,z,t) = j\rho ck e^{j\omega t} U(k_x,k_y) H_u(k_x,k_y,z), \tag{5}$$

where $H_u(k_x,k_y,z)$ is the *spectral propagator*, which is defined as

$$H_u(k_x,k_y,z) = \begin{cases} \dfrac{e^{-jz\sqrt{k^2-k_x^2-k_y^2}}}{j\sqrt{k^2-k_x^2-k_y^2}}, & k_x^2+k_y^2 \leqslant k^2 \\ \dfrac{e^{-z\sqrt{k_x^2+k_y^2-k^2}}}{\sqrt{k_x^2+k_y^2-k^2}}, & k_x^2+k_y^2 > k^2. \end{cases} \tag{6}$$

The Fourier transform decomposes the diffracted wave into the superposition of plane waves, where (k_x,k_y) are the transverse wavenumbers. In the region where $k_x^2+k_y^2 \leqslant k^2$, $H_u(k_x,k_y,z)$ propagates the field in the z direction by applying a complex weight to each plane wave component. In the region where $k_x^2+k_y^2>k^2$, $H_u(k_x,k_y,z)$ exponentially attenuates evanescent waves. The product of the source spectrum $U(k_x,k_y)$ and the propagator function $H_u(k_x,k_y,z)$ describes the spectrum of the propagating wave in an arbitrary plane parallel to the source plane. The pressure distribution is then obtained from the inverse 2D Fourier transform of $P(k_x,k_y,z,t)$.

When the angular spectrum approach calculates pressure fields with the *spatial propagator* $h_u(x,y,z)$,[10–12] the infinite field is truncated in both the x and the y directions so that the extent of each computational plane is $D\times D$. In the simulations that follow, the radiator is a $2a\times 2a$ square piston with $2a<D$. The source plane is illustrated in Fig. 1. The normal particle velocity distribution is defined as $u(x,y)$ within the $2a\times 2a$ area and zero otherwise. With a sample spacing of δ, the $D\times D$ source plane that contains the piston surface is discretized to $N\times N$ grid points, where $N=D/\delta+1$. The field is computed in a discretized grid, and the discrete coordinates are

$$\begin{aligned} x &= m\delta, \quad m = -N/2+1+\phi, \ldots, N/2+\phi, \\ y &= n\delta, \quad n = -N/2+1+\phi, \ldots, N/2+\phi, \end{aligned} \tag{7}$$

$$\phi = \begin{cases} -\dfrac{1}{2} & \text{when } N \text{ is odd}, \\ 0 & \text{when } N \text{ is even}, \end{cases} \tag{8}$$

where ϕ ensures that the grid is symmetric for odd N. For even N, the grid is biased so that points on the central axis

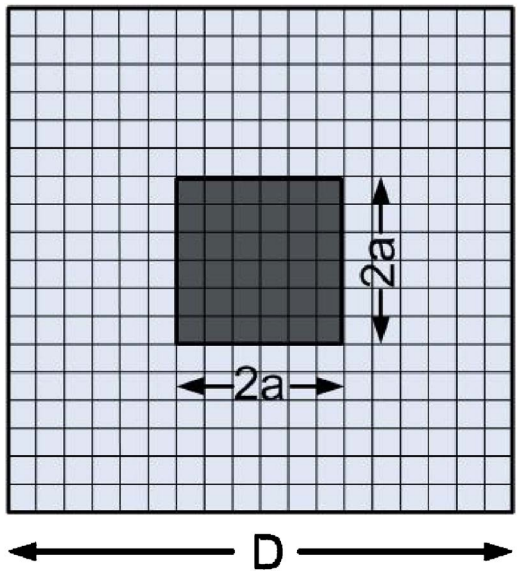


FIG. 1. (Color online) The $D \times D$ source plane consisting of a nonzero normal particle velocity distribution in a $2a \times 2a$ square area on the piston surface. The remaining area is filled with zeros.

are included in the three-dimensional grid. The 2D FFT is then applied to both $h_u(x,y,z)$ and $u(x,y)$, and these are multiplied in the spatial frequency domain. The complex pressure is the inverse 2D FFT of the result.

When the angular spectrum approach calculates pressure fields with the *spectral propagator* $H_u(k_x,k_y,z)$, the spectrum is discretized to an $N \times N$ grid, and the discrete wavenumbers are

$$k_x = m\Delta k, \quad m = -N/2 + 1 + \phi, \ldots, N/2 + \phi,$$

$$k_y = n\Delta k, \quad n = -N/2 + 1 + \phi, \ldots, N/2 + \phi, \tag{9}$$

where the offset ϕ is defined in Eq. (8). The maximum value of k_x is π/δ, the maximum value of k_y is π/δ, and the spectral sample spacing is $\Delta k = 2\pi/(N\delta)$. A large value of N is often required to adequately sample $H_u(k_x,k_y,z)$. The corresponding value of D is therefore significantly larger than $2a$.

D. Angular restriction

The spectral propagator $H_u(k_x,k_y,z)$ encounters problems with rapidly oscillating real and imaginary components as $\sqrt{k_x^2+k_y^2}$ approaches k. In this region, H_u is inherently undersampled, and this undersampling leads to severe aliasing errors.[7] The undersampling can be reduced by increasing N at an expense of increased computational cost. Christopher and Parker[10] and Wu *et al.*[7,8] use a spatial frequency truncation technique as an alternative solution to this problem. This technique reduces the aliasing errors without increasing the size of the computational grid. Spatial frequency truncation is achieved through angular restriction, which applies a spatial low-pass filter to the spectral propagator function H_u, and the cut-off, or angular threshold, is given by[8]

$$k_c = k\sqrt{\frac{D^2/2}{D^2/2 + z^2}}, \tag{10}$$

which specifies a radially symmetric window in 2D. Angular restriction removes the under-sampled angular spectra and prevents the high spatial frequency components from leaking into the propagating field.

E. Attenuation calculations

When the spatial propagator is used in attenuating media, the real-valued wavenumber k in $h_u(x,y,z)$ is replaced by the complex wavenumber $k - j\alpha$, where α is the attenuation coefficient for a particular frequency. For the spectral propagator, $H_u(k_x,k_y,z)$ is multiplied by an exponential term

$$S(k_x,k_y,z) = \exp\left(-\frac{\alpha z k}{\sqrt{k^2 - k_x^2 - k_y^2}}\right), \tag{11}$$

which is a simplified expression that is mathematically equivalent to the trigonometric attenuation term in Eq. (12) of Ref. 10. Within the region where $k_x^2 + k_y^2 \leq k^2$, $S(k_x,k_y,z)$ is analogous to a spatial low-pass filter. As z increases, the peak amplitude of $S(k_x,k_y,z)$ decreases, and the higher spatial frequency components are increasingly attenuated.

F. Error metric

The numerical errors produced by the angular spectrum approach are evaluated with the normalized root mean squared error (NRMSE). This error metric is defined by

$$\text{NRMES} = \frac{\sqrt{\frac{1}{n_x n_y n_z}\sum_{i,j,k} |p^{i,j,k} - p_{\text{ref}}^{i,j,k}|^2}}{\max_{i,j,k}|p_{\text{ref}}^{i,j,k}|}, \tag{12}$$

where the superscripts (i,j,k) represent discrete field points in the computational grid and n_x, n_y, and n_z describe the number of points in each direction. The variable p_{ref} denotes the complex reference pressure, which is computed by the fast near-field method for a piston with uniform normal velocity distribution and by the Rayleigh–Sommerfeld integral for an apodized piston. The variable p is the complex pressure computed by the angular spectrum approach. The root mean squared error is normalized by the global maximum of the 3D reference pressure amplitude.

III. NUMERICAL RESULTS

Simulated pressure fields are generated by a single square piston in a 3D computational grid. The piston, which is 3 cm wide and 3 cm high ($a = 1.5$ cm), is excited at a frequency of 1 MHz. The speed of sound for these simulations is 1500 m/s, so the piston size in wavelength is $20\lambda \times 20\lambda$. The axial grid extends from 0.15 to 30 cm (λ to $2a^2/\lambda$) with an axial sampling interval of 1.5 mm (λ). The maximum transverse extent is $D = 9$ cm (60λ) with a transverse sample spacing of $\lambda = 0.3$ mm ($\lambda/5$), so $N = 301$. The transverse sample spacing is relatively small so that the rapid pressure oscillations are captured in the near field. The spatial propagator and the spectral propagator are both implemented in nonattenuating and attenuating media for a piston

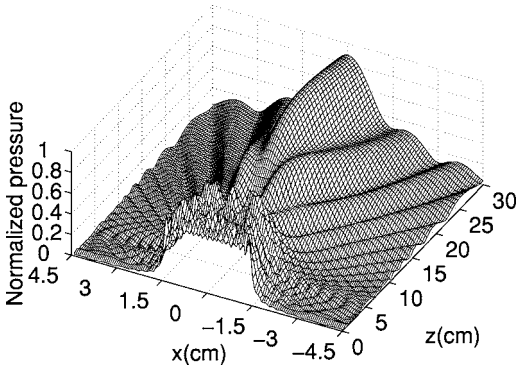


FIG. 2. A two-dimensional cross section of the three-dimensional reference pressure generated by a 3 cm × 3 cm square piston in nonattenuating media. The excitation frequency is 1 MHz, and the normal particle velocity distribution is uniform across the piston surface. The reference pressure is computed with the fast near-field method. The result is normalized to the maximum pressure amplitude computed in the three-dimensional volume.

with uniform normal particle velocity. A quadratically apodized normal velocity distribution is also evaluated in nonattenuating media.

A. Pressure calculations in nonattenuating media for a square piston with uniform normal particle velocity

The reference pressure field generated by the square piston with uniform normal velocity distribution is shown in Fig. 2. The pressure field is evaluated in the $y=0$ plane, and the transverse extent of this plane in both directions is $D=9$ cm. The reference pressure is calculated with the fast near-field method[15] using 100 Gauss abscissas. The reference pressure achieves 11 digits of accuracy throughout the computational grid, where the accuracy is determined by comparing the result to that obtained with 2000 Gauss abscissas. The 3D pressure field is then calculated in successive 9 cm × 9 cm transverse planes using the spectral propagator and the spatial propagator.

Figures 3(a) and 3(b) demonstrate the pressure computed by the spectral propagator without angular restriction and with angular restriction, respectively. The pressure field in Fig. 3(a) contains significant errors due to the aliasing of high spatial frequencies in the spectral propagator, where aliasing is caused by undersampling H_u. In Fig. 3(b), the oscillatory errors are significantly reduced by angular restriction. According to Eq. (10), the angular thresholds at locations $z=10$, 20, and 30 cm are $0.5369k$, $0.3032k$, and $0.2075k$, respectively, where k is the wavenumber. However, ripples still appear in Fig. 3(b) because the spectrum corresponding to intermediate spatial frequencies is also undersampled. Figures 3(a) and 3(b) demonstrate that removing the high spatial frequencies with angular restriction is beneficial for simulations in nonattenuating media, e.g., in water.

Figure 3(c) shows the pressure computed by the spatial propagator without angular restriction. The computed field is smooth everywhere; therefore, angular restriction is not needed for the spatial propagator. In the paraxial region, the pressure field in Fig. 3(c) closely resembles the reference field in Fig. 2. However, starting a certain distance from the

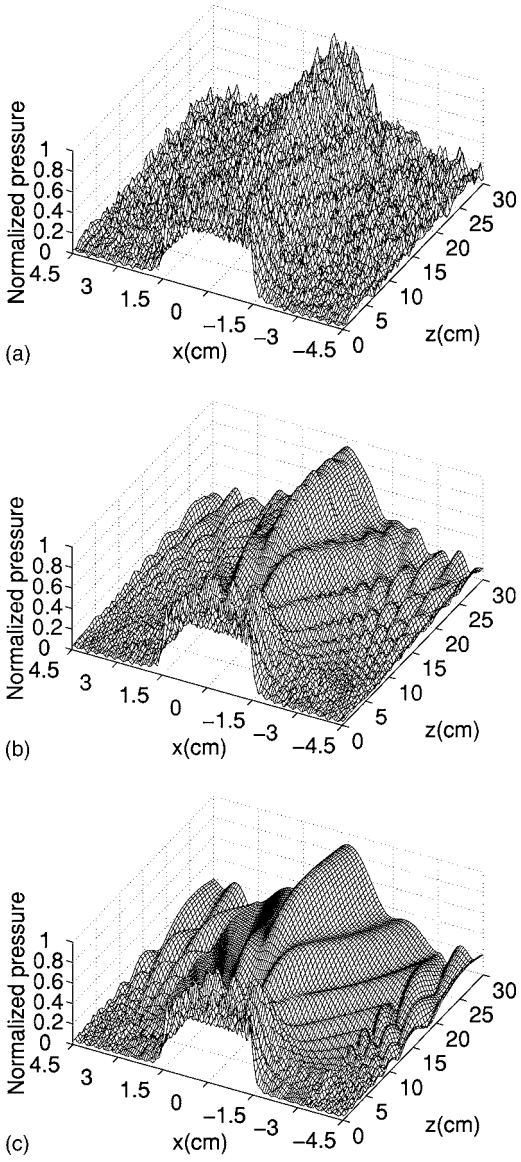


FIG. 3. Simulated pressure generated by a 3 cm × 3 cm square piston in nonattenuating media computed by the angular spectrum approach using (a) the spectral propagator without angular restriction, (b) the spectral propagator with angular restriction, and (c) the spatial propagator. The excitation frequency is 1 MHz. All fields are calculated in successive $D \times D$ transverse planes ($D=9$ cm), so the circular convolution errors generated by the spatial propagator are included.

edge of the computational grid, a discrete jump appears in the computed field. This error is an artifact of circular convolution between the spatial propagator and the velocity source, where the spatial propagator is evaluated on a $D \times D$ grid, and the dimension of the nonzero velocity source is $2a \times 2a$. All convolutions are performed over a $D \times D$ area with 2D FFTs to minimize the number of parameters needed to describe each simulation. If extra zero padding is used to enlarge both the $D \times D$ spatial propagator plane and the source plane, the circular convolution error is replaced with another error. With zero padding, as soon as the nonzero part of the velocity source is shifted to a location that overlaps with a value outside of the $D \times D$ rectangular window, the source is convolved with zero instead of the correct value of the spatial propagator. As a result, the pressure fields com-

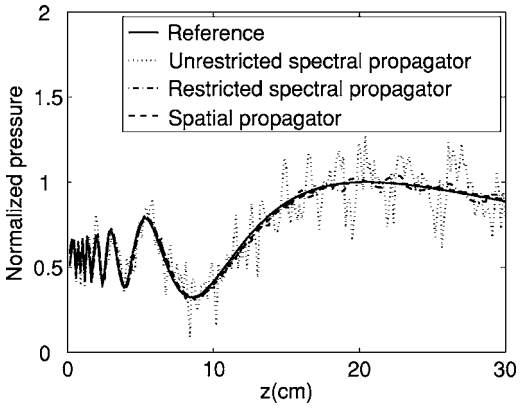


FIG. 4. Axial plots of the absolute value of the simulated complex pressure. Results are shown for the reference, the angular spectrum approach using the spectral propagator without and with angular restriction, and the spatial propagator.

puted with the spatial propagator are erroneous starting from a distance a from the edge. With or without zero padding, the errors due to the finite window applied to the spatial propagator always appear within a peripheral band of width a, but the $L\times L$ area in the center is not affected, where $L=D-2a$. In contrast, no such windowing error is present in the results from the spectral propagator, because the spectral propagator is analytically evaluated in the spatial frequency domain.

Figure 4 provides a more detailed comparison between the reference and the results in Figs. 3(a)–3(c) by highlighting the pressure variations in the axial direction where $x=y=0$. The axial pressure computed using the spectral propagator without angular restriction (dotted line) is corrupted by errors with high spatial frequencies. The field computed from the spectral propagator with angular restriction (dash-dot line) contains some intermediate frequency ripples. The field computed from the spatial propagator (dashed line) closely tracks the reference (solid line). Figure 4 emphasizes that the spatial propagator computes axial pressures more accurately than the spectral propagator in nonattenuating media.

In Fig. 5, the normalized root mean squared errors are calculated in three-dimensional volumes for different D values, where D represents the extent of the computational volume in both the x and the y directions. The results are plotted as a function of D, where the value of D ranges from 6 to 15 cm. The normalized root mean squared error curve for the spatial propagator (dashed line) is between that of the spectral propagator without angular restriction (dotted line) and with angular restriction (solid line). Thus, when the circular convolution errors produced by the spatial propagator are included, the spectral propagator with angular restriction achieves a smaller normalized root mean squared error.

In Fig. 5(b), the volumes are truncated at the edges to exclude the convolution errors, and then the normalized root mean squared errors are computed. The truncated volume has a lateral extent of $L\times L$, so the results are plotted as a function of L with $L=D-2a$, where L ranges from 3 to 12 cm. In Fig. 5(b), the errors from the spatial propagator are 2 to 4 times smaller than those in Fig. 5(a). The errors computed for the spectral propagator with and without angular restriction are influenced less by the truncation. The errors

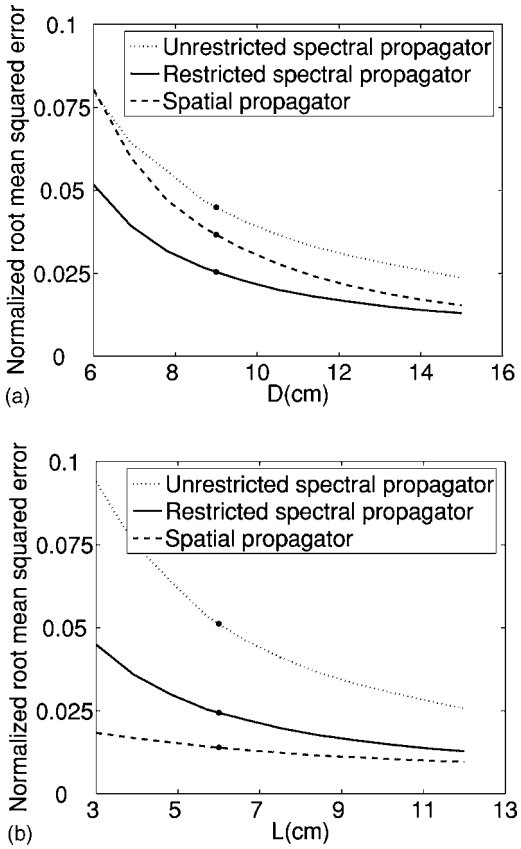


FIG. 5. Normalized root mean squared errors for the pressure generated by a uniform normal particle velocity distribution in nonattenuating media. The errors are evaluated in three-dimensional volumes where the lateral dimensions are (a) $D\times D$ and (b) $L\times L$ with $L=D-2a$. The markers on the curves indicate the corresponding results shown in Fig. 3.

without angular restriction are about 1.5 to 2 times higher than those with angular restriction in both Figs. 5(a) and 5(b). The difference between the dashed curves in Figs. 5(a) and 5(b) reinforces the need to truncate the field computed with the spatial propagator to exclude the errors produced at the edge of the grid.

B. Pressure calculations in attenuating media for a square piston with uniform normal particle velocity

When time-harmonic acoustic waves propagate in attenuating media, e.g., biological tissue, the angular spectrum approach either evaluates the spatial propagator with a complex wavenumber or applies the attenuation term $S(k_x,k_y,z)$ in Eq. (11) to the spectral propagator. Figure 6 shows the reference pressure generated by the same 3 cm$\times$3 cm square piston in attenuating media with $\alpha=1$ dB/cm/MHz. In Fig. 6, the reference pressure field attenuates quickly as z increases. The reference pressure is calculated with the fast near-field method[15] using 100 Gauss abscissas, which achieves 11 digits of accuracy compared to the result obtained with 2000 Gauss abscissas.

The pressures in attenuating media calculated by the angular spectrum approach in Fig. 7 are calculated with $D=9$ cm and then truncated to $L=6$ cm so that circular convolution errors are excluded from the spatial propagator result.

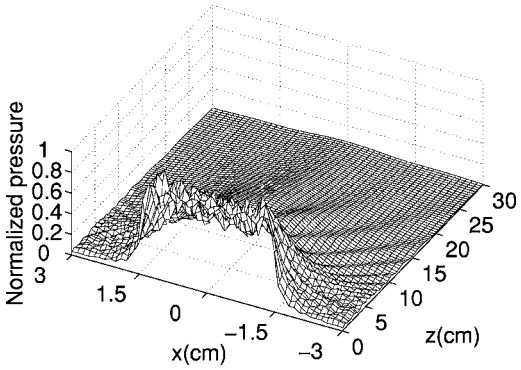


FIG. 6. The absolute value of the complex reference pressure generated by a 3 cm×3 cm square piston in attenuating media (α=1 dB/cm/MHz). The excitation frequency is 1 MHz and the normal velocity distribution is uniform across the piston surface. The reference pressure is computed with the fast near-field method.

Figure 7(a) shows the result for the spectral propagator without angular restriction. The field contains some high spatial frequency ripples due to undersampling in the spatial frequency domain. Figure 7(b) is calculated using the spectral propagator with angular restriction, and the ripples are reduced somewhat. Figure 7(c), which contains the result obtained with the spatial propagator, is relatively smooth. All three results are quite similar to the reference in Fig. 6. Therefore, the spectral propagator is preferable in attenuating media because the spectral propagator produces acceptable results in larger $D\times D$ transverse planes while the spatial propagator only gives accurate results in smaller $L\times L$ transverse planes. Meanwhile, the spectral propagator uses less computation time because fewer FFT computations are required.

Figure 8 shows the normalized root mean squared errors obtained using the spectral propagator (with and without angular restriction) and the spatial propagator. The pressures are calculated in $D\times D$ transverse planes and then truncated to $L\times L$ with $L=D-2a$. The normalized root mean squared error is then computed in the truncated volume. In Fig. 8, L ranges from 3 to 12 cm, and the three curves approximately converge to the same value for large L. The difference between the dotted and the solid curves is smaller in Fig. 8 than in Fig. 5(b), which implies that angular restriction is less important in attenuating media. This occurs because $S(k_x,k_y,z)$ is a low-pass filter that effectively attenuates the high spatial frequency components in $H_u(k_x,k_y,z)$. Since the attenuation reduces the highly oscillatory spectrum near $k_x^2+k_y^2=k^2$, aliasing errors in attenuating media are less severe than those in nonattenuating media, therefore angular restriction is not required in attenuating media.

C. Pressure calculations in nonattenuating media for a square piston with apodized normal particle velocity

The distribution of the normal particle velocity on the piston surface also influences the numerical accuracy of the angular spectrum approach. To demonstrate the effect of apodization, a tapered window is applied to the aperture. The

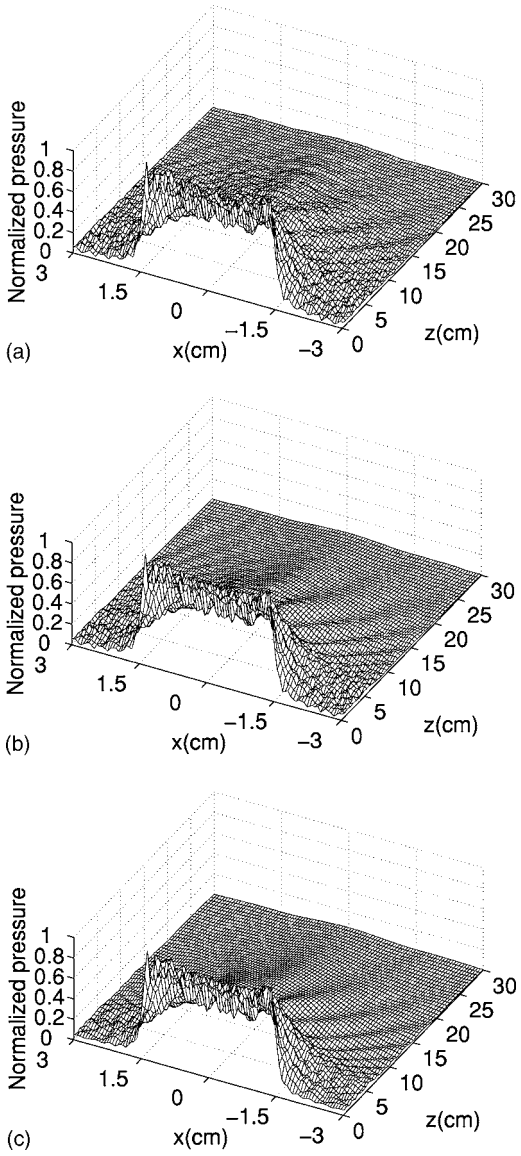


FIG. 7. The absolute value of the simulated complex pressure in attenuating media computed with the angular spectrum approach using (a) the spectral propagator without angular restriction, (b) the spectral propagator with angular restriction, and (c) the spatial propagator. All fields are calculated in successive $D\times D$ transverse planes (D=9 cm) and truncated to $L\times L$ (L=6 cm), where $L=D-2a$.

apodization function evaluated here is a product of quadratic polynomials in both the x and the y directions, where

$$u(x,y)=\begin{cases}\left[1-\left(\frac{x}{a}\right)^2\right]\left[1-\left(\frac{y}{b}\right)^2\right], & |x|<a,\ |y|<b\\ 0, & |x|\geqslant a,\ |y|\geqslant b.\end{cases} \tag{13}$$

This source velocity has a peak at the origin and smoothly decays to zero at the edges of the piston. The pressure from the apodized piston is calculated with point source superposition applied to the Rayleigh–Sommerfeld integral,[16] where each of the point sources is weighted by the value of the normal velocity at the subelement center. In the numerical evaluation of the Rayleigh–Sommerfeld integral, the reference field is computed with 5000×5000 uniform subdivi-

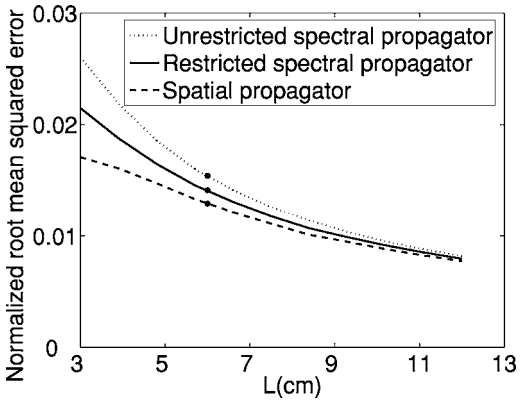


FIG. 8. Normalized root mean squared errors for the pressure generated in attenuating media (α=1 dB/cm/MHz) by a 3 cm×3 cm uniformly excited square piston. The errors are evaluated in three-dimensional volumes where the lateral dimensions are $L\times L$ with $L=D-2a$. The markers on the curves indicate the corresponding results shown in Fig. 7.

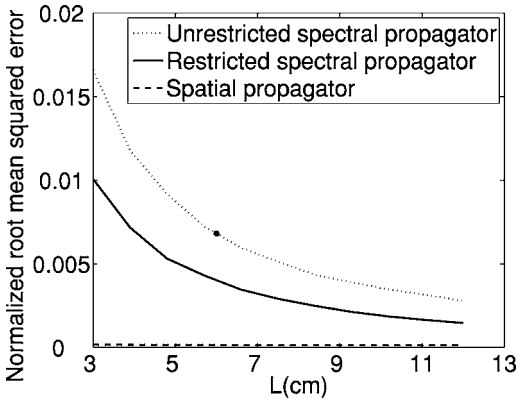


FIG. 10. Normalized root mean squared errors for the pressure generated in nonattenuating media by a 3 cm×3 cm square piston with an apodized normal particle velocity distribution. The errors are evaluated in three-dimensional volumes where the lateral dimensions are $L\times L$ with $L=D-2a$. The marker on the dotted line indicates the result in Fig. 9(b).

sions on the piston surface. This reference is accurate to 7 digits, as determined from a comparison with the same result calculated with 10 000×10 000 subdivisions. Figure 9(a) shows the reference field in the $y=0$ plane computed in nonattenuating media. The angular spectrum approach then computes the 3D field generated by a rectangular piston with an apodized source profile. Figure 9(b) shows the pressure computed by the angular spectrum approach using the spectral propagator without angular restriction. The results obtained using the spectral propagator with angular restriction and from the truncated spatial propagator are very similar to Fig. 9(b) and are therefore not shown.

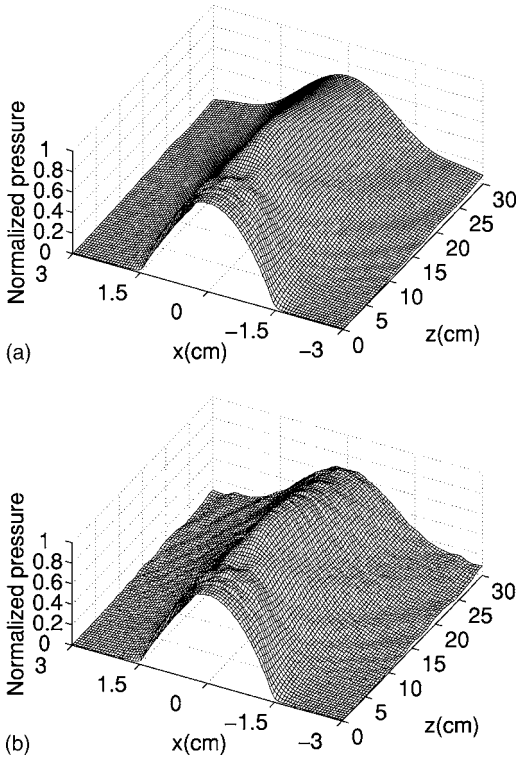


FIG. 9. The absolute value of the simulated complex pressure in nonattenuating media generated by a 3 cm×3 cm square piston with apodized normal particle velocity distribution. (a) The reference pressure and (b) the pressure computed by the spectral propagator without angular restriction. The excitation frequency is 1 MHz.

Figure 10 shows the normalized root mean squared errors of the pressures computed with the spectral propagator (with and without angular restriction) and the spatial propagator. The pressures are calculated in a 3D volume with $D\times D$ lateral extent, then the volume is truncated in the lateral directions to $L\times L$ with $L=D-2a$ and the normalized root mean squared error is evaluated. The errors in all three curves are much lower than those computed with the uniform velocity source in nonattenuating media [Fig. 5(b)] or attenuating media (Fig. 8). The uniform velocity distribution corresponds to a source spectrum that is represented by a 2D sinc function with large sidelobes. The spatial frequencies in the sidelobes of the sinc function leak into the pressure field and cause errors in the near field, as shown in Figs. 3(a), 3(b), 7(a), and 7(b). In contrast, the apodized piston has a tapered velocity distribution, the spectrum of which has lower sidelobes and is therefore less prone to high spatial frequency errors. The source spectrum along the k_x direction where $k_y=0$, contains first sidelobes at −6.8 and −11.2 dB for the uniform velocity distribution and the apodized velocity distribution, respectively. Near $k_x=k$ and $k_y=0$, where the spectrum becomes evanescent outside of this range, the sidelobes are −17.7 and −30.8 dB, for the uniform and the apodized velocity distributions, respectively. Thus, by reducing the sidelobe levels in the source spectrum, apodization achieves significant reduction in the error and eliminates the need for angular restriction.

IV. DISCUSSION

A. Reference pressure calculations

The reference fields for the uniformly excited piston are computed with the fast near-field method.[15] Previous calculations have demonstrated that the fast near-field method, the impulse response method,[17] and other integral methods converge to the same result when the number of abscissas is very large.[15] Overall, the fast near-field method provides the most accurate results in the shortest time for numerical calculations in the near-field region. For example, the fast near-

field method takes 19 min to compute the pressure in a $201\times201\times200$ grid and is accurate to 11 digits. A 500×500 point source superposition calculation with the Rayleigh–Sommerfeld integral for the same uniform velocity distribution on the same grid takes 40 min and is only accurate to 4 digits. These computation times are evaluated on a Pentium 4 PC with 4 Gbytes memory running the Windows XP operating system. All routines are written in C and executed with MATLAB 7.0. The FFT calculations are computed with FFTW library version 3.1.2.

B. Error and time trade-offs

Computations with the spatial propagator take longer than calculations with the spectral propagator because an extra 2D FFT calculation is needed in each plane to convert the spatial propagator into the spatial frequency domain. Simulations of the pressure generated by the 3 cm $\times$ 3 cm piston with N ranging from 301 to 1024 show that the computation time for the spatial propagator is 1.1–1.9 times that for the spectral propagator. This ratio approaches the smaller value as N increases. For example, after the pressure is calculated with $N=301$ ($D=9$ cm), the field is truncated to a $201\times201\times200$ grid (6 cm $\times$ 6 cm $\times$ 30 cm volume) for error evaluations. The spectral propagator computes the complex pressure in 15.22 s with a normalized root mean squared error of 2.45%, and the spatial propagator computes the result in 28.90 s with a normalized root mean squared error of 1.39%. When $N=512$ ($D=15.33$ cm), the angular spectrum approach evaluated with the spectral propagator computes the 3D grid of complex pressures with a normalized root mean squared error of 1.47% in 1.23 min, whereas the spatial propagator takes 1.99 min using the same parameters while achieving a normalized root mean squared error of 1.38%. As the computational grid becomes larger, the difference between the spatial and spectral propagator becomes smaller in terms of both error and time.

Overall, the spectral propagator computes pressures faster than the spatial propagator in the same grid. Furthermore, the spatial propagator evaluated in a $D\times D$ plane yields accurate results only in an $L\times L$ plane, where $L=D-2a$, while the spectral propagator produces useful results in the entire $D\times D$ plane using the same $N\times N$ grid with $N=D/\delta+1$. The spatial propagator generates more accurate results than the spectral propagator in nonattenuating media, whereas the two propagators achieve similar accuracies for simulations in attenuating media or for an apodized velocity source. Therefore, the spectral propagator is preferred if the computational grid is large and time is a limiting factor (for example, when a large number of these calculations are performed in a parametric simulation). When the numerical accuracy is the primary consideration, the spatial propagator is preferred, especially in nonattenuating media.

C. Error sources for the spectral propagator

The spectral propagator is an analytical function, but the numerical implementation discretizes the spectral propagator by multiplying the continuous spectrum with $\mathrm{comb}(k_x/\Delta k, k_y/\Delta k)$, where $\Delta k=2\pi/(N\delta)$. The resulting spatial pressure distribution is therefore shifted and added in blocks of size $D\times D$. Since the analytical inverse Fourier transform of the spectral propagator extends to ± infinity in the spatial domain, the tails from all of the other blocks leak into the central $D\times D$ area and cause aliasing errors as shown in Fig. 3(a).

The nonzero portion of the spectral propagator is mostly confined within $k_x^2+k_y^2\leqslant k^2$, and the undersampling of the spectral propagator can be severe as $\sqrt{k_x^2+k_y^2}$ approaches k. In Eq. (6), two terms contain $\sqrt{k^2-k_x^2-k_y^2}$, where one is found in the exponential phase term, and the other is in the denominator. The real and imaginary parts of the numerator oscillate more rapidly as k_x and k_y approach $k_x^2+k_y^2=k^2$. Meanwhile, the value of the denominator decays quickly until the singularity is encountered at $k_x^2+k_y^2=k^2$. Both of these phenomena lead to numerical difficulties in terms of spectral sampling. The high spatial frequency components are reduced by applying a low-pass filter to $H_u(k_x,k_y,z)$ either through angular restriction or attenuation, and the same effect is achieved by filtering the source spectrum through apodization. In an effort to address the problem with singularities in $H_u(k_x,k_y,z)$ at $k_x^2+k_y^2=k^2$, different values were substituted for the infinite value encountered in this location, including 0, the amplitude of an adjacent point, and an average value of the adjacent points around the singularity.[14] Each of these strategies produces very similar errors. This suggests that the numerical singularities at $k_x^2+k_y^2=k^2$ have little influence on the overall error, especially when evaluated in the contexts of angular restriction, attenuation, or apodization.

D. Error sources for the spatial propagator

The spatial propagator $h_u(x,y,z)$ is smooth everywhere except at locations adjacent to the radiator, i.e., less than 1 wavelength from the piston surface, and the only singularity occurs on the piston surface at (0,0,0). Away from the piston surface, the Nyquist sampling rate of the spatial propagator is easily satisfied when $\delta\leqslant\lambda/2$ is the sampling interval.[3] The analytical representation of the spatial propagator extends to ± infinity in both the x and the y directions. However, the spatial propagator is truncated by a $D\times D$ window in angular spectrum calculations. When this truncated propagator is convolved with a velocity source of size $2a\times2a$ using 2D FFTs, circular convolution errors will appear in a band of width a along each edge as shown in Fig. 3(a). If the $D\times D$ plane is enlarged with zero padding, the circular convolution error is replaced with the windowing error due to the truncation of the spatial propagator. In other words, the pressure field that excludes the erroneous boundary is always $L\times L$, where $L=D-2a$, with or without zero padding.

V. CONCLUSION

The performance of spatial and spectral propagators applied to the angular spectrum approach is evaluated for near-field pressure simulations with a square piston. Calculations with the spatial propagator are performed in a $D\times D$ plane, and results show that the $L\times L$ portion at the center (with $L=D-2a$) contains accurate results, whereas the peripheral band of width a consistently contains errors with or without

zero padding. Therefore, the edge region computed with the spatial propagator should always be discarded. When the spectral propagator is used, undersampling in the spatial frequency domain causes errors in the computed pressure fields, and these are reduced in nonattenuating media by angular restriction. In attenuating media or when the source is apodized, the errors produced by the spatial propagator and the spectral propagator (with and without angular restriction) are small, and the difference between the spatial and spectral propagators is negligible. Meanwhile, the spatial frequency filtering effect of attenuation or apodization reduces the aliased components, and the need for angular restriction is eliminated in these cases. The spatial propagator, which only yields accurate results in planes of size $L \times L$, requires more calculation time than the spectral propagator. Thus, the spatial and spectral propagator each has distinct advantages and disadvantages that depend on the grid size, the attenuation value, and the source velocity distribution, and these determine the trade-offs between the numerical accuracy and the computation time for angular spectrum calculations.

ACKNOWLEDGMENTS

This work was supported in part by NIH R01CA093669, NIH R21CA121235, and NSF Theoretical Foundations Grant No. 0634786.

[1]J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed. (McGraw-Hill, New York, 1996).
[2]G. T. Clement and K. Hynynen, "Field characterization of therapeutic ultra-sound phased arrays through forward and backward planar projection," J. Acoust. Soc. Am. **108**, 441–446 (2000).
[3]D. P. Orofino and P. C. Pedersen, "Efficient angular spectrum decomposition of acoustic sources. I. Theory," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **40**, 238–249 (1993).
[4]D. P. Orofino and P. C. Pedersen, "Efficient angular spectrum decomposition of acoustic sources. II results," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **40**, 250–257 (1993).
[5]P. R. Stepanishen and K. C. Benjamin, "Forward and backward projection of acoustic fields using FFT methods," J. Acoust. Soc. Am. **71**, 803–812 (1982).
[6]P. Wu, R. Kazys, and T. Stepinski, "Analysis of the numerically implemented angular spectrum approach based on the evaluation of two-dimensional acoustic fields. I. Errors due to the discrete Fourier transform and discretization," J. Acoust. Soc. Am. **99**, 1339–1348 (1996).
[7]P. Wu, R. Kazys, and T. Stepinski, "Analysis of the numerically implemented angular spectrum approach based on the evaluation of two-dimensional acoustic fields. II. Characteristics as a function of angular range," J. Acoust. Soc. Am. **99**, 1349–1359 (1996).
[8]P. Wu, R. Kazys, and T. Stepinski, "Optimal selection of parameters for the angular spectrum approach to numerically evaluate acoustic fields," J. Acoust. Soc. Am. **101**, 125–134 (1997).
[9]P. Wu and T. Stepinski, "Extension of the angular spectrum approach to curved radiators," J. Acoust. Soc. Am. **105**, 2618–2627 (1999).
[10]P. T. Christopher and K. J. Parker, "New approaches to the linear propagation of acoustic fields," J. Acoust. Soc. Am. **90**, 507–521 (1991).
[11]D. Liu and R. C. Wagg, "Propagation and backpropagation for ultrasonic wavefront design," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 1–13 (1997).
[12]R. J. Zemp and J. T. Tavakkoli "Modelling of nonlinear ultrasound propagation in tissue from array transducers," J. Acoust. Soc. Am. **113**, 139–152 (2003).
[13]E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic, London, 1999).
[14]E. G. Williams and J. D. Maynard, "Numerical evaluation of the Rayleigh integral for planar radiators using the FFT," J. Acoust. Soc. Am. **72**, 2020–2030 (1982).
[15]R. J. McGough, "Rapid calculations of time-harmonic nearfield pressures produced by rectangular piston," J. Acoust. Soc. Am. **115**, 1934–1941 (2004).
[16]J. Zemanek, "Beam behavior within the nearfield of a vibrating piston," J. Acoust. Soc. Am. **49**, 181–191 (1971).
[17]J. C. Lockwood and J. G. Willette, "High-speed method for computing the exact solution for the pressure variations in the nearfield of a baffled piston," J. Acoust. Soc. Am. **53**, 735–741 (1973).

Comparative measurements of loudspeakers in a listening situation

Mathieu Lavandier,[a] Philippe Herzog,[b] and Sabine Meunier[c]
Laboratoire de Mécanique et d'Acoustique, C.N.R.S. UPR 7051, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

(Received 25 May 2007; revised 28 September 2007; accepted 29 October 2007)

Comparison of loudspeakers is a major concern during design or product selection. There are several standards for the measurement of loudspeaker characteristics, but none of them provides hints for a rigorous comparison between devices. In this study, different ways of evaluating acoustical dissimilarity between loudspeakers were compared. Several methods of signal analysis were used, and for each method a metric evaluating the dissimilarity between two signals was defined. The correlation between the different dissimilarity evaluations over a significant panel of loudspeakers led to identified classes of measurements. A specific aspect of this work is that measurements were performed in a standard listening environment, rather than in an anechoic or reverberant one. It allowed the use of the recorded signals for a simple listening test, providing a perceptual metric which was compared to the acoustical ones. It also allowed the introduction of auditory models in the computation of some acoustical metrics, so defining a new class of measurements which gave results close to the perceptual ones. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2816571]

PACS number(s): 43.38.Md, 43.58.Kr, 43.20.Ye, 43.66.Lj [AJZ] Pages: 77–87

I. INTRODUCTION

There are situations when studying loudspeakers where relatively slight differences must be quantified: parametric studies, classifications, or ranking of experimental or design factors. Dealing with loudspeakers requires that such comparative measurements use a rigorous protocol, especially when dealing with small changes, but also that their results correlate well with the differences which would be perceived by a listener. The need for such a tool might increase nowadays, as there is a trend to build active loudspeakers with embedded amplifiers and dedicated electronics, including linear filtering, adaptation to radiation conditions, and potentially dynamics or nonlinear processing. Studying such complex systems which involve many parameters usually requires the comparison of numerous samples (different loudspeakers and/or tunings). Associated measurements should be as fast as possible, and hopefully require widely available facilities.

This paper focuses on the comparison of loudspeakers by acoustical measurements, with a special interest for differences which can be perceived by average listeners. The long term objective of our work is to build a measurement tool for assessing the relevance of modifications in design or modeling of loudspeakers, in relation to their perceived influence. As a first step, our main interest was the timbre-related dissimilarities between loudspeakers, generally investigated in monophonic reproduction.

Timbre-related accuracy is a multidimensional characteristic of the perception of reproduced sound. It was studied in different ways depending on the aims of the researchers. Some researchers investigated the perceptual dimensions underlying the judgements of listeners.[1–5] They either used dissimilarity evaluations and multidimensional scaling analysis,[6] or absolute evaluations on chosen semantic scales and factor analysis.[7] Other researchers investigated the fidelity judgments or preferences of listeners as a global absolute evaluation of loudspeakers,[8–13] in order to look for their link with individual perceptual dimensions or acoustical measurements (see Toole[14] for a review of this approach). Again, both our acoustical and perceptual approaches were focused on the relative differences between loudspeakers, but not on their absolute quality or the preference of listeners. The study presented in this paper identifies the most relevant acoustical discrimination of loudspeakers regarding perception, without evaluation along a quality scale or any particular perceptual scale.

Researchers and manufacturers have at their disposal different standards to evaluate their loudspeakers, physically by using acoustical measurements,[15,16] or perceptually by using listening tests.[17–21] The goal of a standard is that an absolute evaluation of a loudspeaker performed at a given time and location could be compared to the same measurement performed at another time or elsewhere. For this reason, standards define measurement environments, test signals, and measuring equipment. However, they allow different combinations of those as long as the actual measurement conditions are specified together with the results. This adds to the uncertainty of the measurement results, especially when they must be compared: some of the differences could be related to differences in the measurement protocol rather than to actual differences between devices. When relative

[a]Electronic mail: lavandiermn@cardiff.ac.uk; Presently at: School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom.
[b]Electronic mail: herzog@lma.cnrs-mrs.fr
[c]Electronic mail: meunier@lma.cnrs-mrs.fr

differences between loudspeakers are the main goal of the measurement, it is therefore better to use a single measurement setup so that changing the device under test is, to some extent, the only parameter that varies.

An important aspect of the measurements of loudspeakers is that their indirect goal is to quantify the quality of the sound reproduction, which is a very complicated notion, mixing (at least) physical, psychological, and cultural aspects. Listening tests on loudspeakers require facilities that are close to ordinary rooms. Such facilities are designed so that they provide a usual "look and feel" to the listener, and thus avoid bias listening habits. They must, however, ensure a limited noise and a controlled reverberation environment. More generally, they should not add a specific character to the sound reproduction. Beyond this need for a "natural" reproduction environment, a standardized listening environment ensures that the loudspeaker itself has the same behavior during its test and its normal use. The room may have a significant influence on the vibration of loudspeaker parts, and it has a major influence on the spatial repartition of the acoustic field.[22,23] The frequency and dynamic content of the program excerpts also has some influence on the test results, as it may interact with nonlinear and memory effects.[24] Hence, the test signals have to be speech or musical excerpts chosen with great care. When dealing with a comparison of loudspeakers which are very similar (i.e., prototypes with slight design changes), selecting a realistic environment is more relevant than choosing one of the usual standardized acoustical measurement environments.

Searching for a relationship between measurements and perception is not a new topic. There is a general agreement among researchers on the fact that "the frequency response of the loudspeaker is the most important factor related to perceived sound quality,"[25] but there is less agreement on the most relevant way to measure this frequency response in order to link it with perception.[26] The timbre-related accuracy of a loudspeaker might also be influenced by its nonlinearities and its directivity, which mediates its interaction with the room.[14] 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 to greater or lesser extents and only a posteriori. Measurements were done in listening rooms,[2,27] or this environment was resimulated more or less completely from anechoic measurements.[4,25,28] After taking into account the listening room, some measurements considered the influence of the musical excerpt,[4,5,27] and the properties of the listener auditory system.[2,4,5] In previous studies, either acoustical measurements were compared but their link with perceptual evaluation was not quantified,[2,27,28] or the link between acoustical and perceptual evaluations was quantified but acoustical measurements were chosen a priori among several possibilities without comparing these different possibilities.[4,5,13] The aim of our study was to compare different acoustical discriminations of loudspeakers, and quantify their link with a parallel perceptual evaluation.

In order to find a link between our acoustical and perceptual evaluations, and in front of the difficulty to do it a posteriori, we tried to keep these two approaches as close as possible one to the other. The acoustical measurements had to be done in the same environment as the listening tests, and preferably at the same time, so we could be sure to measure the same sound field along both approaches. To comply with the constraints on listening tests,[17–21] loudspeakers had to be compared in the same room, for the same positions of receiver (microphone/listener) and loudspeaker. To evaluate their relative differences, they also had to be compared one just after the other, due to our short auditory memory.[11] An experimental protocol compatible with both the acoustical and perceptual approaches was designed. It first consisted of recording the sound radiated by loudspeakers in a room, and then submitting the recorded sounds to signal analysis as well as to listening tests using headphones.

Of course this protocol suffers from restrictions. For example, the spatial dimension of reproduced sound and the interaction between loudspeaker and room cannot be reliably investigated using recordings and headphones; neither could an absolute evaluation of the sound reproduction such as preference or fidelity be undertaken with such protocol. We did not intend to evaluate all the potential dissimilarities between the loudspeakers tested, as some of these dissimilarities might not be present in the final recordings. Our aim was to study the dominant remaining dissimilarities, with the advantage of complying with the constraints mentioned above, of being sure that the remaining dissimilarities were associated with the loudspeakers and not another experimental parameter, and of having a direct access to the signals presented to the listeners. Previous studies dealing with the perception of reproduced sound compared listening tests realized live or using headphones.[11,29–31] Even if the studies had various aims, they showed that the two types of listening tests gave very similar results concerning timbre-related accuracy.

Several methods of signal analysis were tested on our recordings, and for each method a metric evaluating the dissimilarity between two signals was defined. These acoustical dissimilarities were computed for the recordings, and were systematically compared to the perceptual dissimilarities resulting from the listening tests, in order to identify the most relevant method for the acoustical discrimination. We chose to base the acoustical approach on signal analysis done directly on the recordings used for the listening tests, rather than considering estimations of the responses of loudspeakers, in order to remain as close as possible to the signals heard and judged by listeners. Instead of measuring separately the frequency response, the directivity, and the nonlinear characteristics of the loudspeakers in the room, we considered a measurement involving a combination of these effects. Once the link with perception is established, separate measurements might be considered. Our approach followed the trend encountered in the literature, which showed that taking into account listening conditions helped to bridge the gap between measurement and perception. The listening conditions were then directly recorded.

This paper first presents 13 ways of calculating the acoustical dissimilarity between two signals. Our acoustical measurements of loudspeakers are then detailed. They are used to compare the different techniques of acoustical dis-

TABLE I. Definition of the acoustical dissimilarities ($\langle\ \rangle_{t,f,b}$: arithmetic mean over time t, frequency f, or Bark scale b, $|||$: modulus of a complex value, Min{ }: minimum value, Max{ }: maximum value, DFT: discrete Fourier transform, STFT: short-time Fourier transform, PSD: power spectral density, PSD_w: weighted power spectral density, Dens_o: overall specific loudness, Dens: time-varying specific loudness, Dens_m: temporal mean of time-varying specific loudness).

Method of analysis	Acoustical dissimilarity between signals $x(t)$ and $y(t)$
(a) Time signal	$\mathrm{Min}\{\langle[x(t)-y(t)]^2\rangle_t, \langle[x(t)+y(t)]^2\rangle_t\}$
(b) Spectrum (complex)	$\mathrm{Min}\{\langle\|\mathrm{DFT}(x)-\mathrm{DFT}(y)\|^2\rangle_f, \langle\|\mathrm{DFT}(x)+\mathrm{DFT}(y)\|^2\rangle_f\}$
(c) Time-frequency transform	$\mathrm{Min}\{\langle\langle\|\mathrm{STFT}(x)-\mathrm{STFT}(y)\|^2\rangle_f\rangle_t, \langle\langle\|\mathrm{STFT}(x)+\mathrm{STFT}(y)\|^2\rangle_f\rangle_t\}$
(d) Modulus of spectrum	$\langle[\|\mathrm{DFT}(x)\|-\|\mathrm{DFT}(y)\|]^2\rangle_f$
(e) Modulus of time-frequency transform	$\langle\langle[\|\mathrm{STFT}(x)\|-\|\mathrm{STFT}(y)\|]^2\rangle_f\rangle_t$
(f) Power spectral density (PSD)	$\langle[\sqrt{\mathrm{PSD}(x)}-\sqrt{\mathrm{PSD}(y)}]^2\rangle_f$
(g) A weighted PSD	$\langle[\sqrt{\mathrm{PSD}_W(x)}-\sqrt{\mathrm{PSD}_W(y)}]^2\rangle_f$
(h) 70 phone contour weighted PSD	$\langle[\sqrt{\mathrm{PSD}_W(x)}-\sqrt{\mathrm{PSD}_W(y)}]^2\rangle_f$
(i) Overall specific loudness	$\langle\mathrm{Max}\{\mathrm{Dens}_o(x),\mathrm{Dens}_o(y)\}/\mathrm{Min}\{\mathrm{Dens}_o(x),\mathrm{Dens}_o(y)\}-1\rangle_b$
(j) Time-varying specific loudness 1	$\langle\langle\mathrm{Max}\{\mathrm{Dens}(x),\mathrm{Dens}(y)\}/\mathrm{Min}\{\mathrm{Dens}(x),\mathrm{Dens}(y)\}-1\rangle_b\rangle_t$
(k) Temporal mean of time-varying specific loudness 1	$\langle\mathrm{Max}\{\mathrm{Dens}_m(x),\mathrm{Dens}_m(y)\}/\mathrm{Min}\{\mathrm{Dens}_m(x),\mathrm{Dens}_m(y)\}-1\rangle_b$
(l) Time-varying specific loudness 2	$\langle\langle\mathrm{Max}\{\mathrm{Dens}(x),\mathrm{Dens}(y)\}/\mathrm{Min}\{\mathrm{Dens}(x),\mathrm{Dens}(y)\}-1\rangle_b\rangle_t$
(m) Temporal mean of time-varying specific loudness 2	$\langle\mathrm{Max}\{\mathrm{Dens}_m(x),\mathrm{Dens}_m(y)\}/\mathrm{Min}\{\mathrm{Dens}_m(x),\mathrm{Dens}_m(y)\}-1\rangle_b$

crimination. The acoustical dissimilarities are finally compared to the perceptual dissimilarities resulting from the listening tests. The multidimensional analysis of the measured dissimilarities is not considered here. It will be described in detail in a following paper.

II. ACOUSTICAL DISSIMILARITIES

A. Definitions

Thirteen evaluations of the acoustical dissimilarity between two signals were defined. Different methods of signal analysis were assessed. The temporal, spectral, and time-frequency domains were investigated. Two spectral weightings were tested. Different auditory models were also used. For each method of analysis, we defined a metric evaluating the acoustical dissimilarity between two signals. The different types of acoustical dissimilarities are presented in Table I.

Except for the analyses involving auditory models, the acoustical dissimilarities were all quadratic regarding the compared variables. The corresponding metrics were all defined in the same way in order to focus on the methods of analysis. Research is still ongoing testing other types of metrics. Each metric was defined to evaluate a global acoustical dissimilarity between two signals, giving a single value from the differences between the analyses of these two signals. To get a single scalar, dissimilarities calculated at each sample period or each frequency were integrated. As we could not suppose that one sample period or frequency region was more important than another, this integration was realized by an arithmetic mean, giving the same weight to all dissimilarities.

1. Time, spectral, and time-frequency domains

The spectral domain was investigated using the discrete Fourier transform (DFT) of the signals (b and d) and their power spectral density (PSD) (f, g, and h). The calculation of the power spectral density involved successive windows of 8192 samples with nonoverlapping Hanning windows. We checked that the dissimilarities obtained were largely independent of the frequency resolution used for the computation. The time-frequency plane was obtained by calculation of the short-time Fourier transform (STFT) of the signals (c and e). It involved successive windows of 1024 samples, with overlapping Hanning windows.

We considered the spectral and time-frequency domains with (b and c) and without (d, e, f) phase information. When phase information was kept, and also in the time domain (a), a potential phase inversion between the two compared signals x and y was taken into account by considering y and $-y$.

2. Spectral weightings

Spectral weightings were applied to the power spectral density of our signals, to model the ear sensitivity. They were the A weighting (g) and a weighting based on the normal equal-loudness contour[32] at 70 phones (h), which is close to the sound level used during our listening tests.

3. Auditory models

Auditory models are sometimes considered as perceptual analyses, because they were built using human perception. However, they are considered here as signal representations, because they are used like any other signal analysis technique, without any listener being involved. The only dissimilarities considered as perceptual will be those directly resulting from the listening tests.

Two auditory models were involved in our evaluations of acoustical dissimilarity. These models were proposed by Zwicker and Fastl[33] to calculate loudness, and are based on auditory masking. They were used to analyze our signals in terms of specific loudness which is the density of loudness along the Bark scale, which models the perceptual frequency scale. The first model[34,35] 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 our signals, which is the specific loudness determined over the entire signal taken at once. The second model was designed for nonstationary sounds.[33,36] It takes into account both frequency and temporal masking, calculating specific loudness every 10 ms. Two

time-frequency patterns of specific loudness were examined for our signals. 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 (j). For the second pattern, the second auditory model was applied. We called the resulting analysis time-varying specific loudness 2 (l). Finally, the temporal means of time-varying specific loudnesses 1 (k) and 2 (m) were considered. Compared to overall specific loudness, they took into account the fact that auditory masking depends on the spectral content of signals, and that this content varied over time for nonstationary musical signals.

Two loudnesses have to be compared by their ratio and not by their difference.[33] Thus, acoustical dissimilarities between two signals were computed by taking the ratio of their specific loudnesses (Table I). 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 ratios, a threshold was introduced in order to avoid artificially high dissimilarities where there was no useful signal but noise. The threshold was arbitrarily chosen at 0.03 sone/Bark. We also subtracted one from each ratio to make sure that two identical values of specific loudness would lead to a dissimilarity equal to zero. Integrations of ratios over time and Bark scales were done by taking the arithmetic mean of the dissimilarities over these two scales. Figure 1 illustrates our calculation of the dissimilarity between two specific loudnesses, taking the example of the overall specific loudness.

The definition of a metric from the output of the auditory models involved the estimation of a single scalar to quantify the difference between two signals. For the stationary analysis (i), the comparison data were straightforwardly integrated over the Bark scale. For the time-varying analyses, a temporal integration had to be performed. The auditory masking effects depend on the spectral content of the analyzed signals, and this content may vary over time. The metric may thus vary significantly when changing the order of the integrations. We tested the two possible orders. The two methods involving the time-varying specific loudnesses therefore differed only by the fact that the temporal integration was done as the first step for the temporal means of time-varying specific loudness (k) and (m), and as the last step for time-varying specific loudnesses (j) and (l). The combination of the two auditory models and two methods thus led to four different metrics, in addition to the stationary one.

B. Loudness equalization

Before computing the acoustical dissimilarities, signals had to be equalized in level, to prevent any sound level difference from creating an artificial dissimilarity between them. In the same way, any loudness differences during the listening tests would create uninteresting perceptual dissimilarities potentially masking more subtle ones.[8] Before the listening tests and the acoustical analyses, the overall loudness of our signals was set to 70 phons. This equalization

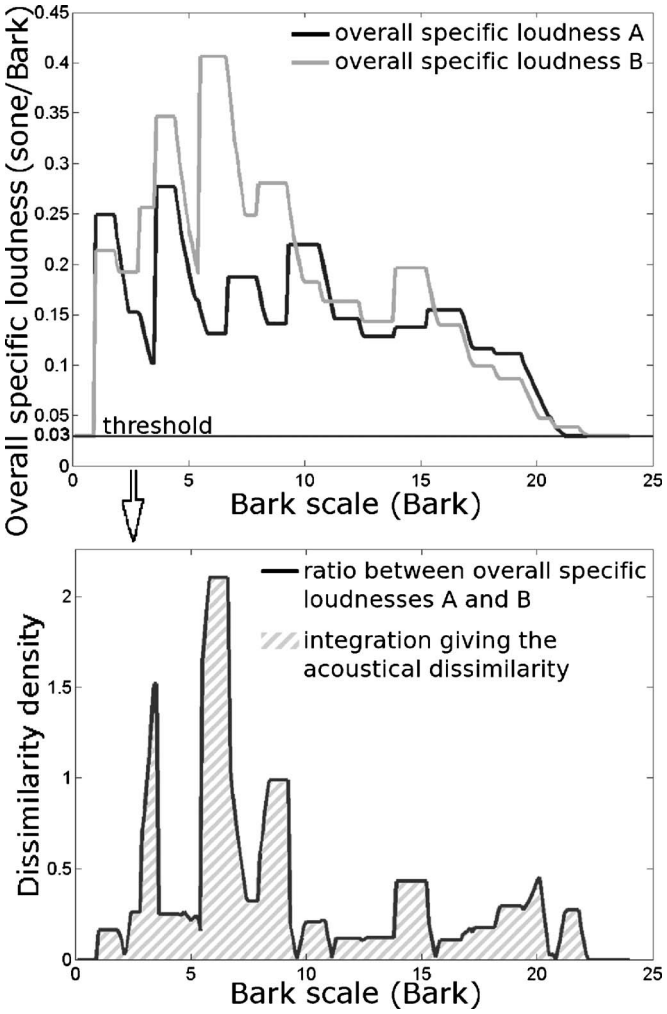


FIG. 1. Calculation of the acoustical dissimilarity between two overall specific loudnesses *A* and *B*.

was realized subjectively by the experimenters, and verified by computing the overall loudness of the equalized stimuli using a loudness estimation model.[33,36]

C. Synchronization

In the same way as loudness was equalized to avoid uninteresting dissimilarities, our signals had to be synchronized. Time delays of a few sample periods between signals, although being crucial for the acoustical analyses, are not perceived during listening tests. Acoustical analyses should copy this behavior, and signals thus had to be synchronized in order to eliminate this irrelevant information. The evaluation of the time delay between two signals was not trivial. Our signals were not identical signals simply delayed in time: the musical stimuli used for the recording were filtered by the different loudspeakers and the room before being recorded by the microphones. The evaluation of the time delay between two signals depended on the frequency range considered, and so on the frequency responses of the loudspeakers involved. The reverberation and the first reflections recorded in the room further complicated the task.

Three methods evaluating the time delay between two signals were tested. The first one evaluated the pure delay

TABLE II. 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

contained in the transfer function between the two signals.[37] The minimum phase component of this transfer function was extracted by considering its complex cepstrum.[38] The time delay between the signals was estimated by a linear regression on the phase excess. The frequency range considered for this regression was restricted to the range containing useful signal, i.e., to the range for which the coherence between the signals was close to unity.[37] Even when the criterion of high coherence was verified, the time delays obtained by this method appeared to be very dependent on the frequency range considered, and the method was abandoned. The second method estimated the time delay between two signals by maximizing their cross-correlation function.[37] The third method consisted of minimizing their dissimilarity in the time domain (a, Table I). This dissimilarity is minimum for synchronized signals. Computing the dissimilarity for the first signal and delayed versions of the second signal allowed identifying the delay minimizing the dissimilarity, and compensating for this delay, thus synchronizing the two signals. The delay applied to the second signal was varied by steps of one sample period, testing positive and negative values of delay. Any potential phase inversion between the two signals was taken into account in the definition of the metric.

None of the tested methods allowed a perfect global synchronization of all our signals at the same time. We found some incoherent results such as the delay A-C between signals A and C being different from the sum of the delays A-B and B-C. The observed differences were limited to a few sample periods for the last two methods mentioned above. We could have chosen one signal as a reference, and synchronized all the signals with this unique reference, but then the signals might not have been perfectly synchronized with all the others signals. So, we chose to synchronize the signals taking them by pairs. The time delay between two signals was determined within one sample period by minimizing their dissimilarity in the time domain for each pair of signals, independently of the other pairs. Acoustical dissimilarities were then evaluated within each synchronized pair.

III. ACOUSTICAL MEASUREMENTS

The acoustical dissimilarities presented in the section above were tested on acoustical measurements of loudspeakers. These measurements consisted of recording the sound radiated by loudspeakers in a room, using different short musical excerpts. As the spatial component of sound reproduction could not be reliably investigated with our protocol, the loudspeakers were used in monophonic reproduction. "Timbre-related accuracy is much more easily heard in single-loudspeaker listening."[18]

A. Loudspeakers

Twelve single loudspeakers were measured. We tried to use a wide range of models coming from different manufactures: high-fidelity column or bookshelves, studio monitoring, and computer monitoring. To get a reference for comparison, two loudspeakers of the same model were measured.

B. Room

The room used for the measurements 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.[39] The microphones were at 1 m from the floor and the point between the medium and tweeter of the loudspeakers was approximately at this same height. The floor was entirely carpeted. The reverberation time was evaluated at the position of the recording microphones, with the source at the loudspeaker position. The measurement microphone was omnidirectional and the source was a Genelec 1031A loudspeaker reproducing pink noise. The reverberation time was evaluated by measuring the time required for the sound level to decrease by 60 dB after the offset of the noise. The measured signals were filtered by third-octave bands and the reverberation time was evaluated for each band individually (Table II). Its value was found around 0.5 s at midrange frequencies, satisfying standard requirements.[17,19]

C. Signals

Three musical excerpts were used for the recordings and we only kept a short part of them for the computation of the acoustical dissimilarities: (1) McCoy Tyner ("Miss Bea", right channel, jazz, 3.3 s), (2) Kan'nida ("Konsyans", left channel, percussion, 1.7 s), and (3) Vivaldi ("L'Europa Galante", left channel, symphonic orchestra, 4.7 s). Such short excerpts were also suitable for the evaluation of perceptual dissimilarities between loudspeakers.[40–42]

D. Procedure

Several recording techniques were tested and three were submitted to listeners: the stereophonic AB ORTF, stereophonic MS, and monophonic omnidirectional techniques.[39] These three types of recordings were chosen for their differences among the available recordings. The binaural technique using an artificial head was tested, but it did not allow a better externalization of our recordings involving frontal monophonic loudspeakers. In order to externalize frontal sources, individualized head related transfer functions (HRTFs) should be used as well as a head tracker to follow head movements. As our goal was to propose a sufficiently

easy way to compare loudspeakers, this technique was abandoned. As our study was focused on the timbre-related accuracy of the sound reproduction rather than its spatial component, the use of binaural techniques seemed less critical. The dissimilarity results from the three recording techniques submitted to listeners were very similar, so that the other techniques were not involved in listening tests, and only the results from the ORTF technique (arbitrarily chosen) are presented here. This technique involved 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 a 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 carried out directly with an audio workstation featuring a RME DIGI9652 sound card and an external Fostex VC-8 analog to digital (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 prior to the acoustical analyses and listening tests (see Sec. II B).

IV. COMPARISON OF ACOUSTICAL DISSIMILARITIES

The 13 types of acoustical dissimilarities (Table I) were computed on the recordings presented above. The right and left channels of these stereophonic recordings were analyzed separately. For each musical excerpt and each channel, we determined the acoustical dissimilarities among the 12 recordings corresponding to the 12 measured loudspeakers, leading to 13 66-dissimilarity vectors. The different acoustical dissimilarities were compared by evaluating the coefficient of correlation between these dissimilarity vectors.

A. Results

Figure 2 presents the coefficients of correlation between the acoustical dissimilarities calculated on the right and left channels of the recordings, for the three musical excerpts tested. The significant correlations are indicated in bold and marked with an asterisk. The individual significance level was fixed at 0.0005, following the conservative Bonferroni correction,[43] in order to keep the family wise significance level below 0.05 for the 78 tests of correlation realized on each recording channel. The right and left channels of the stereophonic recordings did not lead to exactly the same results, but the correlations followed the same trend with both channels. A perfect correlation was observed between the dissimilarities calculated in the time, spectral, and time-frequency domains [(a), (b) and (c); or (d), (e) and (f)], whereas weak or no correlations were obtained when comparing these domains with or without phase information [(a), (b), (c) and (d), (e), (f)]. The dissimilarities using phase information were not correlated with any of the other types of dissimilarity for the excerpt McCoy Tyner. The dissimilarities based on the two spectral weightings (g) and (h) were almost perfectly correlated. They were strongly correlated with the dissimilarities involving the spectral and time-frequency domains without weightings nor phase information [(d), (e), and (f)] for the excerpt Vivaldi, but not for the two others musical excerpts. The dissimilarities based on the different specific loudnesses were strongly correlated [(i), (j), (k), (l) and (m)].

B. Discussion

The comparison of the acoustical dissimilarities showed that some dissimilarity evaluations were identical, but also that there were large differences among the tested methods, leading to identified classes of measurements.

It confirmed that the time, spectral, and time-frequency domains contained the same information. As our metrics were all defined in the same way from these different representations—quadratic regarding the compared variables (Table I)—they led to perfectly correlated dissimilarities. Conversely, taking into account or discarding phase information led to large differences in the discrimination of signals. Again, the calculations based on the power spectral density of signals led to the same results as the calculations based on the modulus of their spectrum or time-frequency transform: these are three representations of the same information. The *A* weighting and the weighting based on the normal equal-loudness contour at 70 phones gave almost identical results. For the excerpt Vivaldi, taking these weightings into account did not seem to change the information contained in the dissimilarities with no weighting. The different metrics using auditory models led to dissimilarities close to each others, but not identical, indicating that they might not contain exactly the same information.

The comparison of the acoustical dissimilarities showed that different types of dissimilarity contained different information, but did not show which one, if any, contained relevant information regarding perception. The acoustical dissimilarities were then compared to perceptual ones, in order to determine the analyses and metrics showing the differences heard by listeners.

V. RELEVANCE OF ACOUSTICAL DISSIMILARITIES REGARDING PERCEPTION

A. Perceptual dissimilarities

1. Procedure

The perceptual dissimilarities between the recordings were gathered during three listening tests, one for each musical excerpt. These tests were paired comparisons. 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." This evaluation of the perceptual dissimilarities between the recordings was completely similar to the evaluation of their acoustical dissimilarities presented above.

Before the listening tests, the overall loudness of the recordings was set to 70 phons. This prevented loudness

McCoy Tyner

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)
(a)		1.00*	1.00*	0.22	0.23	0.21	0.03	0.04	0.07	0.10	0.11	0.13	0.14
(b)	1.00*		1.00*	0.22	0.23	0.21	0.02	0.04	0.07	0.10	0.11	0.13	0.14
(c)	1.00*	1.00*		0.22	0.23	0.21	0.02	0.04	0.07	0.10	0.11	0.13	0.14
(d)	0.03	0.03	0.03		1.00*	1.00*	0.18	0.23	0.42*	0.52*	0.55*	0.51*	0.58*
(e)	0.08	0.08	0.08	0.98*		1.00*	0.21	0.26	0.43*	0.52*	0.54*	0.53*	0.57*
(f)	0.04	0.04	0.04	0.99*	0.98*		0.18	0.23	0.42*	0.50*	0.54*	0.50*	0.57*
(g)	0.08	0.08	0.08	0.65*	0.68*	0.59*		0.97*	0.60*	0.66*	0.48*	0.73*	0.43*
(h)	0.08	0.08	0.08	0.65*	0.68*	0.58*	0.98*		0.61*	0.69*	0.52*	0.75*	0.46*
(i)	0.05	0.05	0.05	0.62*	0.61*	0.56*	0.74*	0.73*		0.95*	0.96*	0.90*	0.92*
(j)	0.06	0.06	0.06	0.71*	0.71*	0.65*	0.80*	0.81*	0.96*		0.94*	0.95*	0.90*
(k)	0.07	0.07	0.07	0.65*	0.63*	0.60*	0.70*	0.69*	0.98*	0.96*		0.87*	0.98*
(l)	0.13	0.13	0.13	0.71*	0.75*	0.66*	0.84*	0.84*	0.89*	0.96*	0.88*		0.83*
(m)	0.06	0.06	0.06	0.67*	0.64*	0.63*	0.71*	0.70*	0.96*	0.94*	0.99*	0.87*	

left channel

right channel

Kan'nida

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)
(a)		1.00*	1.00*	0.54*	0.53*	0.52*	0.35	0.44*	0.57*	0.54*	0.48*	0.61*	0.54*
(b)	1.00*		1.00*	0.54*	0.53*	0.52*	0.35	0.44*	0.57*	0.54*	0.48*	0.61*	0.54*
(c)	1.00*	1.00*		0.54*	0.53*	0.52*	0.35	0.44*	0.57*	0.54*	0.48*	0.61*	0.54*
(d)	0.48*	0.49*	0.49*		1.00*	1.00*	0.23	0.38	0.57*	0.51*	0.51*	0.66*	0.62*
(e)	0.48*	0.48*	0.49*	1.00*		1.00*	0.23	0.38	0.57*	0.50*	0.50*	0.65*	0.61*
(f)	0.47*	0.47*	0.47*	1.00*	1.00*		0.24	0.38	0.60*	0.53*	0.53*	0.67*	0.64*
(g)	0.30	0.30	0.30	0.18	0.18	0.19		0.98*	0.65*	0.79*	0.76*	0.77*	0.68*
(h)	0.39	0.39	0.39	0.34	0.33	0.34	0.98*		0.70*	0.81*	0.79*	0.82*	0.73*
(i)	0.52*	0.52*	0.52*	0.41	0.40	0.43*	0.73*	0.77*		0.92*	0.90*	0.90*	0.87*
(j)	0.50*	0.50*	0.50*	0.39	0.38	0.41	0.73*	0.76*	0.98*		0.99*	0.96*	0.96*
(k)	0.47*	0.47*	0.47*	0.39	0.38	0.40	0.70*	0.73*	0.97*	0.99*		0.94*	0.97*
(l)	0.55*	0.55*	0.55*	0.57*	0.56*	0.58*	0.75*	0.79*	0.94*	0.96*	0.94*		0.95*
(m)	0.48*	0.48*	0.48*	0.53*	0.52*	0.55*	0.65*	0.70*	0.92*	0.96*	0.97*	0.95*	

left channel

right channel

Vivaldi

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)
(a)		1.00*	1.00*	0.36	0.36	0.40	0.36	0.37	0.30	0.34	0.38	0.31	0.37
(b)	1.00*		1.00*	0.36	0.36	0.40	0.36	0.37	0.30	0.34	0.38	0.31	0.37
(c)	1.00*	1.00*		0.36	0.36	0.40	0.36	0.37	0.31	0.34	0.38	0.31	0.37
(d)	0.48*	0.48*	0.48*		1.00*	1.00*	0.96*	0.96*	0.72*	0.83*	0.75*	0.85*	0.76*
(e)	0.49*	0.49*	0.49*	1.00*		1.00*	0.96*	0.96*	0.72*	0.84*	0.76*	0.85*	0.76*
(f)	0.48*	0.48*	0.48*	1.00*	1.00*		0.97*	0.97*	0.71*	0.83*	0.75*	0.84*	0.76*
(g)	0.45*	0.45*	0.45*	0.96*	0.96*	0.97*		0.99*	0.72*	0.80*	0.74*	0.81*	0.74*
(h)	0.48*	0.48*	0.49*	0.97*	0.97*	0.97*	0.99*		0.74*	0.82*	0.76*	0.82*	0.77*
(i)	0.54*	0.54*	0.54*	0.68*	0.69*	0.67*	0.60*	0.63*		0.94*	0.96*	0.92*	0.95*
(j)	0.49*	0.49*	0.49*	0.75*	0.76*	0.73*	0.69*	0.70*	0.96*		0.98*	0.99*	0.97*
(k)	0.52*	0.52*	0.52*	0.63*	0.64*	0.62*	0.56*	0.58*	0.97*	0.97*		0.95*	0.99*
(l)	0.47*	0.47*	0.47*	0.78*	0.79*	0.76*	0.72*	0.73*	0.94*	0.99*	0.94*		0.95*
(m)	0.54*	0.54*	0.54*	0.69*	0.70*	0.68*	0.63*	0.65*	0.96*	0.96*	0.99*	0.94*	

left channel

right channel

FIG. 2. Correlations between the acoustical dissimilarities calculated on the right and left channels of the recordings, for the three musical excerpts (McCoy Tyner, Kan'nida, Vivaldi): (a) time signal; (b) spectrum; (c) time-frequency transform; (d) modulus of spectrum; (e) modulus of time-frequency transform; (f) power spectral density (PSD); (g) *A* weighted PSD; (h) 70 phone contour weighted PSD; (i) overall specific loudness; (j) time-varying specific loudness 1; (k) temporal mean of time-varying specific loudness 1; (l) time-varying specific loudness 2; and (m) temporal mean of time-varying specific loudness 2. The significant correlation coefficients are in bold, marked with an asterisk.

from creating uninteresting dissimilarities, potentially masking more subtle ones. Listening tests were run in an isolated soundproof room, using a Tucker&Davis workstation and Stax SR Lambda Professional headphones.

2. Participants

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. They were all otologically normal. They were members of the laboratory or students. None of them had significant previous experience in loudspeaker comparison.

3. Results

For each 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 the listeners. This verification was realized by a cluster analysis of the similarities between listeners, with the similarity between two listeners being evaluated by the correlation of their judgments.[44] Multidimensional analyses of the perceptual dissimilarities are not presented here, as the topic of this paper is to evaluate the relevance of the acoustical dissimilarities regarding perception. Perceptual dissimilarities are used here as a reference for this evaluation.

Figure 3 presents the perceptual dissimilarities measured for the 66 pairs of recordings and the three musical excerpts. These dissimilarities ranged from 0.22 to 0.83 (McCoy Tyner), 0.13 to 0.90 (Kan'nida), and 0.21 to 0.84 (Vivaldi), indicating that the recordings of the chosen loudspeakers and musical excerpts led to a broad range of dissimilarities. This was further verified for each listening test by comparing the different values of dissimilarity. Paired-sample *t* tests were

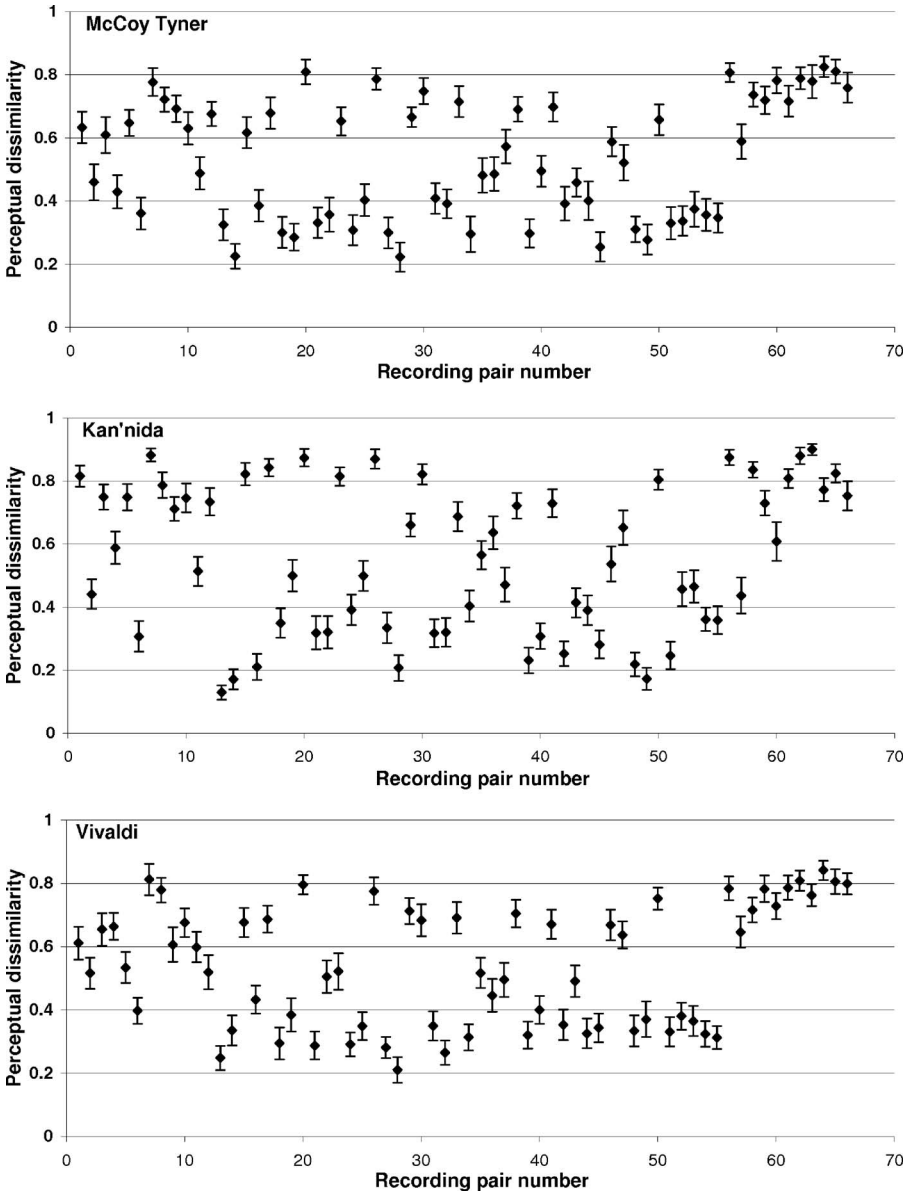


FIG. 3. Mean perceptual dissimilarities with standard errors measured for the 66 pairs of recordings and the three musical excerpts (McCoy Tyner, Kan'nida, Vivaldi).

used to investigate the significance of the differences between these values. The individual significance level of each *t* test was fixed at 0.00002, following the Bonferroni correction, in order to keep the family wise significance level below 0.05 for the 2145 comparisons realized for each musical excerpt. These statistical analyses confirmed that 621 dissimilarities were significantly different for the excerpt McCoy Tyner. The number of significantly different dissimilarities was 876 for the excerpt Kan'nida and 672 for the excerpt Vivaldi.

B. Correlation between acoustical and perceptual dissimilarities

For each musical excerpt, the perceptual dissimilarities averaged over all listeners resulted in a 66-dissimilarity vector. This vector was used as a reference to assess the relevance of the acoustical discrimination methods regarding perception. This assessment was realized by evaluating the coefficient of correlation between the acoustical and perceptual dissimilarity vectors. Although the absolute values of these correlations might be contaminated by noise and nonlinearities, the values obtained with the different acoustical methods could be compared to determine which acoustical dissimilarities were the closest to the judgements of listeners.

1. Results

Figure 4 presents the coefficients of correlation between the mean perceptual dissimilarities resulting from the listening tests and the acoustical dissimilarities calculated on the right and left channels of the recordings using each of the 13 acoustical discrimination methods, for the three musical excerpts involved. The significant correlations are marked with an asterisk. The individual significance level was fixed at 0.001, following the Bonferroni correction, in order to keep the family wise significance level below 0.05 for the 26 tests of correlation realized for each musical excerpt. The correlations obtained with the acoustical dissimilarities calculated on the right and left channels of the recordings followed the same trend. The effects were also similar for the three musical excerpts. Dissimilarities based on the time (a), spectral

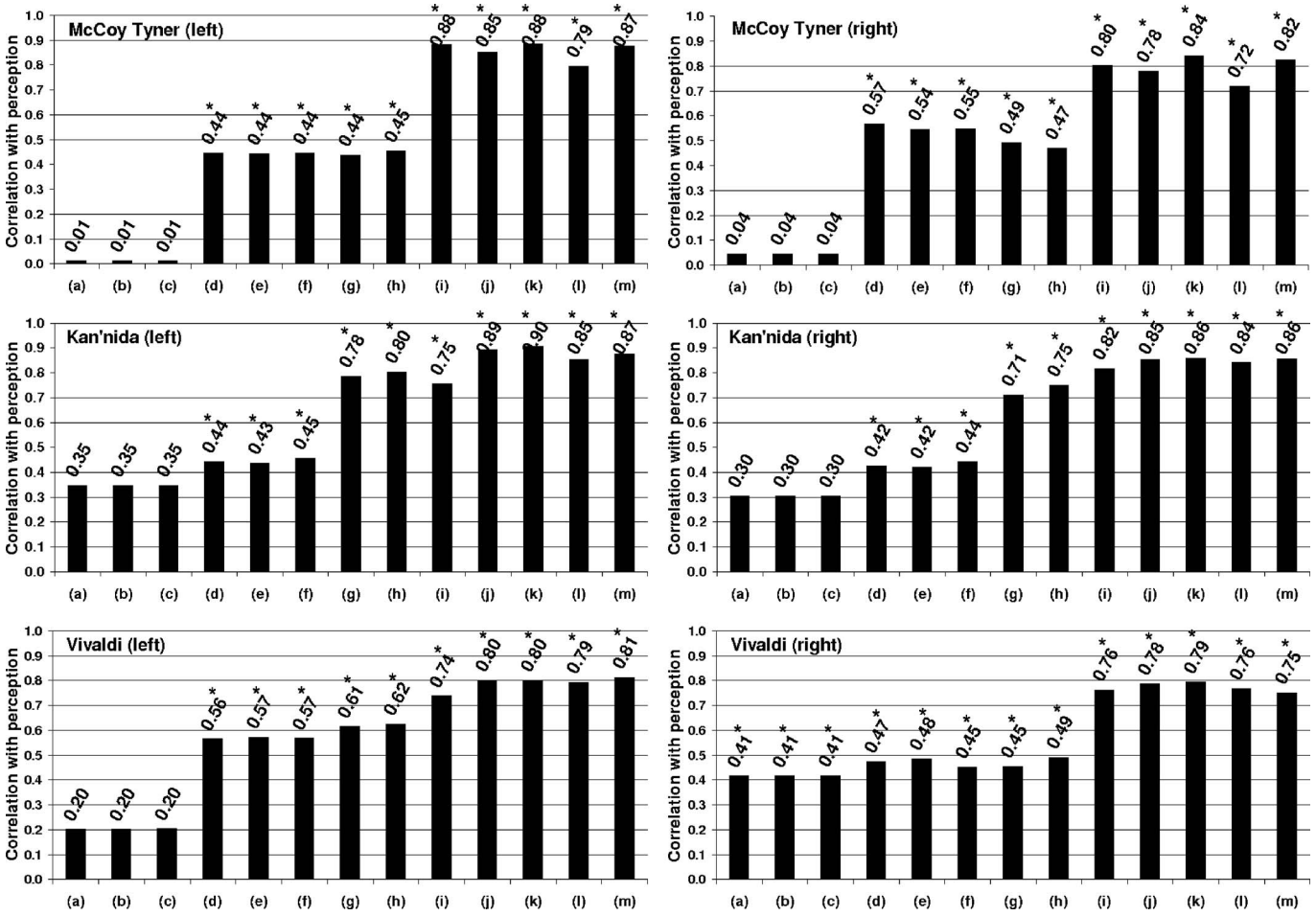


FIG. 4. Correlations between the mean perceptual dissimilarities resulting from the listening tests and the acoustical dissimilarities calculated on the right and left channels of the recordings, for the three musical excerpts (McCoy Tyner, Kan'nida, Vivaldi): (a) time signal; (b) spectrum; (c) time-frequency transform; (d) modulus of spectrum; (e) modulus of time-frequency transform; (f) power spectral density (PSD); (g) *A* weighted PSD; (h) 70 phone contour weighted PSD; (i) overall specific loudness; (j) time-varying specific loudness 1; (k) temporal mean of time-varying specific loudness 1; (l) time-varying specific loudness 2; and (m) temporal mean of time-varying specific loudness 2. The significant correlation coefficients are marked with an asterisk.

(b), and time-frequency (c) domains were never correlated with perceptual dissimilarities, except for the dissimilarities calculated on the right channel of the excerpt Vivaldi. When acoustical dissimilarities were evaluated using the modulus of the spectrum (d), the one of the time-frequency transform (e), or the power spectral density (f), the correlations were significant but weak. The weightings applied to the power spectral density [(g) and (h)] led to an increase of the correlation for the musical excerpt Kan'nida, but this increase was not confirmed for the two other excerpts. Using specific loudnesses to evaluate acoustical dissimilarities greatly enhanced the correlation with perceptual dissimilarities [(i), (j), (k), (l), and (m)].

2. Discussion

The acoustical discrimination methods which were almost perfectly correlated (Fig. 2) gave almost identical results when compared to perceptual dissimilarities. Discarding phase information improved the correlation with perceptual dissimilarities, whereas the tested spectral weightings did not seem to offer any further improvement, except for the excerpt Kan'nida. The high correlations obtained with the methods involving auditory models showed that the corresponding acoustical dissimilarities contained at least part of the relevant information regarding perception. It should be noted that perfect correlations would have meant that acoustical dissimilarities accounted for the noise in the evaluation of perceptual dissimilarities, which would be highly undesirable. The present data did not discriminate between the different methods using auditory models. Overall and mean specific loudnesses seemed to contain as much useful information towards perception as time-varying specific loudnesses. The temporal evolution of information did not seem to be crucial to explain our perceptual dissimilarities.

VI. GENERAL DISCUSSION

By defining metrics associated with several signal analyses, we calculated acoustical dissimilarities between musical signals radiated by loudspeakers. These acoustical dissimilarities were compared to perceptual dissimilarities resulting from listening tests, in order to identify the most relevant method of acoustical discrimination. The correlation between acoustical and perceptual dissimilarities was used to quantify the link between the two types of evaluation. This criterion of correlation was severe, as we could not suppose a priori that a linear link would exist. Yet, the high values of corre-

lation obtained revealed that such a link was established for the cases tested. The acoustical dissimilarities based on auditory models were close to the perceptual ones and contained at least part of the relevant information regarding perception. They were closer to the perceptual dissimilarities than to the other acoustical dissimilarities tested, bridging to some extent the gap between acoustical measurements and perception.

Other signal analyses and metrics should be tested to refine these results. For example, the acoustical dissimilarities considered in this study only used monaural information, as the left and right channels of the recordings were treated separately. The potential phase differences between the two ears/microphones were not incorporated in the dissimilarity evaluation. Such differences might be important to take into account if perceptual dimensions such as "feeling of space"[3,4] or "spaciousness"[5] were used by listeners to discriminate the recordings.

The contribution of auditory models was quantified by an increment in correlation, so we went a little further than Staffeldt,[2,10,45] who justified the use of such a model by visual comparisons between the results of his listening tests and the shape of the frequency responses of his loudspeakers. Bramsløw[5] and Klippel[4] based their acoustical evaluation of sound reproduction on the specific loudness of their signals. They chose this method of analysis a priori, by 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 to test their hypothesis.

The correlation between acoustical and perceptual dissimilarities led to the same trends with three different musical excerpts. They remained relatively independent of the experimental parameters tested in following experiments considering other recording techniques,[39] another room, up to 37 loudspeakers, and another listening task.[46]

Other comparisons of acoustical and perceptual dissimilarities should be investigated. A linear link has been found by evaluating their correlation, but this relationship might be refined. A multidimensional comparison of our acoustical and perceptual dissimilarities was therefore performed,[39,47] and will be fully described in a following paper. It allowed a comparison without an implicit linear relationship, but it only relied on visual comparisons of multidimensional spaces, without quantitative evaluation of their resemblance to one another. This comparison indicated that our acoustical analyses using specific loudnesses led to dimensions very similar to the perceptual ones used by listeners to discriminate our recordings during the listening tests.

Whereas specific loudnesses 1 and 2 and their temporal mean gave similar results, the overall specific loudness did not seem sufficient to fully describe the perceptual dissimilarities.[39] This difference between specific loudnesses determined on the overall signal or as a function of time could not be quantified by the correlations presented in this paper (Fig. 4). Klippel[4] and Staffeldt[2] only considered overall specific loudness, while Bramsløw[5] only used a time-varying specific loudness. More experimental data are required to make a clear difference between these acoustical evaluations. This question is of great importance to assess the necessity of taking into account the temporal dependency of auditory masking effects. When Staffeldt[2,10,45] applied an auditory model directly 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 direct examination of their frequency response might be questioned.

VII. CONCLUSION

Different ways of evaluating the acoustical dissimilarity between loudspeakers were compared. Twelve single loudspeakers were measured in a listening situation. The correlation between the comparative measurements led to identified classes of measurements. The recorded signals were involved in listening tests, providing a perceptual metric used to evaluate the relevance of the acoustical discrimination methods regarding perception. Our first results showed the importance of using auditory models in order to discriminate the loudspeakers in the same way as listeners did.

ACKNOWLEDGMENTS

We wish to thank 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 on a first version of this paper. 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.

[1] A. Gabrielsson, U. Rosenberg, and H. Sjögren, "Judgments and dimension analyses of perceived sound quality of sound-reproducing systems," J. Acoust. Soc. Am. **55**, 854–861 (1974).
[2] H. Staffeldt, "Correlation between subjective and objective data for quality loudspeakers," J. Audio Eng. Soc. **22**, 402–415 (1974).
[3] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," J. Acoust. Soc. Am. **65**, 1019–1033 (1979).
[4] W. Klippel, "Multidimensional relationship between subjective listening impression and objective loudspeaker parameters," Acustica **70**, 45–54 (1990).
[5] L. Bramsløw, "An objective estimate of the perceived quality of reproduced sound in normal and impaired hearing," Acta. Acust. Acust. **90**, 1007–1018 (2004).
[6] I. Borg and P. Groenen, *Modern Multidimensional Scaling. Theory and Applications* (Springer, New York, 1997).
[7] W. R. Dillon and M. Goldstein, *Multivariate Analysis. Methods and Applications* (Wiley, New York, 1984).
[8] A. Gabrielsson and B. Lindstrom, "Perceived sound quality of high-fidelity loudspeakers," J. Audio Eng. Soc. **33**, 33–53 (1985).
[9] F. E. Toole, "Subjective measurements of loudspeaker: sound quality and listener performance," J. Audio Eng. Soc. **33**, 2–32 (1985).
[10] H. Staffeldt, "Differences in the perceived quality of loudspeaker sound reproduction caused by the loudspeaker-room-listener interactions," *Proceedings AES 90th Convention*, 1991, p. 3046.
[11] S. E. Olive, P. L. Schuck, S. L. Sally, and M. E. Bonneville, "The effect of loudspeaker placement on listener preference ratings," J. Audio Eng. Soc. **42**, 651–669 (1994).
[12] S. E. Olive, "Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: a case study," J. Audio Eng. Soc. **51**, 806–825 (2003).

[13]S. E. Olive, "A multiple regression model for predicting loudspeaker preference using objective measurements: Part 1 - Listening test results," *Proceedings AES 116th Convention*, 2004, p. 6113.
[14]F. E. Toole, "Loudspeakers and rooms for sound reproduction—A scientific review," J. Audio Eng. Soc. **54**, 451–476 (2006).
[15]IEC Publication 60268-5, "Sound system equipment - Part 5: Loudspeakers," *International Electrotechnical Commission*, Geneva, Switzerland, 1989.
[16]IEC Publication 60581-7, "High fidelity audio equipment and systems. Minimum performance requirements - Part 7: Loudspeakers," *International Electrotechnical Commission*, Geneva, Swizerland, 1986.
[17]IEC Publication 60268-13, "Sound system equipment - Part 13: Listening tests on loudspeakers," *International Electrotechnical Commission*, Geneva, Swizerland, 1998.
[18]AES20-1996, "AES recommended practice for professional audio - Subjective evaluation of loudspeakers," J. Audio Eng. Soc. **44**, 382–400 (1996).
[19]ITU-R Recommendation BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *International Telecommunication Union*, Geneva, Swizerland, 1997.
[20]ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate quality levels of coding systems," *International Telecommunication Union*, Geneva, Swizerland, 2003.
[21]ITU-R Recommendation BS.1284, "General methods for the subjective assessment of sound quality," *International Telecommunication Union*, Geneva, Swizerland, 2003.
[22]R. V. Waterhouse, "Radiation impedance of a source near reflectors," J. Acoust. Soc. Am. **35**, 1144–1151 (1963).
[23]T. Salava, "Acoustic load and transfer functions in rooms at low frequencies," J. Audio Eng. Soc. **36**, 763–775 (1988).
[24]W. Klippel, "Nonlinear large-signal behaviour of electrodynamic loudspeakers at low frequencies," J. Audio Eng. Soc. **40**, 402–415 (1992).
[25]S. E. Olive, "A multiple regression model for predicting loudspeaker preference using objective measurements: Part 2 - Development of the model," *Proceedings AES 117th Convention*, 2004, p. 6190.
[26]F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences: Part 1," J. Audio Eng. Soc. **34**, 227–235 (1986).
[27]A. Gabrielsson, B. Lindström, and O. Till, "Loudspeaker frequency response and perceived sound quality," J. Acoust. Soc. Am. **90**, 707–719 (1991).
[28]F. E. Toole, "Loudspeaker measurements and their relationship to listener preferences: Part 2," J. Audio Eng. Soc. **34**, 323–348 (1986).
[29]F. E. Toole, "Binaural record/reproduction systems and their use in psychoacoustic investigations," *Proceedings AES 91st Convention*, 1991, p. 3179.
[30]S. Bech, "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 (2002).
[31]J. A. Pedersen and A. Mäkivirta, "Requirements for low-frequency sound reproduction, Part 2: Generation of stimuli and listening system equalization," J. Audio Eng. Soc. **50**, 581–593 (2002).
[32]British Standard ISO 226 2003, "Acoustics - Normal equal-loudness level contours," BSI, London, 2003.
[33]E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer, New York, 1999).
[34]E. Paulus and E. Zwicker, "Programme zur automatischen bestimmung der lautheit austerzpegeln oder frequenzgruppenpegeln (Computer programs for calculating loudness from third octave band levels or from critical band levels)," Acustica **27**, 253–266 (1972).
[35]E. Zwicker, H. Fastl, and C. Dallmayr, "BASIC-Program for calculating the loudness of sounds from their 1/3-oct band spectra according to ISO 532 B," Acustica **55**, 63–67 (1984).
[36]E. Zwicker and H. Fastl, "A portable loudness-meter based on ISO 532B," *Proceedings 11th International Congress on Acoustics*, 1983, pp. 135–137.
[37]J. Max and J.-L. Lacoume, *Méthodes et Techniques de Traitement du Signal (Methods and Techniques for Signal Processing)* (Dunod, Paris, 2000).
[38]A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).
[39]M. Lavandier, P. Herzog, and S. Meunier, "Perceptual and physical evaluation of loud-speakers," *Proceedings AES 117th Convention*, 2004, p. 6240.
[40]S. Bech, "Timbral aspects of reproduced sound in small rooms. 1," J. Acoust. Soc. Am. **97**, 1717–1726 (1995).
[41]S. Bech, "Timbral aspects of reproduced sound in small rooms. 2," J. Acoust. Soc. Am. **99**, 3539–3549 (1996).
[42]B. C. J. Moore and C. T. Tan, "Perceived naturalness of spectrally distorted speech and music," J. Acoust. Soc. Am. **114**, 408–419 (2003).
[43]G. Keppel and T. D. Wickens, *Design and Analysis. A Researcher's Handbook*, 4th ed. (Pearson Prentice-Hall, Englewood Cliffs, NJ, 2004).
[44]M. Lavandier, "Différences entre enceintes acoustiques: Une évaluation physique et perceptive (Differences between Loudspeakers: A physical and perceptual evaluation)," Ph.D. thesis, Université de la Méditerranée - Aix-Marseille II, 2005, ⟨http://tel.archives-ouvertes.fr/tel-00087414⟩.
[45]H. Staffeldt, "Measurement and prediction of the timbre of sound reproduction," J. Audio Eng. Soc. **32**, 410–414 (1984).
[46]M. Lavandier, S. Meunier, and P. Herzog, "Perceptual and physical evaluation of differences among a large panel of loudspeakers," *Proceedings Forum Acusticum 2005*, 2005, pp. 1689–1694.
[47]M. Lavandier, P. Herzog, and S. Meunier, "Physical and perceptual estimation of differences between loudspeakers," C. R. Mec. **334**, 732–736 (2006).

On the reflection of coupled Rayleigh-like waves at surface defects in plates

Bernard Masserey[a] and Paul Fromme
Department of Mechanical Engineering, University College London, WC1E 7JE, United Kingdom

(Received 5 July 2007; revised 10 October 2007; accepted 14 October 2007)

The reflection of coupled Rayleigh-like waves from surface defects in elastic plates is investigated experimentally and analyzed on the basis of an analytical model and finite difference simulations. The propagation of Rayleigh-like waves in plates is characterized by an energy transfer to the opposite plate side and back over a distance called the beat length. Experimental results clearly show this beating effect and its dependency on the frequency-thickness product, and excellent agreement is obtained with existing analytical predictions. The propagation and scattering are modeled separately for the fundamental A_0 and S_0 Lamb modes that constitute the incident Rayleigh-like wave. The reflection coefficients from surface slots are investigated using finite difference simulations and the reflected Rayleigh-like wave is obtained by superposition. The theoretical model reveals strong dependencies of the reflected field on the ratio between excitation distance and beat length and on the cutoff frequencies of specific higher Lamb modes. Standard pulse-echo measurements allow for the detection of small defects from a remote transducer location. Good agreement is obtained between the predicted and measured amplitude spectra of the reflected Rayleigh-like wave. The developed model allows for the evaluation of defect location and damaged plate side using a combination of time-of-flight and frequency measurements.
 [DOI: 10.1121/1.2805668]

PACS number(s): 43.40.Dx, 43.35.Zc, 43.20.Gp, 43.35.Pt [YHB] Pages: 88–98

I. INTRODUCTION

This paper investigates the propagation and reflection of coupled Rayleigh-like waves in aluminum plates with a view toward potential applications for the nondestructive inspection of aircraft and other technical structures. Existing ultrasonic methods adopted for defect detection in this type of structure are bulk wave ultrasonic testing (UT) and the use of guided ultrasonic waves. Bulk wave UT typically works in a frequency region of several megahertz and has a proven track record and sensitivity for the detection and sizing of small cracks.[1] However, UT involves manual or automated scanning over the area of interest and can therefore be time consuming and necessitates local access to the inspected part. Guided ultrasonic waves allow for rapid and cost-efficient inspection and permanent monitoring of large surface areas.[2] Guided waves in plates are often used in a significantly lower frequency range (up to hundreds of kilohertz), below the cutoff frequency of the first higher Lamb mode A_1. The reason for this choice is the resulting limitation of the number of propagating guided wave modes, simplifying the interpretation of signals, and the possible generation of a single Lamb mode.[3] However, the wavelengths of the employed guided waves are significantly larger than in bulk wave UT and the sensitivity for the detection of small defects has to be ascertained.[4]

A significant improvement on the detection sensitivity as compared to standard guided wave techniques can be expected for coupled Rayleigh-like waves, which work in a higher frequency range, resulting in shorter wavelengths.[5] The propagation of Rayleigh-like waves, described theoretically by Viktorov,[6] can be interpreted as the superposition of the first antisymmetric and symmetric Lamb modes. In the frequency range of interest, A_0 and S_0 propagate with little dispersion, but have slightly different phase velocities. This results in a continuous shift in relative phase as the waves travel, causing the transfer of the Rayleigh-like wave to the other plate side and then back. The significant distance L for this energy exchange, the so-called beat length[7] or beat wavelength,[8] is calculated as

$$L = \frac{2\pi}{k_{A0} - k_{S0}}, \tag{1}$$

where k_{A0} and k_{S0} are the wave numbers of the A_0 and S_0 modes, respectively. The real wave number dispersion curves in an aluminum plate are shown in Fig. 1. The beat length mainly obeys the fast decrease in the denominator term $(k_{A0}-k_{S0})$ with increasing frequency thickness. For large frequency-thickness products (small wavelength compared to the plate thickness), the two fundamental modes A_0 and S_0 converge to the wave number of a Rayleigh wave and the beat length tends toward infinity, corresponding to the propagation of a Rayleigh wave in a semi-infinite medium.

Very few experimental assessments of this phenomenon were found in the literature. Li and Thompson (private communications) made measurements of the beating effect by exciting a Rayleigh-like wave on a thin aluminum plate and scanning the generated ultrasonic field on both plate sides using electromagnetic acoustic transducers (EMATs), but did not widely report them. One of the first detailed experimental assessments was reported by Ti *et al.*[7] They performed beat-

[a]Electronic mail: masserey@alumni.ethz.ch

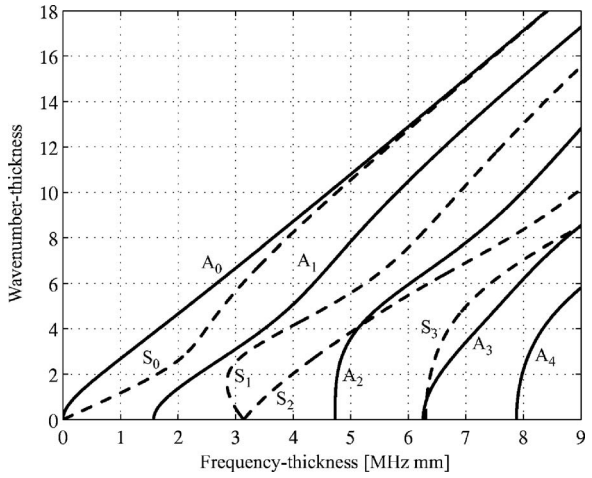


FIG. 1. Dispersion curves for Lamb waves in an aluminum plate (alloy 2014 T6).

length measurements and compared their results to theoretical predictions for brass plates with a thickness slightly smaller than the Rayleigh wavelength. Harris[9] used an asymptotic expansion to describe a Rayleigh-like wave propagating within a curved waveguide. Similar beating effects arising from the interaction between twin modes in double skin features were reported by Dalton *et al.*[10]

On the other hand the scattering of Rayleigh and Lamb waves at surface defects has been widely addressed. Achenbach *et al.*[11] described the interaction of Rayleigh waves with surface cracks theoretically using elastodynamic ray theory. Hirao *et al.*[12] discussed the scattering of Rayleigh waves at surface defects using finite difference (FD) calculations complemented by experimental results. The reflection and transmission coefficients of oblique incident Rayleigh waves from surface cracks were investigated theoretically by Angel and Achenbach[13] and verified experimentally by Dong and Adler.[14] The interaction of Lamb waves with surface features has been mostly investigated below the first Lamb mode cutoff frequency.[15,16] Flores-Lopez and Gregory[17] studied the scattering of Lamb waves at a surface crack using a projection method. Alleyne and Cawley[4] used finite element simulations and experiments to investigate the interaction of the A_0, S_0, and A_1 modes with small surface notches. The sensitivity of Lamb waves for various defects was evaluated from the variation of the reflection and transmission factors by Cho *et al.*[18] The interaction of a multimode signal, generated by a point impact source, with a crack in a plate was investigated by Liu and Datta.[19] Recently, Terrien *et al.*[20] presented a combined finite element and modal decomposition method to study the interaction of Lamb waves with defects for frequency-thickness products up to 5 MHz mm.

In this paper the propagation and the scattering of coupled Rayleigh-like waves in aluminum plates with small surface defects are investigated theoretically and experimentally. Rayleigh-like waves are excited selectively significantly above the first Lamb mode cutoff frequency using standard Rayleigh wave wedge transducers in a frequency-thickness region where the wavelength of the Rayleigh-like wave is approximately half the plate thickness (4–9 MHz mm). The wave propagation of a narrow-band pulse is investigated experimentally by means of laser interferometer measurements. The beating phenomenon is clearly observable in both the time and frequency domain. The beat length is measured precisely in the frequency-thickness range of interest using a standard fitting procedure. Good agreement is found with analytical calculations.

The scattering at a surface defect is investigated with particular attention to the reflected Rayleigh-like wave, measured using a standard broadband pulse-echo technique. Small slots situated either on the excitation side of the plate or on the opposite side can be well detected. A theoretical model is proposed in order to describe the main physical characteristics of a Rayleigh-like wave propagating in a plate and reflected from a small surface defect. Reflection coefficients are calculated on the basis of finite difference simulations of the scattering. The model accurately predicts the measured amplitude spectrum. The investigation in the frequency domain of the reflected Rayleigh-like wave contains information to determine the damaged side of the plate.

II. EXPERIMENTAL SETUP

Large aluminum plates with a width of 200 mm, a length of 1000 mm, and a thickness of 3.05 or 6 mm were used. This allowed for the precise measurement of beat lengths up to approximately $L=500$ mm. The plate material was an aluminum alloy 2014 T6 widely used for aerospace applications, having a Young's modulus E of 73 GPa, a density ρ of 2800 $\mathrm{kg/m^3}$, and a Poisson's ratio ν of 0.33 (all material property and thickness data as from supplier). The corresponding Rayleigh wave velocity is $c_R=2918$ m/s. The 3.05-mm-thick aluminum plate was subsequently cut in half (two pieces, 500 mm long) for the experimental investigation of the reflection of a Rayleigh-like wave from a surface defect. 0.1 mm and 0.5 mm deep slots were manufactured using electrical discharge machining (EDM). The slots were about 0.45 mm wide and 25 mm long with a reasonably flat bottom.

The Rayleigh-like wave was generated at the specimen surface using standard half inch transducers mounted on a 90° angle beam wedge for steel. The angle of incidence in the acrylic wedge was about 4° smaller than the optimal angle for Rayleigh wave generation in aluminum. This slightly reduced the amplitude of the generated Rayleigh-like wave and excited an additional A_1 Lamb mode. Center frequencies of 2.25 and 1 MHz were used on the 3.05- and 6-mm-thick plates, respectively, resulting in Rayleigh wavelengths of 1.3 and 2.9 mm, about half the plate thickness.

The scattering was investigated by means of pulse-echo measurements on the 3.05-mm-thick plate. The 2.25 MHz angle-beam transducer was positioned at 300 mm ($100h$) from the slot and 450 mm from the end of the aluminum plate. The measurements were performed so that the main propagation line crosses the middle of the slot perpendicularly. This geometry was chosen as a good approximation of the plain strain assumption made in the two-dimensional plate model. The transducer was driven using a standard ultrasonic pulser/receiver. The frequency content of the gener-

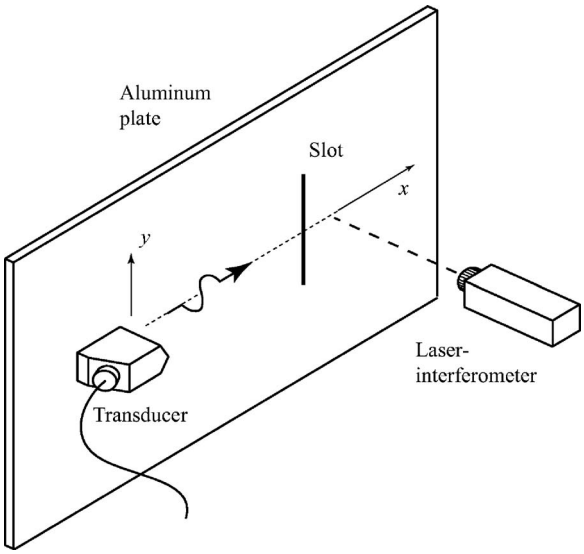


FIG. 2. Schematic diagram of the experimental arrangement.

ated ultrasonic pulse was wide band up to about 3.5 MHz. The received signal was amplified and bandpass filtered.

The beat-length investigation was performed using more specialized equipment to achieve better control of the frequency content of the excited pulse. An N-cycle tone burst (sinusoid multiplied by a Hanning window) was generated in a programmable function generator and amplified using a broadband power amplifier. By varying the number of cycles, either narrow-band or wideband pulses could be excited using the same transducer with good control of the frequency bandwidth. A heterodyne laser interferometer was used for point measurements of the out-of-plane displacement field along lines parallel to the propagation of the ultrasonic pulse, as displayed in Fig. 2. The demodulator output was bandpass filtered around the center frequency. The interferometer head was moved parallel to the plate surface by means of a positioning system. A more detailed description of this setup can be found in a previous publication.[21]

III. WAVE PROPAGATION MEASUREMENT

The beat-length phenomenon was investigated on the 3.05-mm-thick, 1-m-long plate using a 2.25 MHz, 20 cycle tone burst (sinusoid in a Hanning window). A narrow-band signal (width of the main lobe $\Delta f=0.45$ MHz) was selected to minimize dispersion. The out-of-plane displacement component was measured at discrete locations along the x axis (see Fig. 2) on both plate surfaces using a laser interferometer. Figure 3(a) shows the waterfall plot for the excitation side, while Fig. 3(b) shows the corresponding plot for the back side. The theoretical prediction for the corresponding frequency-thickness product gave a beat length L of about 270 mm. As expected, the out-of-plane displacement measured on the upper surface directly in front of the wedge ($x=L/20$) has the largest amplitude. The out-of-plane displacement at the corresponding position on the plate back side is very small. On the upper side the out-of-plane displacement signal decreases in amplitude and reaches a minimum at $x=L/2$. At that position the energy of the Rayleigh-

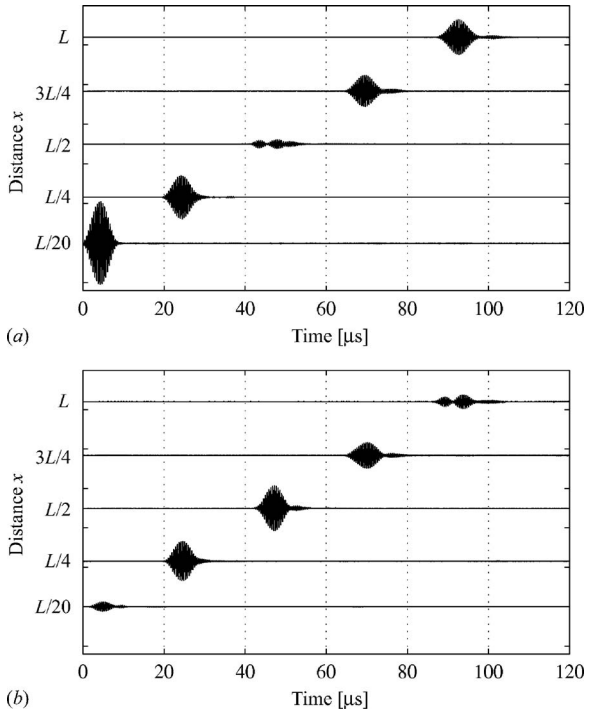


FIG. 3. Time traces of the out-of-plane displacement presented in the form of waterfall plots for discrete locations on the $h=3.05$-mm-thick specimen, $f_0=2.25$ MHz, $n_c=20$: (a) Measurements on the excitation side and (b) measurements on the back side. $L\cong 270$ mm.

like wave has transferred to the back side of the specimen. Moving further away from the transducer shows the energy of the Rayleigh-like wave gradually transferring back to the excitation side of the specimen.

The transducer generates higher Lamb modes, in particular the A_1 mode, with approximately 20 dB smaller amplitude. At 6.75 MHz mm, A_1 propagates with group velocity $c_g\cong 2700$ m/s and dissociates from the main pulse with increasing distance from the excitation source, as can be seen in the tail of the pulse developing in Fig. 3. The amplitude of this mode could be significantly reduced by the use of a nonstandard angle beam wedge designed specifically for aluminum.

The same experiment was performed using a five cycle tone burst with a significantly wider bandwidth (main lobe $\Delta f=1.8$ MHz). The out-of-plane displacement component was measured by moving the laser interferometer along the x axis with step size $\Delta x=1$ mm. A time window was applied to the measured time traces to isolate the incident Rayleigh-like wave. The amplitude values were subsequently extracted at 2.25 and 2.1 MHz using fast Fourier transform (FFT). The resulting amplitude curves, measured on both plate sides, are displayed in Fig. 4. The overall decrease in amplitude with increasing distance from the excitation source is mainly due to the ultrasonic beam spread. The energy transfer between the upper surface and the lower surface is clearly observable with the amplitude maxima and minima alternating between the upper side and the back side of the specimen. The distance between two maxima (or two minima) corresponds to the beat length L. The frequency dependence of the beat length can be observed, with the evaluation at 2.25 MHz

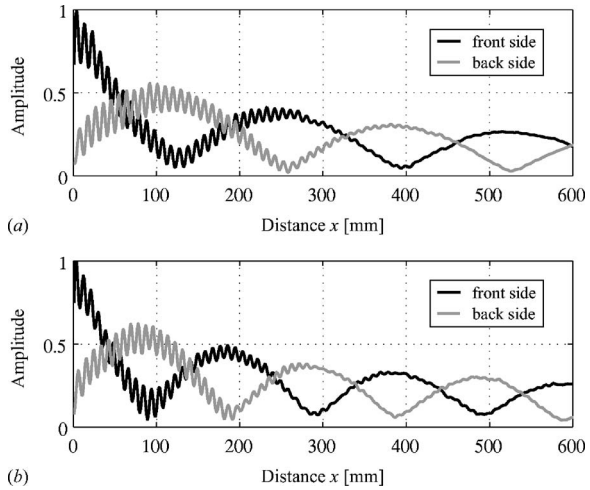


FIG. 4. Out-of-plane displacement amplitude (normalized) vs distance x from wedge tip measured on the excitation side (black line) and on the back side (gray line). Measurements on the $h=3.05$-mm-thick specimen, $f_0=2.25$ MHz, $n_c=5$: Evaluation at (a) 2.25 MHz and (b) 2.1 MHz.

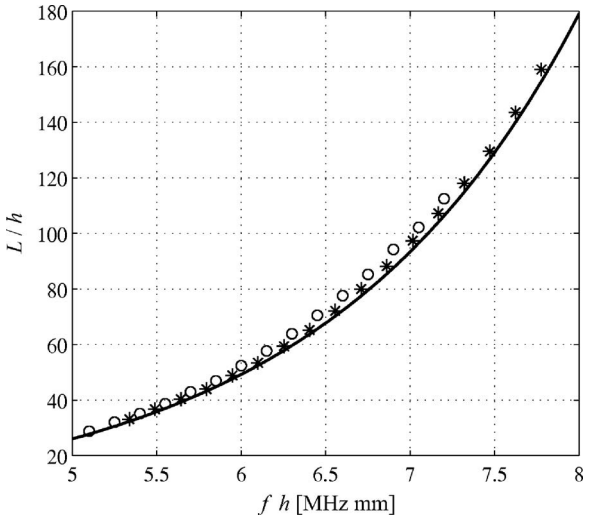


FIG. 5. Normalized beat length L/h vs frequency-thickness product. Experimental data $h=6$ mm (open points) and $h=3.05$ mm (asterisks), and theoretical curve using the nominal material parameters (solid line).

[Fig. 4(a)] showing a beat length significantly larger ($L=270$ mm) than at 2.1 MHz [Fig. 4(b), $L=200$ mm]. The amplitude curves do not completely go to zero at the minima positions. This is due to the small difference in the mode shapes between the A_0 and S_0 Lamb modes around 6.75 MHz mm, as well as to the excitation of higher Lamb modes, as previously mentioned. The slight interference pattern observable up to approximately 300 mm from the excitation position arises from the overlapping in time of the Rayleigh-like wave with the Lamb mode A_1.

IV. BEAT-LENGTH CALCULATION

The determination of the beat length is performed on the basis of the amplitude curves illustrated in Fig. 4 using a fitting procedure to determine the beat length L and the phase shift φ_0. The fitting function is of the form:

$$f_{\text{fit}}(x,L,\varphi_0) = A(x)\left|\cos\left(\pi\frac{x}{L}+\varphi_0\right)\right|, \tag{2}$$

where the independent variable x is the distance from the excitation source. The amplitude function $A(x)$ accounts for the characteristic ultrasonic field attenuation and beam spread. $A(x)$ is a function of the form $A_0/\sqrt{x+x_0}$, where A_0 and x_0 characterize the ultrasonic field in front of the wedge and were determined from a separate fitting procedure. It can be shown that within reason these constants have no significant influence on the determination of the beat length L. The resulting beat-length curves are displayed in Fig. 5. The solid line was calculated using the beat-length formula [Eq. (1)] and the material parameters provided by the supplier. The measurements were performed on both the 3.05- and 6-mm-thick plates using a five cycle excitation signal with 2.25 and 1 MHz center frequency, respectively. The results were normalized with the plate thickness. For each measurement the beat-length values were evaluated in the frequency range where the Rayleigh-like wave carries sufficient energy, yielding a portion of the beat-length curve corresponding to approximately half the bandwidth of the incident wave, i.e., $f\in[1.8,2.7]$ MHz for the 2.25 MHz excitation signal and $f\in[0.8,1.2]$ MHz for the 1 MHz excitation signal.

Very good agreement was obtained between experimental and theoretical curves. The beat length of the 6-mm-thick plate shows a maximum deviation of approximately 7% from the analytical curve. This consistent error can possibly be explained by the high sensitivity of the beat length to the material parameters (phase velocity) and to the plate thickness. In fact, it can be shown from an analytical evaluation that a discrepancy of 1% in terms of plate thickness or phase velocity (i.e., ~2% in terms of Young modulus) generates a change in the beat-length value of approximately 5%.

V. BROADBAND PULSE-ECHO MEASUREMENTS

Section IV has shown that the propagation of a coupled Rayleigh-like wave is characterized by a continuous energy transfer from one side of the plate to the other side. The use of a wideband excitation, corresponding to a large portion of the beat-length curve in Fig. 5, should therefore allow for a distribution of the energy on both plate sides and thus for the detection of defects on both sides of the plate.

Pulse-echo measurements were performed on the 3.05-mm-thick aluminum plate using a standard ultrasonic pulser/receiver as described in Sec. II. The time trace resulting from a measurement at a 0.5-mm-deep slot ($a/\lambda=0.38$) are shown in Fig. 6(a). The wedge was positioned at 300 mm from the slot on the damaged side of the plate. The Rayleigh-like wave reflected from the slot can be seen at about 200 μs. The second pulse at about 300 μs is the reflection from the end of the plate. The location of the defect along the propagation axis can be found on the basis of time-of-flight measurements. Taking the arrival time of the maximum amplitude in the reflected pulse and the theoretical Rayleigh velocity, the distance between the transducer and the slot was measured with an error of less than 1%. The scattering of the Rayleigh-like wave at the slot generates a reflected Rayleigh-like wave as well as other Lamb modes. However, the use of

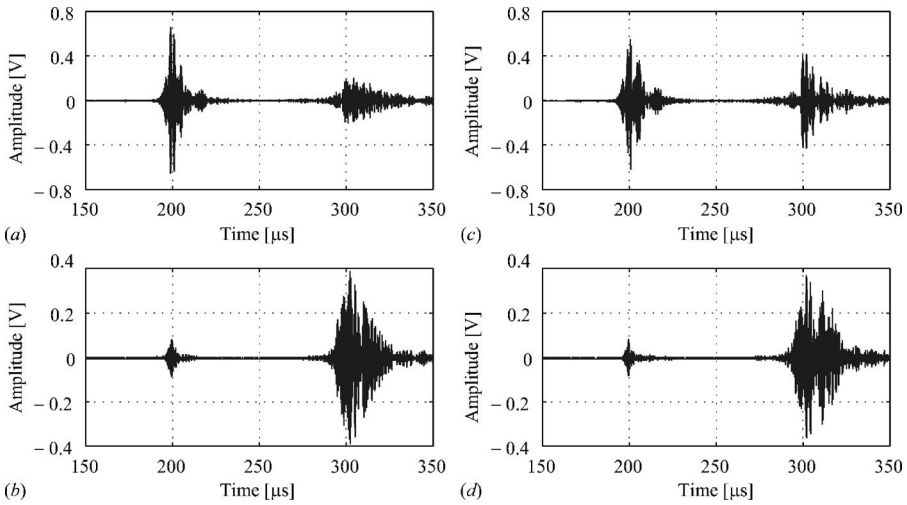


FIG. 6. Time traces from pulse-echo measurements on the 3.05-mm-thick aluminum plate using a 2.25 MHz transducer, $d=300$ mm. Wedge on damaged side: (a) 0.5-mm-deep slot, (b) 0.1-mm-deep slot; wedge on undamaged side: (c) 0.5-mm-deep slot, (d) 0.1-mm-deep slot.

an angle beam transducer as receiver allows for most of these other modes to be filtered out from the measured time trace.

The same measurement was performed at a 0.1-mm-deep slot ($a/\lambda=0.08$). The time trace of the reflected Rayleigh-like wave is displayed in Fig. 6(b). Although the amplitude of the reflected wave is approximately five times smaller than for the reflection at a 0.5-mm-deep slot, the defect is clearly detected and its location found with the same accuracy.

Similar measurements were performed with the wedge positioned on the undamaged side of the aluminum plate. For both 0.5- and 0.1-mm-deep slots the defects could be clearly detected as shown in Figs. 6(c) and 6(d) and the distance from the wedge to the defect was found with the same accuracy. However, the amplitude of the reflected Rayleigh-like waves is very similar to the results obtained with the transducer on the damaged plate side. Therefore, analysis in the time domain alone is not sufficient for the identification of the damaged plate side, as the energy of a broadband pulse is distributed over the whole plate thickness after a minimum propagation distance. The identification of the damaged plate side will need to incorporate an analysis of the reflected pulse in the frequency domain, based on the development of the theoretical model introduced in Sec. VI.

VI. THEORETICAL SCATTERING INVESTIGATION

The model proposed here describes the propagation and reflection of the Rayleigh-like wave in a plate. It uses a combination between an analytical model to describe the wave propagation in a plate and FD calculations of the scattering at a small defect.

A. Analytical description of the reflected Rayleigh-like wave

The present model assumes that the A_0 and S_0 Lamb modes constituting the incident Rayleigh-like wave carry the same amount of energy through the plate thickness. The displacement profiles are defined in such a way that at the excitation position A_0 and S_0 are in phase on the upper surface and out of phase on the plate back side. The validity of these assumptions will be discussed on the basis of simulation results in the following subsection. Choosing $x=0$ as the defect location, $x=-d$ to denote the excitation position, and assuming initially that excitation and defect are on the same plate side ($z=+h/2$), the incident Rayleigh-like wave can then be written as the sum of the following A_0 and S_0 modes:

$$\mathbf{u}_{A0}^{\mathrm{inc}} = \frac{1}{\sqrt{2}} \cdot \mathbf{U}_{A0}(z)\cos(\omega t - k_{A0}x),$$

$$\mathbf{u}_{S0}^{\mathrm{inc}} = \frac{1}{\sqrt{2}} \cdot \mathbf{U}_{S0}(z)\cos(\omega t - k_{S0}x + \Delta\varphi_0), \tag{3}$$

where $\mathbf{U}_{A0}(z)$ and $\mathbf{U}_{S0}(z)$ stand for the displacement profiles of A_0 and S_0, respectively. The constant $1/\sqrt{2}$ ensures that the Rayleigh-like wave carries a unit average power flow through the section of the plate in the x direction. $\Delta\varphi_0$ denotes the phase shift between both modes at defect location $x=0$. Since the two modes are assumed to be in phase at excitation position $x=-d$, the phase shift $\Delta\varphi_0$ can be written as

$$\Delta\varphi_0 = (k_{A0} - k_{S0})d = 2\pi\frac{d}{L}. \tag{4}$$

Each incident mode scattered at the crack generates both A_0 and S_0 Lamb modes, higher Lamb modes, as well as nonpropagating modes. In the present study mostly the propagating Lamb modes are of interest, with particular attention given to the reflected Rayleigh-like wave. Using the reciprocal theorem it can be shown that the fraction of energy reflected (transmitted) as mode n for an incoming mode m is equal to the fraction of energy reflected (transmitted) as mode m for an incoming mode n.[8] Furthermore, in the vicinity of the upper surface the stress fields of A_0 and S_0 are similar. Reflection coefficient and phase shift of the reflected modes for A_0 incidence or S_0 incidence are therefore assumed to be equal. Consequently, the four reflected A_0 and S_0 modes can be expressed as functions of one single reflection coefficient C and one single phase shift term $\Delta\varphi$:

$$\mathbf{u}_{A0\to A0}^{\mathrm{ref}} = \frac{C}{\sqrt{2}} \cdot \mathbf{U}_{A0}(z)\cos(\omega t + k_{A0}x + \Delta\varphi),$$

$$\mathbf{u}^{\mathrm{ref}}_{A0\to S0} = \frac{C}{\sqrt{2}} \cdot \mathbf{U}_{S0}(z)\cos(\omega t + k_{S0}x + \Delta\varphi),$$

$$\mathbf{u}^{\mathrm{ref}}_{S0\to A0} = \frac{C}{\sqrt{2}} \cdot \mathbf{U}_{A0}(z)\cos(\omega t + k_{A0}x + \Delta\varphi + \Delta\varphi_0),$$

$$\mathbf{u}^{\mathrm{ref}}_{S0\to S0} = \frac{C}{\sqrt{2}} \cdot \mathbf{U}_{S0}(z)\cos(\omega t + k_{S0}x + \Delta\varphi + \Delta\varphi_0). \quad (5)$$

Superscript "ref" stands for reflected, the first subscript term stands for the incident mode and the second for the reflected mode. C and $\Delta\varphi$ are functions of the frequency-thickness product and the geometry of the defect. The displacement vector at the specimen surface $z=h/2$ can be substituted by a constant $\mathbf{U}_0=\mathbf{U}_{A0}(h/2)=\mathbf{U}_{S0}(h/2)$. The surface displacement vector $\mathbf{U}_0$ of an A_0 or a S_0 Lamb wave having a unit average power flow and the surface displacement vector $\mathbf{U}_{R0}$ of a Rayleigh wave with unit average power flow are related using the relationship $\mathbf{U}_0=\mathbf{U}_{R0}/\sqrt{2}$. The reflected Rayleigh-like wave, obtained from the sum of the four reflected modes described in Eq. (5), can be written as

$$\mathbf{u}_R \cong 2 \cdot C \cdot \mathbf{U}_{R0}\cos\left(\pi\frac{d}{L}\right)\cos\left(\pi\frac{x}{L}\right) \times\cos\left(\omega t + k_R x + \Delta\varphi + \pi\frac{d}{L}\right). \quad (6)$$

Equation (6) describes the reflected wave propagating with the Rayleigh wave phase velocity c_R in the negative x direction. In the frequency-thickness range considered in this contribution (4–9 MHz mm) the change in the values of k_{A0} and k_{S0} is balanced so that the average $(k_{A0}+k_{S0})/2$ remains close to the Rayleigh wave number. In fact, the maximum discrepancy at 4 MHz mm is smaller than 1%. The wave number $(k_{A0}+k_{S0})/2$ was therefore substituted by the Rayleigh wave number k_R.

The second cosine term in Eq. (6) describes the beating phenomenon between the reflected A_0 and S_0 Lamb modes. It has to be pointed out that the reflected A_0 and S_0 modes start propagating in phase at the crack position ($x=0$), independently from the distance d between excitation and defect. The first cosine term is a function of the ratio between distance d from the excitation source to the defect and beat length L, where L is a function of the frequency-thickness product. If the incident A_0 and S_0 modes are in phase at the defect position they sum into a Rayleigh-like wave on the upper surface of the specimen and the reflected Rayleigh-like wave has maximum energy. If they are out of phase the energy of the Rayleigh-like wave is concentrated on the back side of the plate at the defect position and there is no reflected wave. The distances corresponding to maxima or minima for a given frequency can be expressed as a function of the beat length at that frequency:

$$d_{\max} = nL, \quad d_{\min} = \left(n+\tfrac{1}{2}\right)L, \quad n = 0,1,2,\ldots . \quad (7)$$

Equation (7) holds for the case of excitation and defect on the same side of the specimen. If the defect is situated on the back side of the specimen where $\mathbf{U}_{A0}(z)$ and $\mathbf{U}_{S0}(z)$ have opposite mode shapes there will be a phase shift of π between the reflected Lamb modes for $A0$ or for $S0$ incidence. This phase shift can be taken into account by substituting the cosine terms by sine functions in Eq. (6). In that case the distances corresponding to maxima or minima are given by

$$d_{\max} = \left(n+\tfrac{1}{2}\right)L, \quad d_{\min} = nL, \quad n = 0,1,2,\ldots . \quad (8)$$

This simple model allows for a physical understanding of some important characteristics of the reflected Rayleigh-like wave, especially the dependence of the reflected amplitude on the beat length and hence on the frequency. In order to use this model to predict the amplitude spectrum of the Rayleigh-like wave reflected from the slot, the evaluation of the reflection coefficient C as a function of the slot geometry and the frequency-thickness product is required.

B. Numerical investigation of the reflection coefficient

1. Numerical tools

The two-dimensional scattering at a surface defect was calculated using a finite difference (FD) algorithm, which has been thoroughly validated against analytical and experimental results in an earlier work.[22] The FD algorithm implements an explicit time simulation of the wave propagation and scattering at the defect. The equations of momentum conservation and the stress-strain relations for a two-dimensional (2D), isotropic, linear elastic medium were discretized on a Cartesian, staggered grid.[23] Plane strain wave propagation (2D) was assumed. The defect was implemented as a slot with a flat end and a width $w=0.45$ mm, approximating the machined EDM slots. The incident wave was generated by imposing the complete displacement field in the two initial time steps. For this purpose the input displacements at every node were evaluated using FFT as described by Munasinghe *et al.*[24] The excitation signal was a five cycle tone burst (sinusoid in a Hanning window) to obtain a signal bandwidth (width of main lobe) $\Delta f \cong 0.8 f_0$, where f_0 describes the center frequency. All the numerical results presented in this section were calculated from FD simulations on a 3-mm-thick aluminum plate using an excitation signal with center frequency $f_0=2.25$ MHz. The simulations were performed with approximately 30 nodes per Rayleigh wavelength, using a grid size of $\Delta x=50$ μm. With such a fine spatial sampling the error in the numerical phase velocity of A_0 and S_0 is below 0.3% and the scattering process is simulated accurately.[22]

The FD simulation results were evaluated by means of modal decomposition. Kirmann[25] demonstrated the completeness of the system of fundamental modes in a plate. The functions of this basis are associated with the well-known orthogonality relation:[8]

$$\langle \Phi_m, \Phi_n \rangle = \frac{i\omega}{4}\int\int_S (\underline{\sigma}^m \bar{\mathbf{u}}^n - \bar{\underline{\sigma}}^n \mathbf{u}^m)\mathbf{e}_x dS, \quad (9)$$

where m and n are two fundamental modes, S is the cross section, and x the direction of propagation. The stress field tensor is denoted by $\underline{\sigma}$, the displacement field vector by $\mathbf{u}$, and $\bar{\alpha}$ denotes the complex conjugate of α. For two different fundamental modes ($m \neq n$) the orthogonality relation gives

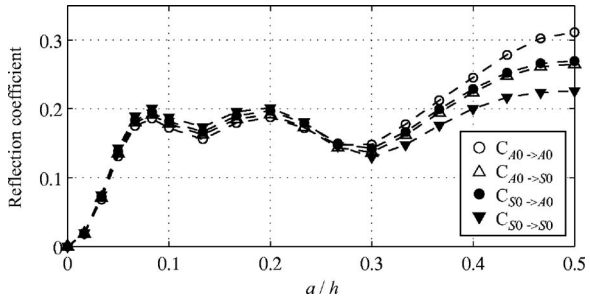


FIG. 7. Reflection coefficients of the fundamental modes for A_0 and S_0 incidence vs normalized crack depth a/h. Values from a scattering simulation at a w=0.45-mm-wide slot in a h=3-mm-thick plate, f_0=2.25 MHz, n_c=5.

$\langle \Phi_m, \Phi_n \rangle = 0$. Any acoustic field distribution Φ in the plate can be written as a superposition of the fundamental modes Φ_m. The coefficient of each fundamental mode can be calculated using the orthogonality relation:[26]

$$C_m = \langle \Phi, \Phi_m \rangle. \tag{10}$$

In the simulations the displacement and stress fields were recorded 20 mm away from the defect at 60 points through the plate thickness. The distance between the slot and the monitoring line was chosen large enough, so that only the propagating modes were monitored. The energy was systematically computed from the reflection and transmission coefficients to ensure the energy balance being correct with an error of less than 2%. The displacement and stress fields were subsequently transformed in the frequency domain by means of FFT and the coefficients of the scattered Lamb modes were determined according to Eq. (10).

2. Reflection coefficients

The reflection coefficients of the first fundamental Lamb modes for A_0 or S_0 incidence are displayed in Fig. 7 as a function of the ratio of slot depth to plate thickness. The coefficient evaluation has been performed at center frequency f_0=2.25 MHz. For all four cases the reflection coefficient shows a steep increase for short cracks up to 7% of the plate thickness (corresponding to a/λ=0.15) followed by nonmonotonic behavior. The four reflection coefficients are nearly identical for slots smaller than 1/3 of the plate thickness. The small deviations in that range arise from the slight differences in mode shape between A_0 and S_0 at 6.75 MHz mm. For cracks deeper than 1/3 of the plate thickness the coefficients diverge with the exception of $C_{A0\rightarrow S0}$ and $C_{S0\rightarrow A0}$, which are nearly identical for any depth, as given by the reciprocity theorem. These observations support the assumption made in the analytical description of the reflected Rayleigh-like wave of a unique reflection coefficient C, valid for slot depths up to about $h/3$. This value depends on the discrepancy between the stress profiles of A_0 and S_0 for the given frequency-thickness product. In the limit case ($\lambda \ll h$) the stress profiles are identical on half the plate thickness and the validity limit value would increase up to $h/2$.

In the case of a Rayleigh wave propagating at the surface of a semi-infinite medium the reflection coefficient

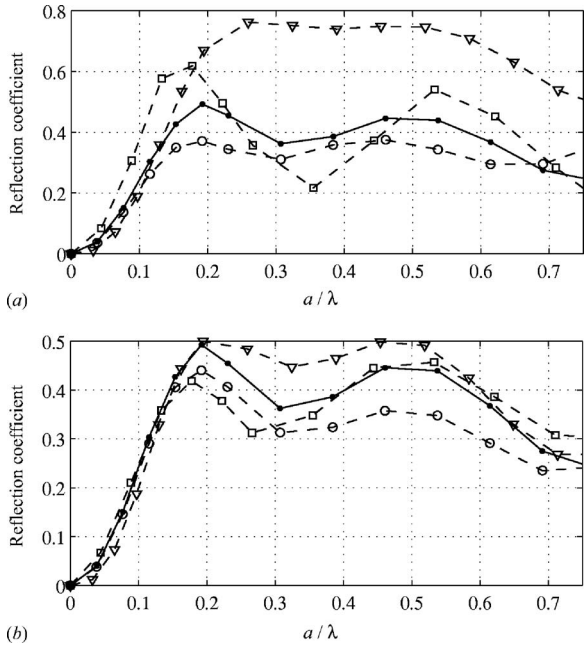


FIG. 8. Reflection coefficient curves of a Rayleigh-like wave vs normalized slot depth a/λ in a 3-mm-thick plate, f=1.9 MHz (triangles), f=2.25 MHz (open points), and f=2.6 MHz (squares) and reflection coefficient of a Rayleigh wave in a semi-infinite medium, f=2.25 MHz (solid points). Values from scattering simulations at a w=0.45-mm-wide slot: (a) No time windowing and (b) time windowing.

curve can be plotted as a function of the ratio of slot depth to Rayleigh wavelength. This curve is shown in Fig. 8(a) (solid line) for a 0.45-mm-wide slot. The values have been evaluated at center frequency f_0=2.25 MHz. As mentioned by Masserey and Mazza,[27] for slots with a width and a depth of the same order of magnitude, this curve differs from the reflection coefficient from a crack with infinitely small opening. The Rayleigh wave curve was compared with the reflection coefficient of a Rayleigh-like wave evaluated at three different frequencies within the bandwidth of the excitation signal. The three curves are displayed in Fig. 8(a) as a function of the ratio of slot depth to wavelength. The reflection coefficient curves for a Rayleigh-like wave show a significant variation with the evaluation frequency, particularly for defects above $a/\lambda \cong 0.15$, corresponding to a defect depth of about a tenth of the plate thickness. Significant differences can also be observed by comparison with the reflection coefficient of a Rayleigh wave, especially for the 1.9 MHz curve. The understanding of these strong discrepancies, which are mainly a consequence of the plate geometry, requires the evaluation of the reflection coefficient of the Rayleigh-like wave in a plate as a function of the frequency for a constant, significant crack depth. Figure 9 shows the reflection coefficient for a 0.5-mm-deep slot evaluated within the frequency bandwidth of the excitation signal. The results demonstrate the strong dependency of the reflection coefficient on the frequency. Three frequencies are visible, where no significant Rayleigh-like wave is reflected. An investigation of the position of the minima in Fig. 9 shows that they correspond to the cutoff frequencies of the A_2, S_3, and A_4 modes (see Fig. 1). For the symmetric case the cutoff frequencies can be written as[28]

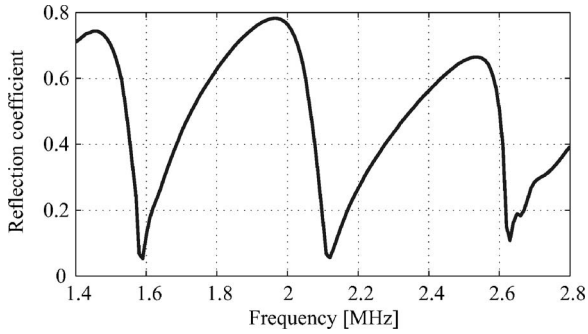


FIG. 9. Reflection coefficient of a Rayleigh-like wave vs frequency. Values from a scattering simulation at a 0.5-mm-deep, 0.45-mm-wide slot in a h=3-mm-thick plate, f_0=2.25 MHz, n_c=5.

$$fh = nc_T, \quad fh = \frac{2n+1}{2}c_L, \tag{11}$$

and for the antisymmetric modes:

$$fh = \frac{2n+1}{2}c_T, \quad fh = nc_L, \tag{12}$$

where $n=0,1,2,\ldots$. All three cutoff frequencies of interest were found to be functions of the transverse wave velocity c_T and the corresponding displacement field shows pure transverse standing waves (group velocity equal to zero) with no out-of-plane particle displacement. Furthermore, by substitution of c_T by $\lambda_T f$ in Eqs. (11) and (12) it can be shown that for A_2, S_3, and A_4 the ratio between double plate thickness $2h$ and wavelength λ_T is an integer.

The physical process accounting for the behavior of the curve in Fig. 9 can be understood in terms of bulk wave scattering. In a semi-infinite medium the scattering of a Rayleigh wave at a surface defect generates semicircular bulk waves propagating into the material, in particular a transverse wave with significant energy concentration below the defect.[29] In the current frequency-thickness range the scattering of an incident A_0 or S_0 Lamb wave in a plate shows locally a similar pattern. However, in the vicinity of the slot the transverse wave is reflected up and down between both sides of the plate.

If the frequency is equal to the cutoff frequency of one of the higher modes mentioned previously, $2h/\lambda_T$ is an integer and, therefore, a quasistanding wave with a mode shape corresponding to the mode shape of the respective higher mode is generated. A significant part of the incident energy is thus converted into that mode, accounting for the minima in Fig. 9. For frequencies significantly higher than a cutoff frequency but lower than the next cutoff frequency, the transverse wave does not generate a standing wave but generates additional reflected and transmitted Rayleigh-like pulses on both sides of the slot each time it propagates upwards and hits the slot (given that the slot exceeds a minimum length). Consequently, the evaluation of the reflection coefficient is affected by the interferences between the many backpropagating Rayleigh-like waves and the coefficient values diverge from the reflection of a Rayleigh wave in a semi-infinite medium.

This statement can be demonstrated by using time windowing to evaluate the reflection coefficients. The reflection coefficient of a Rayleigh-like wave as a function of the ratio of crack depth to wavelength, displayed in Fig. 8(a), was reevaluated at the three cutoff frequencies using time windowing. The duration of the window was set so as to obtain a separation of the direct reflection from the slot from the backpropagating Rayleigh-like pulses arising from the scattering of the transverse wave at the defect. The reflection coefficient curves are illustrated in Fig. 8(b). All reflection coefficient curves obtained using time windowing show a behavior similar to the reflection of a Rayleigh wave in a semi-infinite medium, with a very good agreement for short cracks up to $a/\lambda=0.15$. The discrepancies in the nonmonotonic range are mainly due to the time windowing procedure. Indeed, the direct reflection and the first Rayleigh-like wave arising from the scattering of the transverse wave at the slot slightly overlap in time and the windowing does not allow for an ideal separation of both waves.

It must be stressed that this influence of the cutoff frequencies is only observed for defect depths above a certain threshold. For very short cracks, the transverse wave arising from the scattering at the defect carries very little energy. In addition, the FD simulations show that the semicircular transverse wave generated at the tip of a slot and propagating across the specimen thickness changes for short slots to a quarter-circular wave with most of the energy propagating backwards in the negative x direction. Therefore the energy transfer between the incident Rayleigh-like wave and the standing waves at cutoff frequencies is hardly visible in the amplitude spectrum for short defects, e.g., the 0.1-mm-deep slot.

In summary, the reflection coefficient of a Rayleigh-like wave from a defect in a plate mainly depends on two parameters. Considering only the direct, initial reflection from the defect, the reflection coefficient curve, similar to the coefficient curve of a Rayleigh wave in a semi-infinite medium, depends mainly on the ratio between slot depth, slot width, and wavelength. The plate geometry and, therefore, the multiple reflections through the plate thickness of the transverse wave scattered at the slot, introduce the frequency-thickness product as a second relevant parameter. This parameter yields a strong modification of the reflection coefficient curve and, for frequency-thickness products close to a cutoff frequency, the reflection coefficient is very small ($<5\%$), where one would normally expect a significant reflected signal.

VII. RESULTS AND DISCUSSION

The parameters affecting the reflected Rayleigh-like wave, i.e., excitation distance, excitation frequency, and frequency-thickness product have been discussed in Sec. VI on the basis of the hybrid model. In combination with the beat-length phenomenon described earlier, the model allows for the interpretation of the amplitude spectrum of the pulse-echo measurements presented in Sec. V.

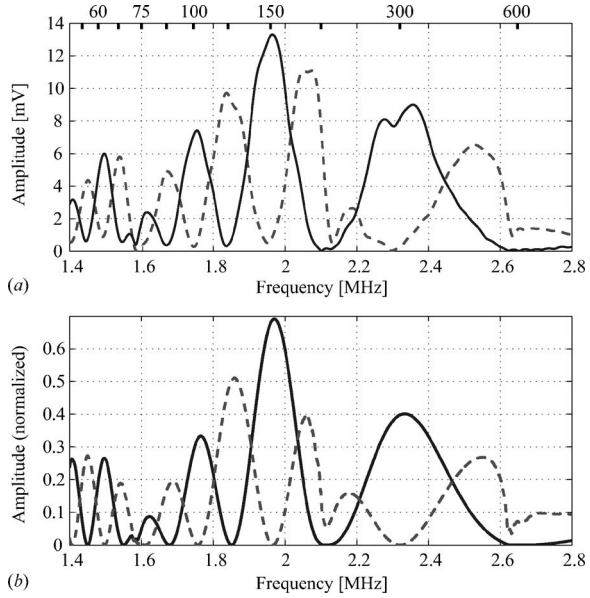


FIG. 10. Amplitude spectrum of the reflected Rayleigh-like wave at a 0.5-mm-deep slot: (a) Measurements on the 3.05-mm-thick plate, f_0=2.25 MHz, d=300 mm; (b) theoretical curve evaluated using the hybrid model. Wave generation on the damaged side (solid line) and on the undamaged side (dashed line).

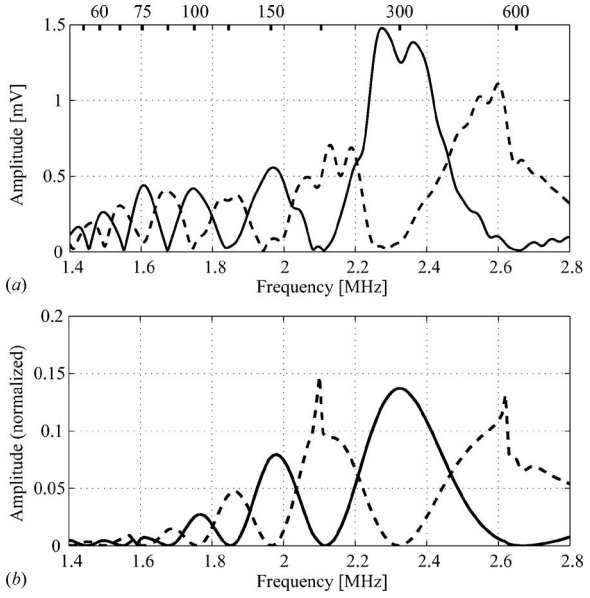


FIG. 11. Amplitude spectrum of the reflected Rayleigh-like wave at a 0.1-mm-deep slot: (a) Measurements on the 3.05-mm-thick plate, f_0=2.25 MHz, d=300 mm; (b) theoretical curve evaluated using the hybrid model. Wave generation on the damaged side (solid line) and on the undamaged side (dashed line).

A. Frequency domain analysis of the pulse-echo measurements

The time traces of the pulse-echo measurements displayed in Fig. 6 were transformed into the frequency domain by means of a FFT. The amplitude spectrum displayed in Fig. 10(a) shows the result for the evaluation of the pulse-echo measurements at the 0.5-mm-deep slot in the 3.05-mm-thick plate. The curve is characterized by a clear maxima and minima pattern. The beat-length values corresponding to the expected extrema can be computed from Eqs. (7) and (8). The corresponding frequencies are obtained by interpolation using the beat-length values measured on the 3.05-mm-thick plate (see Fig. 5). The agreement between the actual maxima and the expected values obtained from the experimental beat-length curve is excellent for most of the peaks. As expected, the comparison of the amplitude curves obtained on the damaged and on the undamaged plate sides shows an alternation between the frequency position of the maxima and minima. Both curves are characterized by common minima at approximately 1.6, 2.1, and 2.6 MHz, which interrupt the maxima and minima pattern resulting from the beat-length phenomenon. These minima correspond to the cutoff frequencies of the A_2, S_3, and A_4 modes, respectively, where the scattered energy is mainly stored in the corresponding higher Lamb mode, as has been shown in Fig. 9.

The corresponding theoretical curves were calculated using the hybrid model described in Sec. VI. For pulse-echo measurements the monitoring distance x in Eq. (6) is the distance d between excitation position and defect. The reflection coefficient was calculated by means of FD simulations and the beat-length values were interpolated from the theoretical data displayed in Fig. 5. The resulting amplitude curves for a defect on the excitation surface or on the back side of the plate after multiplication with the amplitude spectrum of the excitation signal are displayed in Fig. 10(b).

Comparison of the experimental amplitude spectra with the amplitude curves obtained by means of the hybrid model shows a very good agreement. The combination of the analytical model and the coefficients from the FD calculations allows for the calculation of the characteristic behavior of the amplitude spectrum of the reflected Rayleigh-like wave over the entire bandwidth of the generated pulse.

The amplitude spectrum of the Rayleigh-like wave reflected from the 0.1-mm-deep slot is shown in Fig. 11(a) (experiments) and Fig. 11(b) (hybrid model). The measured amplitude extrema show a very good agreement with the predicted values. The distance between wedge and defect is the same as for the 0.5-mm-deep slot; the frequencies corresponding to the extrema are therefore identical. The frequency range displayed in Fig. 11 corresponds to ratios of crack depth to wavelength of 0.05–0.1. As shown in Fig. 8 the reflection coefficient is characterized by a steep increase in that parameter range. This accounts for the overall increase of the amplitude spectrum with increasing frequency in Fig. 11. Comparison of experimental and predicted curves shows a good qualitative agreement. The amplitude discrepancies can be associated with the small size of the slot, making the modeling of the exact geometry of the EDM slot difficult, as the 0.1-mm-deep slot was modeled by only two elements. The secondary oscillations of the experimental curves are due to interferences with higher modes. Indeed, in that frequency range, the group velocity of many higher modes is close to the Rayleigh wave propagation velocity. For instance, at approximately 2.4 MHz the difference between Rayleigh wave velocity and the group velocities of A_1, A_2, and S_3 is less than 7%. These modes interfere with the reflected Rayleigh-like wave and affect the amplitude spec-

trum despite the partial filtering of the wedge. This occurred as well for the 0.5 mm slot (see Fig. 10(a)), but there the energy ratio between the first Lamb modes and the higher Lamb modes was significantly higher.

Comparison of Figs. 10 and 11 confirms the statement made in Sec. VI that the influence of the cutoff frequencies is only observable for defect depths above a threshold. The amplitude spectra in Fig. 11(b) are characterized by an amplitude rise and steep drop just below the cutoff frequencies of the S_3 and A_4 modes (2.1 and 2.6 MHz, respectively). The steep amplitude drop just below 2.6 MHz can as well be observed in the experimental spectrum shown in Fig. 11(a). However, the reflected Rayleigh-like wave still carries significant energy and the amplitude curves do not drop to zero.

The evaluation of the reflected Rayleigh-wave pulse in the frequency domain by means of a simple FFT shows the characteristic amplitude spectrum due to the beat-length phenomenon for both the 0.5- and 0.1-mm-deep slots. By means of a simple comparison of the frequencies of the amplitude maxima with the beat-length predictions for the propagation distance obtained from the time-of-flight measurement, the side of the plate at which the defect occurred can be easily ascertained. An error in the time-of-flight measurement of 1%, or an error in the beat-length curve similar to the difference between experimental and theoretical curves in Fig. 5 would result in an error of less than 1% for the position of the peaks in the amplitude spectrum, which would not be significant for the determination of the damaged plate side.

B. Requirements for defect detection

For the pulse-echo experimental configuration the shortest beat length L, corresponding to the lowest frequency in the spectrum, is approximately 45 mm ($15h$). This does not allow for the detection of defects situated on the back side of the plate at less than 40 mm from the wedge, unless the distance to the defect is increased by shifting the angle-beam transducer. However, this range could easily be inspected using standard shear wave inspection techniques if necessary. For most inspection purposes with Rayleigh-like waves, the use of a wideband signal is advantageous over a narrow-band pulse. This allows as well for the determination of the damaged side on the basis of the amplitude spectrum. If a narrow-band pulse with a center frequency close to a cutoff frequency was used, the defect may not be detected as little energy would be reflected back toward the transducer as a propagating Rayleigh-like pulse. Depending on the distance d between transducer and defect (which would not be known *a priori*), for a narrow-band pulse the incident Rayleigh-like wave could be concentrated on the undamaged plate side at the defect location so that no significant energy would be reflected.

VIII. CONCLUSIONS

The propagation and scattering of Rayleigh-like waves in isotropic, linear elastic plates has been investigated experimentally and theoretically for frequency-thickness products where the wavelength is of the order of about half the plate thickness, i.e., significantly above the cutoff frequencies of the higher Lamb wave modes. Experimentally, coupled Rayleigh-like waves were excited in aluminum plates employing standard Rayleigh wave angle beam transducers. This allowed for the selective excitation of the required modes at frequency-thickness products of about 6.5 MHz mm. For the structures under consideration with a thickness typically found in aircraft components, this is a good experimental setting, allowing the use of standard transducer technology to excite waves with a short enough wavelength for the detection of small surface defects.

The energy transfer occurring between the upper and lower plate surfaces during the propagation of the Rayleigh-like wave was observed in both time domain and frequency domain. The beat length was measured accurately using a fitting procedure. Excellent agreement was found between the measured beat-length curve and theoretical predictions. The results revealed a high sensitivity of the beat length to the material parameters as well as to the plate thickness.

A theoretical model was proposed to describe the main characteristics of the Rayleigh-like wave reflected from a small surface defect. The scattering process was modeled separately for the A_0 and S_0 modes that constitute the incident Rayleigh-like wave. The reflection coefficients of the fundamental modes from a slot were investigated using finite difference simulations. The superposition property of linear wave propagation was subsequently used to describe the reflected Rayleigh-like wave.

The hybrid model revealed a strong dependency of the reflected Rayleigh-like wave on the ratio between excitation distance and beat length. For specific ratios, a narrow-band incident wave would propagate through the defect region without any perturbations. A wideband excitation signal induces a clear minima and maxima pattern in the amplitude spectrum of the reflected Rayleigh-like wave. In addition, a significant energy transfer occurs between the incident wave and specific higher Lamb modes at their cutoff frequencies, generating distinct minima in the amplitude spectrum of the reflected Rayleigh-like wave.

The use of Rayleigh-like waves allowed for the detection of small slots on both plate sides, using a standard pulse-echo technique. The measurements were performed at a large distance from the slot (about 100 times the plate thickness). Good agreement was obtained between the measured and predicted amplitude spectrum of the reflected Rayleigh-like wave. The combination of information in the time domain and frequency domain, i.e., the arrival time and the amplitude spectrum of the reflected Rayleigh-like wave, provided sufficient information for the defect location, including determination of the damaged plate side. The use of a pulse with a wideband spectrum for defect detection avoids a full transmission of the pulse past the defect or the conversion of a significant amount of the reflected energy into standing higher Lamb modes.

For the inspection over larger distances, one needs to compare to other guided wave techniques and consider the trade-off among complexity of the propagating wave modes, achievable propagation distance, and sensitivity for the detection of small defects, governed by the excitation frequency and thus wavelength. This contribution has shown

that a nondispersive, single wave pulse can be excited and propagated along a plate at frequency-thickness products significantly above the cutoff frequency of the A_1 Lamb wave mode. For propagation distances which are significantly larger than those commonly achieved using shear wave inspection, good sensitivity for small defects has been observed due to the short wavelength of the excited Rayleigh-like wave mode.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the Swiss National Science Foundation, Grant No. PBEZ2-114186, and the UK Engineering and Physical Sciences Research Council (EPSRC), Grant No. EP/D065011/1.

[1]P. Bövik and A. Boström, "A model of ultrasonic nondestructive testing for internal and subsurface cracks," J. Acoust. Soc. Am. **102**, 2723–2733 (1997).
[2]P. Fromme, P. D. Wilcox, M. J. S. Lowe, and P. Cawley, "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).
[3]D. N. Alleyne, B. Pavlakovic, M. J. S. Lowe, and P. Cawley, "Rapid long-range inspection of chemical plant pipework using guided waves," Insight **32**, 93–96 (2001).
[4]D. N. Alleyne and P. Cawley, "The interaction of Lamb waves with defects," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **39**, 381–397 (1992).
[5]P. Fromme, B. Masserey, and M. B. Sayir, "On the detectability of fatigue crack growth at fastener holes using guided waves," in *Review of Progress in QNDE*, AIP Conference Proceedings 657, edited by D. O. Thompson and D. E. Chimenti (AIP, New York, 2003), Vol. **22**, pp. 189–196.
[6]I. A. Viktorov, *Rayleigh and Lamb Waves* (Plenum, New York, 1967), pp. 93–96.
[7]B. W. Ti, W. D. O'Brien, and J. G. Harris, "Measurements of coupled Rayleigh wave propagation in an elastic plate," J. Acoust. Soc. Am. **102**, 1528–1531 (1997).
[8]B. A. Auld, *Acoustic Fields and Waves in Solids* (Wiley, New York, 1973), Vol. **2**, pp. 93–94.
[9]J. G. Harris, "Rayleigh wave propagation in curved waveguides," Wave Motion **36**, 425–441 (2002).
[10]R. P. Dalton, P. Cawley, and J. S. Lowe, "The potential of guided waves for monitoring large areas of metallic aircraft fuselage structures," J. Nondestruct. Eval. **20**, 29–46 (2001).
[11]J. D. Achenbach, A. K. Gautesen, and D. A. Mendelsohn, "Ray analysis of surface-wave interaction with an edge crack," IEEE Trans. Sonics Ultrason. **27**, 124–129 (1980).
[12]M. Hirao, H. Fukuoka, and Y. Miura, "Scattering of Rayleigh surface waves by edge cracks: Numerical simulation and experiment," J. Acoust. Soc. Am. **72**, 602–606 (1982).
[13]Y. C. Angel and J. D. Achenbach, "Reflection and transmission of obliquely incident Rayleigh waves by a surface-breaking crack," J. Acoust. Soc. Am. **75**, 313–319 (1984).
[14]R. Dong and L. Adler, "Measurements of reflection and transmission coefficients of Rayleigh waves from cracks," J. Acoust. Soc. Am. **76**, 1761–1763 (1984).
[15]M. J. S. Lowe and O. Diligent, "Low-frequency reflection characteristics of the s0 Lamb wave from a rectangular notch in a plate," J. Acoust. Soc. Am. **111**, 64–74 (2002).
[16]M. J. S. Lowe, P. Cawley, J.-Y. Kao, and O. Diligent, "The low frequency reflection characteristics of the fundamental antisymmetric Lamb wave a_0 from a rectangular notch in a plate," J. Acoust. Soc. Am. **112**, 2612–2622 (2002).
[17]M. A. Flores-Lopez and R. D. Gregory, "Scattering of Rayleigh-Lamb waves by a surface breaking crack in an elastic plate," J. Acoust. Soc. Am. **119**, 2041–2049 (2006).
[18]Y. Cho, D. Hongerholt, and J. Rose, "Lamb wave scattering analysis for reflector characterization," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 44–52 (1997).
[19]S. W. Liu and S. K. Datta, "Scattering of ultrasonic wave by cracks in a plate," J. Appl. Mech. **60**, 352–357 (1993).
[20]N. Terrien, D. Osmont, D. Royer, F. Lepoutre, and A. Déom, "A combined finite element and modal decomposition method to study the interaction of Lamb modes with micro-defects," Ultrasonics **46**, 74–88 (2007).
[21]P. Fromme and M. B. Sayir, "Measurement of the scattering of a Lamb wave by a through hole in a plate," J. Acoust. Soc. Am. **111**, 1165–1170 (2002).
[22]B. Masserey and E. Mazza, "Analysis of the near-field ultrasonic scattering at a surface crack," J. Acoust. Soc. Am. **118**, 3585–3594 (2005).
[23]R. Madariaga, "Dynamics of an expanding circular fault," Bull. Seismol. Soc. Am. **66**, 639–666 (1976).
[24]M. Munasinghe and G. W. Farnell, "Finite difference analysis of Rayleigh wave scattering at vertical discontinuities," J. Geophys. Res. **78**, 2454–2466 (1973).
[25]P. Kirrmann, "On the completeness of Lamb modes," J. Elast. **37**, 39–69 (1995).
[26]M. Castaings, E. Le Clezio, and B. Hosten, "Modal decomposition method for modeling the interaction of Lamb waves with cracks," J. Acoust. Soc. Am. **112**, 2567–2582 (2002).
[27]B. Masserey and E. Mazza, "Ultrasonic sizing of short surface cracks," Ultrasonics **46**, 195–204 (2007).
[28]K. F. Graff, *Wave Motion in Elastic Solids*, (Oxford University Press, London, 1975), pp. 446–458.
[29]R. J. Blake and L. J. Bond, "Rayleigh wave scattering from surface features: Up-steps and troughs," Ultrasonics **30**, 255–265 (1992).

Faxén relations in solids—a generalized approach to particle motion in elasticity and viscoelasticity

Andrew N. Norris[a)]
Mechanical and Aerospace Engineering, Rutgers University, Piscataway, New Jersey 08854, USA

(Received 17 July 2007; revised 29 October 2007; accepted 3 November 2007)

A movable inclusion in an elastic material oscillates as a rigid body with six degrees of freedom. Displacement/rotation and force/moment tensors which express the motion of the inclusion in terms of the displacement and force at arbitrary exterior points are introduced. Using reciprocity arguments two general identities are derived relating these tensors. Applications of the identities to spherical particles provide several new results, including simple expressions for the force and moment on the particle due to plane wave excitation. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2817359]

PACS number(s): 43.40.Fz, 43.20.Rz, 43.20.Tb, 43.80.Ev [AJMD] Pages: 99–108

I. INTRODUCTION

Faxén relations are named after Hilding Faxén who derived several identities for calculating hydrodynamic forces and torques on particles in low Reynolds number flows, e.g., Ref. 1. As an example of a Faxén relation, or law, the force and torque on a rigid sphere of radius a moving with velocity $\mathbf{v}_0$ and spinning with angular velocity $\boldsymbol{\omega}$ in an unbounded fluid of viscosity μ and velocity field $\mathbf{v}(\mathbf{r})$ in the absence of the sphere are [Eqs. (3-2.46) and (3-2.47) of Ref. 2]

$$\mathbf{F} = 6\pi\mu a[\mathbf{v}(\mathbf{0}) - \mathbf{v}_0] + \pi\mu a^3\boldsymbol{\nabla}^2\mathbf{v}(\mathbf{0}), \tag{1a}$$

$$\mathbf{T} = 4\pi\mu a^3[\mathrm{curl}\mathbf{v}(\mathbf{0}) - 2\boldsymbol{\omega}]. \tag{1b}$$

The reader will note that the identities have as a special case the classical Stokes drag law, but they include additional effects caused by spatially variable flow fields. These and other Faxén relations for nonspherical particles are based upon general integral identities relating the force and torque on the particle to the external flow field.[2–4] Although Faxén relations are commonly used in hydrodynamics and microfluidics, they seem to be essentially unknown outside that subject area. For instance, I am aware of only one mention[5] of a Faxén-type relation in elasticity and that one was is in regards to elastostatics.

The objective of this paper is to develop similar ideas in the context of elastodynamics and in the process demonstrate their utility and wide application. Using dynamic reciprocity, a set of relations is first derived between the velocity or force of a particle in a solid matrix and the displacement or force at a distant point in the solid. These equations include but go far beyond the notion of particle impedance, which relates the force on a particle to its velocity. Numerous applications of the general relations are obtained by considering spherical particles. Faxén-like relations are derived for the force and moment on a spherical particle caused by plane wave incidence. Like their hydrodynamic counterparts, the elastodynamic Faxén relations are simple in form.

The analysis here is the second in a series of papers developing a simplified algebra for calculating the radiation and scattering from inclusions in elastic and viscoelastic materials. In the previous paper[6] the forced motion of a spherical particle in an elastic matrix was considered. Although this is a classical problem, originally solved by Oestreicher,[7] it turns out that the dynamic impedance of the inclusion can be represented in a simplified manner. This was achieved[6] through several lumped mass impedances for a spherical inclusion, in terms of which impedance or its inverse, admittance, has a simple form. The purpose of the present paper is to develop these ideas further, focusing on the interaction between the inclusion and remote points in the matrix.

The plan of the paper is as follows. The dynamic properties of an inclusion are defined in Sec. II, as well as some important quantities that are used throughout the paper: the displacement/rotation tensors $\mathbf{U}$ and $\mathbf{W}$, the force/moment tensors $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, and the impedances $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. In Sec. III we prove the symmetry of these impedances, and derive two fundamental relations between the displacement/rotation and force/moment tensors. The remainder of the paper focuses on the special case of spherical inclusions. The fundamental quantities for the spherical particle are presented in Sec. IV in a concise format using lumped mass impedances. Section V is the longest in the paper, as it contains numerous applications, discussion of limiting cases, and the new elastodynamic Faxén relations that are analogous to the classic hydrodynamic identities. The many results and their import are summarized in Sec. VI.

Regarding notation, the time harmonic factor $e^{-i\omega t}$ is omitted but understood. Boldface quantities are either vectors or second order tensors. Vectors are usually denoted by lower case, and tensors are capitalized, with the exceptions $\mathbf{F}$ and $\mathbf{M}$ which indicate force and moment vectors, respectively. The axial tensor $\mathrm{axt}(\mathbf{a})$ of the vector $\mathbf{a}$ is a skew symmetric tensor defined by $\mathrm{axt}(\mathbf{a})\mathbf{b} = \mathbf{a} \wedge \mathbf{b}$.

II. INCLUSIONS AND RIGID BODY MOTION

An inclusion in a solid matrix is defined to be the surface ∂V_p of a finite volume V_p within which there could be a

a)Electronic mail: norris@rutgers.edu

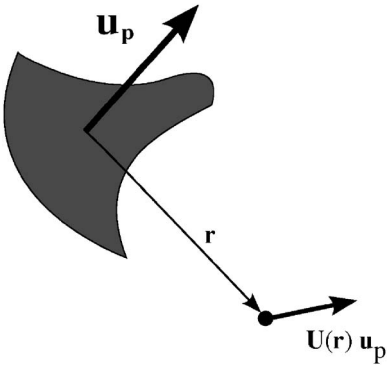


FIG. 1. The inclusion undergoes time harmonic linear displacement $\mathbf{u}_p$, resulting in displacement $\mathbf{u}=\mathbf{U}(\mathbf{r})\mathbf{u}_p$ at position $\mathbf{r}$.

particle, or there could be some complicated "black box" with its own internal dynamics. The key feature of the inclusion is that its boundary ∂V_p undergoes rigid body motion. In this sense the boundary is a rigid interface between the particle, whatever that may be, and the solid matrix. The term inclusion rather than particle is used throughout in order to remind us of this distinction.

A. Tensor functions U, W, Φ, and Ψ

Rigid body motion has six degrees of freedom, which we characterize by two vector quantities: $\mathbf{u}_p$ and $\boldsymbol{\theta}_p$. The term $\mathbf{u}_p$ is the rigid body displacement of the inclusion center of mass. $\boldsymbol{\theta}_p$ describes the rotation of the inclusion about the center of mass. The most general displacement possible for the inclusion is

$$\mathbf{u}_P = \mathbf{u}_p + \boldsymbol{\theta}_p \wedge \mathbf{r}, \quad \forall\, \mathbf{r} \in \partial V_p, \tag{2}$$

where $\mathbf{r}$ is the position relative to the center of mass. We will use the vector $\mathbf{u}_P$ to denote the total rigid body displacement, with $\mathbf{u}_p$ reserved for the linear part (the entire development in this paper applies only to *linear* as distinct from nonlinear motion, so that the term linear is synonymous with rectilinear). Note that $\mathbf{u}_p$ has dimensions of length while $\boldsymbol{\theta}_p$ is dimensionless. For the sake of simplicity it is useful to consider the linear and rotational motions separately. Figure 1 shows the inclusion oscillating back and forth with linear displacement $\mathbf{u}_p$. In the absence of other sources of vibrational energy, the inclusion motion induces motion at every point $\mathbf{r}$ in the exterior region $V=\mathbb{R}^3/V_p$ according to $\mathbf{u}=\mathbf{U}(\mathbf{r})\mathbf{u}_p$, as depicted in Fig. 1. Here $\mathbf{U}$ is a second order tensor defined everywhere in the matrix. In the same way the particle displacement at $\mathbf{r}\in V$ caused by a pure rotation of the inclusion may be defined by a second order tensor $\mathbf{W}$. In short, the tensors $\mathbf{U}$ and $\mathbf{W}$ relate the rigid body displacement of the inclusion to the displacement $\mathbf{u}(\mathbf{r})$ in the exterior solid medium V according to

$$\mathbf{u}(\mathbf{r}) = \mathbf{U}(\mathbf{r})\mathbf{u}_p + \mathbf{W}(\mathbf{r})\boldsymbol{\theta}_p. \tag{3}$$

We next define two dual tensor functions associated with force and moment, respectively.

Consider the situation in which a point force of magnitude times direction equal to $\mathbf{F}$ acts at $\mathbf{s}\in V$, inducing motion

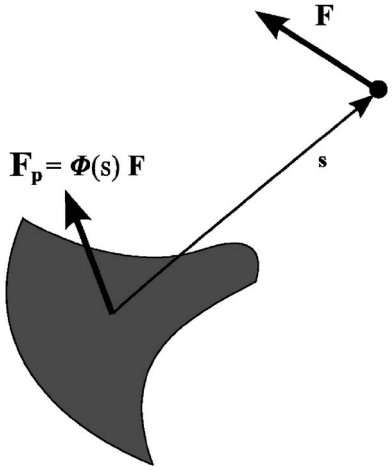


FIG. 2. The time harmonic point force $\mathbf{F}$ is applied at s, resulting in the force $\mathbf{F}_p=\mathbf{\Phi}(\mathbf{s})\mathbf{F}$ on the inclusion.

of the inclusion, Fig. 2. The tensors $\mathbf{\Phi}$ and $\mathbf{\Psi}$ define the net force $\mathbf{F}_p$ and couple $\mathbf{M}_p$ on the inclusion caused by the point force according to

$$\mathbf{F}_p = \mathbf{\Phi}(\mathbf{s})\mathbf{F}, \quad \mathbf{M}_p = \mathbf{\Psi}(\mathbf{s})\mathbf{F}. \tag{4}$$

1. External impedances

We introduce two impedance tensors: $\mathbf{Z}$ and $\widetilde{\mathbf{Z}}$, called external impedances because they depend upon the exterior properties of the solid matrix.

The impedance $\mathbf{Z}$ relates the force acting on the inclusion to the inclusion linear velocity, see Fig. 3, which is similar to the situation in Fig. 4. It is assumed that either force or velocity is controlled and the other is the dependent variable, and that there is no other excitation from sources in V. Thus, let $\mathbf{u}_p$ be the prescribed inclusion linear displacement, then the force $\mathbf{F}_p$ acting on the inclusion is

$$\mathbf{F}_p = -i\omega\mathbf{Z}\mathbf{u}_p. \tag{5}$$

$\mathbf{F}_p$ can be thought of as the resultant of the reactive forces from the solid matrix acting on the inclusion, with zero net moment. The inverse $\mathbf{Z}^{-1}$ exists since we may consider the impedance as defined by an imposed force, resulting in the

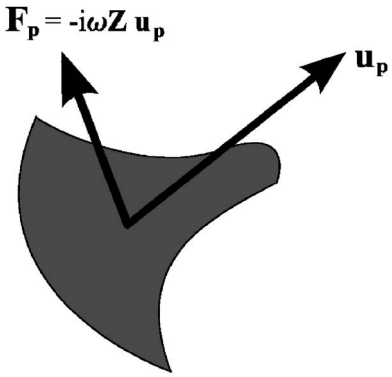


FIG. 3. The inclusion undergoes time harmonic linear displacement $\mathbf{u}_p$, resulting in the net force $\mathbf{F}_p$ acting on the inclusion. Conversely, if the force $\mathbf{F}_p$ is applied then the displacement is $\mathbf{u}_p$.

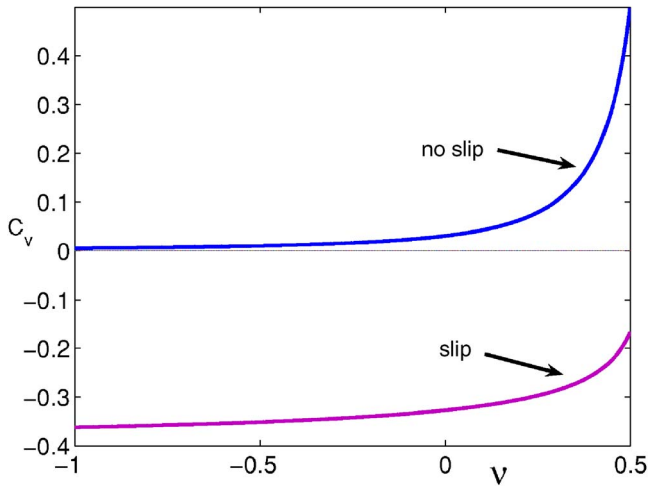


FIG. 4. (Color online) The virtual mass coefficient C_v of the spherical inclusion plotted as a function of the Poisson's ratio. The no-slip and slip curves correspond to $\chi=0$ and $\chi=1$ in Eq. (53), respectively.

inclusion displacement $\mathbf{u}_p=(-i\omega\mathbf{Z})^{-1}\mathbf{F}_p$. We will prove that $\mathbf{Z}$ is symmetric (Lemma 1).

The moment tensor $\tilde{\mathbf{Z}}$ relates the moment of the force on the inclusion with the inclusion angular velocity, $-i\omega\tilde{\mathbf{Z}}\boldsymbol{\theta}_p$,

$$\mathbf{M}_p=-i\omega\tilde{\mathbf{Z}}\boldsymbol{\theta}_p. \tag{6}$$

$\tilde{\mathbf{Z}}$ is also assumed to be invertible, and will be shown to be symmetric (Lemma 1).

2. Internal impedances

The inertial properties of the inclusion are defined by two impedance matrices $\mathbf{Z}_P$ and $\tilde{\mathbf{Z}}_P$ associated with linear and rotational motion, respectively. We call these internal impedances since they depend entirely on the inclusion and are independent of the exterior region.

$\mathbf{Z}_P$ is a mass-like impedance. For a normal solid particle it is defined by the mass m as $\mathbf{Z}_P=i\omega m\mathbf{I}$. We will generally denote $\mathbf{Z}_P$ as a tensor to include the possibility of internal structure, although it may be assumed on general principles that the impedance is symmetric, $\mathbf{Z}_P=\mathbf{Z}_P^t$.

$\tilde{\mathbf{Z}}_P$ is the moment of inertia tensor, and is also symmetric $\tilde{\mathbf{Z}}_P=\tilde{\mathbf{Z}}_P^t$. It has dimensions of a moment of inertia, i.e., mass $\times$ (length)^2, multiplied by frequency.

3. Summary of main results

The first principal result is a pair of relations (i) between the displacement and force tensors and (ii) between the rotation and moment tensors

$$\mathbf{Z}_P^{-1}\boldsymbol{\Phi}(\mathbf{r})=(\mathbf{Z}+\mathbf{Z}_P)^{-1}\mathbf{U}^t(\mathbf{r}), \tag{7a}$$

$$\tilde{\mathbf{Z}}_P^{-1}\boldsymbol{\Psi}(\mathbf{s})=(\tilde{\mathbf{Z}}+\tilde{\mathbf{Z}}_P)^{-1}\mathbf{W}^t(\mathbf{s}). \tag{7b}$$

Thus, $\boldsymbol{\Phi}=\mathbf{U}^t$ and $\boldsymbol{\Psi}=\mathbf{W}^t$ if the inclusion is immovable (infinite impedance), however, the relations (7a) and (7b) are obviously far more general. An immediate corollary is that the motion of the inclusion caused by the remote force at $\mathbf{s}$ is

$$\mathbf{u}_p=(i\omega)^{-1}(\mathbf{Z}+\mathbf{Z}_P)^{-1}\mathbf{U}^t(\mathbf{s})\mathbf{F}, \tag{8a}$$

$$\boldsymbol{\theta}_p=(i\omega)^{-1}(\tilde{\mathbf{Z}}+\tilde{\mathbf{Z}}_P)^{-1}\mathbf{W}^t(\mathbf{s})\mathbf{F}. \tag{8b}$$

Equations (7a) and (7b) are proved in the next section (Lemma 2).

The second set of principal results concern applications to a spherical inclusion in an isotropic matrix. The displacement and rotations tensors $\mathbf{U}$ and $\mathbf{W}$ have particularly simple forms when expressed in terms of some lumped parameter impedances introduced in Ref. 6. Combined with Eqs. (7a) and (7b) these lead to a series of useful identities for the force, moment, displacement and rotation of the sphere under different excitation. For instance, the total force and moment on the sphere caused by a time harmonic longitudinal or transverse plane wave is

$$\mathbf{F}_p=\Lambda\mathbf{u}_0,\quad \mathbf{M}_p=\Gamma\mathbf{u}_0\wedge\mathbf{n}, \tag{9}$$

where $\mathbf{u}_0$ is the wave displacement at the center when the sphere is not present, $\mathbf{n}$ is the propagation direction, and the scalars Λ, Γ, depend upon the wave frequency, particle radius, and other material parameters according to Eqs. (44), (46), and (48). Equations (9) could be called Faxén relations for solids, by analogy with the use of the term in viscous fluid dynamics.

III. RECIPROCITY BASED IDENTITIES

Several identities are derived in this section: (i) the symmetry of the external impedance matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, (ii) the relation (7a) between the force and displacement tensors, $\boldsymbol{\Phi}$ and $\mathbf{U}$, and (iii) Eq. (7b) relating the moment and rotation tensors $\boldsymbol{\Psi}$ and $\mathbf{W}$. The common theme is the use of the dynamic reciprocity relation.

Consider two distinct fields in V, labeled $j=1,2$, each with displacement $\mathbf{u}^{(j)}$, stress $\boldsymbol{\sigma}^{(j)}$ and applied body force density per unit volume $\mathbf{f}^{(j)}$ all in dynamic equilibrium

$$\operatorname{div}\boldsymbol{\sigma}^{(j)}+\rho\omega^2\mathbf{u}^{(j)}+\mathbf{f}^{(j)}=0,\quad \text{in } V. \tag{10}$$

The reciprocity identity (Betti's theorem) follows from standard arguments[8],

$$\int_{\partial V_p}\mathrm{d}s\mathbf{u}^{(1)}\cdot\boldsymbol{\tau}^{(2)}-\int_V\mathrm{d}v\mathbf{u}^{(1)}\cdot\mathbf{f}^{(2)}=\int_{\partial V_p}\mathrm{d}s\mathbf{u}^{(2)}\cdot\boldsymbol{\tau}^{(1)}-\int_V\mathrm{d}v\mathbf{u}^{(2)}\cdot\mathbf{f}^{(1)}, \tag{11}$$

where $\boldsymbol{\tau}$ is the traction vector. The surface integrals in Eq. (11) involve only quantities in the matrix. We assume that the following conditions hold at the interface ∂V_p: (i) continuity of traction, and (ii) the exterior displacement $\mathbf{u}$ is related to the inclusion displacement $\mathbf{u}_P$ by

$$\mathbf{u}-\mathbf{u}_P=\mathbf{A}\boldsymbol{\tau},\quad \mathbf{r}\in\partial V_p, \tag{12}$$

where $\mathbf{A}=\mathbf{A}^t$ is a material parameter. This spring-like interface condition allows for the possibility of, for instance, tangential slip, which we will include in the example of the spherical inclusion later. For the moment we leave $\mathbf{A}$ as arbitrary.

The rigid body motion of the inclusion for each of the two distinct solutions in Eq. (11) is assumed to be a linear

displacement $\mathbf{u}^{(j)}$ and a twist $\boldsymbol{\theta}_p^{(j)}$, such the total displacement is $\mathbf{u}_P^{(j)}=\mathbf{u}_p^{(j)}+\boldsymbol{\theta}_p^{(j)}\wedge\mathbf{r}$, see Eq. (2). Substituting for $\mathbf{u}^{(j)}$ in the surface integrals then gives

$$\mathbf{u}_p^{(1)}\cdot\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(2)}+\boldsymbol{\theta}_p^{(1)}\cdot\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(2)}-\int_V\mathrm{d}v\mathbf{u}^{(1)}\cdot\mathbf{f}^{(2)}$$
$$=-\int_V\mathrm{d}v\mathbf{u}^{(2)}\cdot\mathbf{f}^{(1)}+\mathbf{u}_p^{(2)}\cdot\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(1)}$$
$$+\boldsymbol{\theta}_p^{(2)}\cdot\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(1)}. \tag{13}$$

Note that the interfacial tensor $\mathbf{A}$ does not appear in this identity. We are now ready to derive the fundamental relations, first considering the impedances.

A. Symmetry of the external impedances Z and $\tilde{Z}$

Assume that no force acts in the solid for both solutions, so that $\mathbf{f}^{(1)}=\mathbf{f}^{(2)}=0$. Then Eq. (13) reduces to

$$\mathbf{u}_p^{(1)}\cdot\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(2)}+\boldsymbol{\theta}_p^{(1)}\cdot\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(2)}=\mathbf{u}_p^{(2)}\cdot\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(1)}$$
$$+\boldsymbol{\theta}_p^{(2)}\cdot\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(1)}. \tag{14}$$

The integrals produce the resultant force and moment on the inclusion, which follows from the definition of the impedances $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ as

$$\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(j)}=-i\omega\mathbf{Z}\mathbf{u}_p^{(j)}, \tag{15a}$$

$$\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(j)}=-i\omega\tilde{\mathbf{Z}}\boldsymbol{\theta}_p^{(j)}, \tag{15b}$$

for $j=1,2$. The reciprocity relation becomes

$$\mathbf{u}_p^{(1)}\cdot\mathbf{Z}\mathbf{u}_p^{(2)}+\boldsymbol{\theta}_p^{(1)}\cdot\tilde{\mathbf{Z}}\boldsymbol{\theta}_p^{(2)}=\mathbf{u}_p^{(2)}\cdot\mathbf{Z}\mathbf{u}_p^{(1)}+\boldsymbol{\theta}_p^{(2)}\cdot\tilde{\mathbf{Z}}\boldsymbol{\theta}_p^{(1)}. \tag{16}$$

Since $\mathbf{u}_p^{(1)}$, $\boldsymbol{\theta}_p^{(1)}$, $\mathbf{u}_p^{(2)}$ and $\boldsymbol{\theta}_p^{(2)}$ are arbitrary, we deduce

Lemma 1 *The linear and rotational external impedances are symmetric*,

$$\mathbf{Z}=\mathbf{Z}^t,\quad\tilde{\mathbf{Z}}=\tilde{\mathbf{Z}}^t. \tag{17}$$

B. Relation between the force and displacement tensors

We again take field 1 as the solution for the inclusion undergoing arbitrary rigid body displacement $\mathbf{u}_P^{(1)}=\mathbf{u}_p^{(1)}+\boldsymbol{\theta}_p^{(1)}\wedge\mathbf{r}$ with $\mathbf{f}^{(1)}=0$. Let field 2 be the solution for a point force $\mathbf{F}$ at $\mathbf{s}$:

$$\mathbf{f}^{(2)}(\mathbf{x})=\mathbf{F}\delta(\mathbf{x}-\mathbf{s}),\quad\mathbf{s}\in V. \tag{18}$$

The solution to this, $\mathbf{u}^{(2)}$, is in fact the Green's function in the presence of the movable inclusion. Our objective is to avoid explicit calculation of the Green's function.

The displacement on the inclusion surface is again a rigid body displacement, $\mathbf{u}_P^{(2)}=\mathbf{u}_p^{(2)}+\boldsymbol{\theta}_p^{(2)}\wedge\mathbf{r}$, and therefore the reciprocity identity (13) becomes

$$\mathbf{u}_p^{(1)}\cdot\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(2)}+\boldsymbol{\theta}_p^{(1)}\cdot\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(2)}-\mathbf{F}\cdot\mathbf{u}^{(1)}(\mathbf{s})$$
$$=\mathbf{u}_p^{(2)}\cdot\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(1)}+\boldsymbol{\theta}_p^{(2)}\cdot\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(1)}. \tag{19}$$

The integrals involving $\boldsymbol{\tau}^{(1)}$ again give resultant force and moment according to Eqs. (15) with $j=1$. For field 2, let $\mathbf{F}_p$ and $\mathbf{M}_p$ denote the resultants caused by the point force at $\mathbf{s}$,

$$\int_{\partial V_p}\mathrm{d}s\,\boldsymbol{\tau}^{(2)}=\mathbf{F}_p, \tag{20a}$$

$$\int_{\partial V_p}\mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}^{(2)}=\mathbf{M}_p. \tag{20b}$$

The displacement at $\mathbf{s}$ for field 1 follows from the definition of the tensors $\mathbf{U}$ and $\mathbf{W}$ as $\mathbf{u}^{(1)}(\mathbf{s})=\mathbf{U}(\mathbf{s})\mathbf{u}_p^{(1)}+\mathbf{W}(\mathbf{s})\boldsymbol{\theta}_p^{(1)}$, see Eq. (3).

Elimination of these quantities from Eq. (19) implies

$$\mathbf{F}_p\cdot\mathbf{u}_p^{(1)}+\mathbf{M}_p\cdot\boldsymbol{\theta}_p^{(1)}-\mathbf{F}\cdot\mathbf{U}(\mathbf{s})\mathbf{u}_p^{(1)}-\mathbf{F}\cdot\mathbf{W}(\mathbf{s})\boldsymbol{\theta}_p^{(1)}$$
$$=-i\omega\mathbf{u}_p^{(2)}\cdot\mathbf{Z}\mathbf{u}_p^{(1)}-i\omega\boldsymbol{\theta}_p^{(2)}\cdot\tilde{\mathbf{Z}}\boldsymbol{\theta}_p^{(1)}. \tag{21}$$

But the rigid body displacement $\mathbf{u}_p^{(1)}$ and twist $\boldsymbol{\theta}_p^{(1)}$ are arbitrary, and using the symmetry of the impedances, we deduce

$$\mathbf{F}_p=\mathbf{U}^t(\mathbf{s})\mathbf{F}-i\omega\mathbf{Z}\mathbf{u}_p^{(2)}, \tag{22a}$$

$$\mathbf{M}_p=\mathbf{W}^t(\mathbf{s})\mathbf{F}-i\omega\tilde{\mathbf{Z}}\boldsymbol{\theta}_p^{(2)}. \tag{22b}$$

A second set of independent relations follows from the equilibrium of the inclusion, or Newton's second law applied to a rigid body,

$$\mathbf{F}_p=i\omega\mathbf{Z}_P\mathbf{u}_p^{(2)}, \tag{23a}$$

$$\mathbf{M}_p=i\omega\tilde{\mathbf{Z}}_P\boldsymbol{\theta}_p^{(2)}. \tag{23b}$$

Eliminating the linear displacement $\mathbf{u}_p^{(2)}$ and twist $\boldsymbol{\theta}_p^{(2)}$ between Eqs. (22) and (23) gives

$$\mathbf{F}_p=\mathbf{Z}_P(\mathbf{Z}+\mathbf{Z}_P)^{-1}\mathbf{U}^t(\mathbf{s})\mathbf{F}, \tag{24a}$$

$$\mathbf{M}_p=\tilde{\mathbf{Z}}_P(\tilde{\mathbf{Z}}+\tilde{\mathbf{Z}}_P)^{-1}\mathbf{W}^t(\mathbf{s})\mathbf{F}. \tag{24b}$$

Finally, referring back to the definition of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ in Eq. (4) implies the desired relations.

Lemma 2 *The displacement and force tensors are related by*

$$\boldsymbol{\Phi}(\mathbf{s})=\mathbf{Z}_P(\mathbf{Z}+\mathbf{Z}_P)^{-1}\mathbf{U}^t(\mathbf{s}). \tag{25}$$

The rotation and moment tensors are related by

$$\boldsymbol{\Psi}(\mathbf{s})=\tilde{\mathbf{Z}}_P(\tilde{\mathbf{Z}}+\tilde{\mathbf{Z}}_P)^{-1}\mathbf{W}^t(\mathbf{s}). \tag{26}$$

We are now ready to examine these quantities for a particular case, the spherical inclusion.

IV. SPHERICAL INCLUSION, ISOTROPIC MATRIX

A. Definition of the problem

The inclusion has radius a and is embedded in a uniform isotropic elastic medium of infinite extent with mass density ρ and Lamé moduli λ and μ. The interface conditions at $r=a$ are: (i) continuity of normal displacement, (ii) satisfaction of a slip condition. The latter allows for relative tangential slip between the inclusion and matrix, and is defined by the tangential component of the traction $\boldsymbol{\tau}=\boldsymbol{\sigma}\hat{\mathbf{r}}$ where $\boldsymbol{\sigma}$ is the stress tensor and $\hat{\mathbf{r}}=r^{-1}\mathbf{r}$ denotes the unit radial vector. The tangential component satisfies

$$\boldsymbol{\tau}\cdot\hat{\mathbf{t}}=z_I(\mathbf{v}_P-\mathbf{v})\cdot\hat{\mathbf{t}},\quad r=a, \tag{27}$$

where $\hat{\mathbf{t}}$ is a unit tangent vector, $\mathbf{v}$ the velocity of the elastic medium adjacent to the sphere, $\mathbf{v}_P=-i\omega\mathbf{u}_P$ is the total velocity of the inclusion at the interface $r=a$, and z_I is an interfacial impedance, introduced in Ref. 6. This corresponds to $\mathbf{A}=(i\omega z_I)^{-1}(\mathbf{I}-\mathbf{n}\otimes\mathbf{n})$ in Eq. (12), where $\mathbf{n}$ is the interface normal. The results of Sec III therefore apply for this slip condition.

In summary, the conditions at the surface of the sphere are

$$\left.\begin{aligned}&\mathbf{u}\cdot\hat{\mathbf{r}}=\mathbf{u}_P\cdot\hat{\mathbf{r}}\\&\boldsymbol{\tau}\cdot\hat{\mathbf{t}}=i\omega z_I(\mathbf{u}-\mathbf{u}_P)\cdot\hat{\mathbf{t}}\end{aligned}\right\}\quad r=a. \tag{28}$$

B. External impedances

Symmetry arguments imply that the net force (moment) exerted on the sphere by the surrounding medium and the resulting linear displacement (axis of rotation) are parallel. Hence, the external impedances are isotropic,

$$\mathbf{Z}=Z\mathbf{I},\quad \tilde{\mathbf{Z}}=\tilde{Z}\mathbf{I}. \tag{29}$$

The linear impedance Z has been considered previously[6] while the rotational impedance $\tilde{Z}$ is new. Expressions for both are given next.

1. Linear impedance

The scalar Z can be expressed in a form reminiscent of lumped mass systems[6]

$$\frac{3}{Z+Z_M}=\frac{1}{Z_L+Z_M}+\frac{2}{Z_S+Z_M}, \tag{30}$$

where the additional impedances are

$$Z_M=i\omega\frac{4}{3}\pi a^3\rho, \tag{31a}$$

$$Z_L=(i\omega)^{-1}4\pi a(\lambda+2\mu)(1-ika), \tag{31b}$$

$$Z_T=(i\omega)^{-1}4\pi a\mu(1-iha), \tag{31c}$$

$$\frac{1}{Z_S}=\frac{1}{Z_T}+\frac{1}{4\pi a^2 z_I+(i\omega)^{-1}8\pi a\mu}. \tag{31d}$$

Here k and h are, respectively, the longitudinal and transverse wave numbers, $k=\omega/c_L$, $h=\omega/c_T$ with $c_L=\sqrt{(\lambda+2\mu)/\rho}$ and $c_T=\sqrt{\mu/\rho}$. The impedances in Eqs. (31) depend upon and are defined by the matrix properties, except for Z_S, which involves the interface viscosity term z_I. Thus, Z_M is the mass-like impedance of a sphere of the matrix material of the same size as the inclusion. Note that $Z_S=Z_T$ if the inclusion is perfectly bonded to the matrix ($z_I\to\infty$). See Ref. 6 for further discussion of this and other limits.

2. Rotational impedance

The rotational impedance of a spherical inclusion has not, to our knowledge, been presented in the literature. A derivation is given in Appendix A, with the result that

$$\frac{a^2}{\tilde{Z}}=\frac{3}{8\pi a^2 z_I}+\frac{\frac{1}{2}(1-iha)}{Z_M+Z_T}. \tag{32}$$

The parameters in this identity were defined previously.

C. Displacement, rotation, force, and moment tensors

1. Internal impedance

The internal impedances $\mathbf{Z}_P$ and $\tilde{\mathbf{Z}}_P$ are necessary in order to relate the displacement/rotation tensors with the force/moment tensors via Lemma 2. For the sake of simplicity we restrict consideration in this paper to internal impedances that are isotropic:

$$\mathbf{Z}_P=Z_P\mathbf{I},\quad \tilde{\mathbf{Z}}_P=\tilde{Z}_P\mathbf{I}. \tag{33}$$

For instance, a uniformly solid sphere of mass m has

$$Z_P=i\omega m,\quad \tilde{Z}_P=i\omega\frac{2}{5}a^2m. \tag{34}$$

2. Linear motion

The displacement tensor $\mathbf{U}$ of Eq. (3) is derived in Appendix B as

$$\mathbf{U}(\mathbf{r})=a(Z+Z_M)\left[\frac{-1}{Z_L+Z_M}\nabla\nabla\frac{e^{ik(r-a)}}{k^2r}+\frac{1}{Z_S+Z_M}\frac{Z_S}{Z_T}(\nabla\nabla+h^2\mathbf{I})\frac{e^{ih(r-a)}}{h^2r}\right],\quad r\geqslant a. \tag{35}$$

The force tensor $\boldsymbol{\Phi}$ of Eq. (4) follows from Lemma 2 and the fact that the impedances satisfy Eqs. (29) and (33). Thus,

$$\boldsymbol{\Phi}(\mathbf{r})=\frac{Z_P}{Z+Z_P}\mathbf{U}(\mathbf{r}). \tag{36}$$

The displacement and force tensors satisfy $\mathbf{U}(-\mathbf{r})=\mathbf{U}(\mathbf{r})$, $\boldsymbol{\Phi}(-\mathbf{r})=\boldsymbol{\Phi}(\mathbf{r})$ and are symmetric, $\mathbf{U}=\mathbf{U}^t$, $\boldsymbol{\Phi}=\boldsymbol{\Phi}^t$. We focus on the properties of $\mathbf{U}$ since those of $\boldsymbol{\Phi}$ are easily obtained through Eq. (36).

Equation (35) implies

$$\mathbf{U}=\frac{Z+Z_M}{Z_L+Z_M}\left[\frac{h_1(kr)}{krh_0(ka)}\mathbf{I}-\frac{h_2(kr)}{h_0(ka)}\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}\right]+\frac{Z+Z_M}{Z_S+Z_M}\frac{Z_S}{Z_T}\left(\left[\frac{h_0(hr)}{h_0(ha)}-\frac{h_1(hr)}{hrh_0(ha)}\right]\times\mathbf{I}+\frac{h_2(hr)}{h_0(ha)}\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}\right), \tag{37}$$

where h_n are spherical Hankel functions of the first kind[9] and $\hat{\mathbf{r}}=r^{-1}\mathbf{r}$ denotes the unit radial vector. In particular $h_0(z)=(iz)^{-1}e^{iz}$. In expanded form,

$$\mathbf{U}(\mathbf{r})=(Z+Z_M)\frac{a}{r}\left[\frac{1}{Z_L+Z_M}\left[\left(\frac{1}{(kr)^2}-\frac{i}{kr}\right)\mathbf{I}+\left(1+\frac{3i}{kr}-\frac{3}{(kr)^2}\right)\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}\right]e^{ik(r-a)}+\frac{1}{Z_S+Z_M}\frac{Z_S}{Z_T}\left[\left(1+\frac{i}{hr}-\frac{1}{(hr)^2}\right)\mathbf{I}-\left(1+\frac{3i}{hr}-\frac{3}{(hr)^2}\right)\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}\right]e^{ih(r-a)}\right]. \tag{38}$$

3. Rotational motion

The skew tensor $\mathbf{W}$ relating the rotation to the displacement at a distance follows from Appendix A as

$$\mathbf{W}(\mathbf{r})=-a\Omega\frac{h_1(hr)}{h_1(ha)}\mathrm{axt}(\hat{\mathbf{r}}),\quad \Omega=1-\frac{3\tilde{Z}}{8\pi a^4 z_I}. \tag{39}$$

Then $\boldsymbol{\Psi}$, which relates the moment on the inclusion to an applied force at a distance, is

$$\boldsymbol{\Psi}(\mathbf{r})=\frac{-\tilde{Z}_P}{\tilde{Z}+\tilde{Z}_P}\mathbf{W}(\mathbf{r}). \tag{40}$$

The rotational tensors are odd functions of their arguments, $\mathbf{W}(-\mathbf{r})=-\mathbf{W}(\mathbf{r})$, $\boldsymbol{\Psi}(-\mathbf{r})=-\boldsymbol{\Psi}(\mathbf{r})$, and are skew symmetric, $\mathbf{W}=-\mathbf{W}^t$, $\boldsymbol{\Psi}=-\boldsymbol{\Psi}^t$.

V. APPLICATIONS

This section explores implications of the general theory to the particular case of the spherical inclusion.

A. Force on a particle from plane wave incidence

The force on a particle due to a remote point load is given directly by the tensor $\Phi(\mathbf{r})$. Taking the source point to infinity the effect of the excitation on the particle is equivalent to an incident plane wave, or a combination of two incident plane waves. The far-field form of $\Phi(\mathbf{r})$ follows from Eqs. (36) and (38) as

$$\boldsymbol{\Phi}(\mathbf{r})=Z_P\left(\frac{Z+Z_M}{Z+Z_P}\right)\frac{a}{r}\left[\frac{e^{ik(r-a)}}{Z_L+Z_M}\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}+\frac{e^{ih(r-a)}}{Z_S+Z_M}\frac{Z_S}{Z_T}(\mathbf{I}-\hat{\mathbf{r}}\otimes\hat{\mathbf{r}})\right]+\mathrm{O}(r^{-2}). \tag{41}$$

At the same time, the far-field free space Green's function is (see Eq. (61)),

$$\mathbf{G}^{(0)}(\mathbf{r})=\frac{1}{4\pi\mu r}[\kappa^{-2}e^{ikr}\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}+e^{ihr}(\mathbf{I}-\hat{\mathbf{r}}\otimes\hat{\mathbf{r}})]+\mathrm{O}(r^{-2}). \tag{42}$$

Consider, for instance, a unit point force in the far field at $\mathbf{r}$ in the direction $\mathbf{n}=-\hat{\mathbf{r}}$. This produces a longitudinal plane wave at the origin of the form $\mathbf{u}=u_0\mathbf{n}e^{ik\mathbf{n}\cdot\mathbf{x}}$ where $u_0=e^{ikr}/(4\pi\mu\kappa^2 r)$. The force on the spherical particle due to an incident longitudinal plane wave

$$\mathbf{u}(\mathbf{x})=e^{ik\mathbf{n}\cdot\mathbf{x}}\mathbf{u}_0,\quad \mathbf{u}_0\wedge\mathbf{n}=0 \tag{43}$$

is therefore

$$\mathbf{F}_p=(\lambda+2\mu)\frac{4\pi aZ_P}{Z+Z_P}\left(\frac{Z+Z_M}{Z_L+Z_M}\right)e^{-ika}\mathbf{u}_0. \tag{44}$$

In the same manner, the force on the spherical particle due to an incident transverse plane wave

$$\mathbf{u}(\mathbf{x})=e^{ih\mathbf{n}\cdot\mathbf{x}}\mathbf{u}_0,\quad \mathbf{u}_0\cdot\mathbf{n}=0 \tag{45}$$

is

$$\mathbf{F}_p=\mu\frac{4\pi aZ_P}{Z+Z_P}\frac{Z_S}{Z_T}\left(\frac{Z+Z_M}{Z_S+Z_M}\right)e^{-iha}\mathbf{u}_0. \tag{46}$$

The values of the plane wave induced forces for the rigid immovable particle follow from Eqs. (44) and (46) in the limit as $Z_P\to\infty$. These values actually coincide in the static limit, as discussed below after we consider the quasistatic limit of Z.

Davis and Nagem[10] considered plane wave incidence on an elastic sphere in a compressible viscous fluid, with specific results focused on the rigid immovable limit. This is equivalent to an isotropic elastic medium with shear modulus $\mu=-i\omega\rho\nu_0$, where ν_0 is the kinematic viscosity, and with a viscously damped longitudinal wave. Their expression for the force on the rigid sphere under acoustic plane wave incidence (Eqs. (30) and (31) of Ref. 10) should agree with Eq. (44) in the rigid limit.

B. Moment on a particle from a plane wave

The far-field form of the moment tensor is, from Eqs. (39) and (40),

$$\boldsymbol{\Psi}(\mathbf{r})=\frac{a^2\Omega}{1-(iha)^{-1}}\frac{e^{ih(r-a)}}{r}\frac{\tilde{Z}_P}{\tilde{Z}+\tilde{Z}_P}\mathrm{axt}(\hat{\mathbf{r}}). \tag{47}$$

Based on the discussion for the forcing from plane wave incidence, it is evident that a longitudinal wave produces zero net moment on the spherical particle. A transverse plane wave, does however, exert a moment. It may be shown that the plane wave (45) produces

$$\mathbf{M}_p=\frac{4\pi a^2\mu}{1-(iha)^{-1}}\frac{\Omega\tilde{Z}_P}{\tilde{Z}+\tilde{Z}_P}e^{-iha}\mathbf{u}_0\wedge\mathbf{n}. \tag{48}$$

The rigid and quasistatic limits are discussed below.

C. Rigid body displacement due to a plane wave

The particle displacement under plane wave incidence is a combination of a linear displacement and a rigid body rotation. These follow, respectively, from the forcing $\mathbf{F}_p$ of Eqs. (44) or (46) and the moment $\mathbf{M}_p$ of Eq. (48) as

$$\mathbf{u}_P = (i\omega Z_P)^{-1}\mathbf{F}_p + (i\omega \tilde{Z}_P)^{-1}\mathbf{M}_p \wedge \mathbf{r}, \quad r \leqslant a. \tag{49}$$

Symmetry dictates that the moment tensor $\mathbf{M}_p$ is zero for longitudinal incidence.

D. Quasistatic limit

1. Linear motion: Virtual mass

The quasistatic limit of vanishingly small frequency ($\omega \to 0$) yields

$$Z = \frac{12\pi a\mu}{i\omega}\left[\frac{1}{2+\chi+\kappa^{-2}} - \frac{iha(2+\kappa^{-3})}{(2+\chi+\kappa^{-2})^2} - \frac{C_v}{9}(ha)^2 + \mathrm{O}(h^3a^3)\right], \quad |ha|, \quad |ka| \ll 1, \tag{50}$$

where

$$\kappa = \frac{c_L}{c_T} = \sqrt{\frac{2(1-\nu)}{1-2\nu}}, \tag{51}$$

and ν is the Poisson's ratio. The nondimensional factor χ is related to the interface impedance z_I in this limit, and is chosen so that it takes on the values zero or unity in the limit that the sphere is either perfectly bonded or perfectly lubricated

$$\chi = \begin{cases} 0, & \text{no slip}, \quad z_I \to \infty, \\ 1, & \text{slip}, \quad z_I = 0. \end{cases} \tag{52}$$

The parameter C_v is

$$C_v = 9\frac{(2+\kappa^{-3})^2}{(2+\chi+\kappa^{-2})^3} - 6\frac{\left(2-\frac{5}{4}x+\kappa^{-4}\right)}{(2+\chi+\kappa^{-2})^2} - 1. \tag{53}$$

The expansion (50) goes further than in Ref. 6 (Eq. (30)) which did not contain the C_v term. If the low frequency expansion is of the form $Z = Z^{(-1)}(i\omega)^{-1} + Z^{(0)} + Z^{(1)}(i\omega) + \ldots$ then the coefficient $Z^{(1)}$ determines the extra inertia or added mass caused by the linear motion of the infinite matrix. The virtual mass coefficient is defined as $Z^{(1)}/Z_M$, and is therefore C_v of Eq. (53). As shown in Fig. 4, the coefficient is positive under no slip conditions for all permissible values of Poisson's ratio. It approaches the limiting value of $C_v = 1/2$ in the limit of incompressibility, $\nu \to 1/2$, in agreement with the value for viscous fluids.[11] In contrast, the virtual mass coefficient is always negative when the inclusion is permitted to slip, and is always less than the incompressible limiting value of $C_v = -1/6$.

2. Linear motion: Static displacement

Based on Eq. (50) we have

$$\mathbf{U} = \frac{3a}{2+\chi+\kappa^{-2}}\left[\frac{1}{r}\mathbf{I} + \lim_{h\to 0}\nabla\nabla\frac{1}{h^2 r}\left(\frac{e^{ih(r-a)}}{1-iha-\frac{1}{3}h^2a^2} - \frac{e^{i\kappa^{-1}h(r-a)}}{1-i\kappa^{-1}ha-\frac{1}{3}\kappa^{-2}h^2a^2}\right)\right] = \frac{3a}{2+\chi+\kappa^{-2}}\left[\frac{1}{r}\mathbf{I} + \frac{1}{2}(\kappa^{-2}-1)\nabla\nabla\left(r+\frac{a^2}{3r}\right)\right]. \tag{54}$$

Evaluating the gradients and removing κ in favor of ν gives

$$\mathbf{U}(\mathbf{r}) = \frac{3a}{2r[5-6\nu+2\chi(1-\nu)]}\left[\left(3-4\nu+\frac{a^2}{3r^2}\right)\mathbf{I} + \left(1-\frac{a^2}{r^2}\right)\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}\right], \quad \text{static and immovable}. \tag{55}$$

This can be compared with Walpole's[12] result for the static perfectly bonded immovable spherical inclusion, Eq. (3.21) of Ref. 12. Walpole considered the force tensor $\mathbf{\Phi}$, which as we have seen is equal to $\mathbf{U}^t$ in the limit of a fixed and immovable rigid inclusion, i.e., $Z_P \to \infty$. But $\mathbf{U}$ is symmetric for the sphere, and therefore Eq. (55) represents both $\mathbf{U}$ and $\mathbf{\Phi}$. When $\chi = 0$ this agrees with Walpole. The static result for $\chi = 1$ appears to be new.

A more precise definition of the fixed inclusion limit is that $Z_P/Z \to \infty$. At the same time we are taking the static limit, so the simultaneous static and immovable limit is

$$\lim_{\omega\to 0}\frac{a\mu}{\omega Z_P} = 0. \tag{56}$$

3. Rotational motion: Virtual mass

In the limit of low frequency the rotational impedance $\tilde{Z}$ of Eq. (32) approximates as

$$\tilde{Z} = \begin{cases} \dfrac{8\pi a^3\mu}{i\omega} + i\omega\dfrac{8}{3}\pi a^5\rho + \mathrm{O}(\omega^2), & \text{no slip}, \\ 0, & \text{slip}, \end{cases} \tag{57}$$

where slip and no slip correspond to the limits $z_I = 0$ and $z_I \to \infty$, respectively. The term $i\omega(8/3)\pi a^5\rho = 2a^2 Z_M$ can be identified as the virtual mass due to the rotating solid. The internal rotational impedance of a solid sphere is $\tilde{Z}_P = (2/5)a^2 Z_P$. Hence the virtual mass in rotation is five times the mass of solid matrix material in the volume of the sphere, in agreement with a similar result for Stokes flow.[11]

4. Quasistatic plane wave force on an immovable particle

The forcing on the particle from plane wave incidence is the same, whether longitudinal or transverse waves are incident, in the quasistatic limit for a rigid immovable sphere. It may be checked that both Eqs. (44) and (46) become

$$\mathbf{F}_p = \frac{12\pi a\mu}{2+\chi+\kappa^{-2}}\mathbf{u}_0. \tag{58}$$

This apparently strange result may be reconciled with the physical nature of the limit: the matrix moves by the static displacement $\mathbf{u}_0$, while the particle is stationary. It is therefore sensible that there is a net force on the particle, and that it is in the direction of $\mathbf{u}_0$. This also explains the identical form of the limit for both types of wave incidence. The static limit is one of the unusual features of the immovable sphere. For further discussion see Pao,[13] who quite properly questions the physical validity of the rigid fixed assumption. Among its failings, as Pao notes, this configuration does not display Rayleigh scattering behavior at low frequencies.

5. Quasistatic plane wave moment on a fixed sphere

The moment on the rigid immovable spherical particle has a particularly simple form,

$$\mathbf{M}_p = \frac{-iha}{1-iha}4\pi a^2\mu\Omega e^{-iha}\mathbf{u}_0\wedge\mathbf{n}, \quad \text{immovable}. \tag{59}$$

This is valid at all frequencies, but as $\omega\to 0$ it vanishes, unlike the force on the particle in the same limit.

E. Small inclusion limit

1. Linear motion

As $a\to 0$ we have Z, Z_L, Z_S, $Z_T=\mathrm{O}(a)$, while Z_M, $Z_P=\mathrm{o}(a)$. It may be easily verified that $\mathbf{U}$ and $\mathbf{\Phi}$ reduce in this limit to

$$\mathbf{U}(\mathbf{r}) = i\omega Z\mathbf{G}^{(0)}(\mathbf{r}), \quad \mathbf{\Phi}(\mathbf{r}) = i\omega Z_P\mathbf{G}^{(0)}(\mathbf{r}), \tag{60}$$

where $\mathbf{G}^{(0)}$ is the free space Green's tensor

$$\mathbf{G}^{(0)}(\mathbf{r}) = \frac{1}{4\pi\mu}\left[\frac{e^{ihr}}{r}\mathbf{I}+\nabla\nabla\left(\frac{1}{h^2r}(e^{ihr}-e^{ikr})\right)\right]. \tag{61}$$

In hindsight, the form of $\mathbf{\Phi}$ is obvious based on the dynamic equilibrium of the inclusion: $i\omega Z_P\mathbf{u}_p=\mathbf{F}_p$ with $\mathbf{u}_p=\mathbf{G}^{(0)}\mathbf{F}$, $\mathbf{F}_p=\mathbf{\Phi}F$, and $\mathbf{U}$ then follows from Eq. (36).

2. Rotational motion

The rotational quantities $\mathbf{W}$ and $\mathbf{\Psi}$, on the other hand, become negligible in the small inclusion limit. This follows from the scaling $\mathbf{W}=\mathrm{O}(a)$ in Eq. (39).

F. Surface displacement and traction

1. Linear motion

The displacement tensor $\mathbf{U}$, which is defined in the exterior region, reduces to the following on the interface:

$$\mathbf{U}(a\hat{\mathbf{r}}) = \mathbf{I}+\left(\frac{Z_S-Z_T}{Z_T}\right)\left(\frac{Z+Z_M}{Z_S+Z_M}\right)(\mathbf{I}-\hat{\mathbf{r}}\otimes\hat{\mathbf{r}}). \tag{62}$$

This becomes the identity under no-slip conditions, since then $Z_S-Z_T=0$. Alternatively, the interface conditions (28) can be written

$$\mathbf{u} = \mathbf{u}_p+(i\omega z_I)^{-1}(\hat{\mathbf{t}}\otimes\hat{\mathbf{t}})\boldsymbol{\tau}, \quad r=a. \tag{63}$$

The final term on the right hand side vanishes as $z_I\to\infty$, which is the no-slip limit.

Substituting $\mathbf{u}=\mathbf{U}(a\hat{\mathbf{r}})\mathbf{u}_p$ on $r=a$, in Eq. (63) provides an explicit expression for the interfacial shear traction in terms of the linear displacement $\mathbf{u}_p$,

$$\boldsymbol{\tau}\cdot\hat{\mathbf{t}} = i\omega z_I\left(\frac{Z_S-Z_T}{Z_T}\right)\left(\frac{Z+Z_M}{Z_S+Z_M}\right)\mathbf{u}_p\cdot\hat{\mathbf{t}}. \tag{64}$$

The shear traction vanishes under pure-slip conditions ($z_I=0$), and for a bonded interface it becomes

$$\boldsymbol{\tau}\cdot\hat{\mathbf{t}} = -i\omega\mathbf{u}_p\cdot\hat{\mathbf{t}}\frac{Z_T}{4\pi a^2}\left(\frac{Z+Z_M}{Z_T+Z_M}\right). \tag{65}$$

2. Rotational motion

The displacement on $r=a$ is

$$\mathbf{W}(a\hat{\mathbf{r}}) = -a\Omega\mathrm{axt}(\hat{\mathbf{r}}), \tag{66}$$

where Ω, given in Eq. (39), reduces to unity under no-slip conditions. Conversely, $\Omega=0$ for pure slip, indicating that the solid does not move even as the inclusion rotates.

In this case the traction is pure shear, and

$$\boldsymbol{\tau} = -i\omega\frac{3\tilde{Z}}{8\pi a^3}\boldsymbol{\theta}_p\wedge\hat{\mathbf{r}}, \quad r=a. \tag{67}$$

This is nonzero except under pure-slip conditions, when $\tilde{Z}\to 0$, and there is no rotational interaction between the inclusion and the matrix.

G. Acoustic limit

1. General formulation

Finally, we discuss how the general elastodynamic formulation reduces when the matrix is an acoustic fluid. In this limit shear effects are ignorable and the medium is characterized by density ρ and bulk modulus $K=\rho c^2$, where c is the acoustic wave speed. Taking the displacement $\mathbf{u}$ and pressure p as field variables, the momentum balance and constitutive law are, respectively,

$$\omega^2\rho\mathbf{u} = \nabla p, \quad p = -K\nabla\cdot\mathbf{u}. \tag{68}$$

The acoustic wave number is $k=\omega/c$.

We introduce two vector functions $\mathbf{q}(\mathbf{r})$ and $\boldsymbol{\phi}(\mathbf{s})$ that are analogous to the tensors $\mathbf{U}$ and $\mathbf{\Phi}$. If the inclusion is moved back and forth with the displacement $\mathbf{u}_p$ then the condition on the inclusion surface is that the normal velocity is continuous,

$$\mathbf{u}\cdot\mathbf{n} = \mathbf{u}_p\cdot\mathbf{n} \quad \text{on } \partial V_p. \tag{69}$$

The pressure at a point $\mathbf{r}$ away from inclusion is defined by $\mathbf{q}$ as

$$p(\mathbf{r}) = \mathbf{q}(\mathbf{r})\cdot\mathbf{u}_p\ \mathbf{r} \text{ in } V. \tag{70}$$

Conversely, consider a voluminal source at $\mathbf{s}$:

$$\nabla^2 p + k^2 p = f\delta(\mathbf{x}-\mathbf{s}). \tag{71}$$

The force on the inclusion is

$$\mathbf{F}_p = -\int_{\partial V_p} \mathrm{d}s p\mathbf{n} \equiv f\boldsymbol{\phi}(\mathbf{s}), \tag{72}$$

which defines the vector function $\boldsymbol{\phi}$.

The connection between $\mathbf{q}$ and $\boldsymbol{\phi}$ is given by

Lemma 3. *The acoustic displacement and force vectors are related by*

$$\boldsymbol{\phi}(\mathbf{s}) = (\rho\omega^2)^{-1}\mathbf{Z}_P(\mathbf{Z}+\mathbf{Z}_P)^{-1}\mathbf{q}(\mathbf{s}). \tag{73}$$

This may be derived by application of reciprocity to the acoustic (Helmholtz) equation, in a manner similar to how we derived Lemma 2.

2. Spherical inclusion

Finally, we consider the example of the spherical inclusion. The vector $\mathbf{q}$ follows from the acoustic limit of the elastic result in Eq. (35),

$$\mathbf{q}(\mathbf{r}) = aK\left(\frac{Z+Z_M}{Z_A+Z_M}\right)\boldsymbol{\nabla}\frac{e^{ik(r-a)}}{r}, \quad r \geqslant a, \tag{74}$$

where Z_A, analogous to the longitudinal impedance Z_L in elasticity, is

$$Z_A = (i\omega)^{-1}4\pi aK(1-ika), \tag{75}$$

Z_M is as before, and the sphere impedance Z is now given by Eq. (30) with $Z_S=0$, which implies

$$\frac{1}{Z} = \frac{2}{Z_M} + \frac{3}{Z_A}. \tag{76}$$

VI. CONCLUSION

Starting from the notion of an inclusion with the six degrees of freedom of a rigid body, we introduced displacement/rotation and force/moment tensors relating the motion of the inclusion to the displacement and force at arbitrary exterior points. These can be considered as generalized Green's functions appropriate to the constrained nature of the inclusion. The general relations (7a) and (7b) between the displacement/rotation and force/moment tensors are one of the main contributions of the paper. These identities are extremely useful in providing a means by which one can consider the dynamic properties of particles embedded in a solid matrix.

Useful results have been obtained for the simplest but important configuration of a uniform spherical particle. The linear and rotational impedances, Z and $\tilde{Z}$, are given in Secs. IV B 1 and IV B 2, respectively, the latter for the first time. Explicit expressions are given in Eqs. (36)–(40) for the displacement tensors $\mathbf{U}$ and $\mathbf{W}$ and for the force tensors $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. Perhaps the most practical new results are Eqs. (44), (46), and (48) which provide simple formulas for the force and moment on a particle under plane wave incidence. The associated displacement of the particle is given by Eq. (49). These concise expressions resemble Faxén relations that are frequently used in microhydrodynamics.

Equation (50) extends the quasistatic expansion of Norris[6] to include the virtual mass coefficient, which can be negative if slip occurs. The quasistatic form of $\mathbf{U}$, Eq. (55), which relates the displacement of the inclusion to particle displacement in the matrix, generalizes a recent formula of Walpole.[12] The quasistatic limit of the plane wave force on a sphere, Eq. (58), is reminiscent of Stokes drag law, but is proportional to the displacement vector of the incident plane wave. It also includes the possibility of slip relative to the matrix. However, the moment on a rigid sphere from plane wave incidence is proportional to the incident particle velocity, Eq. (59), and vanishes in the limit of zero frequency. Other limiting cases considered include the small inclusion limit, and the purely acoustic limit.

Taken together the results of this paper offer a consistent means for analyzing wave-particle interaction in elasticity and viscoelasticity. Future applications will look at replacing the solid spherical particle with more complicated, and more interesting, internal structure. This amounts to considering more general forms of the internal impedances. The results developed here can also be used to develop simplified methods for scattering from particles. These issues will be examined in separate papers.

APPENDIX A: ROTATION OF A SPHERE

The sphere $r \leqslant a$ undergoes oscillatory rotation $\mathbf{u}_P = \boldsymbol{\theta}_p \wedge \mathbf{r}$. Let $\mathbf{e}$ be the axis of rotation, so that $\boldsymbol{\theta}_p = \theta_p \mathbf{e}$, and consider the possible solution $\mathbf{u} = \nabla \wedge \mathbf{e}f$ in the matrix $r > a$. This satisfies the equations of motion

$$\mathbf{u} + k^{-2}\nabla\nabla\cdot\mathbf{u} - h^{-2}\nabla\wedge\nabla\wedge\mathbf{u} = 0, \tag{A1}$$

provided that f is a solution of the reduced Helmholtz equation $\nabla^2 f + h^2 f = 0$. The function f must depend only on r in order to match the prescribed rotation on $r=a$. Hence,

$$\mathbf{u} = \beta h_1(hr)\mathbf{e}\wedge\hat{\mathbf{r}}, \quad r > a, \tag{A2}$$

where $\hat{\mathbf{r}} = \mathbf{r}/r$. The traction in an isotropic solid is[7]

$$\boldsymbol{\tau} = \hat{\mathbf{r}}\lambda\,\mathrm{div}\,\mathbf{u} + \frac{\mu}{r}\,\mathrm{grad}\,\mathbf{r}\cdot\mathbf{u} + \mu\left(\frac{\partial}{\partial r} - \frac{1}{r}\right)\mathbf{u}, \tag{A3}$$

from which we obtain

$$\boldsymbol{\tau} = \beta\mu h\left[h_0(ha) - \frac{3}{ha}h_1(ha)\right]\mathbf{e}\wedge\hat{\mathbf{r}}. \tag{A4}$$

The boundary conditions (28) therefore reduce to a single equation for the parameter β of Eq. (A2),

$$\beta\mu h\left[h_0(ha) - \frac{3}{ha}h_1(ha)\right] = i\omega z_l[\beta h_1(ha) - a\theta_p]. \tag{A5}$$

The moment $\mathbf{M}_p = \int \mathrm{d}s\mathbf{r}\wedge\boldsymbol{\tau}$ is obtained from the identity

$$\int_{r=a} \mathrm{d}s\mathbf{r}\wedge(\mathbf{e}\wedge\hat{\mathbf{r}}) = \frac{8}{3}\pi a^3\mathbf{e}, \tag{A6}$$

and the impedance then follows from the definition (6) as $\tilde{\mathbf{Z}} = \tilde{Z}\mathbf{I}$ where $\tilde{Z}$ is given by Eq. (32).

APPENDIX B: DISPLACEMENT OF A SPHERE

The solution to the radiation boundary value problem of the spherical inclusion undergoing linear displacement with boundary conditions defined by Eq. (28) has been solved by Oestreicher[7] for the case of no slip and more recently, by Norris,[6] with slip included. We follow the latter with some slight changes in notation. The solution is based on the following representation for the elastic field outside the sphere, $r \geqslant a$,

$$\mathbf{u} = -C_1 \mathbf{u}_p \cdot \nabla\nabla h_0(kr) + C_2 \mathbf{u}_p \cdot (\nabla\nabla - \mathbf{I}\nabla^2) h_0(hr), \quad \text{(B1)}$$

where $r=|\mathbf{r}|$ is the spherical radius and h_0 is the spherical Hankel function of the first kind, $h_0(z)=(iz)^{-1}e^{iz}$.

In the notation of Ref. 6, $C_1=-A_1/k^2$ and $C_2=3B_1/h^2$. Also, with reference to Norris,[6] the inclusion displacement is of magnitude u_0: $\mathbf{u}_p=u_0\hat{\mathbf{x}}$. Equations (15) and (19a) of Ref. 6 combined with Eq. (29), give (noting that Z_m of Ref. 6 is now Z_M)

$$h_2(ka)A_1 = \left(1 + \frac{Z}{Z_M}\right)u_0, \quad \text{(B2a)}$$

$$6\frac{h_1(ha)}{ha}B_1 = \frac{h_1(ka)}{ka}A_1 - \frac{Z}{Z_M}u_0. \quad \text{(B2b)}$$

Using $h_1(z)=-h_0'(z)$ and $h_2=-h_0+3z^{-1}h_1$ implies the identities

$$\frac{h_1(ka)}{kah_0(ka)} = -\frac{Z_L}{3Z_M}, \quad \frac{h_2(ka)}{h_0(ka)} = -\frac{Z_L+Z_M}{Z_M}. \quad \text{(B3)}$$

Similar identities for arguments ha instead of ka have Z_T instead of Z_L. Combining these results implies

$$C_1 = \left(\frac{Z+Z_M}{Z_L+Z_M}\right)\frac{1}{k^2h_0(ka)}, \quad \text{(B4a)}$$

$$C_2 = \left(\frac{Z+Z_M}{Z_S+Z_M}\right)\frac{Z_S}{Z_T}\frac{1}{h^2h_0(ha)}. \quad \text{(B4b)}$$

The expression (35) for $\mathbf{U}$ then follows.

[1] H. Faxén, "Simplified representation of the generalized Green's equations for the constant motion of translation of a rigid body in a viscous fluid," Arkiv för Matematik, Astronomi och Fysik **20**(8), 1–5 (1927).
[2] J. Happel and H. Brenner, *Low Reynolds Number Hydrodynamics* (Kluwer, Dordrecht, 1991).
[3] S. Kim and J. S. Karilla, *Microhydrodynamics: Principles and Selected Applications* (Butterworth-Heinemann, Boston, 1991).
[4] C. Pozrikidis, *Introduction to Theoretical and Computational Fluid Dynamics* (Oxford University Press, New York, 1997).
[5] N. Phan-Thien and S. Kim, *Microstructures in Elastic Media: Principles and Computational Methods* (Oxford University Press, New York, 1994).
[6] A. N. Norris, "Impedance of a sphere oscillating in an elastic medium with and without slip," J. Acoust. Soc. Am. **119**(4), 2062–2066 (2006).
[7] H. L. Oestreicher, "Field and impedance of an oscillating sphere in a viscoelastic medium with an application to biophysics," J. Acoust. Soc. Am. **23**, 707–714 (1951).
[8] J. D. Achenbach, *Reciprocity in Elastodynamics* (Cambridge University Press, Cambridge, UK, 2004).
[9] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1974).
[10] A. M. Davis and R. J. Nagem, "Curle's equation and acoustic scattering by a sphere," J. Acoust. Soc. Am. **119**(4), 2018–2026 (2006).
[11] A. A. Kendoush, "The virtual mass of a rotating sphere in fluids," ASME J. Appl. Mech. **72**(5), 801–802 (2005).
[12] L. J. Walpole, "The Green functions of an elastic medium surrounding a rigid spherical inclusion," Q. J. Mech. Appl. Math. **58**(1), 129–141 (2005).
[13] Y.-H. Pao and C.-C. Mow, *Diffraction of Elastic Waves and Dynamic Stress Concentrations* (Crane, Russak, New York, 1973).

Approximations of inverse boundary element methods with partial measurements of the pressure field

Nicolas P. Valdivia,[a] Earl G. Williams, and Peter C. Herdic
Code 7130, Naval Research Laboratory, Washington, DC 20375

(Received 29 January 2007; revised 24 October 2007; accepted 29 October 2007)

Boundary element methods (BEMs) based near-field acoustic holography (NAH) requires the measurement of the pressure field over a closed surface in order to recover the normal velocity on a nearby conformal surface. There are practical cases when measurements are available over a patch from the measurement surface in which conventional inverse BEM based NAH (IBEM) cannot be applied directly, but instead as an approximation. In this work two main approximations based on the indirect-implicit methods are considered: Patch IBEM and IBEM with Cauchy data. Patch IBEM can be applied with a continuation procedure, which as its predecessor patch NAH (a well known technique that can be used on separable geometries of the wave equation) continues the pressure field using an iterative procedure, or it can be applied by a direct procedure. On the other hand, IBEM with Cauchy data requires measurements over two conformal patches and it will be shown that this technique will be reliable regardless of the position of the source. The theory behind each method will be justified and validated using a cylindrical surface with numerical data generated by point sources, and using experimental data from a cylindrical fuselage excited by a point force. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816568]

PACS number(s): 43.40.Sk, 43.20.Ye, 43.20.Tb [SFW] Pages: 109–120

I. INTRODUCTION

Near-field acoustical holography (NAH)[1] is a technique that is used to reconstruct the acoustic field (pressure, normal velocity and intensity) on the vibrating surface of a radiating object from pressure measurements on a near concentric surface. This problem is ill posed because of the presence of evanescent waves that decay rapidly from the vibrating to the measurement surface. The measurements are typically made very close to the vibrating surface in order to obtain a high-resolution reconstruction. To treat the ill-posed nature of the problem, when the vibrating surface is planar, cylindrical or spherical (a separable geometry of the wave equation), Fourier decomposition or singular value decomposition are used with the series truncated to remove the very short wavelength components of the field that decay very rapidly in the direction normal to the surface. For an arbitrarily shaped vibrating surface, boundary element methods (BEMs)[2,3] may be applied to discretize the integral equations given by the direct[3,4] or indirect[5–7] representation of the acoustical pressure in order to obtain a relation between the measurements and the acoustic field over the vibrating surface in matrix form. The resultant matrix system is ill posed and requires the use of regularization methods[8] for its solution. The solution to BEM based NAH obtained using regularization is known as inverse BEM (IBEM).

In planar NAH it is required that the measurement plane be considerable larger than the vibrating planar surface. This requirement guarantees that the magnitude of the resultant measurements will decay sufficiently at the measurement plane edges, which yields accurate reconstructions.[1] To obtain accurate reconstructions in NAH for arbitrarily shaped vibrating surfaces it is required that the measurement surface surround the entire vibrating structure (for exterior NAH) or that the measurement surface be a closed surface interior to the vibrating surface (for interior NAH). For that reason large arbitrarily shaped vibrating surfaces require a prohibitive amount of measurements, and the processing is slow and memory intensive. However, there exist approximations that require measurements to be made only on a patch of the original surface. Reconstructions from these patches are restricted to the area directly below/above the measurement. In patch holograms it can not be assumed that the magnitude of the data at the boundaries decays at the edges, and the classical theory must be modified to avoid large aperture errors. A well-known method that addresses this problem for separable geometries of the wave equation is patch NAH.[9–11] This "edge effect" was overcome by an extension of the measurement patch by a process of iterative continuation of the acoustic pressure. Another approach to overcome this edge effect was recently introduced[12–14] called SONAH (statistically optimized near-field acoustical holography) for planar and cylindrical geometries.

For arbitrarily shaped vibrating surfaces Fourier acoustics and SONAH methods do not apply. One approach that avoids the finite aperture effects and can be applied to simple and complex shaped surfaces[15,16] is called Helmholtz equation least squares (HELS). This approach uses spherical harmonic expansions with truncated least-squares coefficients to overcome the Rayleigh hypothesis[17,18] and yield approximate reconstructions. A second very powerful approach is the equivalent source method[19,20] that like the HELS approach is not difficult to implement. Comparisons between the different approaches are, unfortunately, mostly found in conference proceedings[20–22] and the relative success of these meth-

[a]Author to whom correspondence should be addressed. Electronic mail: valdivia@pa.nrl.navy.mil

ods often depends upon the test cases used in the comparison. It is still unknown which methods are the most accurate and the most efficient.

In this paper we confine our attention to BEM methods. The technique of patch NAH can be directly applied using the indirect-implicit approach[7] to create a matrix system that relates the measurement and vibrating patch. In this work we propose two approaches for patch measurements over arbitrary shapes: one that works exactly like patch NAH (iterative continuation of the measurement) and a second that avoids the iterative continuation. We call the first method patch IBEM with continuation procedure and the second patch IBEM with direct procedure. Interior NAH for an arbitrarily shaped vibrating surface possesses the very important problem of back (opposite measurement side) source contamination, which affects the reconstruction of any patch-based technique. For that reason a third reconstruction approach that overcomes this problem is proposed. This approach uses a patch of Cauchy measurements, so it is called IBEM with Cauchy measurements.

We explain in Sec. III patch IBEM and IBEM with Cauchy measurements. Patch IBEM with continuation procedure uses an iterative continuation proposed by Saijyou and modified by Williams.[9,10] We also present a detailed mathematical proof of the convergence for this iterative procedure in the appendixes of this work. Section IV A uses numerically generated data to illustrate the reconstruction difficulties of the residual "edge effect" of patch IBEM. Section IV B shows the effect of the back source contamination in the patch IBEM approaches for interior problems and clarifies the need for the use of Cauchy measurements. Finally, Sec. V validates the proposed approaches using experimental data measured in the interior of a point-driven cylindrical fuselage section.

II. INTEGRAL EQUATION REPRESENTATION

Let G be a domain in $\mathbb{R}$,[3] interior to the closed boundary surface Γ where we assume that Γ is allowed to have edges and corners. Similarly we will denote as G^+ the region outside of G that shares the same boundary Γ. For a time-harmonic ($e^{-i\omega t}$) disturbance of frequency ω the sound pressure $p(\mathbf{x})$ located at $\mathbf{x}=(x_1,x_2,x_3)$ can be determined by the indirect formulation[7,23,24]

$$p(\mathbf{x}) = \int_\Gamma \Phi(\mathbf{x},\mathbf{y})\varphi(\mathbf{y})dS(\mathbf{y}), \tag{1}$$

with $\mathbf{y}=(y_1,y_2,y_3)$ and the free-space Green's function

$$\Phi(\mathbf{x},\mathbf{y}) = \frac{\exp(ik|\mathbf{x}-\mathbf{y}|)}{4\pi|\mathbf{x}-\mathbf{y}|}. \tag{2}$$

We have shown that this formulation is still valid at the so-called interior resonances[7] even though the solution is not unique and that the lack of uniqueness results in only a small increase in the reconstruction error. The normal velocity, derived from Eq. (1), when $\mathbf{x}\in\Gamma$ is given by

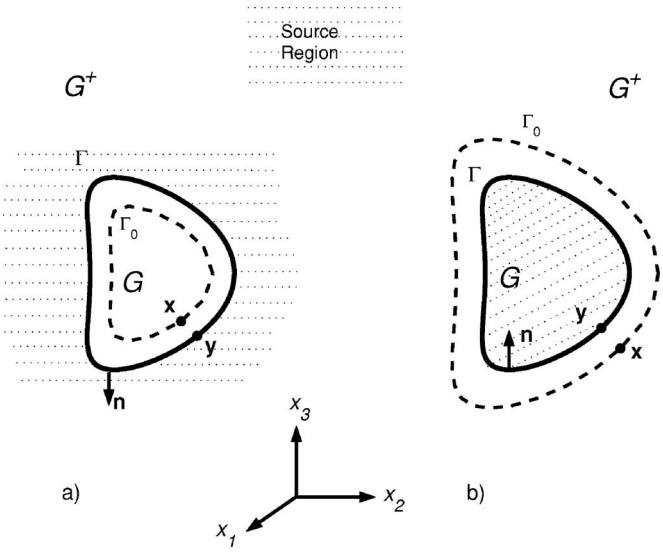


FIG. 1. Formulation for (a) interior NAH and (b) exterior NAH.

$$i\rho\omega v(\mathbf{x})^{\pm} = \int_\Gamma \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{x})}\varphi(\mathbf{y})dS(\mathbf{y}) \mp \Omega(\mathbf{x})\varphi(\mathbf{x}), \quad \mathbf{x}\in\Gamma, \tag{3}$$

where $\mathbf{n}$ is the unit normal with direction shown in Fig. 1 and $\Omega(\mathbf{x})$ is the solid angle coefficient. The superscript sign "+" in Eq. (3) is used for exterior NAH and "−" for interior NAH. This notation will be kept through the rest of this work.

For the numerical solution of NAH we use boundary element methods (BEMs).[2,3,7] In these methods the boundary surface Γ is decomposed into triangular elements with three nodes or quadrilateral elements with four nodes. In this paper, iso-parametric linear functions[7] are selected for interpolating the geometric and acoustical quantities. Given M pressure measurements on Γ_0 (see Fig. 1), represented as $\mathbf{p}$, recover N pressure and normal velocity points on Γ, represented as $\mathbf{p}^s$ and $\mathbf{v}^s$, respectively. When $\mathbf{x}\in\Gamma_0$, Eq. (1) gives the matrix equation

$$[\mathbf{S}]\boldsymbol{\varphi} = \mathbf{p}, \tag{4}$$

where $[\mathbf{S}]$ is a $M\times N$ complex matrix and $\boldsymbol{\varphi}$ is the column vector of N entries that represent values of the density φ on Γ. Similarly, when $\mathbf{x}\in\Gamma$, Eq. (1), and Eq. (3) produce the matrix equations

$$\mathbf{p}^s = [\mathbf{S}^s]\boldsymbol{\varphi},$$

$$\mathbf{v}^s = [\mathbf{K}^{\pm}]\boldsymbol{\varphi}, \tag{5}$$

where $[\mathbf{S}^s]$, $[\mathbf{K}^{\pm}]$ are $N\times N$ complex matrices.

Inverse BEM (IBEM) is known as the numerical solution of NAH using BEM. The implementation of IBEM by the implicit method with indirect formulation[7] requires us to solve Eq. (4) for the density $\boldsymbol{\varphi}$ (using special regularization methods) then use this density in Eq. (5) to produce $\mathbf{p}^s$ and $\mathbf{v}^s$.

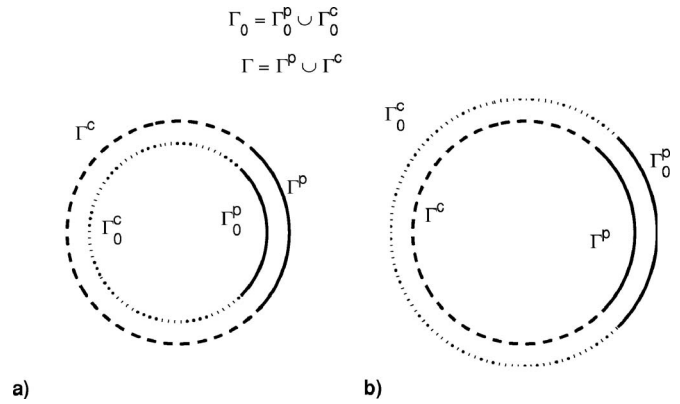


FIG. 2. Approximation setup for (a) interior NAH and (b) exterior NAH.

III. APPROXIMATIONS TO IBEM

In the conventional solution of NAH, the measurement surface Γ_0 is required to be closed and conformal to a nearby reconstruction surface Γ. There are some practical cases in which these requirements over the surface Γ_0 are prohibitive. For example, when the acoustic field needs to be reconstructed from measurements inside the cabin of a commercial airplane, where the physical dimensions of the cabin will require an excessive number of measurements.[25] In these cases instead of the conventional NAH solution we will rely on the approximation called Patch IBEMs that only use an array of measurements.

The setup for patch IBEM is shown in Fig. 2. We assume that from the measurement surface Γ_0 only a patch Γ_0^p is available. It will be expected, when Γ_0^p and Γ are close together, that the reconstruction of φ in Γ will be accurate only over Γ^p.

As observed in patch NAH, the reconstruction of the normal velocity from a patch of pressure measurements will result in large errors distributed over the edges of Γ^p. This problem in patch NAH has been avoided using special continuation iterations that can be easily implemented in patch IBEM for arbitrarily shaped surfaces. We also present a new direct procedure that does not require the continuation iterations, and finally an approach that requires two conformal arrays as in Fig. 4.

A. Patch IBEM with the continuation procedure

As seen in Fig. 3, we consider the extended measurement and reconstruction surfaces $\tilde{\Gamma}_0 = \Gamma_0^p \cup \Gamma_0^e$ and $\tilde{\Gamma} = \Gamma^p \cup \Gamma^e$, respectively. Given M points on $\tilde{\Gamma}_0$ and N points on $\tilde{\Gamma}$ we construct the $M \times N$ complex matrix $[\mathbf{S_e}]$ [as in Eq. (4)]. Then define the singular value decomposition (SVD)

$$[\mathbf{S_e}] = \mathbf{U}_e \underset{e}{\Sigma} \mathbf{V}_e^H,$$

where $\mathbf{U}_e$, $\mathbf{V}_e$ are, respectively, $M \times M$, $N \times N$ unitary matrices and Σ is a diagonal matrix containing the singular values σ_j, $j = 1, \ldots, \min(M, N)$ in decreasing order of magnitude. Here $\mathbf{V}_e^H$ is the conjugate transpose of $\mathbf{V}_e$.

We recall the algorithm for extending the measurement patch Γ_0^p into the surface $\tilde{\Gamma}_0$ given by Williams.[10] Given m pressure measurements represented by the column vector $\mathbf{p}$ spanning the surface Γ_0^p (where $m < M$), set the initial value on $\tilde{\Gamma}_0$ with the vector

$$\mathbf{p}^{(0)} = \begin{cases} \mathbf{p} & \text{on } \Gamma_0^p, \\ 0 & \text{on } \Gamma_0^e. \end{cases} \tag{6}$$

At each iteration step i the regularization parameter $\alpha(i)$ is chosen for $\mathbf{p}^{(i)}$ to obtain

$$\tilde{\mathbf{p}}^{(i)} = \mathbf{U}_e \mathbf{F}_{\alpha(i)} \mathbf{U}_e^H \mathbf{p}^{(i)}. \tag{7}$$

Here $\tilde{\mathbf{p}}^{(i)}$ is a filtered version of $\mathbf{p}^{(i)}$ where the filter $\mathbf{F}_\alpha$ is a diagonal matrix with the filter factors[8] in its diagonal entries. Then the iteration procedure is defined with the updated vector

$$\mathbf{p}^{(i+1)} = \begin{cases} \mathbf{p} & \text{on } \Gamma_0^p, \\ \tilde{\mathbf{p}}^{(i)} & \text{on } \Gamma_0^e. \end{cases} \tag{8}$$

The iteration will be stopped when

$$\|\tilde{\mathbf{p}}^{(i)} - \tilde{\mathbf{p}}^{(i-1)}\|_2 / \|\tilde{\mathbf{p}}^{(i-1)}\|_2 < \epsilon,$$

where $\epsilon > 0$ is a small number (usually $\epsilon = 10^{-5}$).

In this work the filter factors (diagonal entries of $\mathbf{F}_\alpha$) utilized are Tikhonov with highpass filter[26]

$$f_\alpha^j = \frac{\sigma_i^2}{\sigma_i^2 + \alpha \left(\dfrac{\alpha}{\alpha + \alpha_i^2} \right)^2}, \tag{9}$$

As in Sec. II A of the paper of Williams,[10] we estimate the noise variance σ_0 for $\mathbf{p}^{(0)}$. Notice that in this case σ_0 estimates the variance of the measurement noise in $\mathbf{p}$ and the jump between the data in Γ_0^p and 0 in the extension. At each iteration step (i) we choose α using Morozov discrepancy principle for

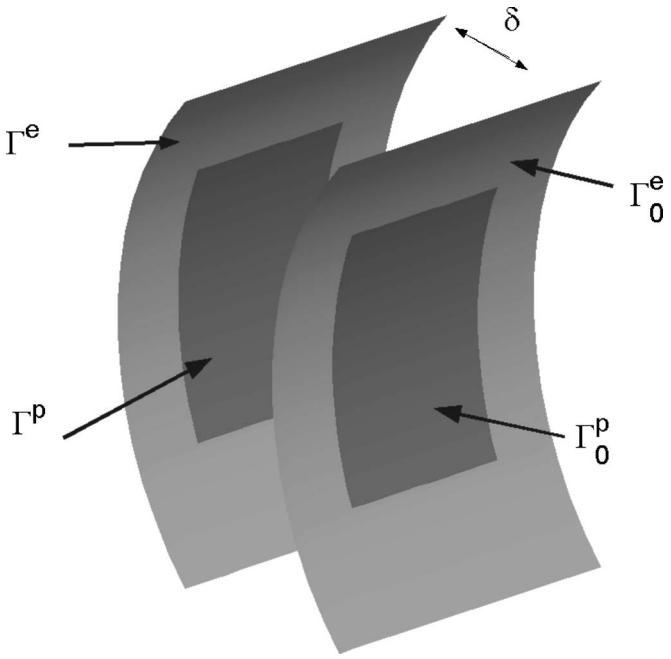


FIG. 3. Setup for the continuation procedures.

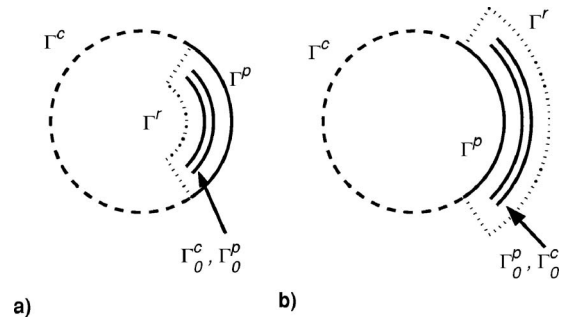


FIG. 4. Approximation setup with double conformal plane measurements for (a) interior NAH and (b) exterior NAH.

$$\|(1-\alpha)\mathbf{U}_e^H\mathbf{p}^{(i)}\| = \sigma_0\sqrt{M}.$$

In Appendix B we provide a mathematical proof of convergence for this iteration procedure. The proof of convergence is still valid if σ_0 is just an estimate of the variance of measurement noise in $\mathbf{p}$ by the use of a Hanning window.

Once the iterative continuation procedure converges for the i_0 iteration, we set

$$\boldsymbol{\phi} = \mathbf{V}_e\mathbf{F}_{\alpha(i_o)}\overset{-1}{\Sigma}_e\mathbf{U}_e^H\mathbf{p}^{(i_0)}, \tag{10}$$

where ϕ is a vector of N complex entries that approximates the density φ in Γ^e. Then the $N\times N$ complex matrices $[\mathbf{S}_\mathbf{p}^s]$, $[\mathbf{K}_\mathbf{p}^\pm]$ [restrictions in Γ^e of the matrices $[\mathbf{S}^s]$, $[\mathbf{K}^\pm]$ in Eq. (5)] are used to obtain

$$\tilde{\mathbf{p}}^i = [\mathbf{S}_\mathbf{p}^s]\boldsymbol{\phi}, \quad \tilde{\mathbf{v}}^s = [\mathbf{K}_\mathbf{p}^\pm]\boldsymbol{\phi}, \tag{11}$$

where $\tilde{\mathbf{p}}^s$, $\tilde{\mathbf{v}}^s$ are, respectively, the pressure and normal velocity in Γ^e.

B. Patch IBEM with the direct procedure

A major difference between the direct procedure and the continuation procedure just discussed is that the *measurement* aperture in not extended in the former. Only the *reconstruction* aperture is extended. Furthermore, no iteration scheme is used for the direct approach. It is quite remarkable that, as we will see in the Sec. IV, using no aperture extension for the measured data greatly decreases the aperture errors as long as the reconstruction aperture is extended.

For the direct procedure we will consider the given measurement patch Γ_0^p and the extended reconstruction surface $\tilde{\Gamma}=\Gamma^p\cup\Gamma^e$ (see Fig. 3). Given m points measurement points on Γ_0^p and N points on $\tilde{\Gamma}$ we construct the system

$$\mathbf{p} = [\mathbf{S}_\mathbf{p}]\boldsymbol{\phi}, \tag{12}$$

where $m\times N$ complex matrix $[\mathbf{S}_\mathbf{p}]$, $\mathbf{p}$ is the vector of measurements and ϕ the density vector of N complex entries.

The SVD

$$[\mathbf{S}_\mathbf{p}] = \mathbf{U}\Sigma\mathbf{V}^H,$$

where $\mathbf{U}$, $\mathbf{V}$ are, respectively, $m\times m$, $N\times N$ unitary matrices and Σ is a diagonal matrix containing the singular values σ_j, $j=1,\ldots,m$ in decreasing order of magnitude. This decomposition is used to invert Eq. (12) to obtain

$$\boldsymbol{\varphi} = \mathbf{V}\mathbf{F}_\alpha\Sigma^{-1}\mathbf{U}^H\mathbf{p}. \tag{13}$$

We use the filter matrix $\mathbf{F}_\alpha$ with filter factors in Eq. (9) and Morozov discrepancy principle for obtaining the optimal α.

As in Eq. (11), the $N\times N$ complex matrices $[\mathbf{S}_\mathbf{p}^s]$, $[\mathbf{K}_\mathbf{p}^\pm]$ are used to obtain $\tilde{\mathbf{p}}^s$ and $\tilde{\mathbf{v}}^s$.

C. IBEM with Cauchy measurements

Cauchy measurements[27] in the mathematical literature are understood to be the measurements of both pressure and velocity at a part of a given surface Γ_0. In practice it is difficult to obtain this type of measurement, but we can approximate it with a dual surface pressure measurement. For our practical purpose we approximate the Cauchy measurements by taking measurements over two conformal arrays Γ_0^p and Γ_0^c, respectively, as shown in Fig. 4. There will be certain conditions to be imposed to the distance between the two arrays that are similar to the conditions imposed for the distance between microphones in the intensity probes (this topic will be discussed in more detail in Sec. V).

For this case we denote as $\tilde{\Gamma}=\Gamma^p\cup\Gamma^r$. Here Γ^r as shown by the dotted line in Fig. 4 is an extension surface used to make $\tilde{\Gamma}$ a closed surface. In general there is no restriction concerning the shape of Γ^r, but in this work we will choose Γ^r to be conformal to Γ_0^c as in Fig. 4. Given $2m$ points on the measurement surfaces Γ_0^p, Γ_0^c and N_c points on the reconstruction surface $\tilde{\Gamma}$ we construct the $2m\times N$ complex matrix $[\mathbf{S}_\mathbf{c}]$. As in the previous subsection, the SVD and a filter matrix with filters as in Eq. (9) is used to create a regularized solution ϕ. Then as in Eq. (11) the $N\times N$ complex matrices $[\mathbf{S}_\mathbf{p}^s]$, $[\mathbf{K}_\mathbf{p}^\pm]$ are constructed to obtain $\tilde{\mathbf{p}}^s$ and $\tilde{\mathbf{v}}^s$.

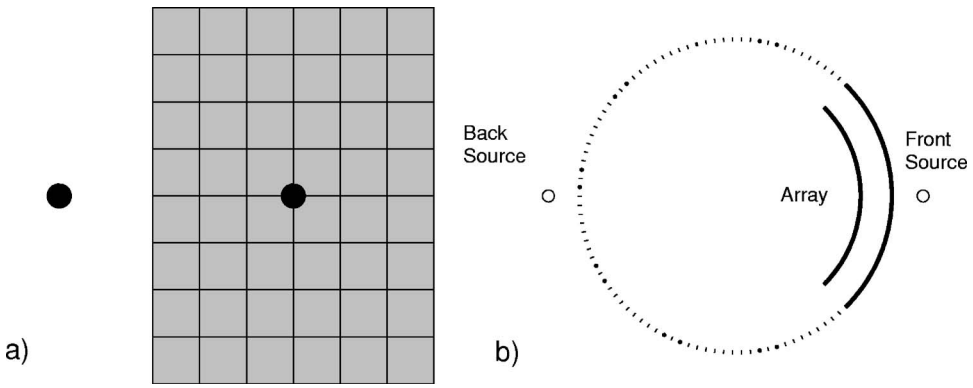


FIG. 5. (a) Planar view of cylindrical array of measurements. The spacing over the horizontal coordinates and the cylindrical arc length is 5 cm. The dots show the position of the front side point sources: in the center of the array and 10 cm to the left of the array. (b) View showing position with respect of cylindrical surface of front sources (two sources appear as one open circle) and two back sources in mirror positions across the axis of the cylinder).

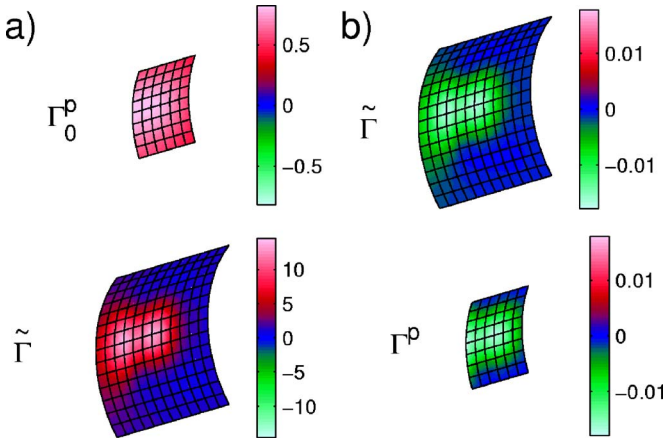


FIG. 6. (Color online) Velocity reconstruction scheme using patch IBEM with direct procedure for 50 Hz and three-extension points. (a) Reconstruction of ϕ in $\tilde{\Gamma}$ from Γ_0^p, (b) $\tilde{\mathbf{v}}_r^s$ is obtained at $\tilde{\Gamma}$ using $[\mathbf{K}_\mathbf{p}^-]$, the reconstruction over the extension Γ^e is discarded, then we obtain the normal velocity at Γ^p.

IV. NUMERICAL EXPERIMENTS

Using a numerical experiment we present results for the three reconstruction methods discussed above. We consider the following *interior* NAH problem. A set of measurements were simulated with the setup illustrated in Fig. 5. Here the measurements simulate a 9×7 cylindrical element microphone array Γ_0^p with radius 45 and 5 cm lattice spacing in both the horizontal and the cylindrical arc-length directions; the microphones are located at the corners of the grid. Two monopole point sources of equal strength are positioned over a cylindrical surface with radius 60 cm. The positions of the sources with respect to the measurement array are indicated in Fig. 5 by black dots. To provide a difficult test case in which the measured field at the array edge is not small, one of the sources is located outside the array. Gaussian noise was added to the measurements to produce a signal to noise ratio of 40 dB. The reconstruction surface Γ^p is a cylindrical element array with radius 50 cm with the same grid distribution as Γ_0^p.

A. Front source reconstructions

In this case we use data from two equal intensity front sources with the back sources turned off [see Fig. 5(b)]. We consider the reconstruction over 27 frequencies uniformly distributed between 50 and 1350 Hz. For the reconstructions using patch IBEM with the direct procedure, the reconstruction surface Γ^p is extended by a length of $h=5$, 10, 15, 20 cm corresponding to the extension of the original surface by 1, 2, 3, 4 points, respectively, on all sides. In the second case, patch IBEM with the continuation procedure, both the measurement and reconstruction surfaces Γ_0^p and Γ^p are extended equally. The reconstructed velocity $\tilde{\mathbf{v}}_r^s$ is compared to the exact velocity $\tilde{\mathbf{v}}^s$ using the relative error

$$100\% \times \frac{\|\tilde{\mathbf{v}}_r^s - \tilde{\mathbf{v}}^s\|_2}{\|\tilde{\mathbf{v}}^s\|_2},$$

taken only over Γ^p (we exclude the computation of error over the extended region Γ^e). Similarly in the case of the Cauchy measurements, discussed later, we do not include the errors over the region Γ^r.

The scheme for patch IBEM with the direct procedure at 50 Hz is shown in Fig. 6. As discussed above, this method does not use measurement aperture extension, and applies the extension only to the reconstruction surface. Also no iteration scheme is used in this method. In (a) we see the real part of the measured pressure (top illustration) and the reconstructed ϕ derived from Eq. (13) in the bottom illustration. The top illustration of (b) shows the reconstructed velocity using from Eq. (11) and the bottom illustration the result after discarding the extended region. The reconstruction errors are plotted in Fig. 7 for the velocity as a function of frequency for different extensions over Γ^p. The importance of aperture extension is evidenced by the fact that the reconstruction without extensions had errors greater than 60% over the entire frequency range and fell outside the range of the plot. The use of extension points reduces the error considerably, but this effect is limited. We clearly observe that using more than three extension points does not reduce the reconstruction errors any further.

The scheme for patch IBEM with the continuation procedure at 50 Hz is illustrated in Fig. 8 and the velocity reconstruction errors versus frequency are shown in Fig. 9 for different extensions over Γ_0^p (and Γ^p). In (a) we see the real part of the extended pressure (top illustration) without continuation procedure [$\mathbf{p}^{(0)}$ in Eq. (6)] and the reconstructed ϕ derived from Eq. (10) in the bottom illustration. In (b) we see the similar picture as in (a) after the iterative continuation steps are applied to the measurements. The top illustration of (b) shows the reconstructed velocity using from Eq. (11) and the bottom illustration the result after discarding the extended region. Figure 9(b) shows the corresponding number of iterations used in the continuation procedure. Similar to Fig. 7 we can observe that more than two extension points will not contribute to a significant decrease of the error, but will increase the number of iterations. It is important to mention that patch IBEM with continuation procedure shows smaller errors than with the direct procedure.

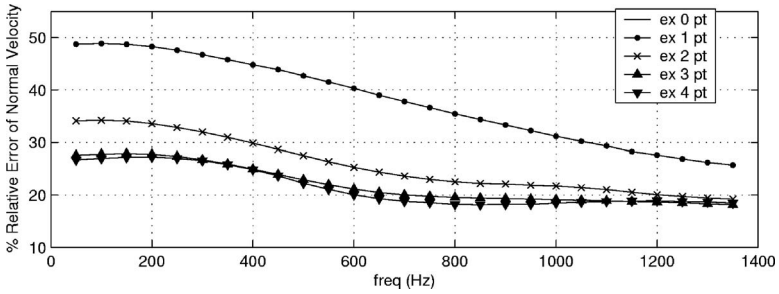


FIG. 7. Relative error from the velocity reconstruction using patch IBEM with the direct procedure for different extensions of the reconstruction surface Γ^p. The zero point extension lies above the plotted region.

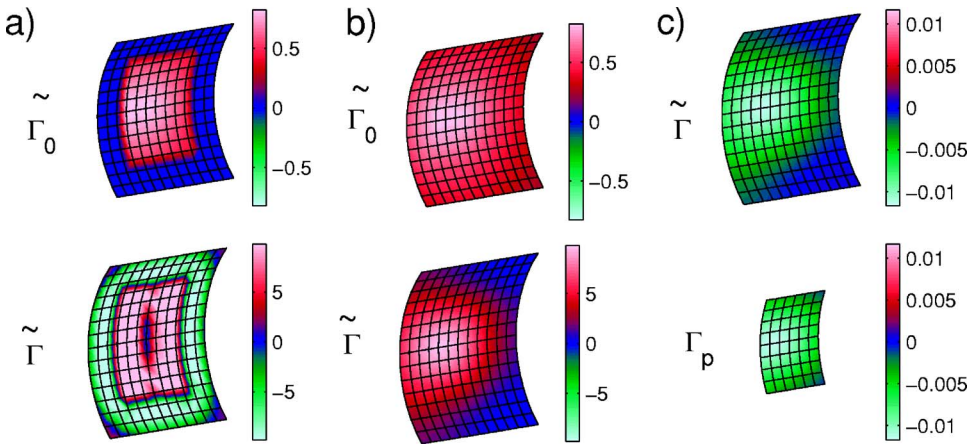


FIG. 8. (Color online) Velocity reconstruction scheme using patch IBEM with continuation procedure for 50 Hz and three extension points. (a) Reconstruction of $\boldsymbol{\phi}$ in $\tilde{\Gamma}$ (zero padded with no iterations) from $\tilde{\Gamma}_0$, (b) reconstruction of $\boldsymbol{\phi}$ in $\tilde{\Gamma}$ (after continuation) from $\tilde{\Gamma}_0$, (c) $\tilde{\mathbf{v}}^s$ is obtained at $\tilde{\Gamma}$ using $[\mathbf{K}_\mathbf{p}^-]$, the reconstruction over the extension Γ^e is discarded, then we obtain the normal velocity at Γ^p.

The scheme for IBEM with Cauchy data at 50 Hz is show Fig. 10 (similar scheme than patch IBEM with direct procedure) and Fig. 11 shows the reconstruction error of the velocity for the same frequency range in the two previous figures for different distances d between the conformal measurement arrays. Notice that $d=5$ cm produces a relatively smaller error up to 1 kHz, $d=1$ cm a slightly smaller error between 250 Hz and 1.25 kHz, and finally $d=0.5$ cm gives smaller errors for frequencies greater than 1.25 kHz. These error results illustrate a criteria for choosing the correct d depending on the frequency range that is similar to the criteria used for the design of intensity probes.

So that comparisons can be made between the three approaches, Fig. 7, Fig. 9(a) and Fig. 11 are plotted using the same scale. The error comparison between the three methods shows that, when measurements are available over a single array, the patch IBEM with the continuation procedure produces a smaller error than with the direct procedure. On the other hand, when Cauchy measurements are available, the reconstructions using IBEM with Cauchy measurements yield significantly smaller reconstruction errors when the optimum spacing d is used.

B. Back source reconstructions

Here we consider data from back sources [see Fig. 5(b)]. Different from the reconstruction of data from the front sources, for this case we encounter a fundamental limitation of patch IBEM with the direct or the continuation procedure: the use of a single array will not make a distinction between front sources or back sources. In order to clarify this situation we will consider the reconstruction of the normal intensity, instead of the normal velocity, since the intensity provides information about the position of the sources. The normal intensity $\mathbf{I}_r^s = \frac{1}{2}\mathrm{re}\{\tilde{\mathbf{p}}^s \bar{\tilde{\mathbf{v}}}_r^s\}$ will be reconstructed over five frequencies uniformly distributed between 50 and 250 Hz. This quantity shows the distribution of the energy in the surface Γ^p. In our surface Γ^p the normals point outward the surface, so a negative normal intensity value means energy entering the surface and a positive normal intensity value will mean energy leaving the surface.

In Fig. 12 we show the reconstructions of the normal intensity when the point sources are located in the back of the measurements (front side sources are turned off). As expected we see in Fig. 12(a) that the exact normal intensity

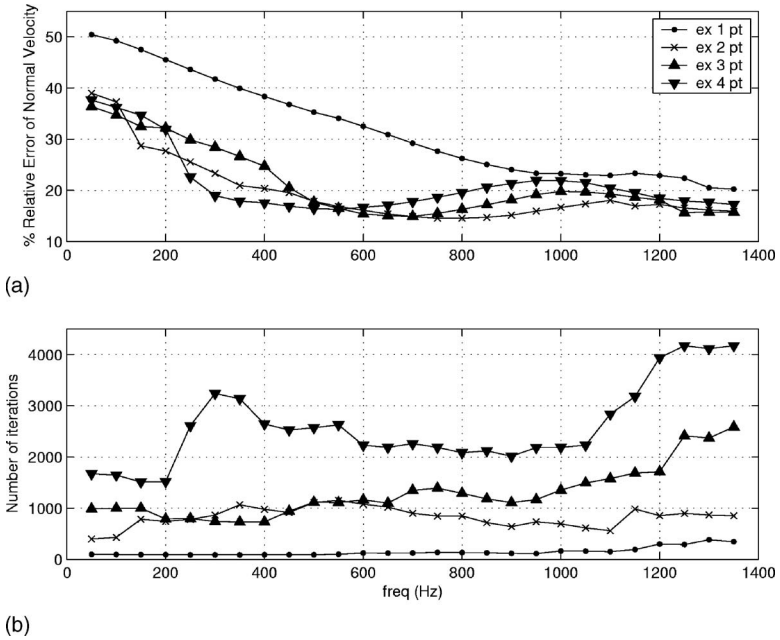


FIG. 9. (a) Relative error from the velocity reconstruction using patch IBEM with continuation procedure for different extensions of the surfaces Γ_0^p, Γ^p. (b) Number of required iterations for the continuation procedure.

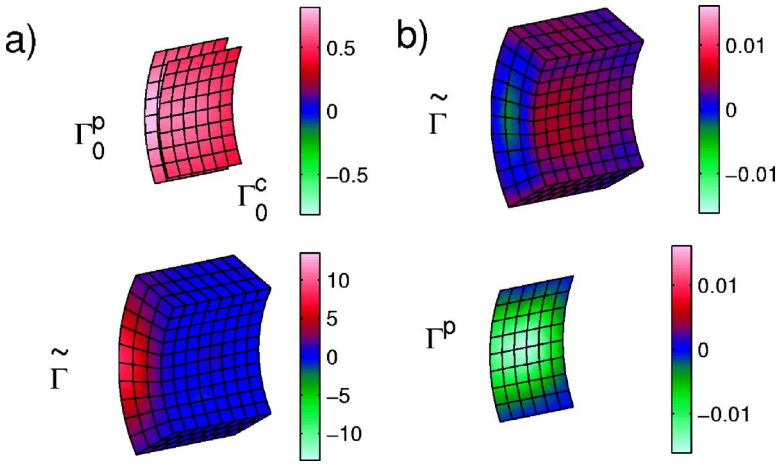


FIG. 10. (Color online) Velocity reconstruction scheme using IBEM with Cauchy data for 50 Hz. (a) Reconstruction of ϕ in $\tilde{\Gamma}$ from Γ_0^p, (b) $\tilde{v}_r^s$ is obtained at $\tilde{\Gamma}$ using $[\mathbf{K}_\mathbf{p}^-]$, the reconstruction over the extension Γ^e is discarded, then we obtain normal velocity at Γ^p.

shows energy leaving the surface. On the other hand, patch IBEM with direct or continuation procedures in b, c, respectively, show the contrary, i.e., energy entering the surface. This particular case is critical, since it shows that patch IBEM with direct or continuation procedures always assumes that the sources are positioned in front of the array. Backside sources are just "turned around" and appear as front side sources. On the other hand, when Cauchy measurements are available the assumption of source direction is not made, and IBEM with Cauchy measurements produces reliable reconstructions.

V. PHYSICAL EXPERIMENTS

The experimental configuration for the holographic measurement performed at the Laboratory for Structural Acoustics, at the Naval Research Laboratory, Washington D.C., is similar to the previous work of Herdic *et al.*[28] The surface Γ is an aluminum stiffened cylindrical shell aircraft fuselage section [0.81 m radius, 2.55 m length and the shell thickness varies between 0.8 and 1.2 mm, see Fig. 13(a)] excited by an exterior point force applied to a rib/stringer intersection near one end of the cylinder. The measurements were conducted using a chirp wave form over a band from 10 to 1000 Hz with 0.61 Hz resolution. The measurement surface Γ_0 is a cylinder of 0.7045 m radius and 2.32 m length as shown in Fig. 13(b).

The pressure on Γ_0 was measured through automated scans on a grid of 32 points over the circumference, 29 points over the length and at the cylinder ends the pressure is taken over five rings with 32 points at each ring (see Figs. 13(c) and 13(d)). Measurements of pressure data are also available over a cylindrical surface with a radius of 0.5765 m, with scan length and distribution of points identical to Γ_0. This measurements will be used for IBEM with Cauchy measurements. The normal velocity on Γ was measured with a vibrometer on a grid of 64 points over the circumference and 32 points over the length.

For the purpose of testing the methods described in the previous sections we will use a measurement array of 9×7 points Γ_0^p from the original 32×29 grid array in Γ_0. Since the exact velocity is known over a different distribution of points in Γ, we use cubic spline interpolation to obtain velocity values over Γ with the same distribution of points as Γ_0. Then the reconstruction of the normal velocity over Γ^p from measurements over Γ_0^p can be compared directly with the vibrometer interpolated measurements.

We consider two positions of the array Γ_0^p shown in Fig. 14. The first position is close to the applied point force and the second position is on the opposite side away from the point source. Figure 15 shows the results from the reconstructions using the methods described in the previous sections. For both positions we can observe that the use of Cauchy measurements (where d the distance between conformal arrays is 12.81 cm) will yield the smallest relative error. The other two methods show mixed error results. For position 1 both the direct and continuation methods produce similar errors. We should notice that in the frequencies 123.6, 134, and 140.7 Hz the error increases dramatically, which coincides with the fact that these frequencies correspond to the acoustic modes[29] or cavity modes. These acoustic modes, which correspond to small surface velocities, are based on rigid boundary conditions, so at these frequencies we will observe a small magnitude of the normal velocity

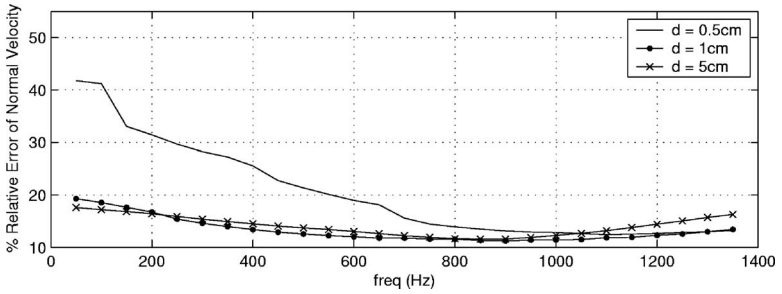


FIG. 11. Relative error from the velocity reconstruction using IBEM with Cauchy measurements for different distances d.

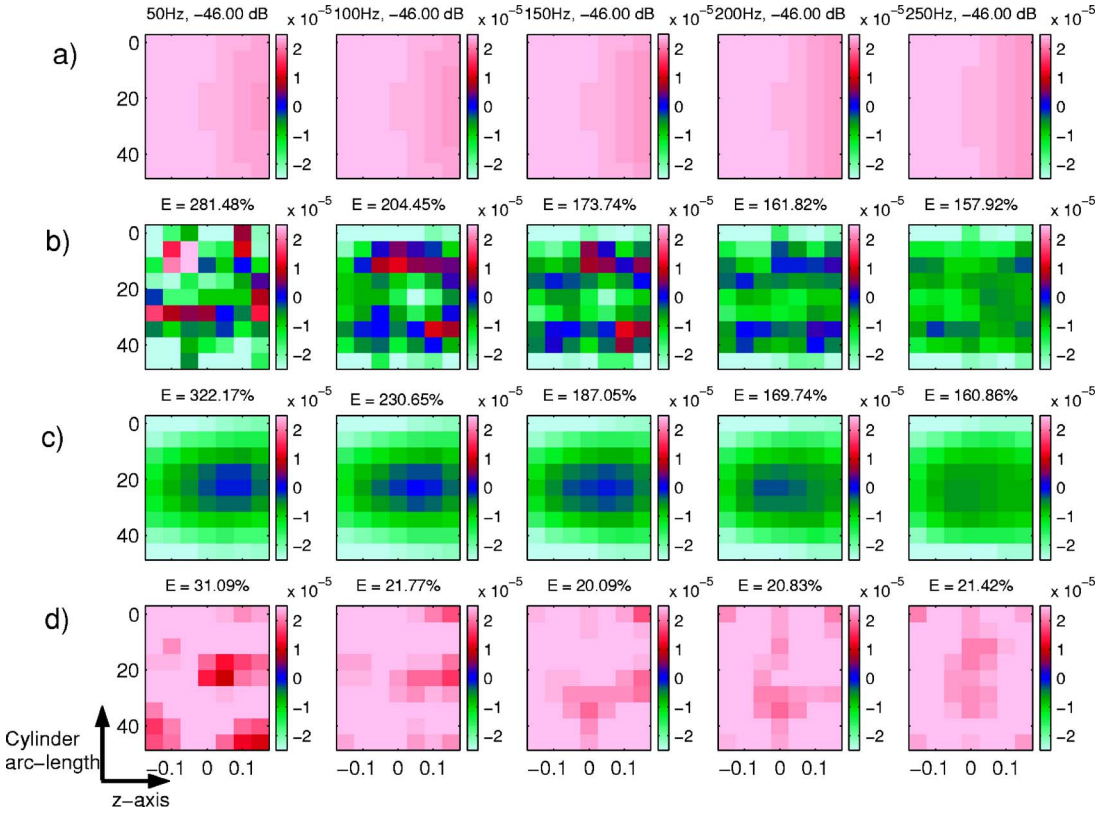


FIG. 12. (Color online) Plot of the reconstructed normal intensity (a) exact normal intensity, reconstructed normal intensity using (b) direct method with three extension points, (c) continuation with 2 extension points and (d) Cauchy measurements with $d=5$ cm.

and an increase in the magnitude of the pressure. We recall the fact that these frequencies will increase the reconstruction error.[7]

For position 2 we observed from the numerical experiments that the reconstructions using the continuation and direct methods produce larger errors than the position 1 reconstructions. Indeed this is the case, and we can clearly observe that even in this case the use of Cauchy measurements produces smaller reconstruction errors.

VI. CONCLUSIONS

In this work we have shown two methods for the reconstruction of the acoustic field when measurements are available over a patch above the structure surface. A third method is suggested for interior NAH where back source contamination is present. This method uses Cauchy measurements, i.e., measurements over two conformal arrays. Measurement probes exist that measure both pressure and velocity at a point. The Cauchy approach is ideally suited for them.

In Sec. III we described the methods for partial measurements. In Sec. III A we discussed patch IBEM with the continuation procedure. This is the direct extension of patch NAH to arbitrarily shaped surfaces. In appendixes A and B we describe the proof of convergence for the iterative continuation procedure used in this approach. In Sec. III B we discussed patch IBEM for the direct procedure. This new method avoids the iterations required in the previous method but appears to have less accuracy. Finally, in Sec. III C we suggested IBEM with Cauchy measurements.

In Sec. IV A, with the help of numerical experiments, we showed that patch IBEM with direct and continuation procedure are accurate when the sources are located in front of the measurement array [front sources as in Fig. 5(b)]. On the other hand, in Sec. IV B we observe that these two methods will fail when the sources are located in the back of the measurement array. IBEM with Cauchy measurements will give reliable reconstructions for this particular case of "back sources" contamination. The Cauchy measurements are approximated using two conformal arrays of measurements separated at a uniform distance d. This distance d plays an important role and depends on the frequency range of the reconstruction. We found that $d=5$ cm for reconstruction of frequencies less than 1 kHz, $d=1$ cm for reconstructions between 250 Hz and 1.25 kHz, and finally $d=0.5$ cm for frequencies greater than 1.25 kHz.

In Sec. V a fuselage section experiment confirmed the numerical results of Sec. IV. The direct and continuation methods are accurate when the sources are located in front of the measurement arrays. For interior measurements, such as inside a car, when there is no information about the sources so that an array cannot be positioned under the dominant sources, then is recommended to take Cauchy measurements instead of a single array of measurement. But even over a suspected source, we also have found in general that the use of Cauchy measurements reduces the reconstruction error compared with single surface measurements.

ACKNOWLEDGMENT

This work was supported by Office of Naval Research.

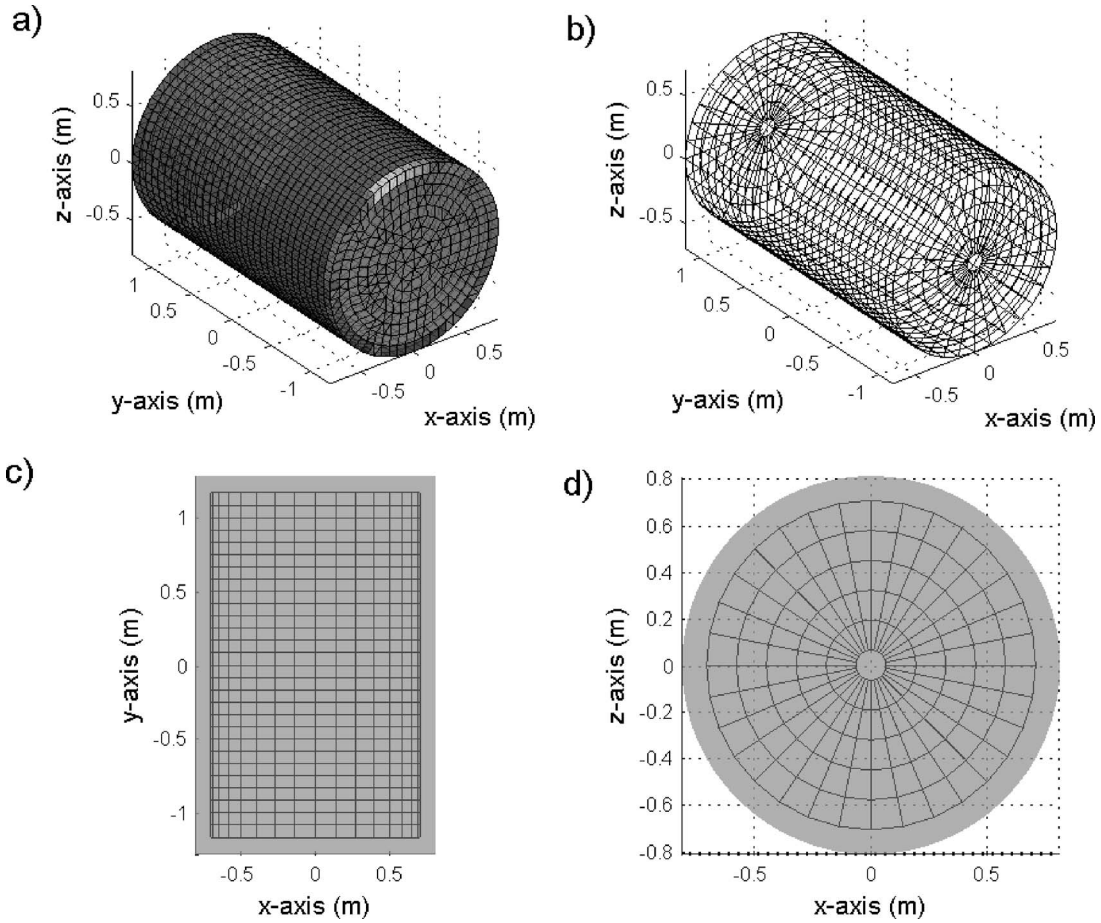


FIG. 13. Setup for physical interior problem experiment. (a) Surface Γ is an aluminum cylindrical shell, (b) measurement surface Γ_0, (c) measurements along the length (top view down on the cylinder) and (d) side-end measurements (side view on the cylinder).

APPENDIX A: HELPFUL RESULTS

We use the $M \times M$ complex unitary matrix $\mathbf{U}_e$ from the SVD of $[\mathbf{S_e}]$ in Sec. III A. Given m pressure measurements $\mathbf{p}$ (where $m < M$) we can assume that the first m rows of $[\mathbf{S_e}]$ coincide with these measurements. Then we can separate

$$\mathbf{U}_e = \begin{bmatrix} \mathbf{U}_{(m)} \\ \hat{\mathbf{U}}_{(m)} \end{bmatrix}. \tag{A1}$$

where $\mathbf{U}_{(m)}$ and $\hat{\mathbf{U}}_{(m)}$ are, respectively, $m \times M$ and $(M-m) \times M$ complex matrices.

Let denote as $\mathbf{I}_{(m)}$ the $m \times m$ identity matrix. The norm $\|.\|$ is the 2 norm and this notation will be assumed thought the rest of the appendices.

Lemma A.1 (orthogonality properties) *For the matrices* $\mathbf{U}_{(m)}$ *and* $\hat{\mathbf{U}}_{(m)}$ *defined in Eq. (A1) we have the properties*

$$\mathbf{U}_{(m)}^H \mathbf{U}_{(m)} + \hat{\mathbf{U}}_{(m)}^H \hat{\mathbf{U}}_{(m)} = \mathbf{I}_{(M)}$$

$$\mathbf{U}_{(m)} \mathbf{U}_{(m)}^H = \mathbf{I}_{(m)}, \quad \hat{\mathbf{U}}_{(m)} \hat{\mathbf{U}}_{(m)}^H = \mathbf{I}_{(M-m)},$$

$$\hat{\mathbf{U}}_{(m)} \mathbf{U}_{(m)}^H = 0, \quad \mathbf{U}_{(m)} \hat{\mathbf{U}}_{(m)}^H = 0. \tag{A2}$$

Proof: Use that $\mathbf{I}_{(M)} = \mathbf{U}_e^H \mathbf{U}_e$ to obtain the first property in Eq. (A2). Similarly $\mathbf{I}_{(M)} = \mathbf{U}_e \mathbf{U}_e^H$ is used to obtain the rest of the properties in Eq. (A2). □

The following lemma is easily proved using the previous lemma.

Lemma A.2 (norm properties) *For the matrices* $\mathbf{U}_{(m)}$ *and* $\hat{\mathbf{U}}_{(m)}$ *defined in Eq. (A1) we have the properties*

$$\|\mathbf{U}_{(m)}^H \mathbf{x}\| = \|\mathbf{x}\|, \quad \|\mathbf{U}_{(m)}\| = \|\mathbf{U}_{(m)}^H\| = 1,$$

$$\|\hat{\mathbf{U}}_{(m)}^H \mathbf{x}\| = \|\mathbf{x}\|, \quad \|\hat{\mathbf{U}}_{(m)}\| = \|\hat{\mathbf{U}}_{(m)}^H\| = 1, \tag{A3}$$

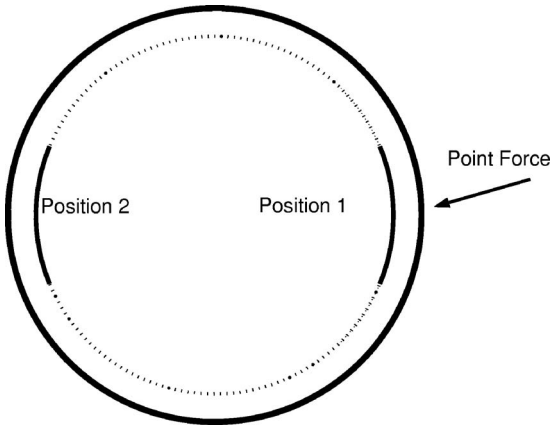


FIG. 14. Positions of the measurements with respect to the applied point force.

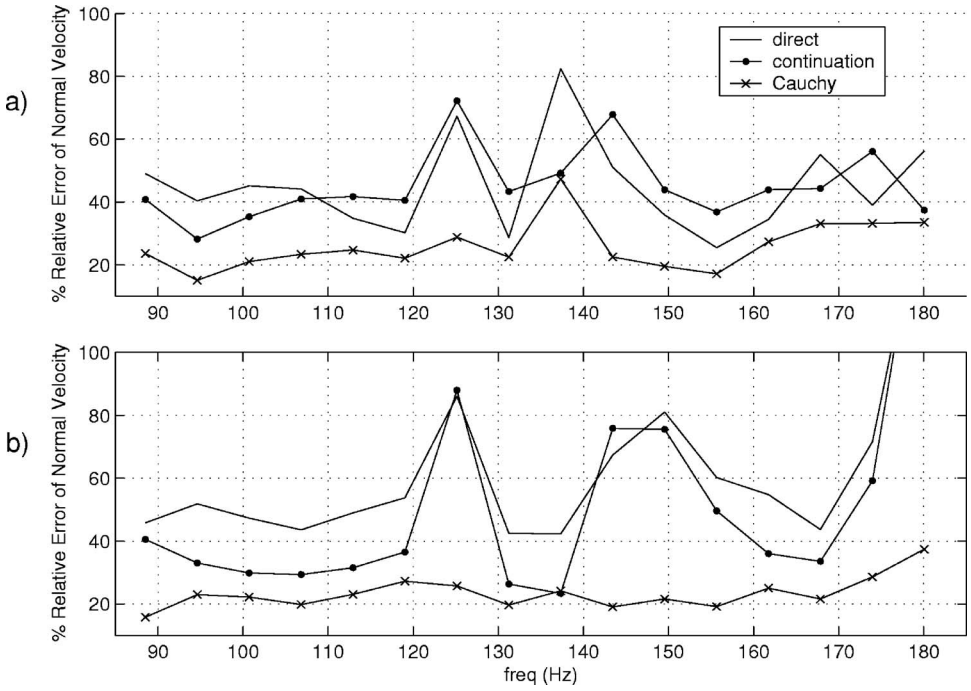


FIG. 15. Error plot for the reconstructed normal velocity when the patch measurements are located (a) in front of shaker (b) away from the shaker.

APPENDIX B: CONVERGENCE THEOREMS

Let $\mathbf{D}$ be a $M\times M$ matrix with 1's in the first m diagonal entries (zeros elsewhere) and $\mathbf{Q}=\mathbf{I}_M-\mathbf{D}$.

The continuation method suggested by Williams[10] is defined by the equation for $i>0$,

$$\mathbf{p}^{(i+1)}=\mathbf{Q}\tilde{\mathbf{p}}^{(i)}+\mathbf{D}\mathbf{p}^{(i)}, \tag{B1}$$

where

$$\tilde{\mathbf{p}}^{(i)}=\mathbf{U}_e\mathbf{F}_{\alpha(i)}\mathbf{U}_e^H\mathbf{p}^{(i)},$$

$\mathbf{p}^{(0)}$ is the initial guess and $\alpha(i)$ is chosen for $\mathbf{p}^{(i)}$ on each iteration step. The iteration will be stop when $\|\tilde{\mathbf{p}}^{(i+1)}-\tilde{\mathbf{p}}^{(i)}\|<\epsilon\|\tilde{\mathbf{p}}^{(i)}\|$ for a small value of ϵ.

For the filter factors f_j^{α}, $j=1,\ldots,M$ we require them (as required in Hansen[8]) to have values between 0 and 1, be nonincreasing with respect to j and finally that all should be 1 when $\alpha=0$. For the iterations in Eq. (B1) we require them to satisfy

$$\|\mathbf{F}_{\alpha(k)}-\mathbf{F}_{\alpha(k-1)}\|\leqslant C|\alpha(k)-\alpha(k-1)|,$$

$$\lim_{k\to\infty}\alpha(k)=\alpha_0,\quad \alpha_0>0, \tag{B2}$$

where C is a constant that depends on the filter $\mathbf{F}_{\alpha}$.

Lemma B.1 *For the iterative procedure given in B1 there exist an $M>0$ such that for all $i>0$*

$$\|\mathbf{p}^{(i)}\|\leqslant M. \tag{B3}$$

Proof: The equation for iteration i gives that

$$\mathbf{U}^H\mathbf{p}^{(i)}=\mathbf{U}_{(m)}^H\mathbf{p}+\mathbf{B}_{(i-1)}\mathbf{U}^H\mathbf{p}^{(i-1)}=\mathbf{U}_{(m)}^H\mathbf{p}+\mathbf{B}_{(i-1)}\mathbf{U}_{(m)}^H\mathbf{p}+\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\mathbf{U}^H\mathbf{p}^{(i-2)},$$

where $\mathbf{B}_{(i)}:=\hat{\mathbf{U}}_{(m)}^H\hat{\mathbf{U}}_{(m)}\mathbf{F}_{\alpha(i)}$. Then we obtain the relation

$$\begin{aligned}\mathbf{U}^H\mathbf{p}^{(i)}&=\mathbf{U}_{(m)}^H\mathbf{p}+\mathbf{B}_{(i-1)}\mathbf{U}_{(m)}^H\mathbf{p}+\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\mathbf{U}_{(m)}^H\mathbf{p}+\cdots\\&\quad+\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\cdots\mathbf{B}_{(1)}\mathbf{U}_{(m)}^H\mathbf{p}\\&\quad+\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\cdots\mathbf{B}_{(0)}\mathbf{U}^H\mathbf{p}^{(0)}.\end{aligned} \tag{B4}$$

Let's assume that $\alpha(i_0)=0$, then Eq. (A2) is used to get the property

$$\mathbf{B}_{(i_0-1)}\mathbf{B}_{(i_0)}\mathbf{B}_{(i_0+1)}=\mathbf{B}_{(i_0-1)}\mathbf{B}_{(i_0+1)}.$$

Then to prove this lemma is sufficient to prove the case where $\alpha(i)>0$ for all $i>0$. Lemma A.1 implies that for a nonzero $\mathbf{y}$, there exists $0<a_i<1$ such that $\|\mathbf{B}_{(i)}\mathbf{y}\|\leqslant\|\mathbf{F}_{\alpha(i)}\mathbf{y}\|\leqslant a_{(i)}\|\mathbf{y}\|$. Similarly it is not difficult to show that for $a:=\max_i a(i)$ we have the estimate

$$\|\mathbf{B}_{(i)}\ldots\mathbf{B}_{(1)}\mathbf{y}\|\leqslant\alpha^i\|\mathbf{y}\|. \tag{B5}$$

Then from Eq. (B4) and Eq. (B5) follows the estimate

$$\begin{aligned}\|\mathbf{p}^{(i)}\|&\leqslant\|\mathbf{p}\|+\|\mathbf{B}_{(i-1)}\mathbf{U}_{(m)}^H\mathbf{p}\|+\|\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\mathbf{U}_{(m)}^H\mathbf{p}\|+\cdots\\&\quad+\|\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\ldots\mathbf{B}_{(1)}\mathbf{U}_{(m)}^H\mathbf{p}\|\\&\quad+\|\mathbf{B}_{(i-1)}\mathbf{B}_{(i-2)}\ldots\mathbf{B}_{(0)}\mathbf{U}_{(m)}^H\mathbf{p}^{(0)}\|\\&\leqslant\|\mathbf{p}\|+a\|\mathbf{p}\|+a^2\|\mathbf{p}\|+\cdots+a^{i-1}\|\mathbf{p}\|+a^i\|\mathbf{p}^{(0)}\|\end{aligned}$$

The previous estimate and the geometric series representation gives that

$$\begin{aligned}\|\mathbf{p}^{(i)}\|&\leqslant(1+\cdots+a^{i-1})\|\mathbf{p}\|+a^i\|\mathbf{p}^{(0)}\|\\&=\left(\frac{1+a^i}{1-a}\right)\|\mathbf{p}\|+a^i\|\mathbf{p}^{(0)}\|.\end{aligned}$$

Then $\|\mathbf{p}^{(i)}\|\leqslant(1-a)^{-1}\|\mathbf{p}\|$ as $i\to\infty$. Equation (B3) follows from this result. □

For the following two lemmas, fix the standard deviation σ_0 used for the calculation of α at each step of the Morozov's discrepancy principle and let

$$\mathbf{p}^{(0)} = \begin{cases} \mathbf{p} & \text{in } \Gamma_0^p, \\ 0 & \text{in } \Gamma_0^c. \end{cases} \tag{B6}$$

Lemma B.2 *Given σ_0 and initial value $\mathbf{p}^{(0)}$ in Eq. (B6), for the iterative procedure given by Eq. (B1), if each $\alpha(i)$, $i>0$ is obtained by the discrepancy principle then we have*

$$\|\mathbf{p}^{(i-1)}\| \leqslant \|\mathbf{p}^{(i)}\|. \tag{B7}$$

Proof: From Eq. (B1) we get the expression

$$\begin{aligned} \mathbf{p}^{(i)} &= \mathbf{Q}\mathbf{U}_e\mathbf{F}_{\alpha(i-1)}\mathbf{U}_e^H\mathbf{p}^{(i-1)} + \mathbf{D}\mathbf{p}^{(i-1)} \\ &= \mathbf{Q}\mathbf{U}_e(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e^H\mathbf{p}^{(i-1)} + \mathbf{p}^{(i-1)} \end{aligned} \tag{B8}$$

We use the previous expression to obtain

$$\begin{aligned} \|\mathbf{p}^{(i)}\|^2 &= \|\mathbf{Q}\mathbf{U}_e(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e^H\mathbf{p}^{(i-1)}\|^2 \\ &\quad + (\mathbf{p}^{(i-1)})^H\mathbf{Q}\mathbf{U}_e(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e^H\mathbf{p}^{(i-1)} \\ &\quad + (\mathbf{p}^{(i-1)})^H\mathbf{U}_e(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e^H\mathbf{Q}\mathbf{p}^{(i-1)} + \|\mathbf{p}^{(i-1)}\|^2 \end{aligned}$$

Using the fact that $\mathbf{Q}$ is a projection matrix and that $\mathbf{U}_e$ is an unitary matrix we finally obtain

$$\begin{aligned} &\|\mathbf{Q}\mathbf{U}_e(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{x}\|^2 + (\mathbf{p}^{(i-1)})^H\mathbf{Q}\mathbf{U}_e(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{x} \\ &\quad + \mathbf{x}^H(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e^H\mathbf{Q}\mathbf{p}^{(i-1)} + \|\mathbf{p}^{(i-1)}\|^2 \\ &= \mathbf{x}^H\{(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e\mathbf{Q}\mathbf{U}_e^H(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M) \\ &\quad + \mathbf{U}_e\mathbf{Q}\mathbf{U}_e^H(\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M) \\ &\quad + (\mathbf{F}_{\alpha(i-1)} - \mathbf{I}_M)\mathbf{U}_e\mathbf{Q}\mathbf{U}_e^H\}\mathbf{x} \\ &= \mathbf{x}^H\{\mathbf{F}_{\alpha(i-1)}\mathbf{U}_e\mathbf{Q}\mathbf{U}_e^H\mathbf{F}_{\alpha(i-1)} - \mathbf{U}_e\mathbf{Q}\mathbf{U}_e^H\}\mathbf{x} \\ &= \|\mathbf{Q}\mathbf{U}_e^H\mathbf{F}_{\alpha(i-1)}\mathbf{x}\|^2 - \|\mathbf{Q}\mathbf{U}_e^H\mathbf{x}\|^2, \end{aligned}$$

where $\mathbf{x}=\mathbf{U}_e^H\mathbf{p}^{(i-1)}$.

Notice that the choice of $\mathbf{p}^{(0)}$ and $\alpha(i-1)$ guarantees that the magnitude of $\mathbf{U}_e^H\mathbf{F}_{\alpha(i-1)}\mathbf{x}$ will be bigger than $\mathbf{x}$ in Γ_0^c. Then we can conclude that

$$\|\mathbf{Q}\mathbf{U}_e^H\mathbf{F}_{\alpha(i-1)}\mathbf{x}\|^2 \geqslant \|\mathbf{Q}\mathbf{U}_e^H\mathbf{x}\|^2$$

and so from the previous estimates applied in Eq. (B8) we get Eq. (B7).

Lemma B.3 *Given σ_0 and initial value $\mathbf{p}^{(0)}$ in Eq. (B6), for the iterative procedure given by Eq. (B1), if each $\alpha(i)$ is obtained by the discrepancy principle then we have*

$$\begin{aligned} &\alpha(i+1) \geqslant \alpha(i), \\ &\alpha(i+1) - \alpha(i) \to 0, \quad \text{as } i \to \infty. \end{aligned} \tag{B9}$$

Proof: From lemma B.2 we can argue that

$$\|(\mathbf{I}_M - \mathbf{F}_\lambda)\mathbf{U}^H\mathbf{p}^{(i)}\| \geqslant \|(\mathbf{I}_M - \mathbf{F}_\lambda)\mathbf{U}^H\mathbf{p}^{(i-1)}\| \tag{B10}$$

The parameter $\alpha(i+1)$ is the solution for the λ such that

$$\sigma_0\sqrt{m} = \|(\mathbf{I}_M - \mathbf{F}_\lambda)\mathbf{U}^H\mathbf{p}^{(i)}\|,$$

since $\|(\mathbf{I}_M-\mathbf{F}_\lambda)\mathbf{U}^H\mathbf{p}^{(i-1)}\|$ increase as λ decreases, then Eq. (B10) implies that $\alpha(i) \leqslant \alpha(i+1)$.

The convergence of $\alpha(i)$ will be obtained by the following arguments. Let i_0 be big enough so that the diagonal terms of $\mathbf{F}_{\alpha(i)}$ are small enough [because $\alpha(i)$ is big enough] so that Eq. (B1) for $i>i_0$ will become

$$\mathbf{p}^{(i+1)} = \mathbf{D}\mathbf{p}^{(i)} + \mathbf{e}.$$

where $\|\mathbf{e}\|$ is vector with arbitrarily small entries. This will imply the convergence of the iterations in Eq. (B1) and so the convergence of $\alpha(i)$.

Theorem B.1 *The iterative procedure, given by Eq. (B1) with a regularization filter $\mathbf{F}_\alpha$ that satisfies property in Eq. (B2), converges.*

Proof: From the definition of $\tilde{\mathbf{p}}^{(i)}$ and previous lemma we get that

$$\begin{aligned} \|\tilde{\mathbf{p}}^{(i)} - \tilde{\mathbf{p}}^{(i-1)}\| &\leqslant \|\mathbf{p}^{(i)} - \mathbf{p}^{(i-1)}\| + \|\mathbf{F}_{\alpha(i)} - \mathbf{F}_{\alpha(i-1)}\|\|\mathbf{p}^{(i-1)}\| \leqslant \|\mathbf{p}^{(i)} \\ &\quad - \mathbf{p}^{(i-1)}\| + C|\alpha(i) - \alpha(i-1)|M \end{aligned} \tag{B11}$$

Using the same notation as in lemma B.1 and relation Eq. (B4), for a fixed i_0 we get that

$$\mathbf{U}^H(\mathbf{p}^{(i_0+i)} - \mathbf{p}^{(i_0+i-1)}) = \mathbf{C}_{(i)} + \mathbf{D}_{(i)} + \mathbf{E}_{(i)}, \tag{B12}$$

where

$$\mathbf{D}_{(i)} = \mathbf{B}_{(i-1)} \ldots \mathbf{B}_{(0)}\mathbf{U}^H\mathbf{p}^{(i)},$$

$$\mathbf{E}_{(i)} = \mathbf{B}_{(i-2)} \ldots \mathbf{B}_{(0)}(\mathbf{U}_{(m)}^H\mathbf{p} - \mathbf{U}^H\mathbf{p}^{(i)}).$$

and

$$\mathbf{C}_{(1)} = 0,$$

$$\mathbf{C}_{(i)} = (\mathbf{B}_{(i-1)} - \mathbf{B}_{(i-2)})\mathbf{H}_{(i-2)} + \mathbf{B}_{(i-2)}\mathbf{C}_{(i-1)}, \quad i \geqslant 2$$

$$\mathbf{H}_{(0)} = \mathbf{U}_{(m)}^H\mathbf{p},$$

$$\mathbf{H}_{(i)} = \mathbf{U}_{(m)}^H\mathbf{p} + \mathbf{B}_{(i)}\mathbf{H}_{(i-1)}, \quad i \geqslant 1.$$

As in lemma B.1 we have a $0<a<1$ for the estimates

$$\begin{aligned} &\|\mathbf{D}_{(i)}\| \leqslant a^i\|\mathbf{p}\| \\ &\|\mathbf{E}_{(i)}\| \leqslant a^{i-1}\|\mathbf{U}_{(m)}^H\mathbf{p} - \mathbf{U}^H\mathbf{p}^{(i_0)}\| \leqslant a^{i-1}(\|\mathbf{p}\| + \|\mathbf{p}^{(i_0)}\|) \end{aligned} \tag{B13}$$

and

$$\begin{aligned} \|\mathbf{H}_{(i)}\| &\leqslant (1 + \ldots + a^{i-1})\|\mathbf{p}\|, \\ &= \left(\frac{1+a^i}{1-a}\right)\|\mathbf{p}\| \leqslant \frac{2}{1-a}\|\mathbf{p}\|. \end{aligned} \tag{B14}$$

Lemma B.1 implies that $\|\mathbf{p}_{(i_0)}\| \leqslant M$. Then it is trivial to observe $\|\mathbf{D}_{(i)}\|$ and $\|\mathbf{E}_{(i)}\|$ converges to 0 as i increases.

For $\|\mathbf{C}_{(i)}\|$ we use the convergence of $\alpha(i)$. We can assume that for $\epsilon>0$, we can choose $i_0>0$ such that $|\alpha(i+i_0)-\alpha(i+i_0-1)| < \epsilon\|p\|^{-1}(1-a)^2/4$. Then from Eq. (B14)

$$\|\mathbf{C}_{(i)}\| \leq \{|\alpha(i_0+i)-\alpha(i_0+i-1)|+a|\alpha(i_0+i-1)-\alpha(i_0+i-2)|+\ \ldots\ +a^i|\alpha(i_0)-\alpha(i_0-1)|\}\left(\frac{2}{1-a}\right)\|p\|$$

$$\leq \frac{\epsilon}{\|p\|}\left(\frac{1-a}{2}\right)^2(1+\ \ldots\ +a^{n-1})\left(\frac{2}{1-a}\right)\|p\| \leq \epsilon. \tag{B15}$$

Equation (B15) proves the convergence of $\|\mathbf{C}_{(i)}\|$. Then we have shown that $\|\mathbf{p}^{(i)}-\mathbf{p}^{(i-1)}\|\to 0$ as $i\to\infty$. Finally, the convergence of $\alpha(i)$ implies that $\|\tilde{\mathbf{p}}^{(i)}-\tilde{\mathbf{p}}^{(i-1)}\|\to 0$ as $i\to 0$. □

Corollary B.1 *The iterative procedure, given by Eq. (B1) with a regularization filter* $\mathbf{F}_{\alpha(i)}$ *chosen as in lemma B.2 converges*

Proof: When the filter $\mathbf{F}_{\alpha(i)}$ is chosen as in lemma B.3, then it satisfies the properties in Eq. (B2). Then we use theorem B.1 to prove the convergence. □

[1] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic, London, 1999).
[2] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind* (Cambridge University Press, New York, 1997).
[3] M. R. Bai, "Application of bem (boundary element method)-based acoustic holography to radiation analysis of sound sources with arbitrarily shaped geometries," J. Acoust. Soc. Am. **92**, 533–549 (1992).
[4] G.-T. Kim and B.-H. Lee, "3-d sound source reconstruction and field reprediction using the helmholtz integral equation," J. Sound Vib. **136**, 245–261 (1990).
[5] Z. Zhang, N. Vlahopoulos, S. T. Raveendra, T. Allen, and K. Y. Zhang, "A computational acoustic field reconstruction process based on a indirect boundary element formulation," J. Acoust. Soc. Am. **108**, 2167–2178 (2000).
[6] A. Schuhmacher, J. Hald, K. B. Rasmussen, and P. C. Hansen, "Sound source reconstruction using inverse boundary element calculations," J. Acoust. Soc. Am. **113**, 114–126 (2003).
[7] N. Valdivia and E. G. Williams, "Implicit methods of solution to integral formulations in boundary element methods based near-field acoustic holography," J. Acoust. Soc. Am. **116**, 1559–1572 (2004).
[8] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems* (Siam, Philadelphia, 1998).
[9] K. Saijyou and S. Yoshikawa, "Reduction methods of the reconstruction error for large-scale implementation of near-field acoustical holography," J. Acoust. Soc. Am. **110**, 2007–2023 (2001).
[10] E. G. Williams, "Continuation of acoustic near fields," J. Acoust. Soc. Am. **113**, 1273–1281 (2003).
[11] E. G. Williams, B. Houston, and P. C. Herdic, "Fast Fourier transform and singular value decomposition for patch nearfield acoustical holography," J. Acoust. Soc. Am. **114**, 1322–1332 (2003).
[12] R. Steiner and J. Hald, "Near-field acoustical holography without the errors and limitations caused by the use of spatial dft," Int. J. Acoust. Vib. **6**, 83–89 (2001).
[13] J. Hald, "Planar near-field acoustical holography with arrays smaller than the sound source," in *Proceedings of the 17th International Congress on Acoustics*, Volume 1-part A, Rome, Italy (2001).
[14] Y. T. Cho, J. S. Bolton, and J. Hald, "Source visualization by using statistically optimized near-field acoustical holography in cylindrical coordinates," J. Acoust. Soc. Am. **118**, 2355–2364 (2005).
[15] S. F. Wu and X. Zhao, "Combined Helmholtz equation-least squares method for reconstructing acoustic radiation from arbitrarilty shaped objects," J. Acoust. Soc. Am. **112**, 179–188 (2002).
[16] S. F. Wu, "On reconstruction of acoustic pressure fields using the Helmholtz equation least squares method," J. Acoust. Soc. Am. **107**, 2511–2522 (2000).
[17] R. F. Millar, "The Rayleigh hypothesis and a related least-squares solution to scattering problems for periodic surfaces and other scatterers," Radio Sci. **8**, 785–796 (1973).
[18] T. Semenova and S. F. Wu, "The Helmholtz equation least-squares method and Rayleigh hypothesis in near-field acoustical holography," J. Acoust. Soc. Am. **115**, 1632–1640 (2004).
[19] G. H. Koopmann, L. Song, and J. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," J. Acoust. Soc. Am. **86**, 2433–2438 (1989).
[20] N. Valdivia and E. G. Williams, "Study of the comparison of the methods of equivalent sources and boundary element methods for near-field acoustic holography," J. Acoust. Soc. Am. **120**, 3694–3705 (2006).
[21] J. Hald and J. Gomes, "A comparison of two patch NAH methods," in *Proceedings Internoise 2006* (in 06-166).
[22] E. G. Williams, "Approaches to patch NAH," in *Proceedings of Internoise 2003*, 2187–2194, Lyon, France (2003).
[23] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory* (Wiley-Interscience, New York, 1983).
[24] T. K. DeLillo, V. Isakov, N. Valdivia, and L. Wang, "The detection of surface vibrations from interior acoustical pressure," Inverse Probl. **19**, 507–524 (2003).
[25] E. G. Williams, B. H. Houston, P. C. Herdic, S. T. Raveendra, and B. Gardner, "Interior NAH in flgiht," J. Acoust. Soc. Am. **108**, 1451–1463 (2000).
[26] E. G. Williams, "Regularization methods for near-field acoustical holography," J. Acoust. Soc. Am. **110**, 1976–1988 (2001).
[27] V. Isakov, *Inverse Problems for Partial Differential Equations* (Springer-Verlag, New York, 1998).
[28] P. Herdic, B. Houston, M. Marcus, E. Williams, and A. Baz, "The vibroacoustic response and analysis of a full-scale aircraft fuselage section for interior noise reduction," J. Acoust. Soc. Am. **117**, 3667–3678 (2005).
[29] B. Houston, P. Herdic, M. Marcus, E. Williams, and J. Bucaro, "High spatial density surface velocity and interior acoustic measurements associated with an aircraft fuselage section under point excitation," in *AIAA Paper No. 96-1765* (1996).

A ray model for hard parallel noise barriers in high-rise cities

Kai Ming Li[a)]
Ray W. Herrick Laboratories, School of Mechanical Engineering, Purdue University, 140 South Intramural Drive, West Lafayette, Indiana 47907-2031, USA

Man Pun Kwok and Ming Kan Law
Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

(Received 6 March 2007; revised 5 September 2007; accepted 11 October 2007)

A ray model is developed and validated for prediction of the insertion loss of hard parallel noise barriers placed in an urban environment either in front of a row of tall buildings or in a street canyon. The model is based on the theory of geometrical acoustics for sound diffraction at the edge of a barrier and multiple reflections by the ground, barrier and façade surfaces. It is crucial to include the diffraction and multiple reflection effects in the ray model as they play important roles in determining the overall sound pressure levels for receivers located between the façade and the near-side barrier. Comparisons of the ray model with a wave-based boundary element formulation show reasonably good agreement over a broad frequency range. Results of scale model experimental studies are also presented. It is demonstrated that the ray model agrees tolerably well with the scale model experimental data. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804944]

PACS number(s): 43.50.Gf, 43.28.En, 43.20.Dk [KA] Pages: 121–132

I. INTRODUCTION

Construction of parallel barriers can help to attenuate noise levels on both sides of a road. Emanating from vehicles, the multiple reflections of sound waves produced by barrier surfaces create a reverberant sound field within the region of interest. These multiple reflected sound waves may travel to a receiver directly or diffract at the top edges of the parallel barriers. This phenomenon, which leads to degradation in barrier performance, is well recognized and was studied in the 1980s by Hutchins *et al.*[1–3] By assuming an incoherent nature of the source, they predicted the sound intensity levels at the top of barriers by summing the contributions from the principal wave propagation paths. Comparisons with upright and inclined parallel barriers were also made in these early studies.

It is of interest to point out that Chew[4] developed a prediction scheme for buildings situated on both sides of an expressway. Chew's model involved direct and reflected energy, both from the ground surface and multiple reflections between the parallel buildings on both sides of the expressway, and diffused energy due to scattering from the rough façade surfaces. His predictions showed that the effect of multiple reflections was significant when the distance between the buildings on opposite sides of the road was small.

Sakurai *et al.*[5] used a time-domain method to investigate the sound field of a façade-barrier system. They used a computer simulation program to evaluate the time averaging sound pressure levels in an outdoor environment. An empirical model[6] was developed to study the acoustic performance of a parallel barrier in front of a building façade. Godinho *et al.*[7] employed a boundary element formulation (BEM) to determine the shielding effects of an infinitely long barrier placed in front of tall building façades.

Despite this widespread interest, there are relatively few studies that have considered the shielding effect of parallel noise barriers in an urban environment. The prime objective of the present paper is to develop an effective numerical model to assess the acoustic performance of parallel barriers in high-rise cities. In particular, we wish to develop a numerical model to predict the sound fields in two urban scenarios: (a) when a pair of parallel barriers is flanked by a row of tall buildings, and (b) when the parallel barriers are placed in a street canyon. The numerical approach is based on the image source model developed by Li and Tang,[8] who investigated the performance of a single noise barrier in the proximity of tall buildings. The ray-based model provided accurate solutions that agreed reasonably well with indoor scaled-model experiments as well as with predictions according to the BEM formulation. On the other hand, Panneton *et al.*[9] and Muradali and Fyfe[10] used the ray model to study the acoustic performance of parallel noise barriers in the absence of other reflecting surfaces except a flat ground. The work presented in this paper is an extension of these earlier studies, but we endeavor to develop the corresponding ray model to study the acoustic performance of hard parallel barriers in high-rise cities.

This paper is organized as follows. Section II discusses the formulation of the ray model for predicting sound fields for two urban cases: (a) when a pair of parallel barriers is placed in front of a building facade, and (b) when the parallel barriers are flanked by two rows of tall buildings forming a street canyon. The image sources and image receivers are addressed in forming the ray series for computing the total sound fields. In Sec. III, we validate the ray model by comparing its predictions with those obtained by a more accurate wave-based numerical method Section IV gives the compari-

[a)]Author to whom correspondence should be addressed. Electronic mail: mmkmli@purdue.edu

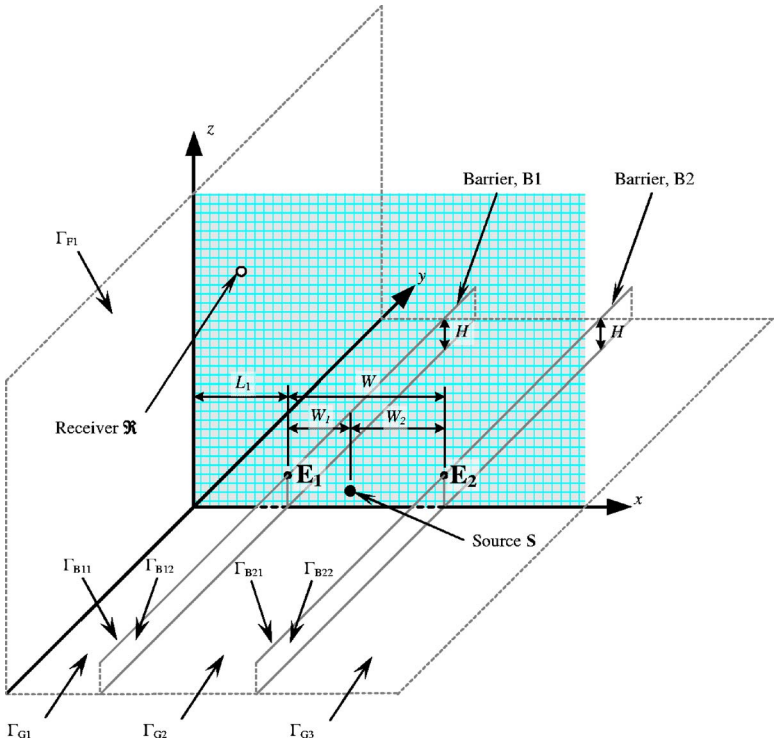


FIG. 1. (Color online) Schematic diagram of the specified problem. A source **S** is located at $\mathbf{S}\equiv(x_S,0,z_S)$ and a receiver $\mathfrak{R}$ at $(x_R,0,z_R)$. A pair of parallel barriers, B1 and B2 of height H is placed at a distance L_1 from a façade surface Γ_{F1} of infinite height. The barriers are separated at a distance W apart and divide the ground surface Γ_G into Γ_{G1}, Γ_{G2} and Γ_{G3}. For the configuration of a street canyon, a further façade Γ_{F2} (not shown in the diagram) is placed at a distance L_2 at the right side of B2.

sons of the ray model with indoor experimental results. Finally, a discussion and concluding remarks are discussion and concluding remarks are offered in Sec. V.

II. THEORETICAL FORMULATIONS

A. Parallel noise barriers in front of tall buildings

In the first case, we study a pair of parallel barriers placed in front of a row of tall buildings, exploring a typical urban situation in a high-rise city. Modeling this scenario, the row of tall buildings is replaced by a flat semi-infinite plane Γ_{F1} perpendicular to the ground Γ_G. To facilitate the numerical analysis, a rectangular coordinate system is chosen, where Γ_{F1} lies at the $x=0$ plane and Γ_G lies at the plane of $z=0$. The origin is located on the ground surface at the bottom of the façade. A pair of barriers is placed parallel to Γ_{F1}, where the near-side barrier B1 is located at a horizontal distance L_1 from the origin. The far-side barrier B2 is located at a horizontal distance W from B2. The distances W_1 and W_2 are, respectively, the horizontal distances measured from the source to B1 and B2. Figure 1 shows a schematic diagram of the problem considered in this case and the coordinate system used in the analysis.

We treat all boundaries in our formulation, i.e., the pair of barriers, the building façade and the ground, as made of perfectly reflecting surfaces. This assumption is justifiable as most surfaces in urban environments are acoustically hard. To simplify the analysis, we do not consider the effect of diffusion from boundary surfaces. We also limit our consideration to the following problem. The source $\mathbf{S}_0$ and receiver $\mathfrak{R}$ are located at the same vertical plane at $y=0$. The respective coordinates of the source and receiver are given by $\mathbf{S_o}\equiv(x_s,0,z_S)$ and $\mathfrak{R}\equiv(x_R,0,z_R)$, where $x_s\equiv(L_1+W_1)$ and the subscripts S and R are used to represent the corresponding parameters for the source and receiver, respectively.

The pair of barriers and the building façades are infinitely long and aligned along the y axis. With these assumptions, the contributions from the top edge of the façade and the vertical edges at the ends of the barriers are excluded from our analysis. Parallel barriers, which are used to shield traffic noise in high-rise cities, are built to the same height in most practical situations. Therefore, we only consider the case when they have the same height, $H\geqslant z_s$. It is straightforward to extend the current work to study the pair of barriers, B1 and B2, with different heights. We point out that the top edges of B1 and B2 are denoted by $\mathbf{E1}\equiv(L_1,y,H)$ and $\mathbf{E2}\equiv(L_1+W,y,H)$, respectively.

B. Image source model

As the source $\mathbf{S}_0$ is placed between the parallel surfaces, a row of image sources are formed as follows. An image source $\mathbf{S}_{-1}$ is created because of the reflection of the source $\mathbf{S}_0(\equiv\mathbf{S}_1)$ on the surface of B1. Image sources $\mathbf{S}_2$ and $\mathbf{S}_3$ are then formed because of the reflection of the image sources $\mathbf{S}_1$ and $\mathbf{S}_{-1}$ on the surface of B2. Next, image sources $\mathbf{S}_{-2}$ and $\mathbf{S}_{-3}$ are created, and so on, see Fig. 2. We see that a series of image sources $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$,… is generated to the right side of B1. The image source $\mathbf{S}_2$ creates the ray that hits B2 before it reaches the receiver $\mathfrak{R}$. The image source $\mathbf{S}_3$ is formed for the ray that hits B1 and then B2 before it reaches $\mathfrak{R}$. The procedure continues for determining the ray paths traced by other image sources, $\mathbf{S}_3$, $\mathbf{S}_4$, and so on. Similarly, image sources $\mathbf{S}_0$, $\mathbf{S}_{-1}$, $\mathbf{S}_{-2}$, etc., are constructed at the left side of B2. We note here that the image source $\mathbf{S}_0(\equiv\mathbf{S}_1)$ is the source itself. It is included in this series for facilitating notations used in the subsequent analysis. Image source $\mathbf{S_{-1}}$ produces a ray that hits B1 before it reaches $\mathfrak{R}$. Image source $\mathbf{S_{-2}}$ generates the ray that hits B2 and then B1 before arriving at $\mathfrak{R}$. Again, image sources $\mathbf{S_{-3}}$, $\mathbf{S_{-4}}$, and so on are

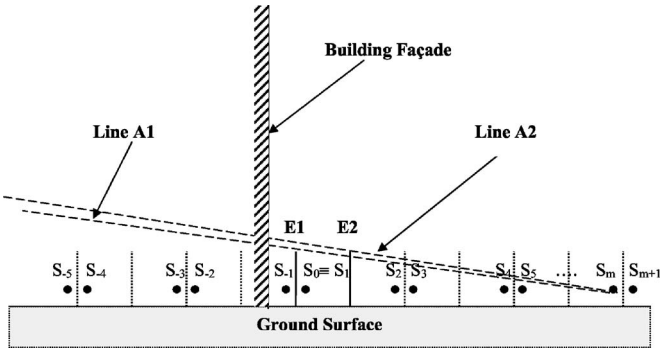


FIG. 2. Schematic diagram of the image sources $\mathbf{S}_{\pm}$ formed on reflections from the pair of parallel barriers. Lines A1 and A2 are drawn, forming a region where an image source $\mathbf{S_m}$ can reach an image receiver. E1 and E2 are the respective edges of the near-side and far-side barriers. A receiver is located in the illumination zone due to the image source S_{m} if it lies above Line A2. It is situated in the penumbra region if the receiver is located between Lines A1 and A2. If the receiver stays below Line A1, then it is located in the shadow zone.

formed in an analogous manner. For convenience, we denote these as $(\mathbf{S_0},\mathbf{S_{-1}},\mathbf{S_{-2}},\ldots)\in\mathbf{S}_-$ and $(\mathbf{S_1},\mathbf{S_2},\mathbf{S_3},\ldots)\in\mathbf{S}_+$.

There are two analogous series of image sources $(\mathbf{\Lambda_{-1}},\mathbf{\Lambda_{-2}},\mathbf{\Lambda_{-3}},\ldots)\in\mathbf{\Lambda}_-$ and $(\mathbf{\Lambda_1},\mathbf{\Lambda_2},\mathbf{\Lambda_3},\ldots)\in\mathbf{\Lambda}_+$ that can also be identified because of the presence of a reflecting ground. These ground reflected image sources are not shown in Fig. 2 for reasons of clarity. The rays of these image sources hit the ground once before they reach $\mathfrak{R}$. The locations of these image sources, $\mathbf{S}_{\pm m}=(x_{\pm m},0,z_s)$ and $\mathbf{\Lambda}_{\pm m}=(x_{\pm m},0,-z_s)$, can be determined from the geometrical configuration of the problem, where the x coordinates of these image sources are given by

$$x_{\pm m}=L_1\pm(mW-W_{(m)})\quad\text{for } m=1,2,3,\ldots, \tag{1a}$$

where

$$W_{(m)}=\begin{cases}W_2 & \text{if } m \text{ is odd}\\ W_1 & \text{if } m \text{ is even}\end{cases} \tag{1b}$$

and $x_0=x_1$. For the image source method, the total sound field at a particular receiver location is a coherent summation of contributions from all these image sources. The main task, however, is to identify all the possible image sources and calculate the corresponding contributions.

A possible sound ray from an image source is determined by establishing a valid transmission path linking the source and receiver. These possible ray paths may be obtained through direct transmission, reflections from the boundary surfaces, and diffractions at the top edge of the barrier. They can also be any combination of reflections and diffractions with different orders. In summary, the total sound field is given by

$$p(\mathbf{S},\mathfrak{R})=P_{\mathrm{direct}}+P_{\mathrm{diffract}}, \tag{2}$$

where P_{direct} takes into account for both direct transmission and the reflections from the boundary surfaces from the source to receiver. The term, P_{diffract}, is the total contributions from the diffracted ray paths. To reduce the complexity in the determination of all possible ray paths, we only consider the diffraction terms with the first order. Any higher order terms are ignored.[9,10] This assumption is justifiable because the contributions from the diffracted sound waves are generally much less than from the direct and reflected waves. By this simplification, the contributions can be classified into primary and secondary groups in the following sections.

C. Illumination zone

The primary group of contributions is generated by image sources where they can be "viewed" by the receiver $\mathfrak{R}$ in a direct line of sight, i.e., the effect of the diffraction at the barrier edges does not play a role in calculating the sound field. The ray paths in this group are established either as a direct transmission or multiple reflections from boundary surfaces.

We are interested in the situation when the receiver is located behind the near-side barrier but in front of the façade, i.e., $0\leqslant x_R\leqslant L_1$ and $z_R\geqslant 0$. With these restrictions, only image sources $\mathbf{S}_m$ and $\mathbf{\Lambda}_m$ (for a positive integral value of m), located at the right side of B1, may "see" the receiver directly. Since the parallel barriers have a finite height of H, further conditions are needed in order to determine whether a ray emanating from image sources $\mathbf{S}_m$ or $\mathbf{\Lambda}_m$ can reach a receiver. This is because the surface of B1 may be too high to shield a direct sight-line contact, and B2 may be too low to provide a surface for the reflection of an incoming ray. By extending a line A_1 joining $\mathbf{S}_m$ to **E1**, and another line A_2 linking $\mathbf{S}_m$ and **E2**, see Fig. 2, we can determine the limit of the regions where the ray path from $\mathbf{S}_m$ can reach the receiver. We can classify that the receiver is situated in the illumination zone due to the image source, S_m if it is located above Line A2. The receiver is positioned in the penumbra region if it is located between Line A1 and A2. Finally, the receiver lies in the shadow zone if it is placed below Line A1.

Because of the reflection from the buildings, additional ray paths can also be constructed for the situation when the sound rays hit the façade surface before reaching the receiver. In this case, it is convenient to consider the reception point consisting of a pair of the image receivers, $\mathfrak{R}_{\pm1}\equiv(\pm x_{\mathrm{R}},0,z_{\mathrm{R}})$, where $\mathfrak{R}_1$ corresponds to the direct arrival of a sound ray and $\mathfrak{R}_{-1}$ is the reflection of sound from the façade. With the introduction of $\mathfrak{R}_{\pm1}$, all possible ray paths can be found by connecting $\mathbf{S}_m$ or $\mathbf{\Lambda}_m$ to $\mathfrak{R}_{\pm1}$. The presence of $\mathbf{S}_m$ in the total sound field can be determined by extending the lines joining $\mathbf{S}_m$ and $\mathfrak{R}_{\pm1}$ to yield the following additional condition:

$$\frac{(z_s-H)(\pm x_R-L_1-W)}{(m-1)W-W_{(m)}}\geqslant z_R-H\geqslant\frac{(z_s-H)(\pm x_R-L_1)}{mW-W_{(m)}}, \tag{3}$$

where $W_{(m)}$ is given in Eq. (1b). These two conditions correspond respectively to lines A_1 and A_2 as shown in Fig. 2. Similarly, the condition for the presence of $\mathbf{\Lambda}_m$ in the ray series can be determined as follows:

$$\frac{(H+z_s)(\pm x_R - L_1 - W)}{W_{(m)} - (m-1)W} \geqslant z_R - H \geqslant \frac{(H+z_s)(\pm x_R - L_1)}{W_{(m)} - mW}. \tag{4}$$

In Eqs. (3) and (4), a positive sign for x_R represents the condition for the receiver $\mathfrak{R}_1$, and a negative sign gives the corresponding condition for the image receiver $\mathfrak{R}_{-1}$.

We remark that there is no primary contribution of the total sound fields when the receiver is located in the shadow zone when $z_R \leqslant H$. The sound field from the primary group can be obtained by summing all possible contributions

$$P_{\text{direct}} = \sum_{m=m_<}^{m_>} G_d(\mathbf{S_m}, \mathfrak{R}_{\pm 1}) + \sum_{m=m_<}^{m_>} G_d(\mathbf{\Lambda_m}, \mathfrak{R}_{\pm 1}). \tag{5}$$

In the above equation, $\mathfrak{R}_1$ and $\mathfrak{R}_{-1}$ are substituted, in turn, obtaining four separate series. The term $\mathfrak{R}_1$ corresponds to the rays linking the image sources to the receiver, and $\mathfrak{R}_{-1}$ represents the contributions of those rays that have an extra reflection from the building façade before they reach the receiver. For a given source and receiver position, a range of m (from m_1 to m_2, say) can be determined from Eqs. (3) and (4). For instance, this information can be obtained for $\mathbf{S}_1$ and $\mathfrak{R}_1$ from Eq. (3) with the positive sign taken. We can see that $m_<$ and $m_>$ are given by

$$m_< = \min(m_1, m_2), \tag{6a}$$

$$m_> = \max(m_3, m_4), \tag{6b}$$

where m_1 is the smallest odd integer just greater than $W_2/W + (H+z_s)(x_R - L_1)/[W(z_R - H)]$, m_2 is the smallest even integer just greater than $W_1/W + (H+z_s)(x_R - L_1)/[W(z_R - H)]$, m_3 is the largest odd integer just smaller than $W_2/W + (z_s - H)(x_R - L_1 - W)/[W(z_R - H)] + 1$, and m_4 is the largest even integer just smaller than $W_1/W + (z_s - H)(x_R - L_1 - W)/[W(z_R - H)] + 1$.

Different ray paths have different $m_<$ and $m_>$ but they can be determined straightforwardly from the corresponding conditions given in Eqs. (3) and (4). For example, suppose that we have the source located at $\mathbf{S}_1 = (7.5, 0, 0.25)$, the near-side barrier is located at $L_1 = 5$ m from the façade, the parallel barrier has a height $H = 2.5$ m, the source is located at $W_1 = 2.5$ m from the near-side barrier, and $W_2 = 7.5$ m from the far-side barrier. If the receiver is located in the illumination zone at $\mathfrak{R} \equiv \mathfrak{R}_1 = (1.0, 0, 10)$, then we can determine that $m_>$ and $m_<$ are all equal to 1 for all series given in the summation of Eq. (5). In other words, only $\mathbf{S}_1$ and $\mathbf{\Lambda}_1$ can see the image receivers, $\mathfrak{R}_{\pm 1}$. On the other hand, if the receiver is located at a lower height at $\mathfrak{R} \equiv \mathfrak{R}_1 = (1.0, 0, 3.5)$, then we can show that the image sources $\mathbf{S}_2 - \mathbf{S}_5$ and $\mathbf{\Lambda_2} - \mathbf{\Lambda_5}$ can see $\mathfrak{R}_{-1}$, but only the image sources $\mathbf{S_2} - \mathbf{S_4}$ and $\mathbf{\Lambda_2} - \mathbf{\Lambda_5}$ can see $\mathfrak{R}_1$. We only sum the contributions from those terms that can establish a sight-line contact between the image sources and image receivers in the series of Eq. (5).

In general, the Green function $G_d(\mathbf{S}, \mathbf{R})$ stands for the sound field radiated directly from an arbitrary point source $\mathbf{S}$ to any reception point $\mathbf{R}$. It is computed by

$$G_d(\mathbf{S}, \mathbf{R}) = \exp(ikd)/4\pi d, \tag{7}$$

where d is the direct distance measured from the source $\mathbf{S}$ to the receiver $\mathbf{R}$. Equations (3) and (4) are applied to ensure a possible direct link between $\mathbf{S}$ and $\mathbf{R}$. No contribution from a particular image source is possible if the corresponding conditions are not satisfied. Nevertheless, only a finite number of rays is summed in Eq. (5), and contributions from image sources located at long distances from the receiver and its image are truncated because of reduced sound levels due to the effect of geometrical spreading of the sound rays.

Next, we wish to discuss the secondary group of contributions. This group of rays consists of diffractions from the edges of the parallel barriers. We reiterate that the higher order diffraction terms are generally ignored in the present analysis, but there may be various orders of reflections both before and after the diffraction at the barrier edges, $(\mathbf{E1}, \mathbf{E2}) \in \mathbf{E}$. We may treat $\mathbf{E}$ as virtual receivers that can be viewed by the image sources. At the source side of the barrier, multiple reflections can then be handled effectively by connecting image sources to $\mathbf{E}$. We shall denote these image sources by $\mathbf{S}_{\pm n}$ and $\mathbf{\Lambda}_{\pm n}$ although, strictly speaking, they are the same image sources situated at the same locations as $\mathbf{S}_{\pm m}$ and $\mathbf{\Lambda}_{\pm m}$ if $m = n$. For clarity, different subscripts are chosen to distinguish the ray paths traced by the same image sources for the direct (the subscript m is used) and diffracted (the subscript n is used) fields.

To handle the multiple reflections at the receiver side of the barrier, we find it convenient to introduce the concept of the image receiver, because E1 and E2 can be seen as two secondary sources from the receiver location. On this side of the barrier, multiple reflections can be established when the sound rays leave $\mathbf{E}$ and hit the outer surface of the near-side barrier and the façade surface before they reach the receiver. The ray paths originating from the barrier edges can then be determined easily by linking $\mathbf{E}$ with the receiver and its images.

If $z_R > H$, there are only two image receivers – the receiver $\mathfrak{R}_1$ and its image $\mathfrak{R}_{-1}$. These locations were discussed in the preceding paragraphs when we addressed the contributions of direct waves from the image sources. Both barrier edges, $\mathbf{E}_1$ and $\mathbf{E}_2$, can see $\mathfrak{R}_{\pm 1}$. The image sources on the right side of B1, $\mathbf{S}_n$ and $\mathbf{\Lambda}_n$, can see $\mathbf{E}_1$. On the other hand, only the image sources on the left side of B2, $\mathbf{S}_-$ and $\mathbf{\Lambda}_-$ can see $\mathbf{E}_2$. In addition to these image receivers, we can also identify another pair of image receivers, $\Psi_{\pm 1}$, because of the presence of the reflecting ground between Γ_{F1} and B1. Only the edge of the near-side barrier, $\mathbf{E1}$, can see $\Psi_{\pm 1}$. The surface of B1 shields the sound rays connecting E2 to $\Psi_{\pm 1}$. The total diffracted field in this region, $z_R > H$ and $L_1 > x_R > 0$, can then be written as

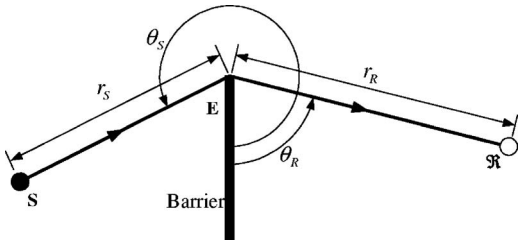


FIG. 3. Schematic diagram of the geometrical configuration for the diffraction of sound by a thin barrier.

$$P_{\text{diffract}} = \sum_{n=1}^{\infty} [G_f(\mathbf{S_n}, \mathfrak{R}_{\pm 1}|\mathbf{E_1}) + G_f(\mathbf{\Lambda_n}, \mathfrak{R}_{\pm 1}|\mathbf{E_1})] + \sum_{n=1}^{\infty} [G_f(\mathbf{S_n}, \mathbf{\Psi}_{\pm 1}|\mathbf{E_1}) + G_f(\mathbf{\Lambda_n}, \mathbf{\Psi}_{\pm 1}|\mathbf{E_1})] + \sum_{n=0}^{\infty} [G_f(\mathbf{S_{-n}}, \mathfrak{R}_{\pm 1}|\mathbf{E_2}) + G_f(\mathbf{\Lambda_{-n}}, \mathfrak{R}_{\pm 1}|\mathbf{E_2})], \tag{8}$$

where $\mathfrak{R}_1$, $\mathfrak{R}_{-1}$, $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_{-1}$ are substituted in turn. We remark that the third series starts with $n=0$ accounting for the ray path that reaches **E2** directly from image sources $\mathbf{S}_0(\equiv\mathbf{S_1})$ and $\mathbf{\Lambda}_0(\equiv\mathbf{\Lambda}_1)$. The Green function $G_f(\mathbf{S},\mathfrak{R}|\mathbf{E})$ is the solution for sound diffracted by a hard noise barrier with edge **E** for a source **S** and receiver $\mathfrak{R}$. It is given by[11]

$$G_f(\mathbf{S},\mathfrak{R}|\mathbf{E}) = \left(\frac{e^{i\pi/4}}{\sqrt{2}}\right)\left[\frac{e^{ik(d_S+d_R)}}{4\pi(d_S+d_R)}\right][A_D(X_+) + A_D(X_-)], \tag{9}$$

where d_S and d_R are the respective distances from the source and receiver to the diffraction point. The function, $A_D(X)$ is the diffraction integral[11] given by

$$A_D(X) = \text{sgn}(X)[f(|X|) - ig(|X|)], \tag{10}$$

where sgn(X) is the sign function, and $f(X)$ and $g(X)$ are the auxiliary Fresnel functions[12] of real argument X. The arguments of the diffraction integral, X_+ and X_-, are determined by

$$X_\pm = X(\phi_R \pm \phi_S), \tag{11a}$$

where

$$X(\Phi) = \left[-2\cos\left(\frac{\Phi}{2}\right)\right]\sqrt{\frac{2\cdot d_S\cdot d_R}{\lambda(d_S+d_R)}}, \tag{11b}$$

λ is the wavelength of the diffracted sound, and the angles ϕ_R and ϕ_S are defined in the surface of the screen as shown in Fig. 3. The argument Φ in Eq. (11b) is $(\phi_R \pm \phi_S)$ for $X_\pm$.

D. Shadow zone

The receiver is located at the shadow zone if $z_R \leqslant H$. In this case, the façade and the outer face of B1 form a pair of reflecting surfaces. It is possible to generate a row of image receivers, $(\mathfrak{R}_{-j}, \mathfrak{R}_j) \in \mathfrak{R}$. The positions of these image receivers, $\mathfrak{R}_{\pm j} \equiv (\xi_{\pm j}, 0, z_R)$ for $j=1,2,3,\ldots$, can be determined readily to give

$$\xi_{\pm j} = \pm jL_1 - D_{(j)}, \tag{12}$$

where

$$D_{(j)} = \begin{cases} x_R' & \text{if } q \text{ is odd} \\ x_R & \text{if } q \text{ is even} \end{cases} \tag{13}$$

and $x_R'(\equiv L_1 - x_R)$ is the horizontal distance of the receiver position measured from B1.

Only those image receivers situated on the left side of B1, …, $\mathfrak{R}_{-3}, \mathfrak{R}_{-2}, \mathfrak{R}_{-1}$ and the receiver $\mathfrak{R}_0 \equiv \mathfrak{R}_1$ itself, can see $\mathbf{E}_1$. Image receiver $\mathfrak{R}_{-1}$ traces the following sound ray: leaves $\mathbf{E}_1$, hits the building façade and arrives at the receiver. Image receiver $\mathfrak{R}_{-2}$ follows the ray path: leaves $\mathbf{E}_1$, hits the building façade, the outer face of B1 and arrives at the receiver. This process repeats for other image receivers $\mathfrak{R}_{-3}, \mathfrak{R}_{-4}, \ldots$ and so on. Since there is a reflecting ground between Γ_{F1} and B1, another row of image receivers, $\mathbf{\Psi}_{\pm j} \equiv (\xi_{\pm j}, 0, -z_R)$ for $j=1,2,3,\ldots$, is formed with $\xi_{\pm j}$ given by Eq. (12). Again, only the image sources $\mathbf{\Lambda}_n$ (for $n=1,2,3$) can see **E1**. It is worth noting that **E2** can see none of these image receivers in the region $z_R < H$ and $L_1 > x_r > 0$, because the near side barrier B1 prevents any contact without a further diffraction at its edge. Hence the total diffracted field is simply

$$P_{\text{diffract}} = \sum_{n=1}^{\infty}\sum_{j=0}^{\infty} [G_f(\mathbf{S_n}, \mathfrak{R}_{-j}|\mathbf{E1}) + G_f(\mathbf{\Lambda_n}, \mathfrak{R}_{-j}|\mathbf{E1})] + \sum_{n=1}^{\infty}\sum_{j=0}^{\infty} [G_f(\mathbf{S_n}, \mathbf{\Psi}_{-j}|\mathbf{E1}) + G_f(\mathbf{\Lambda_n}, \mathbf{\Psi}_{-j}|\mathbf{E1})], \tag{14}$$

where the inner series for the index j starts from 0 to include the terms for the rays connecting from E1 to the receiver and its image on reflection from the ground surface. Again, the Green function $G_f(\mathbf{S},\mathbf{R}|\mathbf{E})$ represents the solution for the diffracted sound field. The diffracted terms are calculated for different combinations of image sources, n and image receivers, j. We also take $\mathfrak{R}_0 \equiv \mathfrak{R}_1$ and $\mathbf{\Psi}_0 = \mathbf{\Psi}_1$ in writing a more compact series given in Eq. (14).

E. Parallel barriers in a street canyon

In the second case, we consider that the pair of parallel barriers, B1 and B2, is placed in a street lined with two parallel rows of tall buildings. The tall buildings are replaced by two flat façade surfaces, Γ_{F1} and Γ_{F2}, perpendicular to the ground. The geometrical configuration of this problem is similar to that described in Sec. II. Again, the separation between the pair of barriers is W and the source is located at W_1 from B1 and W_2 from the far-side barrier B2. The near-side barrier B1 is situated at a distance of L_1 from Γ_{F1}, and B2 is located at a distance of L_2 from Γ_{F2}. The second case is a relatively more complicated problem than the first because of the presence of the additional façade forming the so-called street canyon. However, the total sound field is computed according to Eq. (2) with different P_{direct} and P_{diffract} from those derived in the last section.

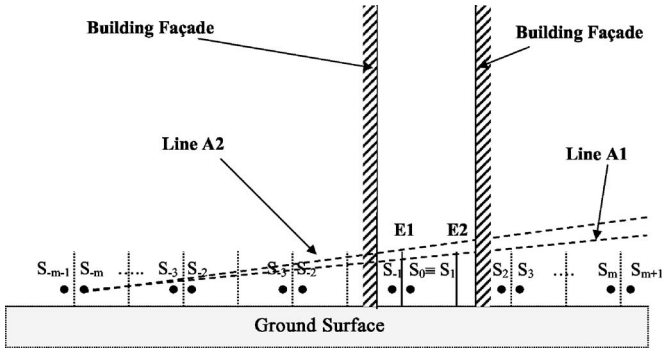


FIG. 4. Schematic diagram of the image sources $\mathbf{S}_{\pm}$ formed on reflections from the pair of parallel barriers. Lines A1 and A2 are drawn, forming a region where an image source $\mathbf{S}_{-\mathbf{m}}$ can reach an image receiver. E1 and E2 are the respective edges of the near-side and far-side barriers. A receiver is located in the illumination zone due to the image source $\mathbf{S}_{-\mathrm{m}}$ if it lies above Line A2. It is situated in the penumbra region if the receiver is located between Lines A1 and A2. If the receiver stays below Line A1, then it is located in the shadow zone.

From the modeling point of view, this problem is the same as the case described in Sec. II A. The only exception is that a further façade surface, Γ_{F2}, is added at the right side of B2. Indeed, when the source is placed at $z_s < H$, and the receiver is located in the region $0 \leqslant x_R \leqslant L_1$ and $z_R \leqslant H$, the sound field in the shadow zone is the same as the first case provided that higher order diffraction terms are ignored. Hence, our focus in this section is the determination of the sound field in the illumination zone.

In this region, where $0 \leqslant x_R \leqslant L_1$ and $z_R > H$, we can again identify a row of image receivers $[(\mathfrak{R}_{-j}, \mathfrak{R}_j) \in \mathfrak{R}$ for $j=1,2,3,\ldots]$ due to the presence of Γ_{F1} and Γ_{F2}. These image receivers are located at $\mathfrak{R}_{\pm j} \equiv (\xi_{\pm j}, 0, z_R)$, with $\xi_{\pm j}$ given by

$$\xi_{\pm j} = \pm j(L_1 + L_2 + W) - D_{(j)}, \tag{15}$$

where

$$D_{(j)} = \begin{cases} \bar{x}_R & \text{if } j \text{ is odd} \\ x_R & \text{if } j \text{ is even} \end{cases}, \tag{16}$$

and $\bar{x}_R$ is the horizontal distance of the receiver position measured from Γ_{F2}, i.e., $\bar{x}_R = L_1 + L_2 + W - x_R$.

The primary group of contributions can be determined by identifying the series of image sources and image receivers that can see each other. The image sources on the right side of B1, i.e. $(\mathbf{S}_m, \mathbf{\Lambda}_m) \in \mathbf{S}_+$, can see the image receivers $\mathfrak{R}_{-j} \in \mathfrak{R}_-$ located on the left side of Γ_{F2}. The additional conditions for establishing these sight lines are given in the analog forms as Eqs. (3) and (4), except that $\pm x_{\mathrm{R}}$ in these two equations is replaced by ξ_{-j} for $j=0,1,2,\ldots$, in turn. Similarly, the image sources located on the left side of B2, $(\mathbf{S}_{-m}, \mathbf{\Lambda}_{-m}) \in \mathbf{S}_-$, can view the image receivers $\mathfrak{R}_j \in \mathfrak{R}_+$ situated on the right side of Γ_{F2}. By considering Fig. 4, the additional conditions for the image sources $\mathbf{S}_{-m}$ and Λ_{-m} are determined, respectively, as follows:

$$\frac{(z_s - H)(\xi_j - L_1)}{mW - W_{(m)}} \geqslant z_R - H \geqslant \frac{(z_s - H)(\xi_j - L_1 - W)}{(m+1)W - W_{(m)}} \tag{17}$$

and

$$\frac{(z_s + H)(\xi_j - L_1)}{mW - W_{(m)}} \geqslant z_R - H \geqslant \frac{(z_s + H)(\xi_j - L_1 - W)}{(m+1)W - W_{(m)}}. \tag{18}$$

We can also determine the ranges of m $[m_1(j)$ and $m_2(j)]$ for different values of j. With this information, we can express immediately the contribution of the total field from this group of image sources as a double summation of direct fields from a combination of image sources $\mathbf{S}_{\pm}$ of the index m to the image receivers $\mathfrak{R}_{\pm}$ of the index j:

$$\begin{aligned} P_{\mathrm{direct}} = & \sum_{j=0}^{\infty} \sum_{m=m_<(j)}^{m_>(j)} G_d(\mathbf{S}_m, \mathfrak{R}_{-j}) + \sum_{j=0}^{\infty} \sum_{m=m_1(j)}^{m_>(j)} G_d(\mathbf{\Lambda}_m, \mathfrak{R}_{-j}) \\ & + \sum_{j=1}^{\infty} \sum_{m=m_<(j)}^{m_>(j)} G_d(\mathbf{S}_{-m}, \mathfrak{R}_j) \\ & + \sum_{j=1}^{\infty} \sum_{m=m_<(j)}^{m_2(j)} G_d(\mathbf{\Lambda}_{-m}, \mathfrak{R}_j). \end{aligned} \tag{19}$$

The first two series start at $j=0$ to include the receiver in the row of image receivers $\mathfrak{R}_-$. The existence of a particular ray path is determined by satisfying the corresponding conditions given in Eqs. (17) and (18). The Green function G_d of the direct field given above can be calculated according to Eq. (7).

The contribution of the diffracted field can also be expressed in a rather similar form except that the pair of receivers $\mathbf{R}_{\pm 1}$ in Eq. (8) is replaced by a series of $\mathfrak{R}_{\pm j}$for $j=1,2,3,\ldots$. An additional sum is needed for the contributions from each of the image sources to each of the image receivers as follows:

$$\begin{aligned} P_{\mathrm{diffract}} = & \sum_{j=1}^{\infty} \sum_{n=1}^{\infty} [G_f(\mathbf{S_n}, \mathfrak{R}_{\pm j} | \mathbf{E_1}) + G_f(\mathbf{\Lambda_n}, \mathfrak{R}_{\pm j} | \mathbf{E_1})] \\ & + \sum_{j=0}^{\infty} \sum_{n=1}^{\infty} [G_f(\mathbf{S_n}, \mathbf{\Psi}_{-j} | \mathbf{E_1}) + G_f(\mathbf{\Lambda_n}, \mathbf{\Psi}_{-j} | \mathbf{E_1})] \\ & + \sum_{j=1}^{\infty} \sum_{n=0}^{\infty} [G_f(\mathbf{S_n}, \mathfrak{R}_{\pm j} | \mathbf{E_2}) + G_f(\mathbf{\Lambda_n}, \mathfrak{R}_{\pm j} | \mathbf{E_2})], \end{aligned} \tag{20}$$

where $G_f(\mathbf{S}, \mathbf{R} | \mathbf{E})$ is the Green function for the diffraction of sound from a source $\mathbf{S}$ to a receiver $\mathbf{R}$ at an edge $\mathbf{E}$. In Eq. (20), the term involving $\mathfrak{R}_{\pm j}$ is substituted with $\mathfrak{R}_j$ and $\mathfrak{R}_{-j}$ in turn. The second ray series of the above equation starts from $j=0$ and includes the image receiver $\mathbf{\Psi}_0 (\equiv \mathbf{\Psi}_1)$.

III. NUMERICAL COMPUTATIONS

The ray (image source) model provides an effective and efficient methodology to assess the acoustic performance of parallel barriers in front of a tall building. Our approach is a direct extension of the earlier developments. In the absence of a building façade, our model is analogous to the prediction methods for parallel barriers proposed by Panneton *et al.*[9] and Muradali.[10] Without the far side barrier, our model is identical to that proposed by Li and Tang[8] for predicting the

acoustic performance of a single barrier placed in front of a building façade. Prior numerical analyses were conducted to compare our model with the published simulation data.[8–10] These comparison results, which show good agreement, serve to confirm the validity of our numerical model and they are not shown here for succinctness.

Next, the numerical results predicted by the image source model are compared with that evaluated by a wave-based numerical formulation. As the geometrical configuration of the present study is an external problem, the boundary element method (BEM) is an appropriate approach for the purpose of validation. In fact, the BEM formulation has been extensively used to study the physical phenomenon of outdoor sound propagation in an irregular terrain and to study the acoustic performance of noise barriers. In the present study, the height of the building façade is taken as 25 m, which should be sufficiently tall to ensure that any contributions due to the diffraction of the sound at the façade's top edge are negligibly small. Generally speaking, more elements are needed at higher frequencies in order to represent the boundary surfaces. They are partitioned with at least ten elements per wavelength in the present study. This requirement ensures a higher degree of accuracy for the numerical results.

A realistic outdoor configuration is used in our analysis. The near-side and far-side barriers, B1 and B2, have identical heights of 2.5 m, and are situated at 5 and 15 m in front of building façade Γ_{F1}, respectively. A further parallel building façade, Γ_{F2}, is placed 20 m from Γ_{F1} in the case of street canyons. In both cases, an omni-directional noise source is located between the parallel barriers at 0.25 m above the ground and 7.5 m in front of the façade, Γ_{F1}. In other words, the source is located 2.5 m from B1 and 12.5 m from B2. The reception points are chosen at 1 m away from Γ_{F1} and with different heights above the ground for the presentation of the numerical results. The choice of these source/receiver geometries allows our numerical models to examine the sound fields in different areas of interest—shadow, penumbra and illumination zones, respectively.

It is of interest to point out that the BEM formulation generally requires relatively higher computational resources especially at high frequencies. To put it in context, let us consider a critical example with a source operating at 5 kHz. These are the highest frequencies we showed in the numerical simulations using the BEM formulations. A FORTRAN program was used to compute the sound field for a pair of parallel barriers locating in a street canyon. A total in excess of 9000 boundary elements (about 3800 for each façade surface and 750 for each barrier) are needed in this case. It took over 30 h for a typical desktop computer (a 2 GHz processor and 1 Giga bytes of RAM memory) to solve the set of simultaneous equations by using a standard matrix method. The computational time for the BEM formulation will increase exponentially for the source frequency extending beyond 5 kHz and for the façade surfaces reaching higher than 25 m. On the other hand, we have developed a suite of MATLAB program for the image source model. The MATLAB program was compiled and was used to predict the sound fields for the two urban scenarios as described in the earlier sections. Using the same desktop computer, the computational time was about 10 min for a source frequency of 5 kHz. It took nearly the same time to compute the sound fields for all other frequencies of interests. The computational time can be reduced further if a FORTRAN program was developed and used to predict the sound fields by the image source model. However, there is no attempt to follow along this route for minimizing the computational time in the present study.

We note that relatively less image sources are required if the receiver is situated in the shadow zone close to the ground. In this situation, only a few reflections from the boundary surfaces are normally required to obtain a set of converged numerical results. On the other hand, more image sources are needed if the receiver is located higher above the ground in the penumbra region and the illumination zone. Although it is possible to optimize the number of image sources required for different receiver heights, we find that variations in the sound pressure level generally become stable after about 60 reflections over 1/3 octave bands varying from 125 Hz to 8 kHz. For simplicity, we set the maximum orders of reflections to be no more than 100 for both sides of the barriers at all receiver locations. This will lead to a simpler program with an acceptable numerical accuracy at the expense of a modest increase in the overall computational time.

Four receiver locations are chosen for comparison in these two cases—parallel barriers in front of building facades and in a street canyon. For the first two locations, Locations 1 and 2, the receiver $\mathfrak{R}$ is placed at points (1, 0, 1) and (1, 0, 2), respectively. Both of these locations are situated in the shadow zone of B1. In Location 3, the receiver $\mathfrak{R}$ is placed at (1, 0, 5) which is close to the direct line of sight of E1 in the penumbra region. The receiver, which is positioned at $\mathfrak{R}=(1,0,10)$, is illuminated directly by the source in Location 4.

We introduce a term known as the insertion loss (*IL*) to facilitate the presentation of the numerical results. It is defined as follows:

$$IL = 20\log_{10}\left(\frac{P_w}{P_{w/o}}\right), \tag{21}$$

which is essentially the difference in sound pressure levels before and after the installation of parallel barriers. Figures 5 and 6 show the predicted *IL* spectra at Locations 1 and 3 for the pair of parallel barriers placed in front of the façade surface. The general trend of the *IL* spectra such as the positions of the peaks and dips predicted by the image source method coincides well with those predicted by the BEM formulation. Although there are noticeable discrepancies for the predicted magnitudes at some frequencies between these two methods, these differences will be less significant if the results are averaged over a frequency band. The comparison of *IL* spectra in 1/3 octave bands will be shown in the next section. It is remarkable as shown in Figs. 5 and 6 that there is a higher level of fluctuation in *IL* at high frequencies as predicted by both methods. This can be explained by using the image source method as follows. The total sound field is computed by summing the contributions coherently from a finite number of image sources produced by multiple reflec-

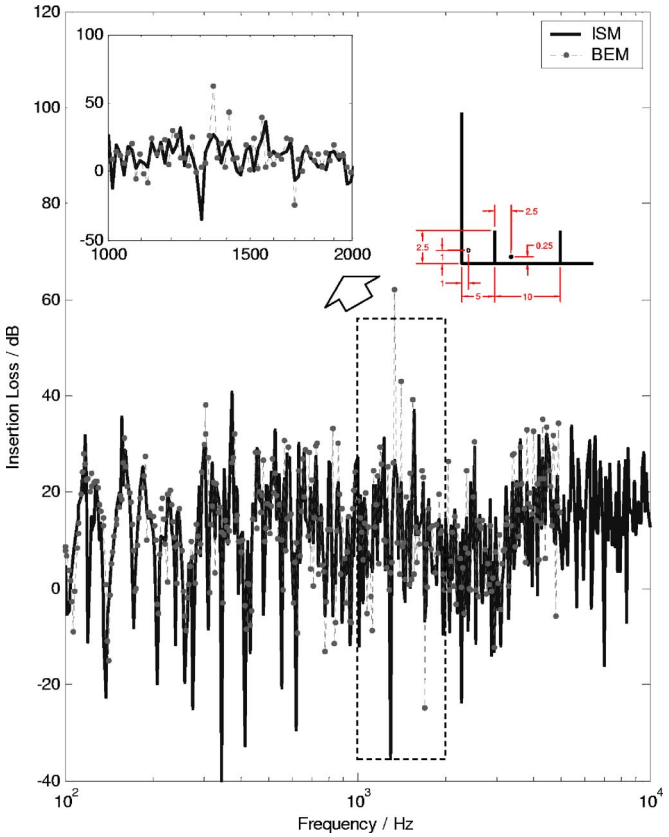


FIG. 5. (Color online) The spectrum of insertion losses at location $\mathfrak{R}=(1,0,1)$, with parallel barriers placed in front of a building façade. The solid line represents predictions by the image source method (ISM), and the dashed line with dots represents numerical predictions based on the boundary element method (BEM).

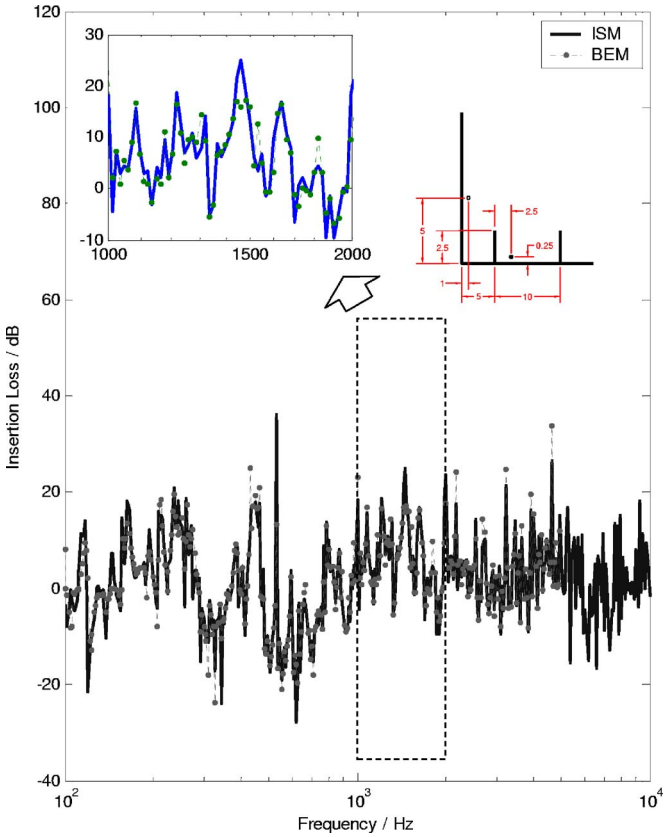


FIG. 6. (Color online) Same caption as Fig. 5 except that the receiver is located at $\mathfrak{R}=(1,0,5)$.

tions of the boundary surfaces. The levels of variation in *IL* reflect the phenomenon of interference due to the contributions from different image sources which is more pronounced at higher frequencies.

Figures 7 and 8 show the corresponding predicted results for the case of the street canyon with the receiver at Locations 2 and 4. The sound fields predicted by the image source method agree reasonably well with those predicted by the BEM formulation. The trends of the peak and dip of the *IL* spectrum predicted by the image source method are generally in agreement with those predicted by the BEM formulation. Again, considerable discrepancies are observed in predicting the magnitudes of *IL* by using the two different numerical methods. In comparison with the cases with a single façade, large fluctuations are observed in the insertion loss due to another set of image sources formed by placing a façade on the other side of the parallel barriers.

The *IL* spectrum is dependent on the geometrical configurations of the problem such as the locations of the parallel barriers and the relative positions of the source and receiver. It is possible to show that similar results can be obtained when the source and receiver are located at other positions. These results are not given here but are shown elsewhere.[13]

IV. EXPERIMENTAL VALIDATIONS

A model pair of parallel noise barriers, which was placed either in front of a façade or in a street canyon, was built at a scale of one-tenth for the present experimental study. Measured data were obtained to validate the image source model. The experimental setup is shown in Fig. 9 which reflects comparable source/receiver configurations discussed in the last section. We note that the setup for the parallel barriers in a street canyon is not shown in Fig. 9. The façade and ground surfaces are made of 8.5-mm-thick wooden boards, which were varnished to prevent sound leakage. Prior measurements[13] were conducted to measure the acoustic characteristic of the varnished wooden boards. We found that they can generally be treated as a perfectly reflecting plane. The heights of the façade and the pair of barriers are 2.44 and 0.5 m, respectively. The pair of barriers consists of two 3-mm-thick aluminum plates with lengths of 4.5 m. They are placed parallel to each other at a distance of 0.75 and 1.5 m in front of the façade.

A Tannoy speaker mounted on a long brass pipe with length of 1.5 m and diameter of 25 mm is used to simulate an omni-directional point source. Preliminary measurements are conducted to examine the directional characteristic of the point source. The measured result, not shown here for brevity, suggests that the deviation in the directivity pattern for all directions is within 1 dB for all frequencies above 250 Hz. Hence, the Tannoy sound source was placed parallel to the barriers as the most significant contributions were expected to be due to the reflections from the boundary surfaces.

In all measurements, the point source was located 1.25 m in front of the façade surface and at a height of 0.125

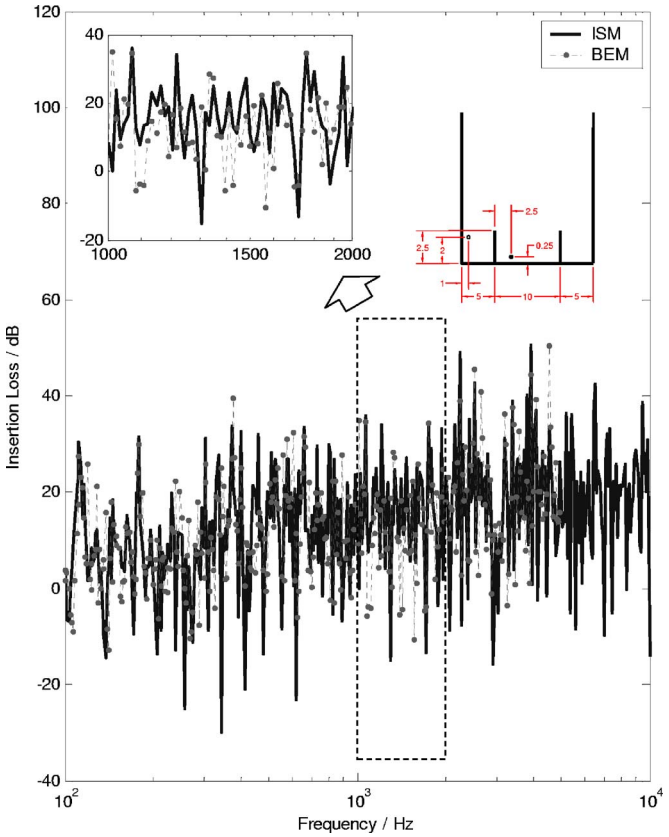


FIG. 7. (Color online) The spectrum of insertion losses (IL) at location $\mathfrak{R}=(1,0,2)$ with parallel barriers placed in a street canyon. The solid line represents predictions by the image source method (ISM) and the dashed line with dots represents numerical predictions based on the boundary element method (BEM).

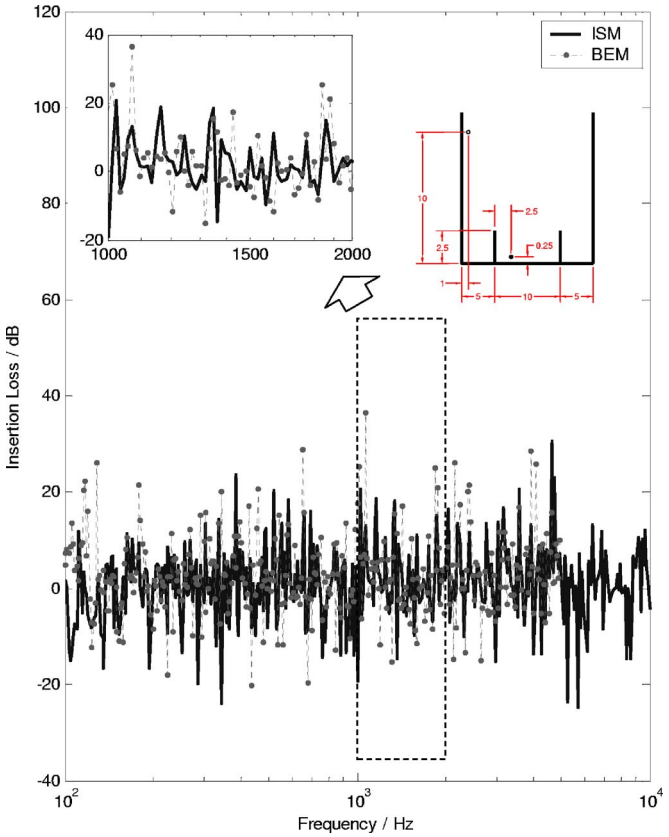


FIG. 8. (Color online) Same caption as Fig. 7 except that the receiver is located at $\mathfrak{R}=(1,0,10)$.

m above the ground. A B&K 4942 microphone, which was connected to a B&K 2671 preamplifier and a B&K NEXUS conditional amplifier, was used as the receiver. The microphone was placed at 0.123 m in front of the vertical wooden boards at various heights.

A special type of test signal called a maximum-length sequence (MLS) was employed to obtain the experimental data. The deterministic nature of the maximum-length sequence provides an excellent signal-to-noise ratio, which is ideal for the current indoor measurements. The MLS signals were generated by the MLSSA 2000 card, transferred via the built-in digital-to-analog-converter and boosted by a B&K 2713 power amplifier. The MLS signals were then connected to the Tannoy speaker, which emitted sound for measurements. As the measurements were recorded in the time domain, manipulation to eliminate unwanted reflections was also possible.

The experiments were conducted inside an anechoic chamber of dimensions 6 m $\times$ 6 m $\times$ 4 m (high). The insertion loss used for comparison was obtained by measuring the transfer function with and without the barrier. In the experimental measurements for the two cases described in Sec. II, the receiver was placed at four locations, A, B, C and D, at (0.123, 0, 0.1), (0.123, 0, 0.5), (0.123, 0, 1) and (0.123, 0, 1.5), respectively. Again, like the numerical validations, Locations A and B were situated in the shadow zone. Location C was placed in the penumbra region along the sight-line contact. Finally, Location D was chosen to be in the illumination region. We wish to point out that valid experimental data are not expected for frequencies below 500 Hz because of the size of the anechoic chamber and the scale model. Nevertheless, we show *IL* spectra with frequencies varying from 100 Hz to 10 kHz in the following plots.

Figures 10 and 11 illustrate the measured *IL* spectra at Locations B and D, respectively, for the model parallel barriers erecting in front of a building façade. Numerical simulations are also shown in these figures. For the narrow band spectra, measured data and numerical simulations show large fluctuations in the insertion loss spectra because of the complex interferences due to the contributions of different image

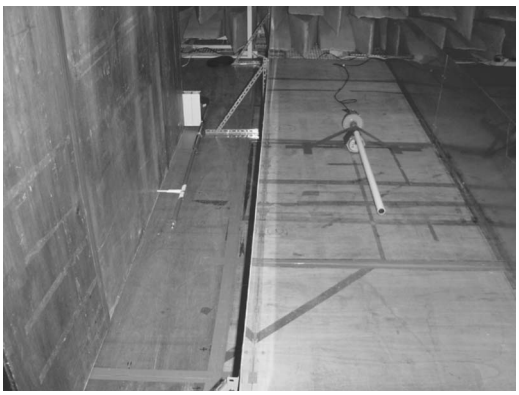


FIG. 9. The experimental setup for measurements in an anechoic chamber for a parallel barrier placed in front of a façade surface. The setup for a street canyon is not shown in the diagram.

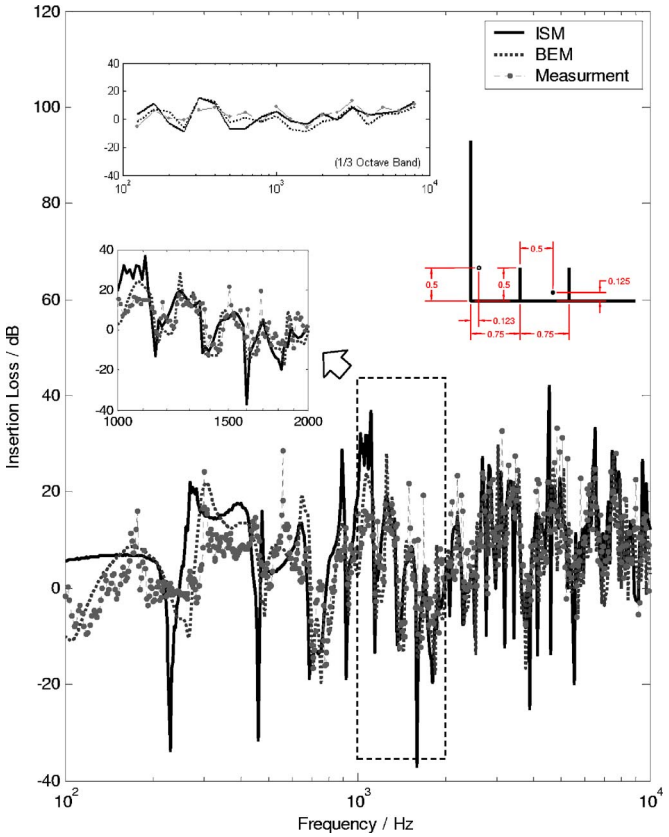


FIG. 10. (Color online) The spectrum of insertion losses at location $\mathfrak{R}=(0.123,0,0.5)$ with parallel barriers placed in front of a building façade. The solid line represents predictions by the image source method (ISM), the dashed line represents numerical predictions based on the boundary element method (BEM), and the dashed line with dots represents results from experimental measurement. Top inset figure: Comparison of the insertion loss between measured and predicted results in one-third octave band for a range of frequencies available from 125 to 8000 Hz. Bottom inset figure: Highlighting the narrow frequency spectrum going from 1000 to 2000 Hz.

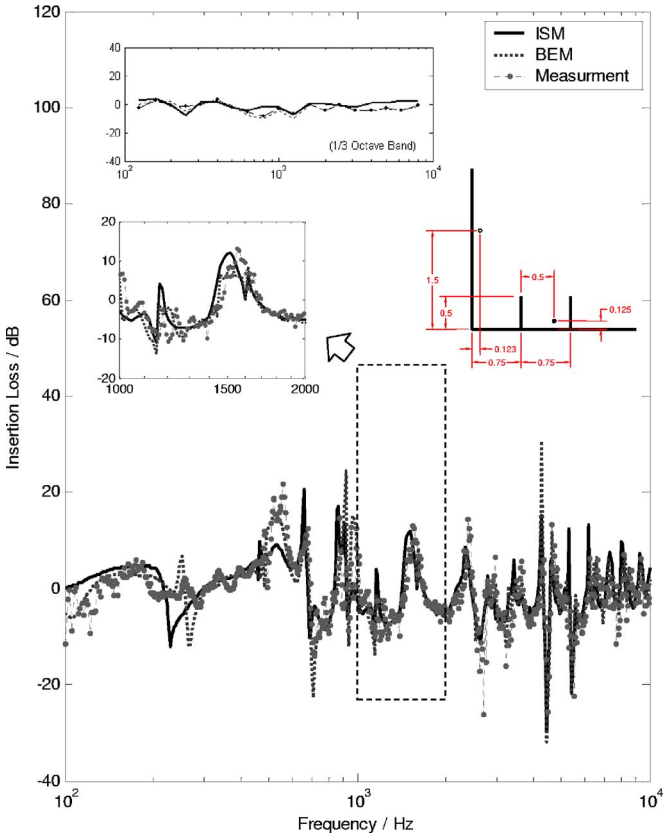


FIG. 11. (Color online) Same caption as Fig. 10 except that the receiver is located at $\mathfrak{R}=(0.123,0,1.5)$.

sources which have comparable magnitudes but with different phases. Numerical and experimental results of the frequency spectrum between 1000 and 2000 Hz are also shown in Fig. 10 in the inset figure. The constructive and destructive interferences of all rays are observed experimentally and they are predicted well by the image source model. This plot highlights the importance for including the information of the magnitudes and phases of each ray in the prediction model. The traditional energy-based ray model cannot be used to predict this wave interference effect.

In order to have a better quantitative comparison, the predictions and measurement of the insertion losses have also been compared in 1/3 octave bands for a range of frequencies varying from 125 to 8000 Hz as shown in the inset of Fig. 10. The predicted results (either by using the image source model or the BEM formulation) and measured data are generally in reasonably good agreement. Compared with the narrowband spectrum, the large fluctuations in the insertion loss are "smoothed" out in the 1/3 octave band spectrum.

Figure 11 shows the insertion loss for location D in which the receiver is located in the illumination region. In comparison with the previous case, the effect of interference is less observable for the frequency range of interest. It is because the receiver is located above the near side barrier which reduces the effect of multiple reflections between the boundary surfaces. Again, the inset figure shows good agreements in terms of constructive and destructive interferences between the measured results and the predictions according to the image source model for frequency ranging from 1000 to 2000 Hz. For the plot of the 1/3 octave bands, the image source model predicts very well the trend of the measured insertion loss.

Next, we show a comparison of the experimental results and theoretical predictions for the case of the parallel barriers placed in a street canyon. In Figs. 12 and 13, we display the comparisons of the experimental results and numerical predictions according to the image source model and BEM formulations at Locations A and C. There is an increase in the number of image sources in this experimental setup because of the presence of an additional building façade at the far side. This leads to an increase in the level of interference which can be observed in the measured data.

The overall compared results for this set of data are not as good as those shown earlier for the sound field of parallel placed in front a building façade. In this case, the finite size of building façades (both height and width) and the finite length of barriers play a significant role in the measured data. It is because extra diffractions at these edges, which have been ignored in the image source model, have comparable magnitudes to the higher-order reflected rays. Nevertheless, good agreements in terms of constructive and destructive interferences are evinced in Figs. 12 and 13 for both receiver

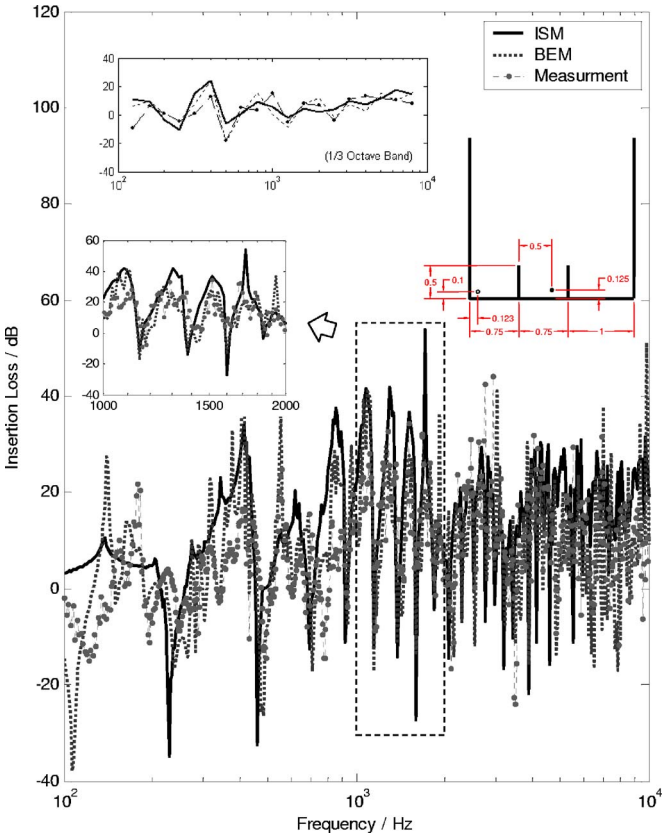


FIG. 12. (Color online) Same caption as Fig. 10 except that the receiver is located at $\mathfrak{R}=(0.123,0,0.1)$ and the parallel barriers are placed in a street canyon.

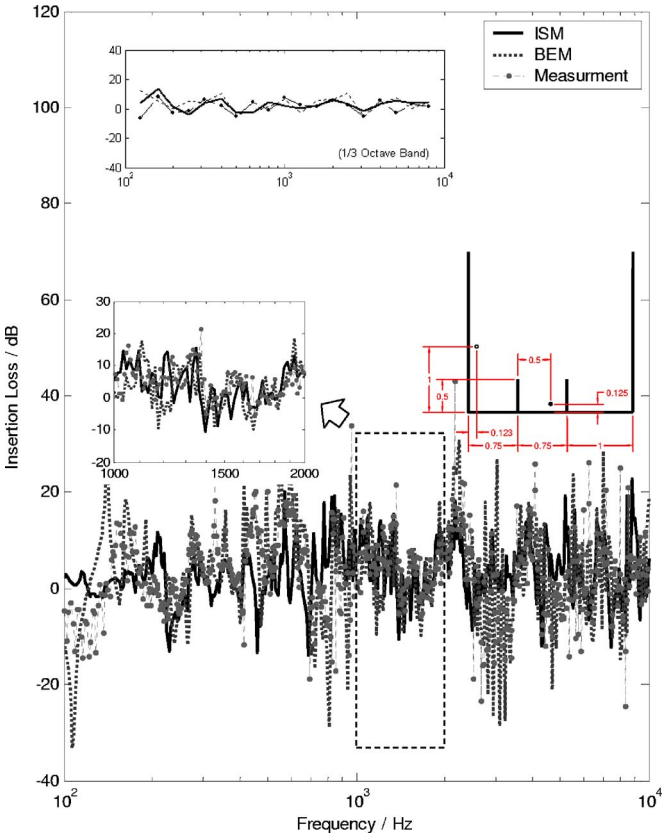


FIG. 13. (Color online) Same caption as Fig. 10 except that the receiver is located at $\mathfrak{R}=(0.123,0,1)$ and the parallel barriers are placed in a street canyon.

locations. Again, the large fluctuations in the sound fields are comparatively smoothed out for the data presented in the 1/3 octave bands. The image source model and BEM predictions are in accord with experimental measurements.

V. DISCUSSIONS AND CONCLUDING REMARKS

The development of a ray model for prediction of the performance of parallel barriers in high-rise cities is presented in this paper. Two typical cases in urban environments, (a) parallel barriers placed in front of a row of tall buildings and (b) parallel barriers in a street canyon, are considered. The contribution of the total sound field due to the effect of multiple reflections in parallel boundary surfaces is elucidated. The proposed ray model is validated by comparing it with indoor scale model experimental data and the BEM formulation, which is a more accurate but computationally intensive method. Compared with the BEM formulation, the ray model uses much fewer computational resources. Moreover, the ray model demonstrates its advantage in handling different frequencies simultaneously as long as all valid ray paths have been determined. This feature of handling different frequencies simultaneously is particularly useful for computing the 1/3 octave band noise levels. The ray model is a very useful numerical technique and computationally efficient in assessing the acoustic performance of parallel barriers in a high-rise city.

ACKNOWLEDGMENTS

This work was conducted while M.P.K. was a graduate student at the Hong Kong Polytechnic University. The research described in this paper was financed jointly by the Innovation and Technology Commission of the Hong Kong Special Administrative Region and the Mass Transition Railway Corporation Limited, through the award of an Innovation and Technology Fund under the category of the University-Industry Collaboration Program (Project No. UIM/39). The authors gratefully acknowledge the Research Committee of the Hong Kong Polytechnic University for the facilities and technical support provided throughout the period of the research. The authors thank Dr. Glenn H. Frommer of the MTR Corporation for his encouragement and many useful discussions.

[1]D. A. Hutchins and D. Pitcarn, "A laser study of multiple reflections within parallel barriers," J. Acoust. Soc. Am. **73**, 2216–2218 (1983).
[2]D. A. Hutchins and H. W. Jones, "Parallel barriers in the presence of ground surfaces," Noise Control Eng. J. **23**, 105–105 (1984).
[3]D. A. Hutchins, H. W. Jones, B. Paterson, and L. T. Russell, "Studies of parallel performance by acoustical modeling," J. Acoust. Soc. Am. **77**, 536–546 (1985).
[4]C. H. Chew, "Prediction of traffic noise from expressways – Part II: Building flanking both sides of expressways," Appl. Acoust. **32**, 61–72 (1991).
[5]Y. Sakurai, E. Walerian, and H. Morimoto, "Noise barrier for a building façade," J. Acoust. Soc. Jpn. **11**, 257–265 (1990).
[6]W. F. Cheng and C. F. Ng, "The acoustic performance of an inclined barrier for high-rise residents," J. Sound Vib. **242**, 295–308 (2001).
[7]L. Godinho, J. Antonio, and A. Tadeu, "3D sound scattering by rigid barriers in the vicinity of tall buildings," Appl. Acoust. **62**, 1229–1248

(2001).
[8]K. M. Li and S. H. Tang, "The predicted barrier effects in the proximity of tall buildings," J. Acoust. Soc. Am. **114**, 821–832 (2003).
[9]R. Panneton, A. L'Espérance, and G. A. Daigle, "Development and validation of a model predicting the performance of hard or absorbent parallel noise barriers," J. Acoust. Soc. Jpn. **14**, 251–258 (1993).
[10]A. Muradali and K. R. Fyfe, "A study of 2D and 3D barrier insertion loss using improved diffraction-based methods," Appl. Acoust. **53**, 49–75 (1998).
[11]W. J. Hadden and A. D. Pierce, "Diffraction of sound around corners and over wide barriers," J. Acoust. Soc. Am. **69**, 1266–1276 (1981).
[12]M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1970), Chap. 7, p. 300.
[13]M. P. Kwok, "Noise barriers in a complex urban environment," M. Phil. Thesis, The Hong Kong Polytechnic University, 2006.

The effects of environmental and classroom noise on the academic attainments of primary school children

Bridget M. Shield[a)]
Faculty of Engineering, Science and Built Environment, London South Bank University, Borough Road, London SE1 0AA, United Kingdom

Julie E. Dockrell[b)]
School of Psychology and Human Development, Institute of Education, 25 Woburn Square, London WC1A 0HH, United Kingdom

(Received 9 November 2006; revised 23 October 2007; accepted 24 October 2007)

While at school children are exposed to various types of noise including external, environmental noise and noise generated within the classroom. Previous research has shown that noise has detrimental effects upon children's performance at school, including reduced memory, motivation, and reading ability. In England and Wales, children's academic performance is assessed using standardized tests of literacy, mathematics, and science. A study has been conducted to examine the impact, if any, of chronic exposure to external and internal noise on the test results of children aged 7 and 11 in London (UK) primary schools. External noise was found to have a significant negative impact upon performance, the effect being greater for the older children. The analysis suggested that children are particularly affected by the noise of individual external events. Test scores were also affected by internal classroom noise, background levels being significantly related to test results. Negative relationships between performance and noise levels were maintained when the data were corrected for socio-economic factors relating to social deprivation, language, and special educational needs. Linear regression analysis has been used to estimate the maximum levels of external and internal noise which allow the schools surveyed to achieve required standards of literacy and numeracy. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2812596]

PACS number(s): 43.50.Qp [NX] Pages: 133–144

I. INTRODUCTION

Children are exposed to many different types of noise while at school. Previous studies have shown that schools may be exposed to high levels of environmental noise, particularly in urban areas.[1,2] Sources include road traffic, trains, aircraft, and construction noise. Inside schools a wide range of noise levels have been measured,[3–7] the levels varying significantly between different types of space and different classroom activities.[1] For much of the day in a primary school classroom, young children are exposed to the noise of other children producing "classroom babble" at levels typically of around 65 dB(A) L_{Aeq},[1] while the typical overall exposure level of a child at primary school has been estimated at around 72 dB(A) L_{Aeq}.[1]

The effects of noise on children and their teachers have been investigated in many studies in the past 40 years. It is generally accepted that noise has a detrimental effect upon the cognitive development of primary school children, and that older children in this age group are more affected than the younger children.[8,9] Two major reviews of previous work in this area, published in the early 1990s, concluded that chronic noise exposure of young children has an adverse effect, particularly upon their reading ability.[10,11]

Most of the previous work has concerned the effects of environmental noise, notably aircraft noise, upon children. Exposure to high levels of aircraft noise has been found to affect memory and reading ability, and to reduce motivation in school children.[11–15] These effects appear to be long term; noise reduction inside a school has been found to have little immediate effect upon children's performance[16] while another study found that when an airport was closed it took several years for the detrimental effects of noise exposure to cease.[13] These results suggest that noise reduces the learning trajectories of the pupils involved so that extended periods of teaching and learning are required for children to reach typical levels of performance.

In addition to aircraft noise other types of environmental noise, including that from railways[17,18] and road traffic,[19] have been found to affect reading. Road traffic noise outside schools, at levels of around 70 dB(A), has also been found to reduce children's attention.[20,21]

While there is a large body of work concerning the effects of external environmental noise upon children at school, there have been far fewer investigations into the effects of typical classroom noise upon children's performance. However in recent years evidence has been found to suggest that noise inside the classroom affects letter, number, and word recognition.[10,22–25]

It is thus now generally accepted that all types of noise exposure at school affect children's learning and academic performance. The majority of the previous studies have com-

a)Electronic mail: shieldbm@lsbu.ac.uk
b)Electronic mail: j.dockrell@ioe.ac.uk

pared the performance of children exposed long term to significant levels of environmental noise with that of children with low noise exposure, or have examined the effects of noise reduction on children's performance. There have been few studies which have demonstrated a dose/response relationship between noise and effects on children's performance, thereby making it difficult to determine threshold levels at which adverse effects occur, which in turn makes it difficult to establish specific guideline values to prevent such effects.[26]

In recent years several countries have introduced standards and guidelines relating to the acoustic design of schools and classrooms. For example, in the United States ANSI standard S12.60,[27] published in 2002, sets out guideline values for noise levels, reverberation times, and sound insulation in schools. Since 2003 new school buildings in England and Wales must comply with the Building Regulations. The acoustic requirements are specified in Building Bulletin 93 (BB93),[28] published in 2003. The requirements of S12.60 and BB93 are similar, for example the maximum noise level specified by both for empty classrooms is 35 dB(A) L_{Aeq}. However, in general the noise specifications for classrooms are based upon speech intelligibility requirements, rather than the levels of noise which have direct detrimental effects upon children's performance in the classroom.

In the study described here noise levels measured outside 142 primary schools in central London (UK), and inside a range of spaces inside 16 schools have been compared with assessment scores of the schools in national standardized tests. The approach taken enables the effects on children at school of different levels and types of noise to be investigated. It is also possible to compare the impact of various types of noise upon different aged children across a variety of academic tasks. In addition, this approach allows the most important property of the noise (for example, its background, maximum, or ambient level) in relation to academic performance to be determined, an issue that has not been considered in previous studies.

A simultaneous study by the authors[29] used experimental testing to investigate the effects of environmental and classroom noise on children's performance on a range of tasks in the classroom. It will be seen that the results of the two investigations are complementary and advance the understanding of the different ways in which children's academic performance and development are affected by noise.

II. MATERIALS AND METHODS

A. Procedure

The study investigated the effects of chronic noise exposure upon children's academic attainments by comparing measured noise levels with recognized standardized measures of children's attainments in primary school. The relationships between attainment scores for individual schools and both external (environmental) and internal noise were examined. The effects of acute exposure to environmental and classroom noise were also investigated in the above-mentioned complementary experimental study.[29]

B. Measures of children's attainments: Standardized assessment tests (SATs)

In the 1990s a standard national curriculum was introduced for all schools in England and Wales. To complement this curriculum, standardized assessment tests (SATs) in various subjects including English, Mathematics, and Science were introduced across the age range at both primary and secondary school level. The majority of children at state schools take these tests at the ages of 7 ("Key Stage 1"), 11 ("Key Stage 2") and 14 ("Key Stage 3") years. Average results for all schools in all subjects are published by the Department for Education and Skills. The published school data consist of the percentages of children in each school who reach a recognized criterion level in each subject at each stage. Average school scores for each stage are also published. Each year the UK government sets targets for literacy and numeracy in primary schools by specifying Key Stage 2 SAT scores which schools must aim to achieve. At the time of the survey the target scores for schools were 75% for Key Stage 2 Mathematics and 80% for Key Stage 2 English.

The study described here concerned children of primary school age. The relevant test data for comparison with noise were therefore Key Stage 1 and Key Stage 2 SAT results. At Key Stage 1 (KS1) the assessment includes both teacher assessments and national standardized tests, which are combined to give a single score for each subject for each child. At Key Stage 2 (KS2) children sit for standard nationwide examinations. Between two and four examinations are taken in each subject, the examination results being averaged to give a single mark for each subject.

The subjects assessed at the two stages at the time of this study were as follows: Key Stage 1 (Year 2 of primary school, 7 years of age on average): Reading; Writing; Spelling; and Mathematics. Key Stage 2 (Year 6 of primary school, 11 years of age on average): English; Mathematics; and Science.

The schools' attainment scores in each subject, plus average scores, at Key Stage 1 and Key Stage 2, were compared with noise levels measured inside and outside the schools.

C. Selection of study areas and schools

The areas chosen for the study were based upon the local government boroughs of London, of which there are 33. It was important for the study that the boroughs chosen should be representative of London as a whole in terms of noise exposure, academic achievements, and demographic characteristics in order to reduce the number of potentially confounding variables.

It was decided that boroughs in which aircraft were the dominant environmental noise source should be excluded from the survey, as there was already a considerable body of research on the effects of aircraft noise on children. There was also a concurrent study of the effects of aircraft noise on children in schools to the west of London, around Heathrow airport.[14] Furthermore, there were fewer detailed studies of the impact of general environmental noise than of aircraft

TABLE I. SAT results, demographic factors, and external noise levels for the three boroughs.

		Borough A		Borough B		Borough C	
Stage	Subject	Mean	s.d.	Mean	s.d.	Mean	s.d.
Key Stage 1 test results	Reading	76.1	14.1	74.7	13.2	78.4	16.9
	Writing	76.8	14.9	74.8	13.9	78.2	16.9
	Spelling	63.8	17.1	59.3	17.2	64.7	18.4
	Maths	86.4	8.9	83.5	12.0	86.4	13.2
Key Stage 2 test results	English	68.5	18.5	69.8	15.7	69.5	16.6
	Maths	66.1	16.2	67.0	15.7	68.2	19.1
	Science	77.9	15.9	81.0	12.6	78.9	17.3
Demographic factors	% FSM	38.8	19.3	41.5	14.2	33.6	10.7
	% EAL	43.9	19.2	35.3	16.8	39.6	17.7
	% SEN	10.3	2.9	28.3	10.0	26.2	7.8
External noise levels	$L_{\mathrm{Aeq,5\ min}}$	57.4	8.8	56.2	9.4	58.9	7.4
	$L_{\mathrm{A10,5\ min}}$	59.4	9.0	58.4	9.9	61.2	7.7
	$L_{\mathrm{A90,5\ min}}$	49.2	7.7	46.5	9.3	50.2	8.2
	$L_{\mathrm{A99,5\ min}}$	47.0	7.4	44.3	9.2	47.8	8.2
	$L_{\mathrm{Amax,5\ min}}$	70.5	10.5	68.3	17.0	72.0	9.0
	$L_{\mathrm{Amin,5\ min}}$	46.0	7.5	41.3	12.4	47.0	8.3

noise. Therefore, in selecting boroughs for the purpose of this study those affected particularly by aircraft noise were excluded.

Remaining boroughs were examined to ensure that their primary school academic attainments and demographic characteristics (see Sec. II D) were typical of London as a whole. The distributions of SAT results in boroughs were studied in order to select boroughs for which (a) test scores displayed an acceptable range, as indicated by the standard deviations of the SAT results in all subjects and (b) the mean scores for reading, writing, and mathematics were not above the mean score of all London boroughs. Of the boroughs selected in this way agreement was obtained from the Directors of Education of three boroughs to participate in the project. Borough A is a suburban London borough, all schools being within approximately 6 miles of central London. Boroughs B and C, on the other hand, are more centrally located, with all schools within a distance of approximately 3 miles from central London. Demographic differences between the boroughs are discussed in Sec. II D.

Means and standard deviations of the subject scores for the three boroughs are shown in Table I. Analysis of variance showed that there was no significant difference between the subject scores for the three boroughs.

It can be seen from Table I that there was in general close agreement between mean subject scores in the three boroughs, while borough C displayed slightly higher standard deviations in most subjects indicating a wider spread of scores in this borough.

D. Demographic characteristics

The socio-economic characteristics of schools in the boroughs were also examined. The data considered were the percentages of children in each school receiving free school meals (FSM); the percentages of children for whom English is an additional language (EAL); and the percentages of children with special educational needs (SEN). The percentage of children receiving free school meals is commonly accepted as a reliable indicator of social disadvantage in an area.[30,31]

The means and standard deviations of these data for the three chosen boroughs are also given in Table I. Analysis of variance showed that there were some differences between the boroughs, particularly in the distributions of children with special educational needs. There were considerably fewer children with special needs in (suburban) borough A while the percentages for the central boroughs were similar and around 2.5 times the percentage in borough A.

A major difference between the boroughs is in the density of population. At the time of the surveys the populations per square kilometer of the three boroughs were approxi-

TABLE II. Internal noise levels.

	School location					Class (age group)							
	Occ teach space	Unocc teach space	Corr/ foyer /stair	Occ hall	Unocc hall	Nurs (3–4)	Rec (4–5)	Yr 1 (5–6)	Yr 2 (6–7)	Yr 3 (7–8)	Yr 4 (8–9)	Yr 5 (9–10)	Yr 6 (10–11)
L_{Aeq}	72.1	47.0	58.1	73.4	53.2	71.9	73.9	74.3	66.3	68.9	69.6	73.2	71.2
L_{A90}	54.1	36.9	44.6	55.1	44.3	57.3	62.3	61.0	51.3	52.5	49.8	53.8	52.9

mately as follows: borough A 7600; borough B 12 200, and borough C 10 100. Boroughs B and C therefore represent the more densely populated inner city areas, while borough A is more typical of suburban boroughs.

E. Noise surveys

Noise levels were measured outside all the state-funded primary schools in boroughs A (N=53) and B (N=50) and outside a majority of the 61 schools in borough C (N=39). Of these, eight schools in boroughs A and B were also selected for internal surveys. The eight schools were chosen to reflect the full range of external noise levels measured, the external L_{Aeq} levels of the 16 schools ranging from 49 to 75 dB(A). The measurement methods, noise levels, and noise sources present have been described elsewhere.[1] The external and internal levels that have been used in examining the impact of noise upon test results are summarized in the following.

1. External levels

Table I also shows the means and standard deviations of various environmental noise parameters measured in the three boroughs. These levels were measured at, or have been normalized to, a distance of 4 m from the school façade during the school day.[1]

It can be seen that the levels were reasonably consistent across the three boroughs, with borough C having slightly higher levels than the other two boroughs. This was to be expected as this borough is the one nearest central London. The mean levels in borough B were slightly lower than might be expected given that this is also an inner city borough. However many of the schools in this area are situated in the middle of housing estates or on side streets, and are thus sheltered to some extent from the noise of road traffic, the main noise source in the areas surveyed.[1] This is illustrated by the larger standard deviations of noise levels in borough B.

2. Internal levels

In the internal school noise survey levels were measured in classrooms and other areas around a school. Most spaces were measured in both occupied and unoccupied conditions. The averaged ambient (L_{Aeq}) and background (L_{A90}) levels for the types of spaces considered in each school are shown in Table II.

Internal levels were also categorized according to the age of the class; the average L_{Aeq} and L_{A90} levels for different age groups in each school are also shown in Table II. For the purposes of analyzing the effects, if any, of noise on SAT results noise levels for Year 2 and Year 6 are the only ones considered in the subsequent discussion.

F. Analyses

In order to study the impact, if any, of noise on children's attainment the noise levels measured inside and outside the schools were correlated with the SAT scores for the academic year in which the noise survey was carried out.

TABLE III. Borough A: Correlation coefficients between test scores and external noise levels.

	L_{Aeq}	L_{Amax}	L_{A90}	L_{A10}
KS1 Reading	−0.34[b]	−0.31[b]	−0.37[a]	−0.33[b]
KS1 Maths	−0.34[b]	−0.27	−0.43[a]	−0.34[b]
KS2 English	−0.37[a]	−0.39[b]	−0.40[a]	−0.33[b]
KS2 Maths	−0.40[a]	−0.46[b]	−0.40[a]	−0.36[a]
KS2 Science	−0.40[a]	−0.45[b]	−0.42[a]	−0.37[a]
KS1 average	−0.36[b]	−0.32[b]	−0.40[a]	−0.36[b]
KS2 average	−0.41[a]	−0.45[a]	−0.43[a]	−0.37[a]

[a]Significant at 1% level.
[b]Significant at 5% level.

For external noise it was found that results for L_{A90}, L_{A99}, and L_{Amin} were very similar, as would be expected and was confirmed by factor analysis. Therefore in the following sections, relationships between SAT results and L_{Aeq}, L_{Amax}, L_{A90}, and L_{A10} only are considered. These are the most commonly cited measures of environmental noise and are generally considered to capture the key features of the noise environment.

Similarly, factor and correlation analysis showed a close relationship among results for KS1 literacy-related tests Reading, Writing, and Spelling, as would be expected. Therefore, in the subsequent analysis and discussion, of these tests, results are presented for KS1 Reading only as being a reliable indicator of the younger children's attainments in literacy.

Correlation and regression analysis were carried out for the noise and test data. The noise levels were correlated with subject and average school SAT scores. Obviously any relationships found between noise and SAT scores in this way could be due to social or other factors rather than representing a direct effect of noise on academic performance. In order to eliminate the effects of socio-economic factors, partial correlations were carried out, in which the schools' data on children with FSM, EAL, and SEN were controlled for.

Current guidance on choosing a site for new school buildings in England and Wales recommends an upper limit of 60 dB $L_{Aeq,30\ min}$ at the boundary of school premises.[28] For this reason, in addition to considering all schools mea-

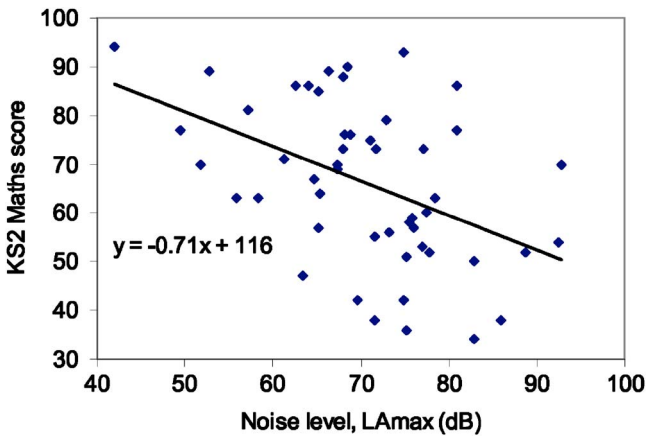


FIG. 1. (Color online) Scatter diagram illustrating relationship between external L_{Amax} and Key Stage 2 Mathematics scores in borough A.

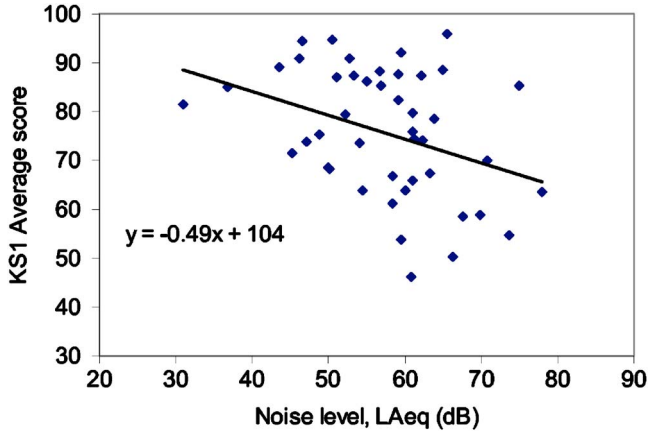


FIG. 2. (Color online) Scatter diagram illustrating relationship between external L_{Aeq} and average Key Stage 1 scores in borough A.

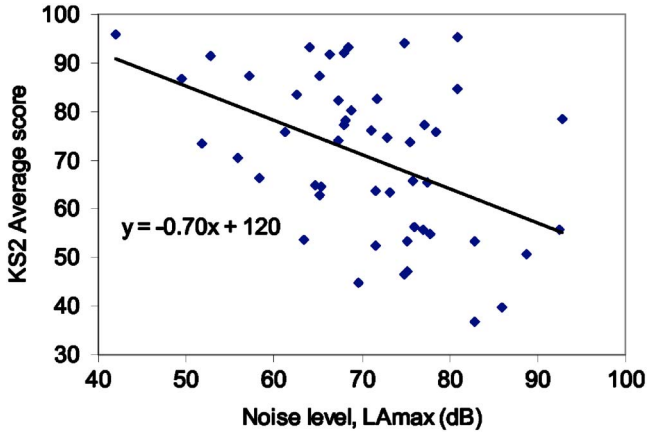


FIG. 3. (Color online) Scatter diagram illustrating relationship between external L_{Amax} and average Key Stage 2 scores in borough A.

sured in each borough, those schools where the measured external L_{Aeq} levels are greater than or equal to 60 dB(A) have been considered separately.

III. RESULTS: RELATIONSHIPS BETWEEN EXTERNAL NOISE AND TEST RESULTS

The values of the noise parameters L_{Aeq}, L_{Amax}, L_{A90}, and L_{A10} measured outside each school were compared with average and subject SAT scores for the younger (aged 7 years) and older (aged 11 years) children.

The Pearson correlation coefficients between average and subject scores and external noise levels were calculated for all schools in boroughs A, B, and C. Table III shows the coefficients for borough A. It can be seen that there were negative relationships between external noise and SATs for all scores, that is, the greater the noise level the lower the school test performance score. Furthermore, all except one of the relationships were significant at the 1% or 5% level. However, for both boroughs B and C the correlation coefficients were very small, varying from −0.15 to 0.28. There were no significant relationships and the coefficients were very similar for the two boroughs. This may be due to the differences between the central and suburban boroughs reflected in the SEN data shown in Table I, and also to the different characteristics of the boroughs as represented by their population densities, discussed in Sec. II D. For this reason the two central boroughs (B and C) are considered together and separately from the suburban borough (A) in the following discussion.

A. Borough A

1. All schools

Table III shows that when all schools in borough A are considered there were significant negative relationships between all SAT scores and all external noise parameters, except for KS1 Mathematics and L_{Amax}. The relationships were stronger for Key Stage 2 subjects, suggesting that noise has more of an impact upon the performance of the older children. A possible explanation for this is that the older children have been exposed to the noise for a longer period of time. This is consistent with the results of previous research demonstrating the effects of long-term noise exposure.[13–16] However, it is also possible that the nature and demands of the tasks for older children differ from those of the younger children and are more vulnerable to the effects of noise.

At Key Stage 1 and for KS2 English the external noise level with the strongest correlation with test scores was the background level, as measured by L_{A90}. For other subjects at Key Stage 2, L_{Amax} was the parameter which had the strongest association with test scores. This suggests that the younger children were affected by general external background noise, while the older children were more affected by individual external noise events such as motorbikes or lorries

TABLE IV. Borough A: Correlation coefficients between test scores and external noise levels corrected for data on FSM, EAL, and SEN.

	L_{Aeq}			L_{Amax}			L_{A90}			L_{A10}		
	FSM	EAL	SEN	FSM	EAL	SEN	FSM	EAL	SEN	FSM	EAL	SEN
KS1 Reading	−0.17	−0.26	−0.32[b]	−0.15	−0.26	−0.29[b]	−0.11	−0.24	−0.35[b]	−0.16	−0.25	−0.31[b]
KS1 Maths	−0.23	−0.28	−0.32[b]	−0.15	−0.22	−0.24	−0.29	−0.35[b]	−0.41[a]	−0.24	−0.28	−0.33[b]
KS2 English	−0.17	−0.27[b]	−0.34[b]	−0.25	−0.38[a]	−0.37[a]	−0.08	−0.23	−0.39[a]	−0.12	−0.22	−0.31[b]
KS2 Maths	−0.23	−0.32[b]	−0.38[a]	−0.36[a]	−0.44[a]	−0.44[a]	−0.10	−0.25	−0.38[a]	−0.19	−0.27	−0.35[a]
KS2 Science	−0.25	−0.32[b]	−0.39[a]	−0.34[b]	−0.42[a]	−0.44[a]	−0.19	−0.30[b]	−0.41[a]	−0.23	−0.29[b]	−0.36[a]
KS1 average	−0.20	−0.29	−0.34[b]	−0.17	−0.27	−0.30[b]	−0.18	−0.29	−0.39[a]	−0.21	−0.28	−0.35[b]
KS2 average	−0.25	−0.33[b]	−0.39[a]	−0.36[a]	−0.45[a]	−0.44[a]	−0.14	−0.28[b]	−0.41[a]	−0.20	−0.28[b]	−0.36[a]

[a]Significant at 1% level.
[b]Significant at 5% level.

TABLE V. Schools in boroughs B and C with external $L_{Aeq} \geq 60$ dB(A): Correlation coefficients between test scores and noise levels.

	L_{Aeq}	L_{Amax}	L_{A90}	L_{A10}
KS1 Reading	−0.40[b]	−0.40[b]	−0.22	−0.36[b]
KS1 Maths	−0.10	−0.09	−0.03	−0.20
KS2 English	−0.39[b]	−0.43[a]	−0.37[b]	−0.38[b]
KS2 Maths	−0.21	−0.31	−0.15	−0.27
KS2 Science	−0.25	−0.36[b]	−0.15	−0.24
KS1 average	−0.31	−0.31	−0.12	−0.28
KS2 average	−0.30	−0.39[b]	−0.24	−0.32

[a]Significant at 1% level.
[b]Significant at 5% level.

passing the school. This is consistent with the findings of previous research,[12–18] which has found that reading is affected by noise caused by individual external sources such as trains or planes. It is also consistent with a questionnaire survey of children carried out by the authors which found that older, Key Stage 2 age, children were more aware of external noise than the younger children at Key Stage 1. The subject showing the strongest negative effect of noise (with background levels at Key Stage 1 and with maximum levels at Key Stage 2) was Mathematics. The mathematics assessment at Key Stage 2 is complex, involving orally presented mental arithmetic, written arithmetic, and word problems. Thus performance at these tasks is vulnerable to the effects of noise on both reading and speeded responses, two areas which have been found to be affected by noise in previous studies.[10–18,29]

Figures 1–3 give examples of scatter diagrams relating external noise levels and SAT scores. Figure 1 shows the relationship between L_{Amax} and Key Stage 2 Mathematics scores; Fig. 2 shows the scatter diagram of L_{Aeq} and average Key Stage 1 score; and Fig. 3 average Key Stage 2 score and L_{Amax}. Regression lines relating external noise levels and SAT scores are also shown in Figs. 1–3. The implications of these relationships are discussed in Sec. V.

Table IV shows the partial correlation coefficients obtained when the data for borough A were controlled for the FSM, EAL, and SEN data. It can be seen that when social deprivation (as measured by FSM data) was taken into account there was still a negative relationship between external noise and test scores, but there were fewer significant relationships than with the uncorrected data. However, L_{Amax} was still significantly correlated with two subject scores (Mathematics and Science) and the average score at Key Stage 2. The strongest relationship was again with the Mathematics scores. When potential language demands (as indicated by EAL data) were accounted for there were still strong associations between L_{Amax} and all subjects at Key Stage 2, with Mathematics again being the subject most strongly related to noise. As with the uncorrected data, KS1 Mathematics scores were most strongly, and significantly, related to the external background noise level. When controlling for SEN, it can be seen that the pattern was very similar to that for the uncorrected data, with KS2 Mathematics and Science again being the subjects most affected by external noise, and L_{Amax} having the strongest negative relationship with test scores at Key Stage 2.

2. Schools with external L_{Aeq} levels of 60 dB(A) or greater

When considering only those schools with external L_{Aeq} levels of 60 dB(A) or more in borough A ($N=22$), KS1 Mathematics was the only subject significantly related to noise, being significantly related at the 5% level to L_{A90}. This significant relationship was maintained when the data were corrected for socio-economic factors, becoming significant at the 1% level when correcting for SEN.

B. Boroughs B and C

1. All schools

As mentioned previously, there were no significant relationships between test scores and external noise for the central London boroughs when all schools in the two boroughs were considered. The reason for the difference between these schools and those in borough A is unclear, but may be related to the discrepancies in the percentages of children with special needs in the central and suburban boroughs, or to the differing population characteristics between the boroughs.

2. Schools with external L_{Aeq} levels of 60 dB(A) or greater

If only those schools where the external level exceeds 60 dB L_{Aeq} in the two boroughs were considered ($N=35$) then there were stronger negative relationships between SAT

TABLE VI. Schools in boroughs B and C with external $L_{Aeq} \geq 60$ dB(A): Correlation coefficients between test scores and noise levels corrected for data on FSM, EAL, and SEN.

	L_{Aeq}			L_{Amax}			L_{A90}			L_{A10}		
	FSM	EAL	SEN	FSM	EAL	SEN	FSM	EAL	SEN	FSM	EAL	SEN
KS1 Reading	−0.35[b]	−0.40[b]	−0.35[b]	−0.40[b]	−0.41[b]	−0.43[a]	−0.13	−0.22	−0.16	−0.23	−0.36[b]	−0.29
KS1 Maths	−0.00	−0.08	−0.02	−0.04	−0.10	−0.10	0.09	0.05	0.07	−0.04	−0.15	−0.10
KS2 English	−0.34[b]	−0.37[b]	−0.32	−0.46[a]	−0.46[a]	−0.48[a]	−0.30	−0.28	−0.29	−0.23	−0.32	−0.29
KS2 Maths	−0.09	−0.18	−0.11	−0.30	−0.32[b]	−0.34[b]	−0.01	−0.06	−0.05	−0.06	−0.21	−0.16
KS2 Science	−0.16	−0.23	−0.20	−0.35[b]	−0.37[b]	−0.37[b]	−0.03	−0.08	−0.09	−0.06	−0.19	−0.17
KS1 average	−0.25	−0.31	−0.25	−0.29	−0.31	−0.33	−0.02	−0.11	−0.04	−0.14	−0.28	−0.21
KS2 average	−0.22	−0.28	−0.23	−0.41[b]	−0.41[b]	−0.43[a]	−0.13	−0.16	−0.16	−0.13	−0.26	−0.22

[a]Significant at 1% level.
[b]Significant at 5% level.

TABLE VII. Internal noise: Correlation coefficients between test scores and Year 2 and Year 6 noise levels.

	Year 2 $N=11$		Year 6 $N=13$	
	L_{Aeq}	L_{A90}	L_{Aeq}	L_{A90}
KS1 Reading	0.01	−0.12		
KS1 Maths	−0.17	−0.33		
KS2 English			−0.45	−0.48
KS2 Maths			−0.04	−0.00
KS2 Science			−0.36	−0.11
KS1 average	−0.15	−0.29		
KS2 average			−0.33	−0.25

scores and noise, as shown in Table V. For most external noise parameters, as with borough A schools, the relationships were stronger for Key Stage 2 results, and in general L_{Amax} was the parameter most closely related to test results. In these boroughs, however, English was the subject showing the greatest effect of noise. Both KS1 Reading and KS2 English scores were significantly related to external L_{Aeq}, L_{Amax}, and L_{A10} levels, while KS2 English was also significantly related to the background L_{A90} level. Unlike the suburban borough, Mathematics scores were not significantly related to any external noise parameter.

Table VI shows the correlations when the data were corrected for socio-economic factors. In all cases the results were very similar to those for the uncorrected data. KS1 Reading and KS2 English were the subjects most affected by external noise, KS2 English being significantly correlated with L_{Amax} at the 1% level and L_{Amax} again being the noise parameter with the strongest correlations with test scores. When correcting for EAL and SEN, all subjects at KS2 were significantly related to L_{Amax}. Relationships between KS2 English and L_{Amax} were significant at the 1% level, and stronger than for the uncorrected data.

IV. RESULTS: RELATIONSHIPS BETWEEN INTERNAL NOISE AND TEST RESULTS

In investigating relationships between internal noise and SATs, average and subject Key Stage 1 and Key Stage 2 SAT scores were correlated with relevant internal noise data. For this analysis, correlations were carried out for the complete set of 16 schools (eight in borough A and eight in borough B) for which internal noise data were available. The internal noise data that were used consisted of the L_{Aeq} and L_{A90} levels for Year 2 and Year 6 (as these are the years in which children sit for SATs); and in the various school locations which were measured.

A. Correlation with year group levels

Table VII shows the correlations between KS1 test scores and Year 2 noise levels, and between KS2 scores and Year 6 levels. It can be seen that there were negative relationships between all scores and noise levels, except for Key Stage 1 Reading; however, none of the correlations were significant, possibly because of the small sample size. The subject showing the strongest effect of internal noise was KS2 English, which was related to both L_{Aeq} and L_{A90} levels. This is consistent with the results of the parallel experimental testing,[29] which showed that classroom babble affected all tasks both verbal and nonverbal.

When the data were corrected for socio-economic factors KS2 English was still the subject most strongly affected by internal noise; when correcting for FSM there was a significant negative relationship ($r=-0.59$, $p<0.05$) between background noise (L_{A90}) in Year 6 classrooms and test scores for this subject.

B. Correlation with location levels

Table VIII shows the correlation coefficients between L_{Aeq} and L_{A90} levels for different school locations and subject test scores. There were negative correlations between all subject scores and all noise levels measured in occupied classrooms, unoccupied classrooms, and corridors and foyers. In general the relationships were strongest for occupied classrooms, with the background (L_{A90}) level being significantly related to test scores for most subjects. The subject most strongly affected by internal noise was again KS2 English, which was significantly correlated at the 1% level with occupied classroom L_{A90}. KS1 Mathematics was significantly related to L_{A90} in both occupied and unoccupied classrooms.

Figures 3–6 show scatter diagrams relating internal noise and KS2 English scores, KS1 average scores, and KS2

TABLE VIII. Internal noise: Correlation coefficients between test scores and school location noise levels.

	Occ class $N=16$		Unocc class $N=14$		Corridor/foyer $N=14$		Occ hall $N=8$		Unocc hall $N=7$	
	L_{Aeq}	L_{A90}	L_{Aeq}	L_{A90}	L_{Aeq}	L_{A90}	L_{Aeq}	L_{A90}	L_{Aeq}	L_{A90}
KS1 Reading	−0.11	−0.60[b]	−0.33	−0.46	−0.38	−0.39	0.32	0.06	0.14	0.18
KS1 Maths	−0.12	−0.57[b]	−0.52	−0.55[b]	−0.38	−0.40	0.36	0.21	0.43	0.34
KS2 English	−0.55[b]	−0.77[a]	−0.08	−0.20	−0.53[b]	−0.62[b]	−0.12	−0.28	0.47	0.49
KS2 Maths	−0.22	−0.46	−0.06	−0.21	−0.47	−0.49	0.18	0.03	0.28	0.36
KS2 Science	−0.41	−0.50[b]	−0.14	−0.32	−0.38	−0.39	−0.09	−0.31	−0.19	−0.04
KS1 average	−0.16	−0.58[b]	−0.41	−0.51	−0.41	−0.39	0.24	0.06	0.15	0.18
KS2 average	−0.43	−0.64[a]	−0.10	−0.46	−0.49	−0.35	−0.00	0.03	0.15	0.35

[a]Significant at 1% level.
[b]Significant at 5% level.

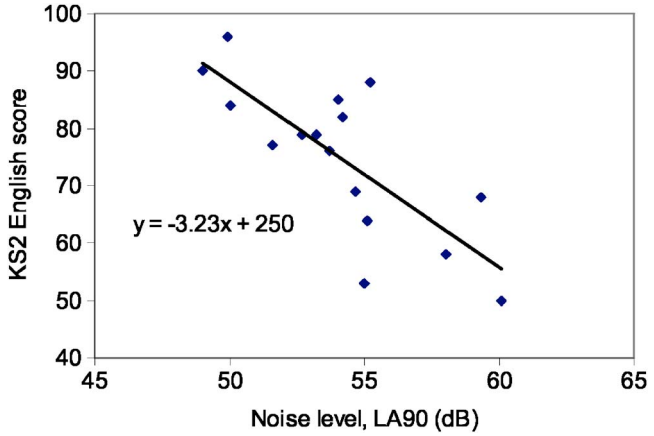


FIG. 4. (Color online) Scatter diagram illustrating relationship between occupied classroom L_{A90} and Key Stage 2 English scores.

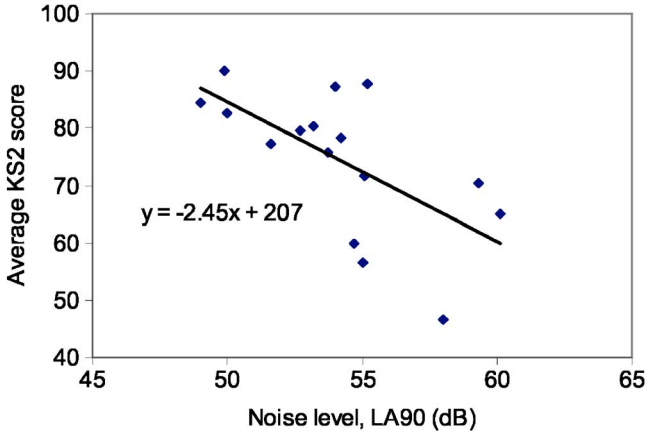


FIG. 6. (Color online) Scatter diagram illustrating relationship between occupied classroom L_{A90} and average Key Stage 2 scores.

average scores, respectively. Regression lines relating internal noise levels and SAT scores are also shown in Figs. 3–6 and are discussed in more detail in Sec. V.

It is interesting to note that there were consistently negative correlations between test scores and all noise levels in corridors and foyers, being significant again for KS2 English. While carrying out internal noise surveys it was subjectively apparent that the noise in such spaces gave a good indication of the general "noise climate" in a school.

It can be seen that there was no relationship between noise levels in school halls, occupied or unoccupied, and test scores. This is as would be expected and validates the fact that there are strong negative relationships between noise in classrooms and test results.

Tables IX and X show the correlation coefficients between test scores and L_{Aeq} and L_{A90} levels, respectively, in classrooms and circulation areas when the data were corrected for socio-economic factors. In general, relationships were slightly less strong when correcting for FSM and EAL but when correcting for SEN correlations coefficients were similar to those for the uncorrected data. KS2 English was still significantly correlated with L_{Aeq} in occupied classrooms and in corridors/foyers. When correcting for all factors there were significant correlations between KS2 English and L_{A90} in occupied classrooms and corridors/foyers.

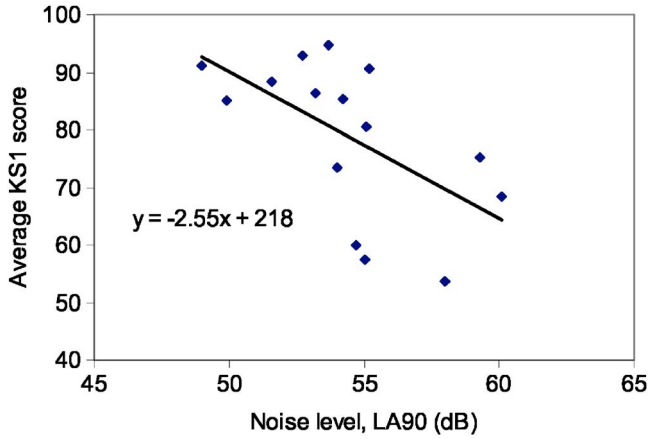


FIG. 5. (Color online) Scatter diagram illustrating relationship between occupied classroom L_{A90} and average Key Stage 1 scores.

V. QUANTIFYING THE EFFECTS OF NOISE

The regression lines relating noise levels and SAT scores for the most significant results have been calculated. In borough A these relationships have been used to investigate the implications of increases in external L_{Aeq}, L_{Amax}, and L_{A90} levels, and to establish the noise levels in this borough which correspond to the UK government targets in numeracy and literacy at the time of the survey (80% of children achieving required level in KS2 English and 75% in KS2 Mathematics). Similar analysis has been carried out for internal background (L_{A90}) levels in occupied classrooms.

A. External noise

The equations of the regression lines relating external noise (L_{Aeq}, L_{Amax}, and L_{A90} levels) and Key Stage 2 English and Mathematics scores in borough A are shown in Table XI. For completeness the relationships between noise and average Key Stage 1 and 2 scores are also shown. These linear relationships have been used to estimate the percentage decreases in the numbers of children achieving the required level for each 10 dB increase in external noise; these are also shown in Table XI. Table XI also shows the external noise levels, derived from the regression lines, which correspond to the UK government targets in English and Mathematics.

It can be seen that an increase of 10 dB(A) in external L_{Aeq}, L_{Amax}, and L_{A90} levels in borough A causes 5%, 4%, and 6% drops, respectively, in the number of children achieving the required levels at Key Stage 1, and drops of 7%, 9% and 9%, at Key Stage 2. This further illustrates the greater detrimental effect of noise on the older children in the primary school age range. The external L_{Aeq}, L_{Amax}, and L_{A90} levels corresponding to the UK government target for literacy are 42 dB(A), 54 dB(A), and 37 dB(A), respectively; for numeracy the corresponding levels are 44, 58, and 38 dB(A). It should be noted that these refer to external levels at a point 4 m from the school façade, and should be interpreted with caution as discussed in Sec. VI.

TABLE IX. Internal noise: Correlation coefficients between test scores and school location L_{Aeq} levels corrected for FSM, EAL, and SEN.

	Occupied classroom $N=16$			Unoccupied classroom $N=14$			Corridor/foyer $N=14$		
	FSM	EAL	SEN	FSM	EAL	SEN	FSM	EAL	SEN
KS1 Reading	0.11	0.13	−0.09	−0.05	−0.19	−0.34	−0.25	−0.33	−0.49
KS1 Maths	0.15	0.18	−0.14	−0.28	−0.42	−0.52	−0.23	−0.33	−0.42
KS2 English	−0.45	−0.44	−0.53[b]	0.32	0.11	−0.10	−0.43	−0.50	−0.71[a]
KS2 Maths	−0.07	−0.09	−0.24	0.23	0.07	−0.05	−0.38	−0.43	−0.51
KS2 Science	−0.33	−0.32	−0.38	0.04	−0.03	−0.15	−0.31	−0.34	−0.53
KS1 average	0.09	0.08	−0.15	−0.12	−0.29	−0.41	−0.27	−0.36	−0.49
KS2 average	−0.32	−0.31	−0.42	0.21	0.05	−0.12	−0.39	−0.45	−0.62[b]

[a]Significant at 1% level.
[b]Significant at 5% level.

B. Internal noise

The regression lines relating internal background L_{A90} levels in occupied classrooms and Key Stage 2 English and Mathematics scores are shown in Table XII. The linear relationships between noise and average Key Stage 1 and 2 scores are also shown. Table XII also shows the percentage decreases in the numbers of children achieving the required level in SATs for each 5 dB increase in internal background noise, plus the internal background noise levels in occupied classrooms, derived from the regression lines, which correspond to the UK government targets in English and Mathematics.

Table XII shows that there is a 13% reduction in the number of children achieving the required level at Key Stage 1 and a 12% reduction at Key Stage 2, for each 5 dB(A) increase in the background noise level in occupied classrooms. The background noise level corresponding to the government target for literacy is 53 dB(A) L_{A90}, while for numeracy it is 50 dB(A) L_{A90}. As with external levels, care is needed in interpreting these figures as discussed in Sec. VI.

VI. DISCUSSION

The study described here has shown that chronic exposure to noise at school has a detrimental effect upon children's academic performance, as measured by standard assessment testing in schools in England and Wales. These are consistent with the findings of previous studies and with the results of experimental testing of children carried out by the authors, as will be discussed in the following. Both external environmental noise heard inside a school and noise generated within a school have an impact upon children's test scores, but affect children in different ways. In addition to different subjects being affected by external and by school noise, the particular characteristics of the noise which impact upon children's performance differ between the two types of noise.

A. External noise

It was seen that different results were obtained for the suburban (A) and central (B and C) boroughs. For borough A there were strong relationships between all noise parameters and all test scores when all schools were considered, but for the other boroughs significant relationships were found when only the schools on the noisier sites were considered. The reasons for the discrepancies are not fully understood but may relate to differences in demographic, population, and/or noise characteristics between the boroughs. There may be "floor" effects for the inner city boroughs in that, however low the noise levels, the overall school test scores would not improve above a certain level. As was noted earlier the two central boroughs considered had high levels of children with SEN. The parallel experimental study carried out by the authors[29] showed that children with SEN were particularly vulnerable to the effects of noise so it is possible that this factor limits the overall achievements of these schools.

TABLE X. Internal noise: Correlation coefficients between test scores and school location L_{A90} levels corrected for FSM, EAL, and SEN.

	Occupied classroom $N=16$			Unoccupied classroom $N=14$			Corrifor/foyer $N=14$		
	FSM	EAL	SEN	FSM	EAL	SEN	FSM	EAL	SEN
KS1 Reading	−0.44	−0.47	−0.60[b]	−0.21	−0.30	−0.45	−0.26	−0.30	−0.40
KS1 Maths	−0.36	−0.40	−0.60[b]	−0.30	−0.40	−0.57[b]	−0.25	−0.29	−0.40
KS2 English	−0.66[a]	−0.69[a]	−0.76[a]	0.19	0.03	−0.17	−0.55[b]	−0.58[b]	−0.64[b]
KS2 Maths	−0.30	−0.36	−0.49	0.06	−0.07	−0.22	−0.40	−0.43	−0.48
KS2 Science	−0.42	−0.42	−0.48	−0.18	−0.21	−0.29	−0.31	−0.33	−0.40
KS1 average	−0.38	−0.44	−0.59[b]	−0.24	−0.36	−0.51	−0.26	−0.31	−0.41
KS2 average	−0.51[b]	−0.54[b]	−0.63[a]	0.01	−0.10	−0.26	−0.44	−0.47	−0.54

[a]Significant at 1% level.
[b]Significant at 5% level.

TABLE XI. Borough A: Regression lines relating external noise levels and SAT scores.

	L_{Aeq}			L_{Amax}			L_{A90}		
	Regression equation	% drop ≈10 dB increase	Level≈target	Regression equation	% drop ≈10 dB increase	Level≈target	Regression equation	% drop ≈10 dB increase	Level≈target
KS2 English	$y=-0.76x+112$	8	42	$y=-0.70x+118$	7	54.2	$y=-0.95x+115$	10	36.8
KS2 Maths	$y=-0.72x+107$	7	44.4	$y=-0.71x+116$	7	57.7	$y=-0.82x+106$	8	37.8
KS1 average	$y=-0.49x+104$	5	⋯	$y=-0.37x+102$	4	⋯	$y=-0.63x+107$	6	⋯
KS2 average	$y=-0.73x+113$	7	⋯	$y=-0.70x+120$	7	⋯	$y=-0.87x+114$	9	⋯

In general, for the suburban borough and for the noisier schools in the inner city boroughs correlations between noise and test scores were stronger for Key Stage 2 scores than for those at Key Stage 1 suggesting that external noise has more of an effect on the older children. It has previously been found that the negative effects of environmental noise are long term.[13,16] The greater effect upon the older children may therefore reflect the fact that these children have been exposed to noise at school for a longer period than the younger children. It may also be due to the higher task demands required of the older children in their tests.

In general, over all boroughs, the noise parameter with the highest and most significant correlations with test scores was L_{Amax}, implying that noise of individual events may be the most important in affecting children's performance. However, in the suburban borough external background noise levels, L_{A90}, were also significantly related to test scores.

Significant relationships between tests scores and noise were maintained when the data were corrected for factors relating to social deprivation, non-native speaking, and additional educational needs. In particular in all boroughs (considering just the noisier schools in the inner city boroughs) all KS2 subjects remained significantly related to L_{Amax} while KS1 Reading was also significantly related to some noise parameters.

The dominant external noise source in the schools considered was road traffic.[1] These findings are thus consistent with the findings of other studies which have found that road traffic noise has an impact upon children's performance at school.[19–21] Furthermore, although schools exposed to aircraft noise were not included in the study, the close relationships between L_{Amax} and test scores suggest that the noise of individual events has an impact upon children's performance. This is thus consistent with the results of other studies which have found that both aircraft[12–16] and railway[17] noise affect children's performance.

TABLE XII. Regression lines relating L_{A90} in occupied classrooms and SAT scores.

	Occupied classrooms L_{A90}		
	Regression equation	% drop ≈5 dB increase	Level≈target
KS2 English	$y=-3.23x+250$	16	52.6
KS2 Maths[a]	$y=-1.87x+169$	9	50.3
KS1 average	$y=-2.55x+218$	13	⋯
KS2 average	$y=-2.45x+207$	12	⋯

Correlation ($r=-0.46$) not significant.

The results also complement the findings of a questionnaire survey of children carried out by the authors which found that the older (Year 6) children were more aware of external noise than the younger children.[32] This is consistent with the finding that the test results of these children were more affected by noise than those of the younger children. Furthermore, annoyance caused by external noise among children was significantly related to external maximum noise levels, the levels that are found to have the most effect upon test scores.

Regression analysis has been used to estimate the noise levels corresponding to UK government targets in English and Mathematics in the suburban borough. In this borough those schools where the external L_{Amax} level 4 m from the school façade exceeds 54 dB(A), or L_{Aeq} exceeds 42 dB(A), fail to meet literacy and numeracy targets. These levels are considerably lower than those recommended in current guidelines,[28] and should be interpreted with caution. As can be seen from Figs. 1–3 there is considerable scatter around the regression lines; many schools with levels greater than these do achieve the SAT targets. Furthermore, there are many other factors apart from noise which may affect children's attainments; the regression analysis was carried out for uncorrected data where additional factors which may impact upon learning are not accounted for. These results may therefore not apply to schools in general.

B. Internal noise

There were consistent negative relationships between test scores and L_{Aeq} and L_{A90} levels measured in occupied and unoccupied classrooms and corridors and foyers. The internal noise levels which had the strongest relationships with test scores were the background (L_{A90}) levels in occupied classrooms. All subjects except KS2 Mathematics were significantly correlated with these levels. KS1 Mathematics was also significantly correlated with L_{A90} measured in unoccupied classrooms and KS2 English with L_{Aeq} and L_{A90} measured in corridor and foyer areas. Many of the relationships, particularly those for KS2 English, were maintained when the data were corrected for socio-economic factors.

These results complement the results of the controlled experimental testing of children carried out by the authors in which children performed various tasks in different class-

room noise conditions.[29] Classroom babble was found to decrease performance on both verbal and nonverbal tasks, with verbal tasks of reading and spelling being particularly affected. This is consistent with the finding that KS2 English test scores are strongly and significantly related to the ambient and background noise levels in classrooms.

Regression analysis showed that of the schools surveyed, in general those in which background (L_{A90}) levels in occupied classrooms exceed 50 dB(A) failed to meet government targets in literacy and numeracy. Current guidelines specify internal levels in classrooms in terms of ambient L_{Aeq} when both classrooms and the whole school are unoccupied. It is difficult, without further extensive noise surveys in schools both empty and occupied, to compare the occupied classroom background noise level with those in current standards. Furthermore, as with the external levels there is considerable scatter around the regression lines as can be seen in Figs. 4–6; therefore care should be taken when interpreting these results.

VII. CONCLUSION

This study has shown that chronic exposure to both external and internal noise has a detrimental impact upon the academic performance and attainments of primary school children. For external noise it appears to be the noise levels of individual events that have the most impact while background noise in the classroom also has a significant negative effect. Older primary school children, around 11 years of age, appear to be more affected by noise than the younger children.

In order to minimize the impact of noise upon children at school it is therefore necessary to consider two factors. The siting and the internal layout of a school should be such that classrooms are not exposed to high levels of noise from external sources such as road traffic. In addition it is essential to minimize background noise levels in the classroom to ensure that optimum conditions for teaching and learning are achieved.

Further field and experimental studies are required to determine the levels at which different types of external and internal noise affect children's academic performance in different circumstances.

ACKNOWLEDGMENTS

This research was funded by the Department of Health and Department for Environment, Food, and Regional Affairs (DEFRA). The authors would like to thank research assistants Rebecca Asker and Ioannis Tachmatzidis for collecting the data in this study, and the London boroughs and schools that participated in the study.

[1]B. Shield and J. E. Dockrell, "External and internal noise surveys of London primary schools," J. Acoust. Soc. Am. **115**, 730–738 (2004).
[2]E. Celik and Z. Karabiber, "A pilot study on the ratio of schools and students affected from noise," Proceedings of the International Symposium on Noise Control and Acoustics for Educational Buildings, Turkish Acoustical Society, Istanbul, May 2000, pp. 119–128.
[3]M. Hodgson, R. Rempel, and S. Kennedy, "Measurement and prediction of typical speech and background noise levels in university classrooms during lectures," J. Acoust. Soc. Am. **105**, 226–233 (1999).
[4]M. Picard and J. S. Bradley, "Revisiting speech interference in classrooms," Audiology **40**, 221–224 (2001).
[5]A. Moodley, "Acoustic conditions in mainstream classrooms," J. British Association of Teachers of the Deaf **13**, 48–54 (1989).
[6]B. Hay, "A pilot study of classroom noise levels and teachers' reactions," J. Voice **4**, 127–134 (1995).
[7]D. MacKenzie, "Noise sources and levels in UK schools," Proceedings of the International Symposium on Noise Control and Acoustics for Educational Buildings, Turkish Acoustical Society, Istanbul, May 2000, pp. 97–106.
[8]B. Berglund and T. Lindvall, "Community Noise," Archives of the Center for Sensory Research, Stockholm University and Karolinska Institute, **2**, 1–195 (1995).
[9]Institute for Environment and Health, "The non-auditory effects of noise," Report No. R10, 1997.
[10]R. Hetu, C. Truchon-Gagnon, and S. A. Bilodeau, "Problems of noise in school settings: A review of literature and the results of an exploratory study," J. Speech-Language Pathology and Audiology **14**, 31–38 (1990).
[11]G. W. Evans and S. J. Lepore, "Nonauditory effects of noise on children: A critical review," Children's Environments **10**, 31–51 (1993).
[12]S. Cohen, G. W. Evans, D. S. Krantz, and D. Stokols, "Physiological, motivational, and cognitive effects of aircraft noise on children. Moving from the laboratory to the field," Am. Psychol. **35**, 231–243 (1980).
[13]S. Hygge, G. W. Evans, and M. Bullinger, "The Munich Airport noise study: Cognitive effects on children from before to after the change over of airports," Proceedings Inter-Noise'96, Liverpool, UK, pp. 2189–2192.
[14]M. M. Haines, S. A. Stansfeld, J. Head, and R. F. S. Job, "Multi-level modelling of aircraft noise on performance tests in schools around Heathrow Airport London," J. Epidemiol. Community Health **56**, 139–144 (2002).
[15]C. Clark, R. Martin, E. van Kempen, T. Alfred, J. Head, H. W. Davies, M. M. Haines, B. Lopez, M. Matheson, and S. Stansfeld, "Exposure-effect relations between aircraft and road traffic noise exposure at school and reading comprehension: The RANCH project," Am. J. Epidemiol. **163**, 27–37 (2006).
[16]S. Cohen, G. W. Evans, D. S. Krantz, D. Stokols, and S. Kelly, "Aircraft noise and children: Longitudinal and cross-sectional evidence on adaptation to noise and the effectiveness of noise abatement," J. Pers Soc. Psychol. **40**, 331–345 (1981).
[17]A. L. Bronzaft and D. P. McCarthy, "The effect of elevated train noise on reading ability," Environ. Behav. **7**, 517–527 (1975).
[18]A. L. Bronzaft, "The effect of a noise abatement program on reading ability," J. Environ. Psychol. **1**, 215–222 (1981).
[19]J. S. Lukas, R. B. DuPree, and J. W. Swing, "Report of a study on the effects of freeway noise on academic achievement of elementary school children, and a recommendation for a criterion level for a school noise abatement program," J. Exp. Psychol. Learn. Mem. Cogn. **20**, 1396–1408 (1981).
[20]S. Sanz, A. M. Garcia, and A. Garcia, "Road traffic noise around schools: A risk for pupils' performance?," Int. Arch. Occup. Environ. Health **65**, 205–207 (1993).
[21]J. Romero and D. Lliso, "Perception and acoustic conditions in secondary Spanish schools," Proceedings of the 15th International Congress on Acoustics, Trondheim, Norway, 1995, pp. 271–274.
[22]F. Berg, J. Blair, and P. Benson, "Classroom acoustics: The problem, impact and solution," Language, Speech and Hearing Services in Schools **27**, 16–20 (1996).
[23]S. Airey and D. Mackenzie, "Speech intelligibility in classrooms," Proc. Institute of Acoustics **21**, 75–79 (1999).
[24]L. Maxwell and G. Evans, "The effects of noise on pre-school children's pre-reading skills," J. Environ. Psychol. **20**, 91–97 (2000).
[25]P. Lundquist, K. Holmberg, and U. Landstrom, "Annoyance and effects on work from environmental noise at school," Noise Health **2**, 39–46 (2000).
[26]World Health Organisation, *Guidelines for Community Noise* (Geneva, 1999).
[27]American National Standards Institute, "ANSI S12.60-2002: Acoustical performance criteria, design requirements, and guidelines for schools," New York, 2002.
[28]Department for Education and Skills, "Building Bulletin 93: Acoustic design of schools," The Stationery Office, London, 2003.
[29]J. E. Dockrell and B. M. Shield, "Acoustical barriers in classrooms: The impact of noise on performance in the classroom," British Educational Research Journal **32**, 509–525 (2006).
[30]W. Williamson and D. D. Byrne, "Educational disadvantage in an urban

setting," in *Social Problems and the City*, edited by D. T. Herbert and D. M. Smith (Oxford University Press, Oxford, 1977).
[31]P. Sammons, A. West, and A. Hind, "Accounting for variations in pupil attainment at the end of Key Stage 1," British Educational Research Journal **23**, 489–511 (1997).
[32]J. E. Dockrell and B. M. Shield, "Children's perceptions of their acoustic environment at school and at home," J. Acoust. Soc. Am. **115**, 2964–2973 (2004).

On boundary conditions for the diffusion equation in room-acoustic prediction: Theory, simulations, and experiments[a)]

Yun Jing and Ning Xiang[b)]
Graduate Program in Architectural Acoustics, School of Architecture, Rensselaer Polytechnic Institute, Troy, New York 12180, USA

(Received 3 July 2007; revised 12 October 2007; accepted 12 October 2007)

This paper proposes a modified boundary condition to improve the room-acoustic prediction accuracy of a diffusion equation model. Previous boundary conditions for the diffusion equation model have certain limitations which restrict its application to a certain number of room types. The boundary condition employing the Sabine absorption coefficient [V. Valeau *et al.*, J. Acoust. Soc. Am. **119**, 1504–1513 (2006)] cannot predict the sound field well when the absorption coefficient is high, while the boundary condition employing the Eyring absorption coefficient [Y. Jing and N. Xiang, J. Acoust. Soc. Am. **121**, 3284–3287 (2007); A. Billon *et al.*, Appl. Acoust. **69**, (2008)] has a singularity whenever any surface material has an absorption coefficient of 1.0. The modified boundary condition is derived based on an analogy between sound propagation and light propagation. Simulated and experimental data are compared to verify the modified boundary condition in terms of room-acoustic parameter prediction. The results of this comparison suggest that the modified boundary condition is valid for a range of absorption coefficient values and successfully eliminates the singularity problem. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2805618]

PACS number(s): 43.55.Br, 43.55.Ka [EJS] Pages: 145–153

I. INTRODUCTION

Room-acoustic predictions have been actively researched in the area of architectural acoustics for decades. A diffusion equation-based method has recently drawn attention in this field, partially due to its ability to combine accuracy and efficiency. The purpose of this work is to present a modified boundary condition for diffusion equation-based room-acoustics prediction.

In 1969, Ollendorff[1] first proposed the diffusion model to describe sound fields in enclosures. Later, based on an analogy between sound propagation in rooms with diffusely reflecting walls and particle propagation in a diffusing fluid,[2] Picaut[3] and his co-workers[4–8] demonstrated this diffusion model and extended its application to a variety of enclosure types, including elongated enclosures such as street canyons, single-volume enclosures, fitted rooms, and coupled-volume rooms. For room-acoustics modeling, Valeau *et al.*[5] provide a generalization of the diffusion model that includes two boundary conditions. These studies[3–8] have shown that the diffusion model is computationally efficient compared to the geometrical-acoustics model. The diffusion model also provides more satisfactory results than does statistical room-acoustics theory since it is capable of modeling the nonuniformity of sound fields in enclosures of a wide variety of shapes. However, the diffusion model has been found to be suitable only for use with low absorption coefficients.[5–7] Most recently, another boundary condition exploiting the Eyring absorption coefficient was proposed independently by Jing and Xiang[9] and Billon *et al.*,[10] and was shown to be able to improve the accuracy of the diffusion model when the absorption coefficient of room surfaces is high. The use of the Eyring absorption coefficient is also justified theoretically in this paper.

In this paper, a modified boundary condition is derived from a boundary condition used in solving analogous optical diffusion problems.[11] Incorporating the modified boundary condition in the room-acoustic diffusion equation, this work compares the diffusion models associated with various boundary conditions with a geometrical-acoustic model as well as experimental results obtained from a physical scale model. Predicted reverberation times (RTs) and sound pressure levels (SPLs) are calculated for cubic rooms with both uniformly and nonuniformly distributed absorbing surfaces. For a flat, long room, the geometrical-acoustics model is used for the comparison of the SPL distribution. Acoustical measurements are conducted in a scale-model flat room to validate the modified boundary condition in terms of the predicted SPLs and the RTs.

This paper is structured as follows: Sec. II presents the diffusion equation model for room-acoustic prediction including the interior equation and different boundary conditions. Section III discusses simulation results obtained using the diffusion equation model with different boundary conditions, as well as those obtained using the geometrical-acoustic model. Section IV describes the scale model experiments, and compares the acoustical measurement results with the diffusion equation model for various boundary conditions. Section V concludes the paper.

[a)]Aspects of this work have been presented at the 153rd meeting of the Acoustical Society of America [J. Acoust. Soc. Am. **121**, 3174 (A) (2007)].
[b)]Author to whom correspondence should be addressed. Electronic mail: xiangn@rpi.edu

II. DIFFUSION EQUATION MODEL FOR ROOM-ACOUSTIC PREDICITION

A. Interior diffusion equation

This section begins with a review of the analogy drawn by Picaut *et al.*[3] between sound propagation in rooms with diffusely reflecting boundaries and gas particle propagation in a diffusing fluid. The motion of a sound particle[12] in a room is equivalent to the movement of a particle in a gas, assuming that there are numerous spherical scattering objects[2] in the volume having scattering cross section Q_s and absorption cross section Q_a. Furthermore, in order to respect the same surface distribution between the enclosure and the diffusing fluid, the scattering objects should have mean free path length $4V/S$, (for volume V and interior surface area S of the enclosure under investigation), which is consistent with classical room-acoustic theory when the walls are considered diffusely reflecting. The scattering objects take the place of the walls and reflect the sound diffusely in the volume. In this case, the sound energy flow vector $\mathbf{J}$ caused by the gradient of the sound energy density w, in the room under investigation at position $\mathbf{r}$ and time t, can be expressed according to Fick's law as[2,3]

$$\mathbf{J}(\mathbf{r},t) = -D \operatorname{grad} w(\mathbf{r},t), \tag{1}$$

with diffusion coefficient D

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

where c is the speed of sound and $\lambda = 4V/S$ is the mean free path of the enclosure under investigation. Here we assume that the rate of change involved in diffusion is slow and the walls are diffusely reflecting. The sound energy density w, in a region (domain V) excluding sound sources, changes per unit time as

$$\frac{\partial w(\mathbf{r},t)}{\partial t} = -\operatorname{div} \mathbf{J}(\mathbf{r},t) = D\nabla^2 w(\mathbf{r},t). \tag{3}$$

In the case of a spatially distributed omni-directional sound source within a region with time-dependent energy density $q(\mathbf{r},t)$ Eq. (3) is replaced by

$$\frac{\partial w(\mathbf{r},t)}{\partial t} = q(\mathbf{r},t) + D\nabla^2 w(\mathbf{r},t), \tag{4}$$

while absorption at the room boundaries leads to energy loss per unit volume (σw) in the absence of sound sources, so the energy balance can be written as

$$\begin{aligned}\frac{\partial w(\mathbf{r},t)}{\partial t} &= D\nabla^2 w(\mathbf{r},t) - \sigma w(\mathbf{r},t) \\ &= D\nabla^2 w(\mathbf{r},t) - Q_a n_t c w(\mathbf{r},t),\end{aligned} \tag{5}$$

where $\sigma = \bar{\alpha} c/\lambda = Q_a n_t c$, with $\bar{\alpha}$ being the mean room-surface absorption coefficient according to Ref. 3. The validity of Eq. (5) will be discussed in Sec. II B.

Taking into account both absorption of room boundaries and sound source excitations in the room under investigation, the combination of Eqs. (4) and (5) yields the diffusion equation

$$\frac{\partial w(\mathbf{r},t)}{\partial t} - D\nabla^2 w(\mathbf{r},t) + \sigma w(\mathbf{r},t) = q(\mathbf{r},t) \quad \text{in } V. \tag{6}$$

It is also possible to include air absorption inside the enclosure using the diffusion equation[13]

$$\begin{aligned}&\frac{\partial w(\mathbf{r},t)}{\partial t} - D\nabla^2 w(\mathbf{r},t) + \sigma w(\mathbf{r},t) + mcw(\mathbf{r},t) \\ &= q(\mathbf{r},t) \quad \text{in } V,\end{aligned} \tag{7}$$

where m is the coefficient of air absorption. All absorption coefficients are frequency dependent, so the frequency dependence of sound energy density of enclosures may be included in the diffusion equation model.[5] For simplicity, however, the following discussion will not consider the air absorption.

B. Boundary conditions

1. Previous boundary conditions

In an enclosure bounded by surfaces (denoted by S), the boundary condition

$$\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n} = -Dw(\mathbf{r},t)\cdot\mathbf{n} = 0, \quad \text{on } S \tag{8}$$

states that sound energy cannot escape from the room boundaries, with $\mathbf{n}$ denoting the surface outgoing normal. Equation (8) is the so-called *homogeneous Neumann boundary condition*.[2,5] The diffusion equation model with the homogenous Neumann boundary condition considers only an overall mean absorption coefficient $\bar{\alpha}$ of the enclosure under investigation. Furthermore, it assumes that the absorption occurs in the volume rather than on the room surfaces, which is nonphysical for the current application. Therefore, this boundary condition, mentioned here for completeness, will not be considered in the following investigations.

The boundary condition[2,5]

$$\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n} = -Dw(\mathbf{r},t)\cdot\mathbf{n} = A_X c w(\mathbf{r},t), \quad \text{on } S, \tag{9}$$

allows energy exchanges on the boundaries S, where A_X is an exchange coefficient denoted in the following as the *absorption factor*. This boundary condition is established to include the absorption at the walls: The local differential equation in Eq. (6) then needs to be modified to remove the absorption term $\sigma w(\mathbf{r},t)$ (the following boundary conditions all require that the absorption term in the interior equation drop out), while the term must be introduced into Eq. (9) to account for energy exchanges on the boundaries.[5]

Assuming the sound energy density is uniform in a proportionate room while supposing that the energy density only varies along the long dimension in a disproportionate room [5], the absorption factor can be expressed as

$$A_X = A_S = \frac{\alpha}{4}, \tag{10}$$

where α is the absorption coefficient of the wall under consideration. The subscript S of A_S is used to denote Sabine absorption. Equation (10) implies the Sabine absorption is assigned to the absorption factor A_X. The resulting system of equations is formulated as

$$\frac{\partial w(\mathbf{r},t)}{\partial t} - D\nabla^2 w(\mathbf{r},t) = q(\mathbf{r},t) \quad \text{in } V, \tag{11}$$

$$D\frac{\partial w(\mathbf{r},t)}{\partial n} + cA_X w(\mathbf{r},t) = 0 \quad \text{on } S, \tag{12}$$

which describe the diffusion equation model [Eq. (11)] with *mixed boundary conditions* [Eq. (12)] for more general situations. The absorption factor A_X assumes different expressions with different subscript X as elaborated in the following. It will then be possible to assign different absorption coefficients to each of the individual walls.

The diffusion equation model with this boundary condition is, however, found to be accurate only for modeling rooms with low absorption.[5–7] To improve the accuracy of the mixed boundary condition associated with high absorption, a new boundary condition was conceived independently by these authors[9] and Billon *et al.*[10] It simply replaces the Sabine absorption coefficient in the absorption factor by the Eyring absorption coefficient

$$A_X = A_E = \frac{-\log(1-\alpha)}{4} \tag{13}$$

to account for high wall absorption in the system of equations in Eqs. (11) and (12).

A justification of the replacement of Eyring absorption coefficient is detailed here. Under the assumption that there are scattering objects in the enclosure, the number dN of phonons absorbed by the scattering objects, between the points x and $x+dx$ along the path of the beam, should be equal to the product of N, the number of phonons penetrating to depth x, n_t, the number of scattering objects per unit volume and the absorption cross section Q_a[2]

$$dN/dx = -Nn_tQ_a, \tag{14}$$

where the negative sign indicates the reduction of the phonon due to absorption. Then, a simple integral over λ (the mean free path length) gives

$$N' = N\exp(-n_tQ_a\lambda), \tag{15}$$

where N' is the number of phonons after penetrating to depth λ.

In room acoustics, the phonon number is reduced after each wall collision according to the absorption coefficient of the wall, which leads to

$$N' = N(1-\bar{\alpha}), \tag{16}$$

where $\bar{\alpha}$ is the mean room-surface absorption coefficient. The combination of Eq. (15) and Eq. (16) yields (see also Ref. 14)

$$-\ln(1-\bar{\alpha}) = n_tQ_a\lambda. \tag{17}$$

A substituting of Eq. (17) into Eq. (5) yields a diffusion equation which is similar to Eq. (5), but with $\sigma = -\ln(1-\bar{\alpha})c/\lambda$. Finally, the Eyring absorption coefficient is applied to each boundary by using an exchange coefficient,[5,9] which gives the boundary condition, i.e., Eq. (12) along with Eq. (13).

Accordingly, the diffusion equation model with the Sabine coefficient in the absorption factor ($A_X = A_S$) is denoted *diffusion-Sabine* model in the following while the model which utilizes the one with the Eyring coefficient ($A_X = A_E$) is denoted *diffusion-Eyring* model.[10] Although the simulation results suggest that the diffusion-Eyring model can improve the accuracy of the diffusion-Sabine model to a certain extent, particularly for cases of high absorption,[9,10] application of this model requires caution as the absorption coefficient $\alpha \to 1.0$. As discussed in the following, the absorption factor A_E will become singular (infinite) in this case, which is problematic.

2. Modified boundary condition

This section presents a modified boundary condition with theoretical justification. The derivation of this boundary condition relies heavily on a well established boundary for light diffusion in media.[11] The modified boundary condition remedies weaknesses in the original diffusion-Sabine model under high absorption and eliminates the singularity of the diffusion-Eyring model. In the following, the diffusion equation model with the modified boundary condition is denoted *modified diffusion model*.

The statement of Picaut *et al.*,[3] which introduces the diffusion interior equation, describes an infinite scattering medium.[5] To bound the scattering medium, the boundary, on the one hand, absorbs a portion of the sound energy at a given absorption coefficient on the corresponding wall. On the other hand, since the boundary is not considered as the wall but artificially imposed to bound the energy, it can be considered to be completely smooth and to reflect specularly the other portion of the sound energy to prevent this portion from leaving the volume. Similar to the previous boundary conditions, different absorption coefficients could be assigned to individual boundaries.

Based on a similar boundary condition in light diffusion, the derivation of the modified boundary condition is presented. Emphasis is given to differences from optical diffusion, and to an explicit formulation of the acoustic boundary condition.

A partial-current boundary condition[11] in light diffusion describes the scenario in which a portion of the energy will be specularly reflected back into the scattering medium when the refractive indices of the strongly scattering medium differ substantially from those of the bounding transparent medium, while the other portion will be refracted and leave the scattering medium.

This boundary condition can be expressed using an optical term fluence rate ϕ in unit W m^{-2}.[11] Because sound intensity has the same unit, in the following room-acoustic discussion, notation ϕ represents sound intensity. The relationship between the sound intensity and the sound energy density can be expressed as,[15]

$$\phi(\mathbf{r},t) = c\mathbf{n}'w(\mathbf{r},t), \tag{18}$$

where $\mathbf{n}'$ is the direction of energy propagation.

Furthermore, the boundary condition is written as

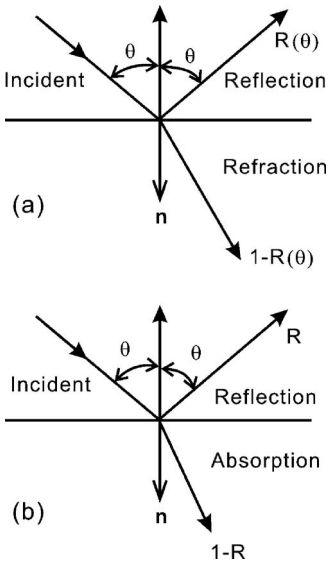


FIG. 1. (a) Light reflection and refraction on a boundary. (b) Sound reflection and absorption on a boundary.

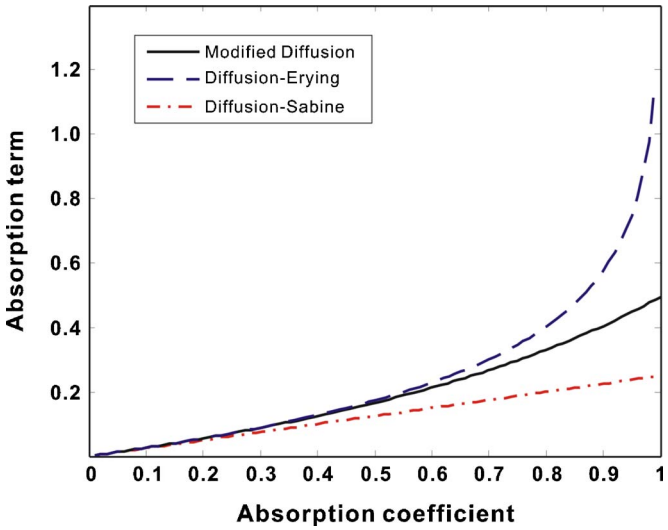


FIG. 2. (Color online). Comparison of the absorption terms of the diffusion-Sabine, diffusion-Eyring, and modified diffusion models versus absorption coefficient.

$$\phi(\mathbf{r},t)=\frac{1+R_j}{1-R_\phi}2\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n}\quad\text{on } S, \tag{19}$$

where

$$R_\phi=\int_0^{\frac{\pi}{2}}2\sin\theta\cos\theta R(\theta)d\theta, \tag{20}$$

$$R_j=\int_0^{\frac{\pi}{2}}3\sin\theta\cos^2\theta R(\theta)d\theta, \tag{21}$$

$$\mathbf{J}(\mathbf{r},t)\cdot\mathbf{n}=-D'\frac{\partial\phi(\mathbf{r},t)}{\partial n}, \tag{22}$$

with $D'=\lambda/3$ being one third of the mean free path length in the medium. Figure 1a schematically illustrates the boundary condition. $R(\theta)$ is the probability of the light being reflected, also so-called Fresnel reflection coefficient which is a function of the angle of incidence θ,[11] $\mathbf{n}$ is the outward-drawn normal to the boundary. $\mathbf{J}$ is the energy flow as explained earlier [see Sec. II A].

The same boundary condition may also be applied to room acoustics since the room-acoustic diffusion equation is essentially the same as the light diffusion equation, and the specular reflection is also desired for the modified acoustic boundary condition proposed in this paper. However, some significant differences should be noted: (1) In the light diffusion boundary condition, the light refracts at the boundary, while in room-acoustics boundary condition, the sound energy is usually considered to be absorbed at the boundary. Nevertheless, since the focus in the current application is on the energy inside the volume, this work disregards the way of sound transmissions beyond the boundary. (2) In room acoustics, the reflectivity R which represents the probability that the sound will be reflected (energy-based reflectivity) can be expressed using the absorption coefficient as $1-\alpha$, so that it is completely independent of incident angle. (Figure 1b illustrates this concept.) Thus,

$$R_\phi=(1-\alpha)\int_0^{\frac{\pi}{2}}2\sin\theta\cos\theta d\theta=1-\alpha, \tag{23}$$

$$R_j=(1-\alpha)\int_0^{\frac{\pi}{2}}3\sin\theta\cos^2\theta d\theta=1-\alpha, \tag{24}$$

A substitution of Eqs. (18) and (22)–(24) into (19) yields

$$cw(\mathbf{r},t)=-\frac{2-\alpha}{\alpha}2D\frac{\partial w(\mathbf{r},t)}{\partial n}\quad\text{on } S, \tag{25}$$

where $D=\lambda c/3$, leading to the modified boundary condition

$$D\frac{\partial w(\mathbf{r},t)}{\partial n}+\frac{c\alpha}{2(2-\alpha)}w(\mathbf{r},t)=0\quad\text{on } S. \tag{26}$$

A comparison of the boundary conditions of the diffusion-Sabine and diffusion-Eyring models reveals that the only difference between them is the absorption factor associated with absorption coefficient α,

$$A_X=A_M=\frac{\alpha}{2(2-\alpha)}. \tag{27}$$

In order to enable a direct comparison of these boundary conditions, Fig. 2 illustrates the three absorption factors in Eqs. (10), (13), and (27) when varying the absorption coefficient. When the absorption coefficient is lower than 0.2, the difference between the absorption factors A_S, A_E, A_M is negligible. With increasing absorption coefficient but below 0.6, the difference between absorption term A_E and A_M is still negligible, while the absorption term A_S differs quite significantly from the other two. As the absorption coefficient increases further, the discrepancy between A_E and A_M is considerable, warranting in-depth investigation. This behavior is easily explained by the Taylor expansion of the three absorption factors

$$A_S = \frac{\alpha}{4}, \tag{28}$$

$$A_E = \frac{-\log(1-\alpha)}{4} = \frac{1}{4}\alpha + \frac{1}{8}\alpha^2 + \frac{1}{12}\alpha^3 + \cdots, \tag{29}$$

$$A_M = \frac{\alpha}{2(2-\alpha)} = \frac{1}{4}\alpha + \frac{1}{8}\alpha^2 + \frac{1}{16}\alpha^3 + \cdots. \tag{30}$$

Note that the Eyring and modified absorption factors share their first two terms, and the Sabine absorption factor is the first order approximation of both of these. (That Sabine absorption coefficient is the first order approximation of the Eyring absorption coefficient is already well known in room acoustics.)

In addition, as the value of the absorption coefficient approaches one, A_E differs substantially from A_M: A_E increases unbounded while A_M maintains a finite value of 0.5. While the Eyring reverberation time equation[16] accounts reasonably well for the resulting infinitely short reverberation time, the Eyring-absorption factor A_E in the diffusion equation becomes singular. This is the failing point of the diffusion-Eyring model, which limits its applicability in often-encountered situations where walls, portions of walls, or opening in them have an absorption coefficient of 1.0.

III. SIMULATIONS

In this section, three diffusion equation models (the diffusion equation with each of the three different boundary conditions) are compared to the geometrical-acoustic method implemented by a commercial software (CATT acoustics®) for three types of rooms: cubic rooms with both uniform and nonuniform absorption distribution, and a flat, long room, in terms of the RTs and SPLs. The diffusion equation models are solved by a commercially available finite element solver. The size of the mesh elements is chosen to be on the order of or smaller than one mean free path $4V/S$ of the room.[5] The time step is chosen as 0.01 s for every case involving time-dependent calculations.

Equations (11) and (12) are solved for the initial condition

$$w(\mathbf{r},0) = 0 \quad \text{in } V, \tag{31}$$

$$w(\mathbf{r},0) = w_0 \quad \text{in } V_s, \tag{32}$$

where V_s is the volume occupied by the sound source.[5] With a time-dependent solution $w(\mathbf{r},t)$, the sound energy-time function can be expressed as[15,17]

$$L_p(\mathbf{r},t) = 10\log\left(\frac{w(\mathbf{r},t)\rho c^2}{P_{\text{ref}}^2}\right), \tag{33}$$

where P_{ref} equals 2×10^{-5} Pa. The sound energy decay functions as well as the RTs can then be obtained.

To calculate the steady state sound field, Eqs. (11) and (12) are solved for a given sound power W_s of the source, and then $q(\mathbf{r},t)$ is set to be equal to W_s/V_s. With a stationary solution $w(\mathbf{r})$, the total SPL $L_p^{\text{tot}}(\mathbf{r})$, including the direct field attributable to the point source, can be expressed as

$$L_p^{\text{tot}}(\mathbf{r}) = 10\log\{\rho c[W_s/(4\pi r^2) + w(\mathbf{r})c]/P_{\text{ref}}^2\}. \tag{34}$$

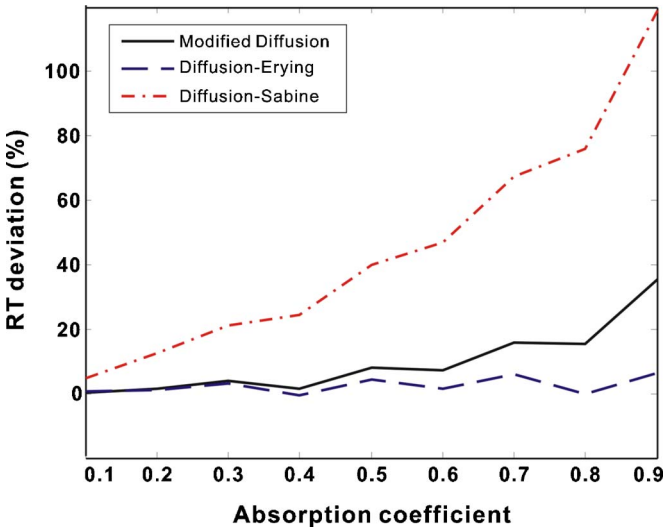


FIG. 3. (Color online). Deviations of the reverberation times calculated by the geometrical-acoustic model and three diffusion models.

A. Cubic rooms with uniformly distributed absorption coefficients

A cubic room with dimensions 5 m* 5 m*5 m is modeled. The source is in the center of the room at coordinate (0, 0, 0) m. The absorption coefficient is assigned uniformly to all room surfaces and ranges from 0.1 to 0.9. The RTs are obtained at the position (1, 0, 0) by three diffusion equation models, and by the geometrical-acoustic model. The number of rays and the ray truncation time in CATT acoustics® are chosen as 5×10^4 and 2000 ms, respectively, where the latter is much longer than the expected RT. The scattering coefficients on the walls are set at 100%, being compatible with the implicit requirement of the diffusion equation model, which is the configuration used throughout all simulations.

Figure 3 illustrates the difference, in terms of predicted RTs, between the results from the geometrical-acoustic model and those obtained via the other methods. Both the diffusion-Eyring and the modified diffusion model improve the simulation accuracy with respect to the diffusion-Sabine model, especially when the absorption coefficient is high. In addition, the diffusion-Eyring and modified diffusion models only show noticeable difference when the absorption coefficient exceeds approximately 0.6.

B. Cubic rooms with nonuniformly distributed absorption coefficients

The interior surfaces of another cubic room with the same dimensions are assigned two different absorption coefficients. One of the walls ($x=-2.5$ m) is given an absorption coefficient of 1.0 while the other walls are assigned an absorption coefficient 0.5. The source is still in the center of the room at (0,0,0) m while the receivers are distributed along $x=-2.5$ m to $x=2.5$ m, ($y=z=1$ m). The number of rays is

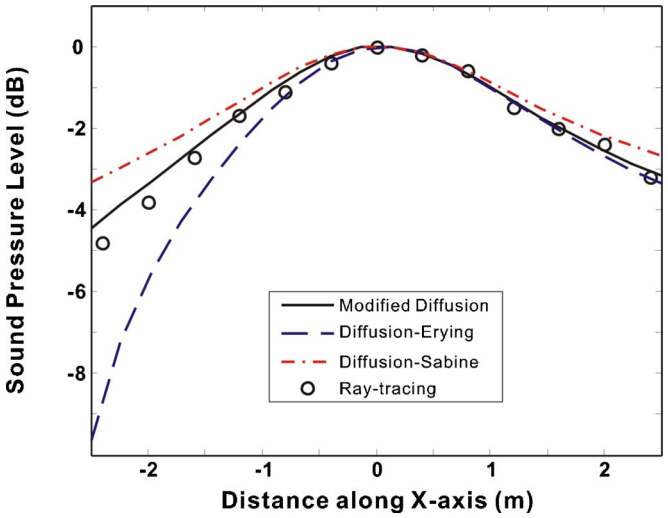


FIG. 4. (Color online). Comparison of sound pressure level distributions along $y=z=1$ m in four different models of a cubic room with two different absorption coefficients, 1.0 and 0.5.

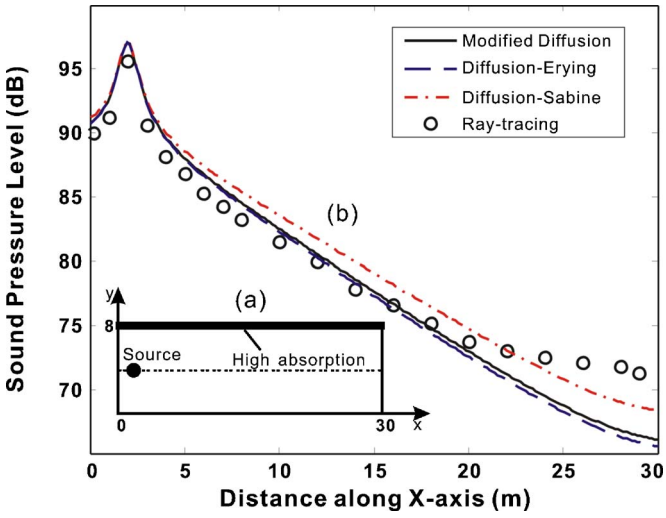


FIG. 5. (Color online). (a) Top view of a long, flat room with dimension 30 m*8 m*3 m, the sound source is at (2,4,2) m, the sound pressure levels are calculated along $y=4$ m (dotted line). (b) Comparison of sound pressure level distributions along $y=4$ m (dotted line) by four different models in a long, flat room with two different absorption coefficients, 0.8 and 0.2.

enlarged to 1×10^6 and the truncation time is set to 2000 ms. The number of rays is much larger than in the last case because many rays are lost when they impinge upon the wall with $\alpha=1.0$. Using the diffusion-Eyring model, since one wall is featured with the absorption coefficient 1, $-\log(1-\alpha)$ becomes an infinitely large value which cannot be implemented in the available finite element solver, thus, we use 1e10 instead of the infinitely large value. The results of SPLs shown in Fig. 4 indicate that:

1. The modified diffusion model predicts SPLs more accurately than does the diffusion-Sabine model.
2. When the room has one open side ($\alpha=1.0$), the diffusion-Eyring model yields incorrect results, especially near the open side. The reverberant sound energy density is nearly zero at the open side calculated by the diffusion-Eyring model which is nonphysical. The very big value of 1e10 is used to represent the infinitely large value for the absorption term $-\log(1-\alpha)$ in the simulations. Increasing this value towards infinity only causes it to deviate more from the results of the geometrical-acoustic model. Since the modified diffusion model does not have this singularity, it can realistically predict the sound field near highly absorptive boundaries.

When using CATT acoustics to obtain the correct late decay, the randomized tail-corrected cone-tracing method needs to extrapolate the reflection density growth (assuming it is quadratic) and is not accurate in an open room. Therefore, this work compares the predicted RTs by giving only the results of three diffusion equation models. In this case, three diffusion equation models have the following RTs at position (1 m, 1 m, 1 m): (diffusion-Sabine) 0.22 s, (diffusion-Eyring) 0.14 s, (modified diffusion) 0.19 s. Without any benchmark for comparison, no further conclusion is given at this time. However, an experiment involving open spaces is planned to provide the means for comparison and verification.

C. Flat and long room

In this example, the three diffusion equation models are compared to the geometrical-acoustics model for a flat, and long room with dimensions 30 m*8 m*3 m. The source is located at (2,4,2) m, and the sound power level W_s is 100 dB. One side wall is assigned absorption coefficient 0.8, while the other five walls share absorption coefficient 0.2. Figure 5(a) illustrates a top view of this room.

Figure 5(b) illustrates the SPL distribution along the line $y=4$ m (x is from 0 to 30 m) at height $z=1.5$ m. In CATT acoustics® software the number of rays and the ray truncation time are chosen as 5×10^4 and 2000 ms, respectively. Overall, the diffusion-Eyring model and the modified diffusion model agree better with the geometrical-acoustics model, although several points near the end of the room appear to be closer to the diffusion-Sabine model results.

IV. EXPERIMENTAL VERIFICATION

This section describes acoustical measurements in a scale-model flat room to verify the diffusion equation model and further compare different boundary conditions for room-acoustic prediction in terms of SPL distributions and RTs. For a flat room, the sound energy density is known to be nonuniform. Statistical room-acoustic theory is unable to handle spatial distributions in highly irregular room shapes. This is the reason such a room was chosen for the experimental verification.

A. Experimental setup

Physical scale modeling is a versatile research tool which has recently been applied in a variety of acoustic investigations.[18–22] In this study, a 1:8 scale model is used throughout the entire experiment. The dimensions of the scale model are 1 m*1 m*0.2 m, corresponding to 8 m*8 m*1.6 m in full scale. Hardwood along with one

FIG. 6. (Color online). Photograph of the flat room scale model showing the speaker, microphone, QRDs, foam, and rocks. A detailed photo of the scale-model sound source is shown in the left corner.

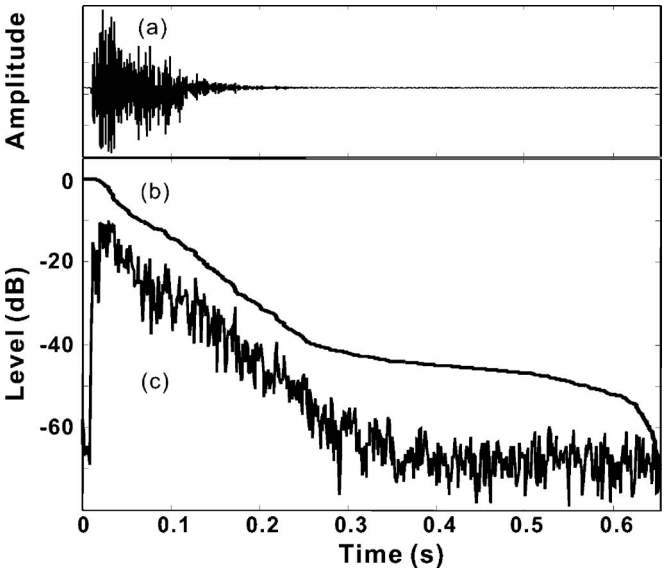


FIG. 7. Segment of one of the impulse responses, its energy-time curve, and its Schroeder decay function in the 1 kHz octave band. (a) Relative amplitude of the impulse response. (b) Schroeder decay curve. (c) Energy-time curve.

sheet of 5/4 in. plywood in thickness are used as wall materials to construct the room model. Diffusers on the walls are quadratic-residue diffusers (QRDs) and rocks, providing strong scattering at high frequency. In addition, such QRDs are known to be able to provide fairly good scattering for a wide range of frequencies. The QRDs are randomly distributed throughout the sidewalls (1.6 m high in full scale) of the scale model. The scattering frequency of the rocks will be discussed in the next section. In reality, diffuse reflections from the walls can be obtained even without such extensive treatment.[23] Two sidewalls near the sound source are of highly absorptive foam. Figure 6 shows a photograph of the scale model.

A 1/4 in. omni-directional microphone is used in the measurements. A miniature dodecahedron loudspeaker system is used as the sound source as shown in the left corner of Fig. 6. Measurements of directivities indicate that the dodecahedron loudspeaker system is omni directional in the frequency range of interest up to 32 kHz. Maximal-length sequence signals are used as an excitation signal. Room impulse responses are measured at chosen locations and the measurement procedure is controlled by a trigger mechanism throughout the entire measurement session to avoid any additional uncertainties across measurements. For frequency-dependent room-acoustic analysis the measured room impulse responses are filtered in octave bands. Figure 7 illustrates a segment of one of the room impulse responses, the corresponding energy-time function, and the Schroeder decay function at 1 kHz octave band.

The absorption coefficients of the materials are measured separately. Two proportionate scale-model rooms of which one is made of the rock-lined panels and the other is made of highly absorptive foam, are constructed and measured. The absorption coefficients of the rock panels and of the foam are estimated from the RTs by inverting the Eyring equation. To estimate the absorption coefficient of the QRD, all the foam in the scale model was first replaced by QRDs. Next, impulse responses are taken to obtain the RTs, and the absorption coefficient of the QRD is iteratively adjusted to match predictions of RTs to experimental results.[21] Of the measured coefficients, that of the QRD is the primary uncertainty. However, the room-acoustic parameters in this case are determined primarily by the absorption coefficient of the rock panels simply because their areas are much greater than those of the QRDs or the foam in the scale model. Note that the scale model testing in this study is not to model a specific existing room, but rather for validation purpose, so no extra measures, such as nitrogen or drying air, are conducted for correcting air absorption in the scale model. However, all of the absorption coefficients of the scale-model wall materials are estimated from measurements made under normal atmospheric conditions. The excess air absorption is considered to be included in the estimated absorption coefficients of the wall materials. Table I lists the estimated absorption coefficients.

B. Experimental verification of the diffusion equation model

The previous section compared the simulation results of three diffusion equation models with the geometrical-acoustic model. This section will compare experimental results from the scale-model flat room with the predictions of the diffusion equation model with the modified boundary condition.

Since the dimensions of the rocks are around 3–4 cm, they diffusely reflect the sound between 8.5 and 11.5 kHz,

TABLE I. Absorption coefficient of the materials used in the scale model at 1 and 2 kHz octave band (8 and 16 kHz octave band in the 1:8 scale model).

Material	1 kHz	2 kHz
QRD	0.10	0.05
Wooden panel with rocks	0.18	0.20
Foam	0.95	0.96

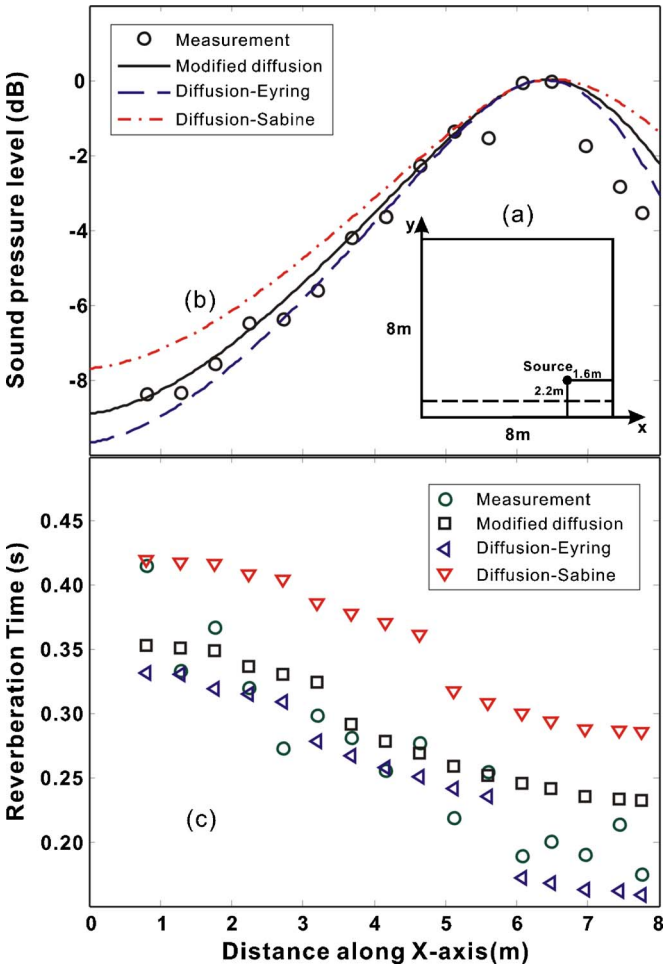


FIG. 8. (Color online). Comparison of three different diffusion models with experimental results in a scale-model flat room in the 1 kHz octave band along the line at y=0.64 m (full scale). (a) A top-view of the flat room scale model. The source is located at the bottom right. The measurements are conducted along the dotted line. (b) Sound pressure level distributions. (c) Reverberation times.

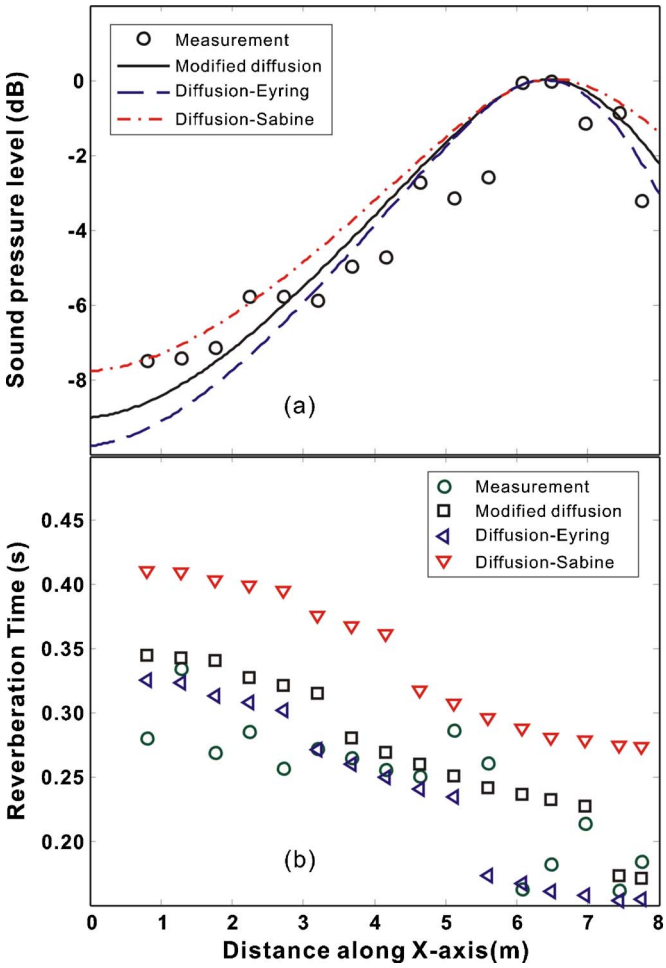


FIG. 9. (Color online). Comparison of three different diffusion models and experimental results in a scale-model flat room in the 2 kHz octave band along a line at y=0.64 m (full scale). (a) Sound pressure level distributions. (b) Reverberation times.

which lies mostly within 1 kHz octave band (and partially in 2 kHz octave band) given the chosen scale factor 1:8. Therefore, the room boundaries meet the inherent requirement of the diffusion equation model best in the 1 kHz octave band, while they qualify only partially in the 2 kHz octave band. Therefore, this section will discuss only the results obtained in these two octave bands.

The measurements are conducted at 16 receiving points along a line near the source. Figure 8(a) illustrates a top view of the scale model. Figure 8(b) and Fig. 9(a) illustrate the SPLs measured in the scale-model flat room for 1 and 2 kHz octave bands, respectively. The SPLs are relative values which are referenced to the value predicted in simulation at x=6.4 m (full scale), where all the diffusion equation models give the maximum value.

At 1 kHz, the diffusion-Eyring model and the modified diffusion model agree with the experimental results very well. The maximum deviation (2 dB) occurs near the foam, probably due to the fact that the foam is less diffusely reflecting. At 2 kHz, as shown in Fig. 9(a), although the diffusion-Sabine model shows better results for several points at the beginning, overall the diffusion-Eyring model and the modified diffusion model are closer to the experimental results. The worse agreement in comparison with those at 1 kHz should be expected, probably due to the lower degree of scattering of the rocks in this octave band.

At both 1 and 2 kHz, the original diffusion-Sabine model overestimates the RTs, while the diffusion-Eyring model and the modified diffusion model well predict the RTs as shown in Fig. 8(c) and Fig. 9(b). Moreover, the three diffusion equation models all predict the "drop" trend of the RT as the receiver location approaches the highly absorptive foam covered wall. The reason for this trend might be that the highly absorptive foam draws the energy much faster than the QRDs do. Note that the RTs predicted by the three diffusion equation models are not changed linearly along the almost linearly increasing x-axis distance. Some pronounced dips are found at different places for different diffusion equation models, while the experimental results also show several dips. For example, in Fig. 9(b), both experimental and the modified diffusion model results show a pronounced dip of the RT between x=7 m and x=8 m. No explanation for this has yet been found.

It seems plausible that the diffusion-Eyring model has slightly better agreement with the experimental results than

does the modified diffusion model for RT prediction. However, it is hard to conclude that in this specific case, the diffusion-Eyring model is more accurate at this moment, because of the uncertainty of the input parameter of the three diffusion equation models,[21] including absorption coefficients, directivities of the microphone and the loudspeaker, etc. Nevertheless, in extending this model to an uncovered room, the diffusion-Eyring model is expected to be problematic due to the singularity mentioned above. Such measurements should be done in the near future for further validations.

In the 250 and 500 Hz octave bands, the agreement between the predicted results and the experimental results in terms of SPLs and RTs drops markedly, due to the decreased diffuse reflections of the walls. It remains for future work to scrutinize the validity of each diffusion equation model with regard to different degrees of diffusion on the walls.

V. CONCLUSION

This work introduces extensions of the diffusion equation model recently applied in room-acoustic predictions by proposing a modified boundary condition. A direct comparison of this modified boundary condition with two previous boundary conditions: diffusion-Sabine and diffusion-Eyring boundary conditions, suggests that these three boundary conditions behave similarly in low absorption region while a considerable discrepancy will be seen in high absorption region. Particularly, there exists a singularity within the diffusion-Eyring model when the absorption coefficient becomes 1.0. The diffusion equation model using the modified boundary condition overcomes the singularity problem. To further compare them, a geometrical-acoustic model is employed as a reference, with RTs and SPLs investigated in cubic rooms. For uniformly distributed absorption coefficients, both the diffusion-Eyring model and the modified diffusion model show good agreement with the geometrical-acoustic model. For nonuniformly distributed absorption coefficients, if the wall has one side open or an absorption coefficient of 1.0, the diffusion-Eyring model fails. Simulations of sound pressure levels (SPLs) in a long, flat room with an average absorption coefficient below 0.8, indicate that the diffusion-Eyring and the modified diffusion model yield very similar results being closer to those estimated by the geometrical-acoustics method.

The experimental results obtained from a scale-model flat room are discussed to further examine the modified diffusion model. Both SPL and RT distributions indicate that the diffusion equation model is capable of modeling the sound field in disproportionate rooms for the nonuniformity of sound energy density as well as the sound energy decay. Within frequency ranges where diffuse surface reflections can be achieved in the scale model, the modified diffusion model and the diffusion-Eyring model agree well with the scale-model experimental results.

The diffusion-Eyring and the modified diffusion model both exceed the diffusion-Sabine model in terms of room-acoustic prediction accuracy. The modified diffusion model provides comparable results with the diffusion-Eyring model in most cases investigated within this study. More importantly, the modified diffusion model can handle an absorption coefficient as high as 1.0, while the diffusion-Eyring model fails due to the singularity.

The discussion of the illustrative examples is limited so far to the cases where the numerical simulations and, particularly the experimental verifications, are conducted using a flat room scale model. More systematic investigations using other room types, including long room scale models, real-size rooms, and coupled spaces, are expected in the near future.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Bengt-Inge Dalenbäck, Dr. Edward W. Larsen, and Professor Paul Calamia for useful discussions. They also thank Aaron Catlin and Xiaohu Chen for their help during the acoustic measurement and Stephen Olson for a critical review of the final draft.

[1] F. Ollendorff, "Statistical room-acoustics as a problem of diffusion (a proposal)," Acustica **21**, 236–245 (1969).
[2] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw–Hill, New York, 1953).
[3] J. Picaut, L. Simon, and J. D. Polack, "A mathematical model of diffuse sound field based on a diffusion equation," Acustica **83**, 614–621 (1997).
[4] V. Valeau, M. Hodgson, and J. Picaut, "A diffusion-based analogy for the prediction of sound fields in fitted rooms," Acustica **93**, 94–105 (2007).
[5] 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).
[6] J. Picaut, L. Simon, and J. Hardy, "Sound field modeling in streets with a diffusion equation," J. Acoust. Soc. Am. **106**, 2638–2645 (1999).
[7] J. Picaut, L. Simon, and J. D. Polack, "Sound field in long rooms with diffusely reflecting boundaries," Appl. Acoust. **56**, 217–240 (1999).
[8] 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).
[9] Y. Jing and N. Xiang, "A modified diffusion equation for room-acoustic predication (L)," J. Acoust. Soc. Am. **121**, 3284–3287 (2007).
[10] 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).
[11] R. C. Haskell, L. O. Svaasand, T. Tsay, T. Feng, M. S. McAdams, and B. J. Tromberg, "Boundary conditions for the diffusion equation in radiative transfer," J. Opt. Soc. Am. A **11**, 2727–2741 (1994).
[12] W. B. Joyce, "Classical-particle description of photons and phonons," Phys. Rev. D **9**, 3234–3256 (1974).
[13] J. Picaut, A. Billon, V. Valeau, and A. Sakout, "Sound field modeling in architectural acoustics using a diffusion equation," *Proc. Inter-Noise* 2006.
[14] U. J. Kurze, "Scattering of sound in industrial spaces," J. Sound Vib. **98**, 349–364 (1985).
[15] A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications* (Acoustical Society of America, Melville, New York, 1981).
[16] H. Kuttruff, *Room Acoustics* 4th ed. (Spon, New York, 2000).
[17] M. R. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am. **37**, 409–412 (1965).
[18] W. Yang and M. Hodgson, "Ceiling baffles and reflectors for controlling lecture-room sound for speech intelligibility," J. Acoust. Soc. Am. **121**, 3517–3326 (2007).
[19] N. Xiang and J. Blauert, "Binaural scale modeling for auralisation and prediction of acoustics in auditoria," Appl. Acoust. **38**, 267–290 (1993).
[20] K. M. Li and K. K. Lu, "Evaluation of sound in long enclosures," J. Acoust. Soc. Am. **116**, 2759–2770 (2007).
[21] J. E. Summers, R. R. Torres, Y. Shimizu, and B. L. Dalenback, "Adapting a randomized beam-axis-tracing algorithm to modeling of coupled rooms via late-part ray tracing," J. Acoust. Soc. Am. **118**, 1491–1502 (2005).
[22] J. Jeon and M. Barron, "Evaluation of stage acoustics in Seoul Arts Center Concert Hall by measuring stage support," J. Acoust. Soc. Am. **117** 232–239 (2005).
[23] M. Hodgson, "Evidence of diffuse surface reflections in rooms," J. Acoust. Soc. Am. **89**, 765–771 (1991).

Spatial correlation and coherence in reverberant acoustic fields: Extension to microphones with arbitrary first-order directivity[a)]

Martin Kuster[b)]
SARC, School of Electronics, Electrical Engineering & Computer Science, Queen's University Belfast, BT7 1NN Belfast, United Kingdom

(Received 26 July 2007; revised 19 October 2007; accepted 22 October 2007)

Spatial correlation and coherence functions in reverberant sound fields are relevant to the acoustics of enclosed spaces and related areas. Theoretical expressions for the spatial correlation and coherence functions between signals representing the pressure and/or the components of the particle velocity vector in a reverberant sound field are established in the literature and most of these have also been corroborated with measurements [F. Jacobsen, J. Acoust. Soc. Am. **108**, 204–210 (2000)]. In the present paper, these expressions are generalized to microphones of first-order directivity, whereby the directivity can be expressed in terms of pressure and pressure gradient. It is shown that the resulting spatial correlation and coherence functions can be expressed in terms of the established spatial correlation and coherence functions. The derived theoretical expression for the spatial coherence function is validated with a modeled diffuse sound field. Further, it is compared with the experimental coherence obtained from the reverberant tails of room impulse responses measured with two common surround sound microphone setups in a concert and a lecture hall.
 [DOI: 10.1121/1.2812592]

PACS number(s): 43.55.Cs, 43.55.Br [NX] Pages: 154–162

I. INTRODUCTION

The model of a diffuse sound field is regularly used in the analysis of the reverberant sound field in enclosed spaces. The most prominent example is the reverberation chamber but the model is also applied to the acoustics of performances spaces. The main advantage of the model is that, due to the assumed homogeneity and isotropy, analytical expressions for several sound field quantities can be derived.

Two examples of such field quantities are the spatial correlation and spatial coherence functions between two measurement positions. In 1954, Cook *et al.*[1] presented the spatial correlation function between the sound pressure measured at two points in a diffuse sound field. The spatial correlation functions between pressure and/or components of the particle velocity have been derived more recently.[2] Jacobsen and Roisin have introduced the spatial coherence functions as more effective measures and verified the resulting theoretical expressions experimentally in a reverberation chamber.[3] For broadband noise excitation, Rafaely[4] has derived theoretical expressions for the spatial correlation functions that have subsequently been verified experimentally by Ingyu *et al.*[5] Also, the diffuse field interaural correlation coefficient as a function of frequency was measured by Lindevald and Benade and found to be well described by a modified $\sin(x)/x$ function.[6]

The spatial correlation or coherence between microphone signals is also relevant in the context of multichannel audio recordings. Assuming pressure-sensitive microphones, a number of authors have considered the correlation between microphones in a surround sound recording setup.[7,8] But in multichannel audio recordings, virtual acoustic simulations and not least generic acoustic applications the employed microphones are often neither purely sensitive to pressure nor to pressure gradient,[9,10] and the spatial correlation and coherence functions for these types of microphones are not yet established.

The current paper thus presents a derivation of the diffuse field spatial correlation and coherence functions for microphone pairs whose directivities can be expressed as a combination of pressure and particle velocity components. The analytical results are validated with a simulated diffuse field. As an application example, the results are further illustrated by and compared with the coherence estimates obtained from room impulse responses measured with a spaced and a coincident surround sound microphone setup.

II. DERIVATION OF SPATIAL CORRELATION AND COHERENCE FUNCTION

The spatial correlation coefficient function between signals x and y measured a distance r apart at time lag τ is defined as[11]

$$\rho_{xy}(r,\tau) = \frac{R_{xy}(r,\tau)}{\sqrt{R_{xx}(0,0)R_{yy}(0,0)}}, \tag{1}$$

where R_{xx} and R_{yy} are autocorrelation functions and R_{xy} is the cross-correlation function.

[a)]Portions of this work were presented in "Spatial coherence between microphones with arbitrary first-order directivity in reverberant acoustic fields," Proceedings of the 19th International Congress on Acoustics, Madrid, September 2007.

[b)]Electronic mail: kuster_martin@hotmail.com

Similarly, the spatial coherence function between signals x and y measured a distance r apart at frequency ω is defined here as[11]

$$\gamma_{xy}^2(r,\omega) = \frac{|S_{xy}(r,\omega)|^2}{S_{xx}(0,\omega)S_{yy}(0,\omega)}, \tag{2}$$

where S_{xx} and S_{yy} are autospectral densities and S_{xy} is the cross-spectral density. To be precise, the coherence function in this form should be referred to as the magnitude-squared coherence function.

The cross-correlation function $R_{xy}(r,\tau)$ and the cross-spectral density function $S_{xy}(r,\omega)$ are related by the Fourier transform

$$S_{xy}(r,\omega) = \int_{-\infty}^{\infty} R_{xy}(r,\tau)e^{-j\omega\tau}d\tau. \tag{3}$$

A. Definition of microphone signals

The spatial correlation and coherence functions will be derived for the microphones M_χ and M_ψ. For notational convenience, the derivation in this paper is limited to the two-dimensional (2D) case where the vector joining the two microphone locations and the vectors pointing in the directions of the microphone main lobes all lie in the horizontal plane. An analogous derivation can be performed for the more general three-dimensional (3D) case.

The two microphone signals at positions $\mathbf{r}_\chi$ and $\mathbf{r}_\psi$ and time t can be written as

$$M_\chi(\mathbf{r}_\chi,t) = b_{w,\chi}\mathcal{W}(\mathbf{r}_\chi,t) + b_{xy,\chi}\cos(\theta_{M_\chi})\mathcal{X}(\mathbf{r}_\chi,t) + b_{xy,\chi}\sin(\theta_{M_\chi})\mathcal{Y}(\mathbf{r}_\chi,t), \tag{4a}$$

$$M_\psi(\mathbf{r}_\psi,t) = b_{w,\psi}\mathcal{W}(\mathbf{r}_\psi,t) + b_{xy,\psi}\cos(\theta_{M_\psi})\mathcal{X}(\mathbf{r}_\psi,t) + b_{xy,\psi}\sin(\theta_{M_\psi})\mathcal{Y}(\mathbf{r}_\psi,t), \tag{4b}$$

where the b coefficients define the directivity of the microphone and θ_{M_χ} and θ_{M_ψ} are the directions of the microphone main lobes. The signals $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ are related to the pressure p and Cartesian components v_x and v_y of the particle velocity as follows:

$$\mathcal{W}(\mathbf{r},t) = Gp(\mathbf{r},t), \tag{5a}$$

$$\mathcal{X}(\mathbf{r},t) = G\rho c v_x(\mathbf{r},t), \tag{5b}$$

$$\mathcal{Y}(\mathbf{r},t) = G\rho c v_y(\mathbf{r},t), \tag{5c}$$

with G an arbitrary scaling parameter and ρc the specific acoustic impedance of air. Therefore, the directivities of the signals $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ are given by

$$\mathcal{W}(\phi,\theta) = 1, \quad \mathcal{X}(\phi,\theta) = \sin\phi\cos\theta,$$

$$\mathcal{Y}(\phi,\theta) = \sin\phi\sin\theta. \tag{6}$$

Since the derivation is performed in the horizontal plane, $\phi = \pi/2$. As an example, a cardioid microphone pointing in direction θ_{M_χ} is obtained by setting $b_{w,\chi}=b_{xy,\chi}=1/2$. Further note that $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ (together with $\mathcal{Z}$) are the B-format signals as used in Ambisonics[12] and as recorded by, e.g., the SoundField microphone system.[13]

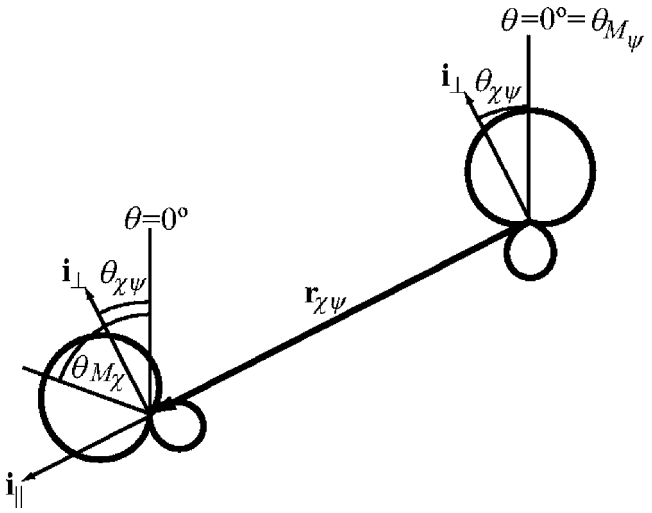


FIG. 1. Microphone setup with the definition of the vector $\mathbf{r}_{\chi\psi}$ and the angles $\theta_{M\chi}$, $\theta_{M\psi}$, and $\theta_{\chi\psi}$.

B. Decomposition into pressure and the two velocity components

In order to derive the spatial correlation or coherence function between M_χ and M_ψ, it is necessary to decompose the microphone signals into terms proportional to the pressure and particle velocity components parallel and perpendicular to the vector $\mathbf{r}_{\chi\psi}$ joining the two microphone positions as shown in Fig. 1. From Fig. 1, it can be seen that the main lobe of microphone M_χ in the coordinate system $(\mathbf{i}_\perp,\mathbf{i}_\parallel)$ defined by the orientation of the vector $\mathbf{r}_{\chi\psi}$ is rotated not by θ_{M_χ} but by $\theta_{M_\chi}-\theta_{\chi\psi}$. Note that the directions $(\mathbf{i}_\perp,\mathbf{i}_\parallel)$ and the value of $\theta_{\chi\psi}$ vary with the relative position and direction of the two microphones.

From Fig. 2, the following relationships with the b coefficients of Eq. (4a) hold:

$$b_{p,\chi} = b_{w,\chi},$$

$$b_{v_\perp,\chi} = b_{xy,\chi}\cos(\theta_{M_\chi}-\theta_{\chi\psi}),$$

$$b_{v_\parallel,\chi} = b_{xy,\chi}\sin(\theta_{M_\chi}-\theta_{\chi\psi}),$$

and equivalent expressions hold for the b coefficients of Eq. (4b). Using these relationships, the signals in Eq. (4) can be rewritten in the coordinate system $(\mathbf{i}_\perp,\mathbf{i}_\parallel)$ as

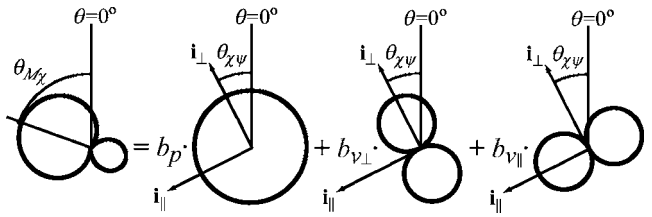


FIG. 2. Decomposition of the microphone direction and directivity into the three orthogonal components with the coefficients b_p, $b_{v_\perp}$, and $b_{v_\parallel}$ in the coordinate system $(\mathbf{i}_\perp,\mathbf{i}_\parallel)$.

$$M_\chi(0,t) = b_{p,\chi}\mathcal{W}_{\theta_{\chi\psi}}(0,t) + b_{v_\perp,\chi}\mathcal{X}_{\theta_{\chi\psi}}(0,t) + b_{v_\parallel,\chi}\mathcal{Y}_{\theta_{\chi\psi}}(0,t), \tag{7a}$$

$$M_\psi(r,t) = b_{p,\psi}\mathcal{W}_{\theta_{\chi\psi}}(r,t) + b_{v_\perp,\psi}\mathcal{X}_{\theta_{\chi\psi}}(r,t) + b_{v_\parallel,\psi}\mathcal{Y}_{\theta_{\chi\psi}}(r,t). \tag{7b}$$

In Eqs. (7a) and (7b), the subscript $\theta_{\chi\psi}$ of $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ denotes that their basis is rotated by $\theta_{\chi\psi}$ and therefore $\mathcal{X}_{\theta_{\chi\psi}}$ and $\mathcal{Y}_{\theta_{\chi\psi}}$ are aligned with $\mathbf{i}_\perp$ and $\mathbf{i}_\parallel$, respectively. Also, since only the distance between the two microphones is relevant in the further derivation, $\mathbf{r}_\chi$ and $\mathbf{r}_\psi$ have been replaced for notational convenience by the distance r, which is the length of the vector $\mathbf{r}_{\chi\psi}$.

Using the relationship between signals $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ and the pressure and Cartesian components of the particle velocity in Eq. (5), the microphone signals, written in terms of pressure and particle velocity components parallel and perpendicular to the vector $\mathbf{r}_{\chi\psi}$, are then finally given by

$$M_\chi(0,t) = b_{p,\chi}p(0,t) + \rho c b_{v_\perp,\chi}v_\perp(0,t) + \rho c b_{v_\parallel,\chi}v_\parallel(0,t), \tag{8a}$$

$$M_\psi(r,t) = b_{p,\psi}p(r,t) + \rho c b_{v_\perp,\psi}v_\perp(r,t) + \rho c b_{v_\parallel,\psi}v_\parallel(r,t). \tag{8b}$$

C. Derivation of spatial correlation function

It follows from Eq. (1) that the spatial correlation coefficient function $\rho_{M_\chi M_\psi}(r,\tau)$ can be calculated from $R_{M_\chi M_\psi}(r,\tau)$, $R_{M_\chi M_\chi}(0,0)$, and $R_{M_\psi M_\psi}(0,0)$.

The cross-correlation function $R_{M_\chi M_\psi}(r,\tau)$ can be calculated through the expected value operator. From the linearity of the expected value operator, it follows that $R_{M_\chi M_\psi}(r,\tau)$ can be written as

$$\begin{aligned} R_{M_\chi M_\psi}(r,\tau) = {} & b_{p,\chi}b_{p,\psi}R_{pp}(r,\tau) + \rho c b_{p,\chi}b_{v_\perp,\psi}R_{pv_\perp}(r,\tau) \\ & + \rho c b_{p,\chi}b_{v_\parallel,\psi}R_{pv_\parallel}(r,\tau) \\ & + \rho c b_{v_\perp,\chi}b_{p,\psi}R_{v_\perp p}(r,\tau) \\ & + (\rho c)^2 b_{v_\perp,\chi}b_{v_\perp,\psi}R_{v_\perp v_\perp}(r,\tau) \\ & + (\rho c)^2 b_{v_\perp,\chi}b_{v_\parallel,\psi}R_{v_\perp v\parallel}(r,\tau) \\ & + \rho c b_{v_\parallel,\chi}b_{p,\psi}R_{v_\parallel p}(r,\tau) \\ & + (\rho c)^2 b_{v_\parallel,\chi}b_{v_\perp,\psi}R_{v_\parallel v_\perp}(r,\tau) \\ & + (\rho c)^2 b_{v_\parallel,\chi}b_{v_\parallel,\psi}R_{v_\parallel v_\parallel}(r,\tau). \end{aligned} \tag{9}$$

The terms involving $R_{v_\perp p}(r,\tau)$, $R_{v_\perp v\parallel}(r,\tau)$, $R_{v_\parallel v_\perp}(r,\tau)$, and $R_{pv_\perp}(r,\tau)$ are zero, see Ref. 2. The remaining correlation functions are defined in Eq. (A2). Further, it can be shown that $R_{pv_\parallel}(r,\tau) = R_{v_\parallel p}(r,\tau)$ and therefore

$$\begin{aligned} R_{M_\chi M_\psi}(r,\tau) = {} & \rho c(b_{p,\chi}b_{v_\parallel,\psi} + b_{v_\parallel,\chi}b_{p,\psi})R_{pv_\parallel}(r,\tau) \\ & + (\rho c)^2 b_{v_\perp,\chi}b_{v_\perp,\psi}R_{v_\perp v_\perp}(r,\tau) \\ & + (\rho c)^2 b_{v_\parallel,\chi}b_{v_\parallel,\psi}R_{v_\parallel v_\parallel}(r,\tau) \\ & + b_{p,\chi}b_{p,\psi}R_{pp}(r,\tau). \end{aligned} \tag{10}$$

For the calculation of the correlation coefficient function $\rho_{M_\chi M_\psi}(r,\tau)$, Eq. (10) has to be normalized by the respective values for the autocorrelation functions at $r=0$ and $\tau=0$. Using again results from Eq. (A2) yields the following expression for $R_{M_\chi M_\chi}(0,0)$:

$$\begin{aligned} R_{M_\chi M_\chi}(0,0) = {} & b_{p,\chi}^2 R_{pp}(0,0) + (\rho c)^2 b_{v_\perp,\chi}^2 R_{v_\perp v_\perp}(0,0) \\ & + (\rho c)^2 b_{v_\parallel,\chi}^2 R_{v_\parallel v_\parallel}(0,0) \\ = {} & G^2\left(\frac{b_{p,\chi}^2}{2} + \frac{b_{v_\perp,\chi}^2}{6} + \frac{b_{v_\parallel,\chi}^2}{6}\right), \end{aligned} \tag{11}$$

and an analogous expression for $R_{M_\psi M_\psi}(0,0)$. For notational convenience, the following two variables are now introduced:

$$D_\chi = b_{p,\chi}^2 + \frac{b_{v_\perp,\chi}^2}{3} + \frac{b_{v_\parallel,\chi}^2}{3}, \tag{12a}$$

$$D_\psi = b_{p,\psi}^2 + \frac{b_{v_\perp,\psi}^2}{3} + \frac{b_{v_\parallel,\psi}^2}{3}. \tag{12b}$$

In terms of the spatial correlation coefficient functions between pressure and/or components of the particle velocity listed in Eq. (A1), the spatial correlation coefficient function $\rho_{M_\chi M_\psi}(r,\tau)$ between microphones M_χ and M_ψ in a diffuse field is then finally given by

$$\begin{aligned} \rho_{M_\chi M_\psi}(r,\tau) = {} & \frac{1}{\sqrt{D_\chi D_\psi}}[b_{p,\chi}b_{p,\psi}\rho_{pp}(r,\tau) \\ & + (b_{p,\chi}b_{v_\parallel,\psi} + b_{v_\parallel,\chi}b_{p,\psi})\rho_{pv_\parallel}(r,\tau)/\sqrt{3} \\ & + b_{v_\perp,\chi}b_{v_\perp,\psi}\rho_{v_\perp v_\perp}(r,\tau)/3 \\ & + b_{v_\parallel,\chi}b_{v_\parallel,\psi}\rho_{v_\parallel v_\parallel}(r,\tau)/3]. \end{aligned} \tag{13}$$

By inserting the respective values for the b coefficients, the expressions in Eq. (A1) can be recovered from Eq. (13).

D. Derivation of spatial coherence function

Jacobsen derived the spatial coherence function $\gamma_{xy}^2(\omega)$ from a relationship with the spatial correlation coefficient function $\rho_{xy}(\tau)$.[3] In the current paper, the coherence function is derived directly from the spectral densities, which is slightly more insightful but ultimately leads to the same result. From Eq. (2), the coherence function $\gamma_{M_\chi M_\psi}^2(r,\omega)$ is defined by the spectral densities $S_{M_\chi M_\psi}(r,\omega)$, $S_{M_\chi M_\chi}(0,\omega)$, and $S_{M_\psi M_\psi}(0,\omega)$, and using Eq. (3), these can be found from the correlation functions $R_{M_\chi M_\psi}(r,\tau)$, $R_{M_\chi M_\chi}(0,\tau)$, and $R_{M_\psi M_\psi}(0,\tau)$.

From Eq. (A2), the correlation function $R_{pv_\parallel}(r,\tau)$ contains a $\sin(\omega\tau)$ time-dependent factor while $R_{pp}(r,\tau)$, $R_{v_\parallel v_\parallel}(r,\tau)$, and $R_{v_\perp v_\perp}(r,\tau)$ contain $\cos(\omega\tau)$ time-dependent factors. The Fourier transform $\mathcal{F}$ of these factors is given by[14]

$$\mathcal{F}\{\sin(\omega_0\tau)\} = j\pi[\delta(\omega+\omega_0) - \delta(\omega-\omega_0)], \tag{14a}$$

$$\mathcal{F}\{\cos(\omega_0\tau)\} = \pi[\delta(\omega+\omega_0) + \delta(\omega-\omega_0)]. \tag{14b}$$

Invoking Eq. (3) and interpreting the delta functions as defining the spectral densities at the frequency ω_0, the cross-spectral density $S_{M_\chi M_\psi}(r,\omega)$ can be written for positive frequencies as

$$S_{M_\chi M_\psi}(r,\omega) = \frac{\pi G^2}{2}\left[b_{p,\chi}b_{p,\psi}\frac{\sin(kr)}{kr} + b_{v_\perp,\chi}b_{v_\perp,\psi}\frac{\sin(kr)-kr\cos(kr)}{(kr)^3} + b_{v_\parallel,\chi}b_{v_\parallel,\psi}\frac{(kr)^2\sin(kr)+2kr\cos(kr)-2\sin(kr)}{(kr)^3}\right]$$
$$- j\frac{\pi G^2}{2}(b_{p,\chi}b_{v_\parallel,\psi} + b_{v_\parallel,\chi}b_{p,\psi})\frac{\sin(kr)-kr\cos(kr)}{(kr)^2}. \tag{15}$$

Note the imaginary term due to $R_{pv_\parallel}(r,\tau)$. From the expressions for $R_{pp}(0,0)$, $R_{v_\parallel v_\parallel}(0,0)$, $R_{v_\perp v_\perp}(0,0)$ in Eq. (A2) and the comment below it, the two autospectral densities are simply given by

$$S_{M_\chi M_\chi}(0,\omega) = \frac{\pi G^2}{2}D_\chi, \tag{16a}$$

$$S_{M_\psi M_\psi}(0,\omega) = \frac{\pi G^2}{2}D_\psi, \tag{16b}$$

with D_χ and D_ψ defined in Eq. (12).

In terms of the spatial coherence functions between pressure and/or components of the particle velocity listed in Eq. (A3), the spatial coherence function $\gamma^2_{M_\chi M_\psi}(r,\omega)$ between microphones M_χ and M_ψ in a diffuse field then follows as

$$\gamma^2_{M_\chi M_\psi}(r,\omega) = \frac{1}{D_\chi D_\psi}[b_{p,\chi}b_{p,\psi}\gamma_{pp}(r,\omega) + b_{v_\perp,\chi}b_{v_\perp,\psi}\gamma_{v_\perp v_\perp}(r,\omega)/3 + b_{v_\parallel,\chi}b_{v_\parallel,\psi}\gamma_{v_\parallel v_\parallel}(r,\omega)/3]^2$$
$$+ \frac{1}{D_\chi D_\psi}[(b_{p,\chi}b_{v_\parallel,\psi} + b_{v_\parallel,\chi}b_{p,\psi})\gamma_{pv_\parallel}(r,\omega)/\sqrt{3}]^2. \tag{17}$$

By inserting the respective values for the b coefficients, the expressions in Eq. (A3) can be recovered from Eq. (17).

III. VALIDATION WITH MODELED DIFFUSE FIELD

The derived analytical expression for the coherence is now validated with the result from a modeled diffuse field.

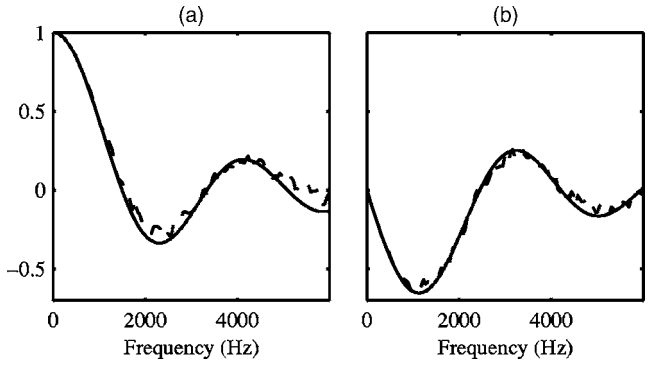


FIG. 3. (a) Real and (b) imaginary part of $S_{M_\chi M_\psi}(r,\omega)/\sqrt{S_{M_\chi M_\chi}(0,\omega)S_{M_\psi M_\psi}(0,\omega)}$ from theory (—) and modeled diffuse field (---).

The model is the same as used by Jacobsen[2] and Rafaely[4] and in the current case generates $N=1145$ plane waves with a uniform directional distribution. Each plane wave carries broadband random noise $s_{i,j}(t)$ drawn from the same statistical distribution. The pressure at the two simulated microphone positions is then given by

$$p(0,t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{I}\sum_{j=1}^{J} s_{i,j}(t), \tag{18a}$$

$$p(r,t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{I}\sum_{j=1}^{J} s_{i,j}\left(t - \frac{r}{c}\cos\phi_i\right), \tag{18b}$$

where $I=\sqrt{\pi N/4}$ and $J=2I\sin\phi_i$ with rounding to the nearest integer when required. The direction of incidence of plane wave (i,j) in spherical coordinates is given by

$$\phi_i = \frac{i\pi}{I}, \quad \theta_j = \frac{2j\pi}{J}. \tag{19}$$

For the simulation of the particle velocity components $v_\parallel(t)$ and $v_\perp(t)$, the signal $s_{i,j}(t)$ has to be weighted inside the

double sum of Eq. (18) by $\cos\phi_i$ and $\sin\phi_i \cos\theta_j$, respectively.

Instead of calculating the magnitude-squared coherence, the complex-valued quantity

$$S_{M_\chi M_\psi}(r,\omega)/\sqrt{S_{M_\chi M_\chi}(0,\omega)S_{M_\psi M_\psi}(0,\omega)} \tag{20}$$

is calculated from the simulated signals. This quantity is also referred to as the complex coherence[15] and it offers the advantage that the correctness of the real and imaginary part of Eq. (15) can be verified independently.

Figure 3 shows a comparison of this quantity between theory and model for $r=0.1$ m and using the following values for the b coefficients

$$b_{p,\chi}=b_{p,\psi}=b_{v_\parallel,\chi}=b_{v_\parallel,\psi}=1, \quad b_{v_\perp,\chi}=b_{v_\perp,\psi}=0. \tag{21}$$

From Fig. 3, the agreement between theory and model is very good. Note that averaging was employed in the estimation of the spectral densities from the model. Due to the noise excitation, some deviation is to be expected at higher frequencies and small magnitude values.

IV. EXPERIMENTAL RESULTS WITH ROOM IMPULSE RESPONSES

As an illustrative application of the derived coherence functions, the coherence between the microphones used for the recording and reproduction of surround sound through a standard ITU-R BS.775-1[16] five-channel loudspeaker setup is considered. At the time of writing, a number of both coincident and spaced recording microphone arrays are in use and one particular configuration of each was investigated.

A. Experimental procedures

In two vastly different rooms, room impulse responses have been measured with a SoundField MKV microphone from which the required microphone directivity is synthesized. The first room is a concert hall with a volume of 24 000 m^3 and a reverberation time of approximately 2.2 s. One hundred forty receiver position pairs have been measured at distances from the source ranging between 14.0 and 14.3 m. The room impulse responses were sampled at 16 kHz and had an effective noise-free length of 2.2 s. The second room is a lecture hall with a volume of 180 m^3 and a reverberation time of approximately 0.9 s. One hundred forty receiver position pairs have been measured at distances from the source ranging between 4.1 and 5.4 m. The room impulse responses were sampled at 14 980 Hz and had an effective noise-free length of 0.8 s.

The spectral densities required for the calculation of the coherence have been estimated from the room impulse responses using Welch's periodogram method.[17] With this method, the signal, in this case the room impulse response, is split into time blocks and the product of the discrete Fourier transforms (DFTs) from all blocks is averaged to give an unbiased estimate of the spectral density. This procedure implicitly results in spectral averaging and in both the concert and lecture hall effectively occurred over 90 DFT frequency bins. For the concert hall, Fig. 4 shows the coherence $\gamma^2_{pp}(\omega)$ between two pressure signals at a 0.05 m distance. In Fig. 4(a), the coherence has been estimated with the periodogram method for a single receiver pair. Whilst the agreement with the theoretical prediction is fair, it can be improved by additional spatial averaging as shown in Fig. 4(b). The frequency resolution in both graphs is 45 Hz.

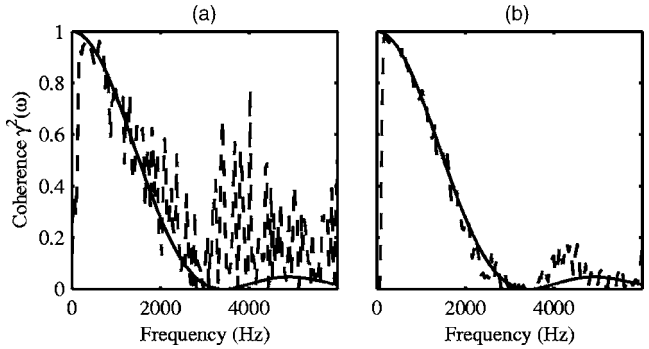


FIG. 4. Theoretical coherence $\gamma^2_{pp}(\omega)$ (—) and estimated coherence (---) from measurements between two pressure signals at a 0.05 m distance in a concert hall using (a) one receiver pair and periodogram averaging and (b) both periodogram and spatial averaging over 140 receiver pairs.

Because neither the direct sound nor strong, discrete early reflections are consistent with diffuse field theory, only the part of the room impulse response where the amplitudes are exponentially decaying has been used for the estimation of the spectral densities. For this purpose, the time varying statistics of the room impulse response are estimated in a sliding time window of 20 ms width as illustrated in Fig. 5. From Fig. 5, the statistics do not change after 110 ms and fluctuate around the known values for a Gaussian distribution. This time limit has been estimated for each room impulse response pair before calculating the spectral densities.

Note that the averaging when estimating the spectral densities is also required for another reason. The room reverberation due to a single source is a linear and time invariant system characterized by the room impulse response. Since the coherence is a measure of linearity and time-invariance,

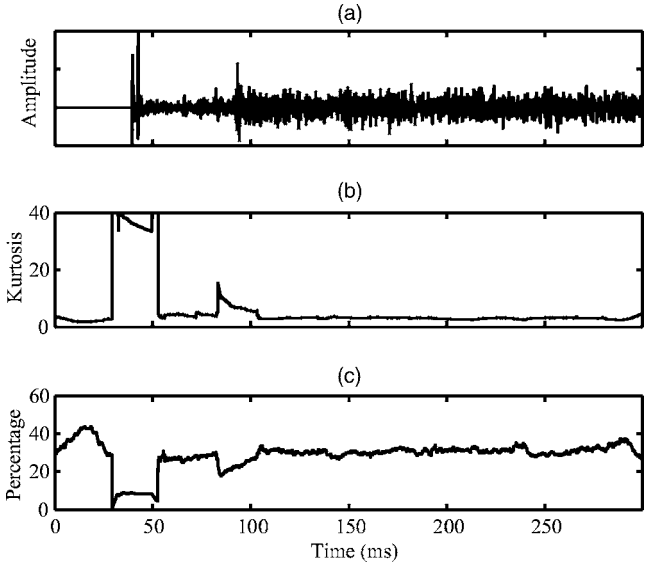


FIG. 5. (a) Example room impulse response in a concert hall, (b) kurtosis, and (c) percentage of samples outside the mean ± 1 s.d. estimated in a sliding time window of 20 ms.

TABLE I. Position $\mathbf{r}_\chi$, direction θ_{M_χ}, and directivity $b_{w,\chi}, b_{xy,\chi}$ of the microphones for the spaced microphone setup.

χ	$\mathbf{r}_\chi$(m)	$b_{w,\chi}$	$b_{xy,\chi}$	θ_{M_χ} (°)
1	(0,0.44)	0.5	0.5	70
2	(0,−0.44)	0.5	0.5	290
3	(0.23,0)	0.5	0.5	0
4	(−0.23,0.28)	0.5	0.5	156
5	(−0.23,−0.28)	0.5	0.5	204

TABLE II. Position $\mathbf{r}_\chi$, direction θ_{M_χ}, and directivity $b_{w,\chi}, b_{xy,\chi}$ of the microphones for the coincident microphone setup.

χ	$\mathbf{r}_\chi$(m)	$b_{w,\chi}$	$b_{xy,\chi}$	θ_{M_χ} (°)
1	(0,0)	0.10	0.32	43
2	(0,0)	0.10	0.32	317
3	(0,0)	0.07	0.31	0
4	(0,0)	0.36	0.57	134
5	(0,0)	0.36	0.57	226

its value for such a system will always be unity. The theoretical expressions for the diffuse field spatial coherence functions can therefore only be observed in a reverberant sound field if either spectral or spatial averaging is employed. The reader is referred to Refs. 3 and 18 for a more extensive discussion.

B. Spaced surround sound microphone array

A surround sound microphone setup with intermicrophone distances on the order of decimeters is considered. A variety of such setups have been suggested by Williams.[9,10] The position, direction, and directivity of each microphone in the chosen setup are given in Table I. Note that all the microphones have cardioid directivity. For the concert hall, Fig. 6 shows the theoretical and estimated coherence $\gamma_{12}^2(r,\omega)$, $\gamma_{13}^2(r,\omega)$, $\gamma_{14}^2(r,\omega)$, and $\gamma_{45}^2(r,\omega)$, where the subscripts denote the microphone pairs between which the coherence is considered. Note that the frequency axis is logarithmic.

This particular microphone setup is aimed for a seamless linking between adjacent microphones with the minimum amount of acoustic overlap. Therefore, the coherence is essentially zero for frequencies above 500 Hz and only reaches

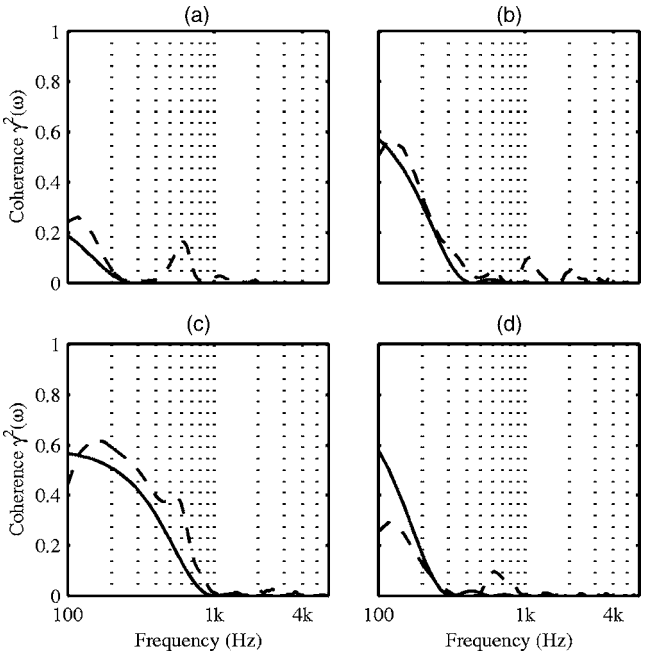


FIG. 6. Theoretical coherence (—) and estimated coherence (---) from measurements averaged over receiver 140 positions for the spaced microphone setup in a concert hall, (a) $\gamma_{12}^2(\omega)$, (b) $\gamma_{13}^2(\omega)$, (c) $\gamma_{14}^2(\omega)$, and (d) $\gamma_{45}^2(\omega)$.

significant values at lower frequencies. The largest values for the coherence occur between microphones 1 and 4. The agreement between theoretical prediction and measurement is good although some deviation can be observed at the lower frequency end.

C. Coincident surround sound microphone array

In the case of coincident microphones, $r=0$ and Eq. (17) can be solved analytically by applying l'Hôpital's rule repeatedly. This results in

$$\gamma_{M_\chi M_\psi}^2(0,\omega) = \frac{\left(b_{p,\chi}b_{p,\psi} + \frac{b_{v_\perp,\chi}b_{v_\perp,\psi}}{3} + \frac{b_{v_\parallel,\chi}b_{v_\parallel,\psi}}{3}\right)^2}{D_\chi D_\psi}. \tag{22}$$

Contrary to the general case of $r \neq 0$, this expression is independent of frequency (unless the microphone directivities change with frequency). An intuitive explanation might be that there is no characteristic distance to be compared with the varying acoustic wavelength.

For the coincident surround sound microphone setup, a least-squares solution for the decoding of the B-format signal into the five microphone signals corresponding to the loudspeaker feeds is considered. This method has been suggested by Daniel *et al.*[19] and Jot *et al.*[20] The resulting direction and directivities of the microphones are listed in Table II. Note that all microphones have hypercardioid directivity. Inserting the values in Table II into Eq. (22), the following theoretical values for the frequency-independent coherence result:

$$\gamma_{12}^2 = 0.09, \quad \gamma_{13}^2 = 0.61, \quad \gamma_{14}^2 = 0.11, \quad \gamma_{45}^2 = 0.28, \tag{23}$$

where the subscript denotes again the microphone pairs between which the coherence is considered. Note that the value of the coherence between microphones 1 and 3 is substantial and suggests that there is considerable acoustic overlap between the two microphones.

For the concert hall, Fig. 7 shows the theoretical and estimated coherence $\gamma_{12}^2(r,\omega)$, $\gamma_{13}^2(r,\omega)$, $\gamma_{14}^2(r,\omega)$, and $\gamma_{45}^2(r,\omega)$. The agreement between theory and measurement is not particularly good for $\gamma_{12}^2(\omega)$ but is fairly good for all other coherence functions. Also, it appears that none of the estimated coherence functions from measurements are exactly constant with frequency. Instead, the values decrease steadily and to a varying degree with frequency.

It seems plausible that the reason for this behavior is the nonideal measurement microphone. In this context, it is important to know that the SoundField microphone consists of

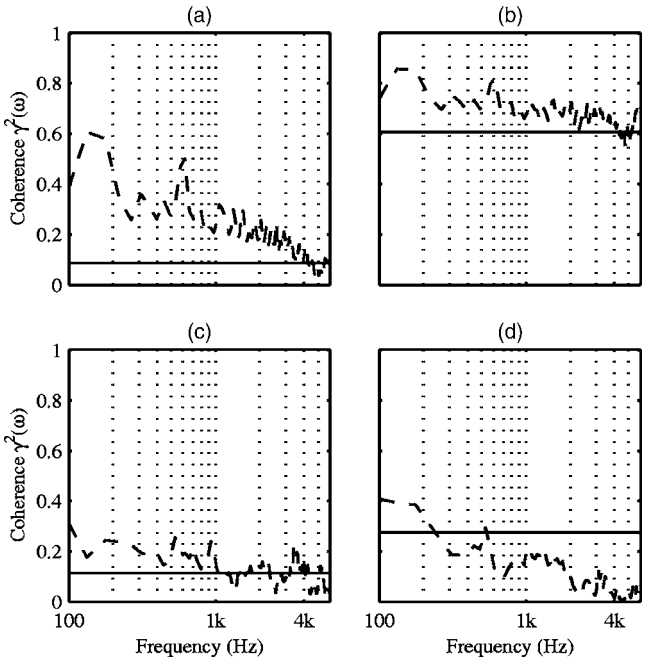


FIG. 7. Theoretical coherence (—) and estimated coherence (---) from measurements averaged over receiver 140 positions for the coincident microphone setup in a concert hall, (a) $\gamma^2_{12}(\omega)$, (b) $\gamma^2_{13}(\omega)$, (c) $\gamma^2_{14}(\omega)$, and (d) $\gamma^2_{45}(\omega)$.

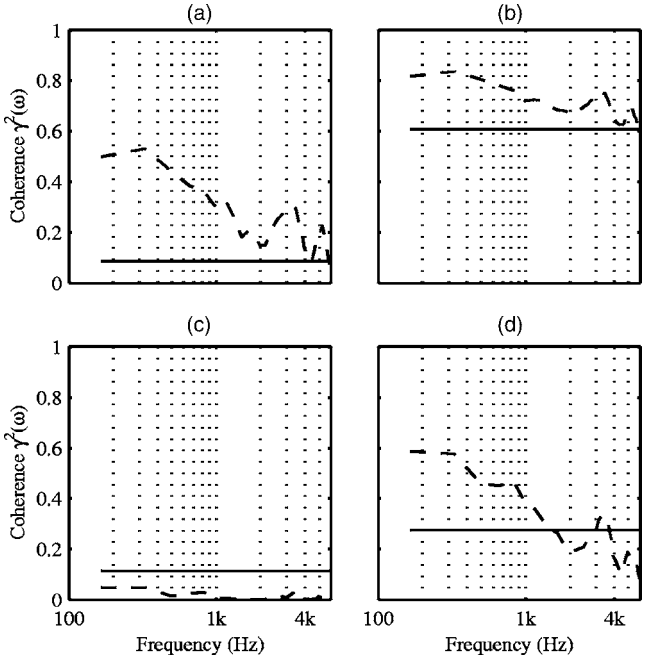


FIG. 9. Theoretical coherence (—) and estimated coherence (---) from measurements averaged over receiver 140 positions for the coincident microphone setup in a lecture hall, (a) $\gamma^2_{12}(\omega)$, (b) $\gamma^2_{13}(\omega)$, (c) $\gamma^2_{14}(\omega)$, and (d) $\gamma^2_{45}(\omega)$.

four sub-cardioid capsules mounted in a tetrahedral arrangement from which the B-format signals $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ are synthesized through electronic summation and subtraction. Experiments have shown that the functions $\gamma^2_{pv_\parallel}(r,\omega)$ and $\gamma^2_{pv_\perp}(r,\omega)$ estimated with the SoundField microphone correlate well with theoretical predictions for $r=0.1$ m but exhibit an anomaly for $r=0$ as shown in Fig. 8. The nonzero values for $\gamma^2_{pv_\parallel}(0,\omega)$ in Fig. 8(a) are an indication that the experimental overestimation of the coherence function values in Fig. 7 is due to the SoundField microphone.

For the lecture hall, Fig. 9 shows the analogous graphs to Fig. 7. A few differences between the results from the concert and lecture halls are present but the same global trend is evident in both rooms. It is also worth noting that similar results have been obtained in a variety of rooms.

V. CONCLUSION

Theoretical expressions for the diffuse field spatial correlation and coherence functions between microphone signals whose directivities can be expressed as a combination of pressure and particle velocity components have been derived. For notational simplicity, the derivation was limited to the 2D situation where the main lobes of the microphones and the vector joining the two microphone positions all lie in the same plane but the derivation can be extended to the more general 3D case. It was shown that the resulting spatial correlation and coherence functions can be expressed in terms of the known spatial correlation and coherence functions between pressure and/or components of the particle velocity.

The theoretical results have been verified with a diffuse field model and with coherence estimates obtained from room impulse response measurements in a concert hall and a lecture hall and a fair to very good agreement between theory and model/measurement was observed. For the measurements, the microphone directivities and distances investigated were taken from a spaced and coincident surround sound microphone setup. In the coincident case, the theoretical coherence is independent of frequency but, possibly due to nonideal microphones, the estimated coherence from measurements did decrease with frequency to a varying degree. Also, for certain microphone pair combinations, the value of the coherence was found to be substantial. With the spaced microphone setup, the values for the coherence from theory and measurements were only significant below 500 Hz and decrease rapidly to zero at higher frequency.

As a final remark, it is noted that it is conceivable to derive analytical expressions for the spatial correlation and coherence between more directional microphones.

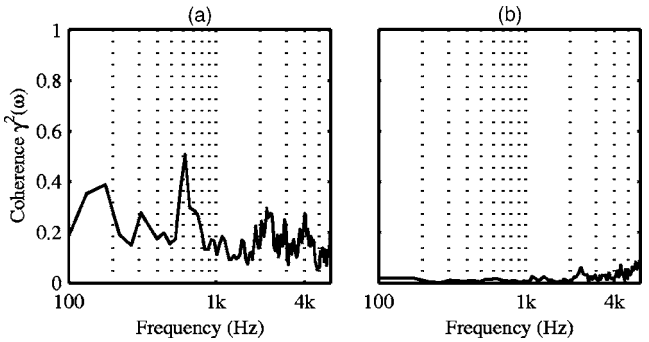


FIG. 8. Estimated coherence (a) $\gamma^2_{pv_\parallel}(0,\omega)$ and (b) $\gamma^2_{pv_\perp}(0,\omega)$ obtained from the same data and method as in Fig. 7. The value for both functions should theoretically be zero.

ACKNOWLEDGMENTS

Maarten van Walstijn and other colleagues at Queen's University Belfast are acknowledged for discussions and re-

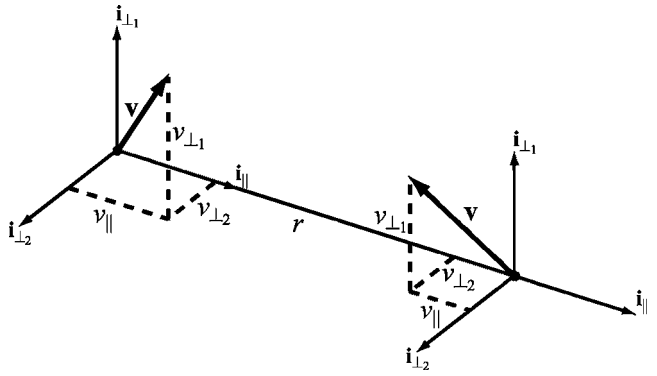


FIG. 10. Decomposition of the velocity vector $\mathbf{v}$ into the three orthogonal components $v_{\perp_1}$, $v_{\perp_2}$, and $v_{\parallel}$ in the coordinate system $(\mathbf{i}_{\perp_1},\mathbf{i}_{\perp_2},\mathbf{i}_{\parallel})$.

marks. The array room impulse responses in the concert and lecture hall have been kindly provided by Diemer de Vries.

APPENDIX: SPATIAL CORRELATION AND COHERENCE BETWEEN PRESSURE AND/OR PARTICLE VELOCITY COMPONENTS

In this appendix, the spatial correlation and coherence function established in the literature are listed. They are used in the main body of the paper for the derivation of the spatial correlation and coherence functions between the directional microphones.

The spatial correlation coefficient function between two signals representing acoustic pressure measured at a distance r and a time τ apart in a pure tone diffuse field of frequency ω_0 was first derived by Cook *et al.*[1] and is given by

$$\rho_{pp}(r,\tau) = \frac{\sin(kr)}{kr}\cos(\omega_0\tau), \tag{A1a}$$

where $k=\omega_0/c$ is the acoustic wave number. Similarly, from Ref. 3 the following expressions can be quoted for the spatial correlation coefficient functions between (i) the pressure and a component of the particle velocity in the direction of the line joining the two measurement points

$$\rho_{pv_{\parallel}}(r,\tau) = \sqrt{3}\frac{\sin(kr)-(kr)\cos(kr)}{(kr)^2}\sin(\omega_0\tau), \tag{A1b}$$

(ii) the two particle velocity components in the direction of the line joining the two measurement points

$$\rho_{v_{\parallel}v_{\parallel}}(r,\tau) = 3\frac{(kr)^2\sin(kr)+2kr\cos(kr)-2\sin(kr)}{(kr)^3} \times \cos(\omega_0\tau), \tag{A1c}$$

and (iii) the two particle velocity components perpendicular toe the line joining the two measurement point but parallel to each othor

$$\rho_{v_{\perp}v_{\perp}}(r,\tau) = 3\frac{\sin(kr)-(kr)\cos(kr)}{(kr)^3}\cos(\omega_0\tau). \tag{A1d}$$

The definition of the various velocity components is illustrated in Fig. 10. Evidently, these expressions have first been derived by Kuno and Ikegaya.[21]

Jacobsen derived the correlation coefficient functions in Eq. (A1) from the following correlation functions for a diffuse field model with plane waves of mean square magnitude G^2,

$$R_{pp}(r,\tau) = \frac{G^2}{2}\frac{\sin(kr)}{kr}\cos(\omega_0\tau), \tag{A2a}$$

$$R_{pv_{\parallel}}(r,\tau) = \frac{G^2}{2\rho_0 c}\frac{\sin(kr)-(kr)\cos(kr)}{(kr)^2}\sin(\omega_0\tau), \tag{A2b}$$

$$R_{v_{\parallel}v_{\parallel}}(r,\tau) = \frac{(kr)^2\sin(kr)+2kr\cos(kr)-2\sin(kr)}{(kr)^3} \times \frac{G^2}{2(\rho_0 c)^2}\cos(\omega_0\tau), \tag{A2c}$$

$$R_{v_{\perp}v_{\perp}}(r,\tau) = \frac{G^2}{2(\rho_0 c)^2}\frac{\sin(kr)-(kr)\cos(kr)}{(kr)^3}\cos(\omega_0\tau), \tag{A2d}$$

$$R_{pp}(0,0) = \frac{G^2}{2}, \tag{A2e}$$

$$R_{v_{\parallel}v_{\parallel}}(0,0) = R_{v_{\perp}v_{\perp}}(0,0) = \frac{G^2}{6(\rho_0 c)^2}. \tag{A2f}$$

It can be shown that $R_{\chi\psi}(0,\tau)=R_{\chi\psi}(0,0)\cos(\omega_0\tau)$ for a pure tone, where χ and ψ are placeholders for the variables on the last two lines of the above presented equations. It can also be shown that the cross-correlation and correlation coefficient functions between all other possible combinations of pressure and/or components of the particle velocity are zero for any distance r, see Ref. 2.

From Eq. (A1), Jacobsen and Roisin derived the following expressions for the spatial coherence functions between pressure p and/or particle velocity components parallel $v_{\parallel}$ and perpendicular $v_{\perp}$ to the vector joining the two measurement points,[3]

$$\gamma^2_{pp}(r,\omega) = \left[\frac{\sin(kr)}{kr}\right]^2, \tag{A3a}$$

$$\gamma^2_{pv_{\parallel}}(r,\omega) = 3\left[\frac{\sin(kr)-(kr)\cos(kr)}{(kr)^2}\right]^2, \tag{A3b}$$

$$\gamma^2_{v_{\perp}v_{\perp}}(r,\omega) = 9\left[\frac{\sin(kr)-(kr)\cos(kr)}{(kr)^3}\right]^2, \tag{A3c}$$

$$\gamma^2_{v_{\parallel}v_{\parallel}}(r,\omega) = 9\left[\frac{(kr)^2\sin(kr)+2kr\cos(kr)-2\sin(kr)}{(kr)^3}\right]^2. \tag{A3d}$$

Note that in the last expression, the two perpendicular velocity components must be parallel to each other. As with the spatial correlation functions, the spatial coherence functions for any other combination of pressure and/or components of the particle velocity are zero for any distance r.

[1]R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, Jr., "Measurement of correlation coefficients in reverberant sound fields," J. Acoust. Soc. Am. **27**, 1072–1077 (1955).
[2]F. Jacobsen, "The diffuse sound field," Technical Report No. 27, The Acoustics Laboratory, Technical University of Denmark, Lyngby, 1979.
[3]F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," J. Acoust. Soc. Am. **108**, 204–210 (2000).
[4]B. Rafaely, "Spatial-temporal correlation of a diffuse sound field," J. Acoust. Soc. Am. **107**, 3254–3258 (2000).
[5]I. Chun, B. Rafaely, and P. Joseph, "Experimental investigation of spatial correlation in broadband reverberant sound fields," J. Acoust. Soc. Am. **113**, 1995–1998 (2003).
[6]I. M. Lindevald and A. H. Benade, "Two-ear correlation in the statistical sound fields of rooms," J. Acoust. Soc. Am. **80**, 661–664 (1986).
[7]J. Usher, "Design criteria for high quality upmixers," in Proceedings of the 28th International Conference of the Audio Engineering Society, Piteå, Sweden, 2006.
[8]G. Martin, "The significance of interchannel correlation, phase and amplitude differences on multichannel microphone techniques," in Proceedings of the 113th Convention of the Audio Engineering Society, Los Angeles, 2002.
[9]M. Williams and G. L. Dû, "Microphone array analysis for multichannel sound recording," in Proceedings of the 107th Convention of the Audio Engineering Society, New York, 1999.
[10]F. Rumsey, *Spatial Audio*, 1st ed. (Focal, Oxford, 2001).
[11]J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Analysis*, 1st ed. (Wiley, New York, 1980).
[12]M. A. Gerzon and G. J. Barton, "Ambisonic decoders for HDTV," in Proceedings of the 92nd Convention of the Audio Engineering Society, Vienna, 1992.
[13]K. Farrar, "Soundfield microphone," Wireless World, October, 48–50 (1979).
[14]G. James, *Advanced Modern Engineering Mathematics* (Addison Wesley, Harlow, 1999), Chap. 5.
[15]J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Analysis*, 1st ed. (Wiley, New York, 1980), Chap. 7.
[16]I.-R. BS.775-1, "Multichannel stereophonic sound systems with and without accompanying picture," Technical Report, International Telecommunication Union Radiocommunication Assembly, Geneva, 1992–1994.
[17]A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, London, 1975), Chap. 11.
[18]F. Jacobsen and T. G. Nielsen, "Spatial correlation and coherence in a reverberant sound field," J. Sound Vib. **118**, 175–180 (1987).
[19]J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in Proceedings of the 105th Convention of the Audio Engineering Society, San Francisco, 1998.
[20]J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-D audio encoding and rendering techniques," in Proceedings of the 16th International Conference of the Audio Engineering Society, Rovaniemi, 1999.
[21]K. Kuno and K. Ikegaya, "A statistical consideration on models for sound fields composed of random plane and spherical wave elements," J. Acoust. Soc. Jpn. **30**, 65–75 (1974), (in Japanese).

Subjective and objective assessment of acoustical and overall environmental quality in secondary school classrooms

Arianna Astolfi[a)]
Politecnico di Torino, Department of Energetics, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

Franco Pellerey
Politecnico di Torino, Department of Mathematics, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy

(Received 6 October 2006; revised 26 October 2007; accepted 26 October 2007)

A subjective survey on perceived environmental quality has been carried out on 51 secondary-school classrooms, some of which have been acoustically renovated, and acoustical measurements were carried out in eight of the 51 classrooms, these eight being representative of the different types of classrooms that are the subject of the survey. A questionnaire, which included items on overall quality and its single aspects such as acoustical, thermal, indoor air and visual quality, has been administered to 1006 students. The students perceived that acoustical and visual quality had the most influence on their school performance and, with the same dissatisfaction for acoustical, thermal and indoor air quality, they attributed more relevance, in the overall quality judgment, to the acoustical condition. Acoustical quality was correlated to speech comprehension, which was correlated to the speech transmission index, even though the index does not reflect all the aspects by which speech comprehension can be influenced. Acoustical satisfaction was lower in nonrenovated classrooms, and one of the most important consequences of poor acoustics was a decrease in concentration. The stronger correlation between average noise disturbance scores and $L_{A\,max}$ levels, more than L_{Aeq} and L_{A90}, showed that students were more disturbed by intermittent than constant noise. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816563]

PACS number(s): 43.55.Hy, 43.71.Gv, 43.55.Gx, 43.50.Qp [NX] Pages: 163–173

I. INTRODUCTION

The environmental quality of a building is its suitability to provide health and comfort for occupants. It includes four main aspects: acoustical, thermal, indoor air and visual quality. The beneficiaries of good environmental conditions in classrooms are the teachers and learners and, as a first consequence, this will lead to an increase in school performance of the students and in productivity of teachers. This paper focuses on the subjective and objective evaluation of the acoustical quality in secondary-school classrooms, and on the subjective evaluation of the other environmental aspects and their influence on the overall quality. The main purposes are: (1) to assess acoustical quality by means of questionnaires and in-field measurements and to discuss the results of changes due to acoustical renovation; (2) to correlate subjective and measured data to identify the correspondence between the perception scales and the main acoustical factors; (3) to investigate the main factors that also affect the thermal, visual and indoor air quality and which environmental aspect is most correlated to overall environmental quality perception.

Only a few studies have dealt with how users perceive acoustical quality during typical classroom use. Speech intelligibility tests and measurements have been performed in classrooms of different grades.[1–3] Hétu *et al.*[4] carried out a study on the effect of noise and reverberation in primary and high-school classrooms, based on questionnaires and measurements. Dockrell and Shield[5] administered questionnaires to primary-school children in order to assess their ability to discriminate in different listening conditions and found relationships between the children's perceptions of awareness and annoyance and objective measures of noise. Hagen *et al.*[6] used questionnaires to evaluate whether adding sound-absorption and/or sound-field amplification systems in classrooms would improve the acoustic comfort for primary-school children, and investigated educational possibilities to improve the listening abilities during lessons. Kennedy *et al.*[7] administered questionnaires to university students to investigate the factors that influence the perceived listening quality. In their work, a measure of perceived classroom-listening quality during typical classroom use, called PLE (perception of listening ease), was identified by means of a response analysis, and correlations among PLE and items regarding classrooms environment, courses, teachers, and individual factors were analyzed.

II. OBJECTIVE ASSESSMENT OF THE ACOUSTICAL ENVIRONMENT

Bad acoustic conditions in classrooms decrease the quality of speech communication, reducing the school performance of students and causing the teachers to suffer from fatigue. According to the ISO 9921:2003 standard,[8] the quality of speech communication can be expressed in terms of speech intelligibility, which is quantified as the percentage of a message that is understood correctly. Speech intelligibility at a listener's position in a classroom depends on the speech-

[a)]Author to whom correspondence should be addressed. Electronic mail: arianna.astolfi@polito.it

TABLE I. Main characteristics of the eight classroom types.

	S1	M1	M2a/M2b	M3	M4	L1	L2	EL1
Location	courtyard	courtyard	street/square	street	street	street	street	street
Floor	first	ground	first	first	first	second	first	ground
No. of classrooms for each type	3	7	5/7	3	1	6	1	7
Sound absorption intervention	partial	full	absent	partial	full	full	partial	full
No. of students in the classroom during measurements (apart from teacher's vocal effort and background noise)	14	18	11/n.c.	16	16	13	15	18
Percentage of students present in the classroom during measurements (apart from teacher's vocal effort and background noise) compared to the full occupancy	70%	86%	65%/n.c.	75%	70%	55%	63%	72%
No. of administered questionnaires	36	120	61/89	59	19	126	17	149
Ceiling treatment	no	yes	no	no	yes	yes	no	yes
Acoustic reflector	no	yes	no	no	no	yes	no	yes
Vaulted ceiling	yes	no	yes	yes	yes	no	yes	yes
Floor area (m^2)	40.0	62.1	42.0	50.3	51.8	78.0	70.0	73.9
Mean height (m)	4.3	3.1	4.5	3.9	3.9	3.2	4.2	6.3
V (m^3)	160.0	189.4	190.0	201.2	207.2	250.4	296.0	465.8

signal-to-noise-ratio and the reverberation and can be predicted by the speech transmission index, STI,[9,10] which varies from 0 to 1. STI combines the two above-mentioned factors in a single quantity and is related to a five-point intelligibility scale:[8,10] "Bad" for STI values lower than 0.30, "Poor" between 0.30 and 0.45, "Fair" between 0.45 and 0.60, "Good" between 0.60 and 0.75, and "Excellent" for STI values higher than 0.75. In situations of a relaxed type of communication, such as during lectures, a "Good" level of intelligibility is recommended, considering a "Normal" vocal effort.[8] Vocal effort refers to the exertion of the speaker. It is quantified by the A-weighted speech level at a distance of 1 m in front of the speaker's mouth and subjectively as Very Loud, Loud, Raised, Normal and Relaxed. Free-field normal vocal efforts are given by Pavlovic[11] and Byrne *et al.*,[12] while typical vocal efforts in classrooms are reported by Houtgast,[1] Sato and Bradley[13] and Picard and Bradley.[14]

Speech intelligibility in a noisy environment with low reverberation, as in the case of a small occupied secondary-school classroom (e.g., 300 m^3), can also be approximately investigated with the reverberation time and A-weighted speech-signal-to-noise ratio, SNR_A.[2] According to Picard and Bradley,[14] the optimal values of the mid-frequency reverberation time and the minimum value of the SNR_A for 12 + years old students, in occupied classrooms, are estimated to be 0.5 s and 15 dB, respectively. As far as the noise level is concerned, an upper level of 33 dB(A) is indicated as the *ideal* condition, restricted to more vulnerable groups, which can rise to 40 dB(A) for an *acceptable* condition, to be used for more general purposes. Research on the effects of noise and poor acoustics in schools[15] has recently led many countries to write or revise a series of guidelines on classroom acoustics. For example, the S12.60 ANSI standard[16] and the UK Building Bulletin 93,[17] in unoccupied classrooms, require a maximum ambient noise level of 35 dB(A), $L_{Aeq,1\ h}$ and $L_{Aeq,30\ min}$, respectively, plus a maximum reverberation time, quoted in terms of the average in the 500 Hz, 1 kHz and 2 kHz octave bands, of 0.6 and 0.8 s, respectively.

III. CASE STUDY

The subject of the study is a 19th Century secondary school in a small town near Turin (Italy). It consists of two different buildings next to each other. The main building, a three story square-court building, contains 39 classrooms which face onto a quiet street or the internal courtyard. The classrooms differ in volume and shape, and were renovated or partially renovated with special acoustical design features. The second building is part of an old two-story building and contains 12 nonrenovated classrooms which face onto a quiet street or a large quiet square.

A subjective survey on perceived environmental quality was carried out on all 51 secondary-school classrooms, and acoustical measurements were carried out in eight of the 51 classrooms, these being representative of all the types of classrooms that were the subject of the survey. The main characteristics of the eight chosen types are shown in Table I. These can be divided into four groups in relation to volume: S1, M1, M2, M3, M4, L1, L2 and EL1 (where S stands for small, M for medium, L for large and EL for extra large). A full acoustical sound-absorption treatment was carried out in four of the eight classrooms (M1, M4, L1, EL1). It consisted of placing holed plaster-board panels filled with mineral wool on the ceiling, on the upper part of the lateral walls and on the back wall. An acoustic reflector was inserted into the flat absorbing ceilings in rooms M1 and L1 in order to increase the first reflections of speech sound to the rear part of the room. In the classroom with the highest ceiling (EL1) two large slightly convex rectangular panels were suspended at a height of 3 m from the floor, in order to reduce the useful volume. The three classrooms S1, M3 and L2 were only partially renovated. Sound absorption material was applied to the upper part of the back wall, and, in M3, also to the upper part of the lateral walls. M3 and M4 are identical and with the same sound absorbing treatment, with the exception of the vault, which in M3 was plastered and in M4 was completely covered with absorbing material. Only one

classroom type, which is in the second building, M2, was not renovated at all. It was divided into two groups, M2a, looking onto a street, and M2b, facing onto a square. The external walls of the buildings are thick and made of masonry and the windows are double glazed, apart from M2 which have a single glass. The sound insulation intervention mainly concerned the walls between adjacent classrooms, while the sound insulation from the corridors was not optimized. The floors were covered with ceramic tiles without a floating floor. The classrooms did not have any speech-reinforcement or ventilation systems.

IV. MEASUREMENTS

The following quantities were obtained from the in-field measurements in each classroom type: the teacher's vocal effort and noise level during regular lessons; the reverberation time in unoccupied and occupied conditions; the speech level, the SNR_A and the STI for six positions in the occupied classrooms. The classrooms chosen for the measurements were representative of the eight selected types, but not all the types were used for all the analyses. As M4 was very similar to M3, it was excluded for the measurements of all the quantities, with the exception of reverberation time. As far as M2a and M2b are concerned, the reverberation time was the same, and no significant difference in noise level was perceived. For these reasons only M2a was considered. Apart from the teacher's vocal effort and noise level, the measurements were all carried out when the building was empty, in order to have low noise from inside the building, and only the classroom under measurement was occupied. It should be pointed out that the classrooms were not fully occupied, as they are during lectures, for this set of measurements. As reported in Table I, the percentages of occupation ranged from between 55% and 86%, compared to the average occupancy during regular class time obtained from the subjective survey data.

A. Measured and calculated quantities

1. Teacher's vocal effort ($L_{spA1\ m}$)

From three to five teachers were asked to speak without pausing during a regular lesson in each classroom type, first speaking directly to the students as they do during a lesson, without dealing with any particular topic, and then reading a text from a book (the same text for all the teachers). Both female and male teachers were tested; they were asked to stand facing the student-seating area. Equivalent continuous speech levels of the teacher's voice, based on 20–60 s recordings, were measured for each type of speech at 1 m in front of the teacher's mouth, obtaining the octave band levels ($L_{\mathrm{sp1\ m}}$) and the overall A-weighted speech levels ($L_{\mathrm{spA1\ m}}$). A total of 26 teachers were tested (20 females and six males), but only five of them agreed to perform both types of experiments. The mean difference of $L_{\mathrm{spA1\ m}}$ values between the lectures and texts was 0.9 dB. Since lectures are more common during lessons only the lecture level was considered for these five teachers in the averaging with the speech levels of the other teachers, in order to obtain the average octave band and the average overall A-weighted speech levels for each classroom type. During the measurements it was checked that at 1 m teacher's voice level exceeded the noise level, in the same position, by more than 10 dB over the entire frequency range. The noise level, even when recorded immediately after the teacher's speech, was representative of the noise that occurred during the voice-level measurements, with quiet students, and there was no significant noise in the classrooms being tested.

2. Background-noise level (L)

This included noise from traffic and other external sources and noise due to student activity in the corridors or adjacent classrooms. It was based on a 3–6 min recording in the center of the room, in the occupied classrooms during regular lessons, immediately after the teachers had spoken. The students were asked to remain quiet and there was no significant noise in the classrooms being tested. The following quantities were obtained for each classroom type: the equivalent continuous octave band level, L_{eq}, the A-weighted equivalent continuous noise levels, L_{Aeq}, the A-weighted noise level that is exceeded by 90% of each sample period, L_{A90}, and the maximum A-weighted level, $L_{\mathrm{A\ max}}$, where maximum levels quantify intermittent sounds.

3. Reverberation time in occupied and unoccupied conditions (RT_o and RT_u)

Octave band reverberation time measurements were carried out in both occupied (RT_o) and unoccupied (RT_u) conditions by means of the interrupted noise method using an omni-directional sound power source, with a pink noise test signal. The results from two source-receiver combinations gave a spatial average value for each classroom type as a whole. The RT_o was also obtained from the impulse response measurements using a sine-sweep signal generated by the 4128 Brüel & Kjær head and torso simulator placed in the same way as for the speech signal measurements, as described in Sec. IV A 4. The octave band classroom RT_o values were then obtained by averaging the results from one source and seven microphone positions distributed over the seating area. At medium and high frequencies the results from the two measurement techniques were coincident for all but two classrooms, where the small differences were due to the slightly different numbers of students present in the classrooms during the two sets of measurements. The results from the sweep technique were then used for the analyses.

4. Spatial distribution of the average speech level (L_{spA})

A 4128 type Brüel & Kjær head and torso simulator was used as a speech source to obtain a spatial distribution of the speech signal in the occupied classrooms. The source, emitting a test signal shaped like a male spectrum,[10] was calibrated in an anechoic chamber, where an output level of 68 dB(A) was set at a distance of 1 m in front of the mouth. It was located at the teacher's position and oriented towards the student-seating area. The receiver positions were placed 1 m from the source's mouth, at mouth height, and at six other representative students' seats uniformly distributed

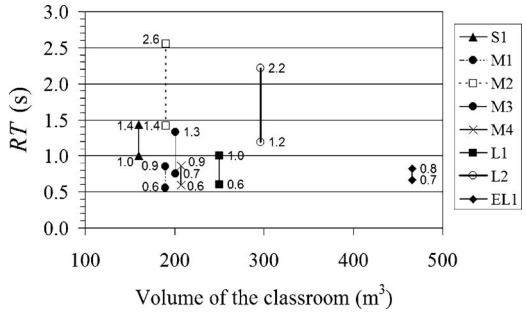


FIG. 1. Single number frequency averaging between 500 Hz, 1 kHz and 2 kHz, for reverberation times as a function of the room volume in unoccupied (upper value) and occupied (lower value) classrooms.

over the seating area, at seated ear height. It was checked that the source level at the measurement locations exceeded the noise level by at least 10 dB, over the entire frequency range, so as to minimize the influence of noise. In order to obtain the speech level distribution throughout the classroom, the source level reductions, with respect to the level measured at 1 m in front of the source's mouth, were determined in octave bands for each microphone position and the same reductions were applied to the 1 m average octave band speech levels for each classroom type. The overall A-weighted speech levels (L_{spA}) in the various positions in each classroom type were then obtained from the octave band values.

5. A-weighted speech-signal to noise ratio (SNR_A) and speech transmission index (STI)

The SNR_A were obtained as the L_{spA} minus the level of noise, at each of the six representative positions used to assess the spatial distribution of the speech signal. The noise was measured in occupied condition, with no student-activity noise, at the center of the room (L_{Aeq}).

The STI was obtained from the octave band filtered squared impulse response and the average speech-signal-to-noise ratio.[9,10] AURORA 4.1 was used for the analyses. The impulse response measurements were obtained from a sine sweep signal generated by the head and torso simulator placed in the same manner as for the speech level measurements. The STI values for the six student positions in each classroom were calculated for the occupied condition with the contribution of noise measured during lessons.

B. Results

1. Reverberation time

In Fig. 1 the average reverberation times at 500 Hz, 1 kHz and 2 kHz of the eight chosen classrooms are presented versus classroom volumes, for unoccupied and occupied conditions. A shorter RT_u in M1, M4, L1 and EL1, for which a full sound-absorption treatment was carried out, can be observed. Among these, only EL1 satisfies the UK regulations[17] requirements, but none satisfies the ANSI requirements.[16] In order to check the reverberation time in fully occupied conditions, corrected RT_o values were calculated applying the Sabine formula, in which the total acoustic absorptions, obtained from measured occupied reverberation time, were increased by an amount equal to the average absorption per student[18] multiplied by the difference in the numbers of students for full and partial occupancy. After the correction the average RT_o reduced from 0.55 to 0.53 s in M1, from 0.59 to 0.54 s in M4 and from 0.64 to 0.56 s in L1, thus approaching the 0.50 s limit required by Picard and Bradley.[14] In the other classrooms, most of them with poor or inexistent acoustical treatment, the corrected values were 0.68 s in M3, 0.67 s in EL1, 0.85 s in S1, 1.01 s in L2, and 1.13 s in M2. All the values are higher than 0.50 s, confirming that acoustical treatment is necessary also in small occupied classrooms.

2. Teachers' vocal effort and background noise level

The measurements were made for each classroom type, with the exception of M2b and M4 (because they were very similar to the M2a and M3 classrooms, respectively). Table II shows the teachers' vocal efforts measured for each teacher in the classroom types with the indication of the teacher's gender and the type of speech (text or lecture), the average values for each classroom type, and the corresponding free-field values based on the averages, $L_{spA1\ m,free\ field}$. The free-field values were calculated applying Barron and Lee's theory.[19]

The average value of the in-field data shown in Table II was 65.3 dBA (standard deviation=3.9 dB), almost all the

TABLE II. Individual teachers' vocal efforts, $L_{spA1\ m}$, measured in seven occupied classroom types with the indication of the teacher's gender (f/m) and the type of speech (t=text/l=lecture), average values for each classroom type and corresponding free-field values based on the averages, $L_{spA1\ m,free\ field}$. Measured background noise level L_{Aeq} and L_{A90}.

	Vocal effort												Noise	
	$L_{spA1\ m}$ dB(A)													
Classroom	Individual teachers' values										Average (st.dev.)	$L_{spA1\ m,free\ field}$ dB(A)	$L_{A,eq}$ dB(A)	L_{A90} dB(A)
S1	68.5	(f,l)	63.9	(f,t)	70.7	(f,t)					67.7 (3.5)	63.1	38.6	33.8
M1	69.0	(f,t)	62.4	(m,l)	68.1	(f,t)	67.2	(f,l)			66.7 (2.9)	64.0	35.2	28.9
M2a	69.8	(f,l)	68.0	(m,l)	65.8	(f,t)	67.2	(f,t)			67.7 (1.7)	62.5	44.3	39.0
M3	59.3	(f,t)	64.1	(m,l)	63.2	(m,l)					62.2 (2.5)	59.0	41.2	31.4
L1	60.4	(f,t)	69.1	(f,t)	71.3	(f,l)	58.2	(f,t)	63.2	(f,l)	64.5 (5.6)	61.5	38.4	28.7
L2	64.2	(m,t)	58.6	(f,t)	60.5	(f,t)					61.1 (2.9)	57.6	37.9	32.1
EL1	70.1	(f,l)	68.8	(f,l)	63.2	(f,t)	64.2	(f,t)			66.6 (3.4)	65.1	32.6	28.2

vocal efforts were above 60 dB(A), and half of the values fell above 66 dB(A). No significant differences were observed between males and females, while the average value for the text reading, 64.2 dBA (s.d.=4.1), was about 3 dB lower than those for the lecture, that is 67.0 dBA (s.d=3.2). As far as the free-field value is concerned a mean value, referred to the same sample, of 62.0 dBA (s.d.=4.0) denotes a vocal effort of between "Normal" (60 dBA) and "Raised" (66 dBA), according to the ISO 9921:2003 standard.[8] For a free-field "Normal" vocal effort Pavlovic[11] and Byrne *et al.*[12] reported 63.0 and 58.0 dB, respectively, which, minus 2.5 dB for conversion to an A-weighted value,[14] gives 60.5 dB(A) and 55.5 dB(A), respectively. Houtgast[1] found a $L_{\text{spA1 m,free field}}$ of 57.0 dB(A) in a 200 m^3 occupied classroom with students exposed to traffic noise. Picard and Bradley[14] indicate 60.1 dB(A) at 2 m from the teacher's mouth, as a mean value over a large set of data from kindergarten to university. If this value were to be measured in an average classroom of 300 m^3, with a reverberation time of 0.7 s, a $L_{\text{spA1 m,free field}}$ of 60.5 dB(A) would be obtained using Barron and Lee's theory.[19] Sato and Bradley[13] found a $L_{\text{spA1 m,free field}}$ of 68.8 dB(A) in noisy primary schools. The present result of 62.0 dB(A) is slightly higher than the literature data, apart from that by Sato and Bradley, but it should be considered that most of the previously indicated vocal efforts were calculated values or obtained from measurements in controlled fields.

Table II shows also the comparison between the vocal efforts of the teachers and the noise levels L_{Aeq} and L_{A90}. Most of the L_{Aeq} values were lower than the *acceptable* target of 40 dB(A) as indicated by Picard and Bradley,[14] but only one is lower than the *ideal* target of 33 dB(A). The L_{A90} noise levels were lower for fully renovated classrooms (M1, L1 and EL1) than for partially and nonrenovated ones, and most of them were lower than 33 dB(A). All the classrooms look onto a quiet street or square, except S1 and M1, which look onto a courtyard, but no marked differences were observed between the two types of classrooms in this respect, which means that the noise comes mainly from inside the building. In a comparison with literature data, all measured in urban area with quiet students, Shield and Dockrell[20] found an average L_{Aeq} of 56.3 dB(A) in primary schools, Houtgast[1] of 47.4 dB(A) (s.d.=3.1) with 8–15-year-old students and Bradley[2] of 41.9 dB(A) (s.d.=2.1) with 12–13-year-old students. The L_{Aeq} values in Table II are similar to those reported by Hétu *et al.*[4] which in empty classrooms in occupied buildings located far from traffic arteries measured 37.2 and 37.8 dB(A).

3. Speech intelligibility

Figure 2 shows the mean SNR_A and STI values and the min-max range bars for each occupied classroom type, with the exception of M2b and M4 (see Sec. IV B 2). These measures were obtained for six positions uniformly distributed over the seating area, and then averaged. The SNR_A values varied from 15.4 to 27.0 dB(A), but no marked differences were observed between nonrenovated and renovated classrooms. High values of SNR_A were found in the classrooms,

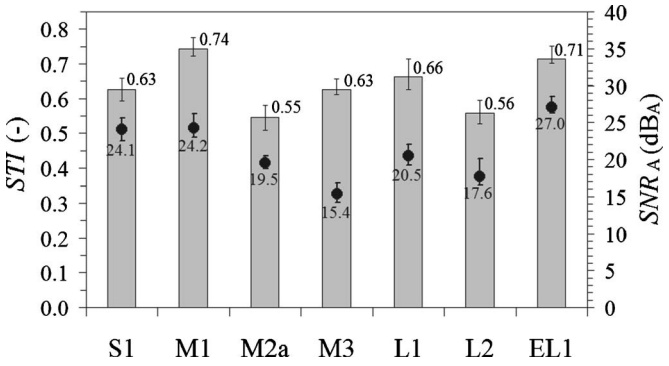


FIG. 2. Mean STI (gray blocks) and SNR_A (black circles) values and min-max range bars for seven of the eight classroom types.

which signifies that the teachers tend to compensate for noise with a greater vocal effort in order to ensure better student-speech comprehension. In the non- or poorly renovated classrooms, M2a and L2, the STI values were 0.55 and 0.56, respectively, 0.63 in both of the partially renovated S1 and M3, and 0.74, 0.66 and 0.71, respectively, in the fully renovated M1, L1 and EL1.

All the SNR_A values are higher than the optimal target of 15 dB(A), while, due to high reverberation, the STI values in M2a and L2 do not meet the minimum criterion of 0.60. The STI values were also mathematically derived following the lines of statistical room acoustics, according to the overall nonfrequency-specific approach reported in Houtgast *et al.*[9] After the satisfactory correspondence between the measured and calculated STI values had been checked for the partial occupancy, the new values for the fully occupied condition were obtained. Even though this method only provides approximations, no relevant differences were observed between the original and corrected values, confirming what has been stated previously.

V. SUBJECTIVE SURVEY

A subjective survey on perceived environmental quality has been carried out on the 51 classrooms by means of questionnaires. The main objectives were to investigate the relevance of the four environmental aspects in the overall environmental quality perception and to analyze the factors that affect the acoustical quality in secondary-school classrooms. All statistical analyses were carried out with the support of the SPSS® package. Subjective data related to acoustical quality were also correlated with the objective values, as described in Sec. VI.

A. Questionnaire

The questionnaire was drawn up following a methodology based on specific literature.[21] Experts in thermo-fluid dynamics and lighting have contributed to acquire all the relevant components of subjective perception concerning each environmental aspect. It was validated after numerous pilot tests with individual classes of different ages with the aim to test the readability and comprehension of the text and the ease of administration. The final version, which is available from the authors, contained 55 questions in six sections:

TABLE III. Influence of the four different environmental aspects on the students' school performance: mean scores of the answers and *t*-test significances for the differences of the mean scores between the renovated and nonrenovated classrooms. The five-point scales range from "very little" (1) to "very much" (5).

	Mean scores attributed to the influence of each aspect on the students' school performance (1–5 scale)				
	Renovated classrooms (702 ind.)		Non-renovated classrooms (150 ind.)		
Environmental quality aspect	Mean	95% confidence interval	Mean	95% confidence interval	*t* test for the difference of the means (*p* value)
Acoustical	**3.47**	[3.38, 3.56]	**3.32**	[3.13, 3.51]	0.16
Thermal	**3.08**	[2.99, 3.18]	**3.13**	[2.92, 3.34]	0.70
Indoor air	**2.92**	[2.83, 3.02]	**3.00**	[2.80, 3.20]	0.50
Visual	**3.59**	[3.50, 3.68]	**3.35**	[3.12, 3.57]	0.05

the first two sections were on general information and overall environmental quality, while the last four sections were on acoustical, thermal, indoor air and visual quality. Most of the answers referred to a 5-point scale, in which each step was labeled from 1 to 5, and the extremes with semantic descriptors.

The general information section was related, among others, to the influence of the four environmental aspects on students' school performances. The overall quality section consisted of one single question on the satisfaction of all the environmental aspects together.

The acoustical quality section covered: intensity and disturbance to lessons due to the average noise in the classroom; intensity, disturbance and frequency of occurrence from some different noise sources in the classroom; reverberation of the teachers' and students' voices; how well students comprehend the spoken words by the teacher; perceived vocal effort of the teacher; frequency of a list of consequences caused by bad classroom acoustics; satisfaction with the classroom acoustics. Only the students who attended the school before the renovation were asked to indicate the degree of improvement or deterioration with respect to the previous condition.

The thermal quality section, according to EN ISO 10551:2001 standard,[22] basically concerned: perception of the thermal environment on a symmetrical 7-point two-pole scale (from "very cold" to "very hot"), frequency of annoyance due to sun rays through the window, frequency of drafts, satisfaction with the thermal conditions. The indoor air quality section covered: frequency of perception of the air as dry, frequency of perception of the classroom as dirty or dusty, frequency of opening the windows, intensity of odors, satisfaction with the indoor air quality. The section on visual quality covered: quantity of light (natural+artificial) over the desks and on the blackboard, annoyance due to glare from windows, lighting and from the overall brightness of the room, frequency of using artificial lighting systems, satisfaction with the lighting conditions.

Questionnaires were filled in during one day of February, about one year after the acoustical treatment in the classrooms had been carried out, so that the students had passed a sufficiently long period of time in the renovated classrooms to make subjective assessments. The students were asked to answer with reference to the winter period, when the typical weather was cold and sunny, with a daily average external temperature of 3.0 °C. In order to obtain coherent and realistic answers, the questionnaire was explained to the students before they filled it in.

B. Sample

The questionnaires were administered to 1006 students in 51 classes. Those containing missing answers, referring to subjects with hearing or visual problems and by non native Italian speakers, were disregarded from the full sample. After this, an analysis of the consistency of the answers was developed by means of the Kolmogorov-Smirnov normality test and using Mahalanobis and Cook distances. A final sample of 852 questionnaires was used for the subjective analyses. The students had an average age of 16.1, with a majority of females (88.5%, as this type of school is predominantly attended by females), and 99.9% were Italian. A reduced sample of 676 students, corresponding to the 40 representative classrooms of the eight chosen types, were also used for the correlation between the subjective and objective acoustical data.

C. Relevance of the single aspects in the overall environmental quality assessment

The relevance of each single aspect of the perceived quality (acoustical, thermal, indoor air and visual) to the overall environmental quality assessment was investigated from the final sample, subdividing the answers between renovated (702) and nonrenovated classrooms (150).

Four questions on the supposed influence of the four aspects on students' school performance, on a five-point scale from "very little" to "very much," were included in the survey. The mean scores the students attributed to the influence of each aspect are shown in Table III. Almost the same importance was awarded to the four aspects by the two groups of students, with a prevalence of influence of visual quality and acoustical quality, followed by thermal and indoor air quality. Apart from visual quality, there are no significant differences between the mean values for the renovated and nonrenovated classrooms.

The correlations of the different aspects with the overall satisfaction scores are shown in Table IV. In the renovated classrooms, the overall satisfaction was more closely corre-

TABLE IV. Correlation of the overall environmental quality satisfaction with the satisfaction of each of the four environmental aspects.

Environmental quality aspect	Correlation with overall environmental quality satisfaction (Pearson's coefficient)	
	Renovated classrooms (702 ind.)	Nonrenovated classrooms (150 ind.)
Acoustical	0.39	0.50
Thermal	0.50	0.28
Indoor air	0.32	0.31
Visual	0.29	0.25

lated to thermal satisfaction, while in the nonrenovated ones, the highest correlation is to acoustical satisfaction (significant with a p value equal to 0.00).

Table V reports the mean scores, the 95% confidence intervals and t-test significances of the mean differences, for the overall and for each environmental aspect in the acoustically renovated and nonrenovated classrooms. The five-point scales range from "very dissatisfied" to "very satisfied." In the renovated classrooms, the students perceived a fair level of satisfaction for acoustical and visual quality, with very similar scores, and lower values for thermal and indoor air quality. Lower values of satisfaction, for all the aspects, were reported in the nonrenovated classrooms, where the only significantly higher aspect than the others was the visual quality. In particular, it can be seen that the mean satisfaction score for acoustical quality increased from 2.21 to 3.48 after renovation. Even the overall quality satisfaction increased, from 2.17 to 3.09, but less than for the acoustical quality aspect, probably because the visual satisfaction (whose influence on the overall judgment in the students' school performances is more relevant, e.g. Table III) remains almost constant.

Some considerations can be made from a comparison of Tables IV and V. In the renovated classrooms, where a fair satisfaction level of acoustical quality was achieved, the overall quality satisfaction closely depended on the thermal quality, one of the aspects the students were less satisfied with. In the nonrenovated ones, where the acoustical quality was poor, this is the aspect that was mainly correlated to the almost negative overall quality judgment. With a parity of dissatisfaction concerning the acoustical, thermal and indoor air quality conditions, it seems that students attribute more relevance, in the overall quality judgment, to the acoustical condition, an aspect they considered more important for their school performance.

D. Results for the acoustical environment

1. Intensity, disturbance and frequency of occurrence of different noise sources

The mean values and standard deviations of the classroom mean values (used instead of the mean value of the total number of answers because of the differences in number of students in the classes) of the intensity, disturbance and frequency of occurrence of different noise sources in the classrooms, are shown in Fig. 3. The 5-point scales were from "very low" to "very high." The highest mean values were attributed to "Students talking in the classroom" (STC), with mean scores of more than 3 on the scale, while lower mean scores of about 2.2 were attributed to "Students moving in the classroom" (SMC). As far as the high mean scores of about 2.6 assigned to "Students talking and moving in the corridor" (STMCO) are concerned, the reason is the low sound insulation of the doors, while the absence of floating floors was probably the reason for the scores (about 2.0) assigned to "Students moving or shuffling in the neighboring classrooms" (SMNC). Sometimes open windows could have been the cause of the mean scores of about 2.1 and 1.8 for "Traffic" (TR) and "Other noise outside the building" (ONOB), respectively, while the lowest mean scores of about 1.6 and 1.3 were assigned to "Students talking in the neighboring classrooms" (STNC) and "Other noise inside the

TABLE V. Satisfaction scores for the overall environmental quality and the four environmental aspects on 1–5 discrete scales from "very dissatisfied" (1) to "very satisfied" (5): mean scores of the answers and t-test significances for the differences of the mean scores between the renovated and nonrenovated classrooms.

Environmental quality aspect	Renovated classrooms (702 ind.)		Nonrenovated classrooms (150 ind.)		t test for the difference of the means (p value)
	Mean	Confidence interval	Mean	Confidence interval	
Overall quality satisfaction	**3.09**	[3.04, 3.15]	**2.17**	[2.06, 2.28]	0.00
Acoustical quality satisfaction	**3.48**	[3.42, 3.55]	**2.21**	[2.08, 2.33]	0.00
Thermal quality satisfaction	**2.81**	[2.73, 2.88]	**1.95**	[1.80, 2.09]	0.00
Indoor air quality satisfaction	**2.55**	[2.49, 2.61]	**2.17**	[2.04, 2.31]	0.00
Visual quality satisfaction	**3.31**	[3.25, 3.39]	**2.87**	[2.73, 3.00]	0.00

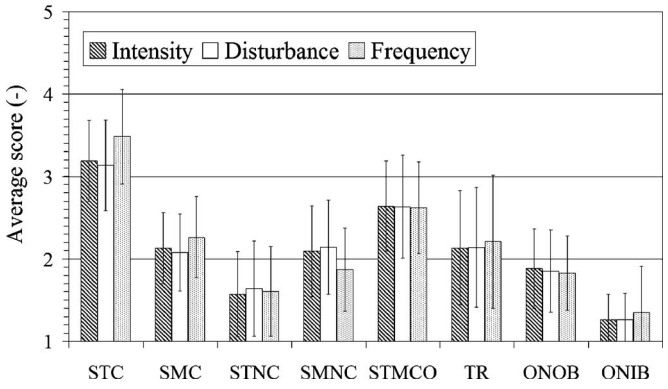


FIG. 3. Mean values and standard deviation of the mean classroom values of intensity, disturbance and frequency of occurrence of different noise sources in the classrooms. The five-point scale is bounded by the words "very low" (1) and "very high" (5). The following abbreviations are used for the noise sources: STC for "Students talking in the classroom," SMC for "Students moving or shuffling in the classroom," STNC for "Students talking in the neighboring classrooms," SMNC for "Students moving or shuffling in the neighboring classrooms," STMCO for "Students talking and moving in the corridor," TR for "Traffic," ONOB for "Other noise outside the building," and ONIB for "Other noise inside the building."

building" (ONIB), respectively. Different results are shown between the mean answer scores of the renovated and nonrenovated classrooms. For example, the mean disturbance scores for most sources in M3 are slightly higher than those in M4 (p value of t test lower than 0.10), with the exception of STNC, SMNC and ONIB. These results can be explained by considering that a higher reverberation time in M3 can amplify the noise inside the classrooms and make it seem more disturbing. The exceptions could be due to the fact that they are distant sources from outside the classroom, and hence more difficult to distinguish. On the other hand, when the occupied reverberation time is almost the same, as for S1 and M1, the mean scores are more similar (the difference is rejected with a p value higher than 0.40 for all the sources with the exception of STC and ONOB).

A correlation and a factorial analysis were performed, and they showed that the intensity, disturbance and frequency of each noise source are closely correlated. Exactly eight factors were singled out from the factorial analysis, each one corresponding to one of the above-mentioned noise sources. For this reason, in subsequent sections, when carrying out the data analyses, the scores attributed to these questions were replaced by the scores of these eight resulting factors.

2. Acoustical quality satisfaction

Noticeable differences between renovated and nonrenovated classrooms on the perception of some acoustical factors were observed. For the renovated classrooms the mean scores and 95% confidence intervals of speech comprehension (on a 5-point scale from "very badly" to "very well"), teachers' vocal effort (5-point scale from "very low" to "very raised") and voice reverberation (5-point scale from "very dry" to "very reverberant") are 3.88 [3.81, 3.95], 2.86 [2.81, 2.92] and 2.06 [1.99, 2.12], respectively, while for the nonrenovated ones the same mean scores are 3.07 [2.90, 3.23], 3.43 [3.31, 3.54] and 3.69 [3.52, 3.87]. In all the three cases the t tests strongly reject (with p values lower than 0.01) the hypothesis of no differences between the perceptions of the two groups. One of the questions on the acoustic environment, which was only answered by those students who were in the renovated rooms, was about the improvement in classroom acoustics after renovation. The arithmetic mean of these answers is 4.17, with a standard deviation equal to 0.93 on a 1 ("much worse") to 5 ("much better") discrete scale, thus it can be stated that the improvements after renovation were noticed by the students.

Table VI shows the most significant part of the correlation matrix for acoustic answers related to the renovated and nonrenovated classrooms. An arbitrary limit of the correlation coefficient $|r| \geq 0.25$ was chosen and only the coefficients with $p \leq 0.01$ are shown. Some correlations are only present for the nonrenovated classrooms with poor acoustic conditions. From the analysis, it seems that the poorer the acoustics, the more the acoustical quality satisfaction is af-

TABLE VI. Correlation matrix between the acoustic answers for the renovated classrooms (RC) and for the nonrenovated ones (NRC). An arbitrary limit of the correlation coefficient $|r| \geq 0.25$ was chosen and only the coefficients with $p \leq 0.01$ are shown.

	Acoustical quality satisfaction (AQS)		Speech comprehension (SC)		Teachers' vocal effort (TVE)		Voice reverberation (VR)		Noise intensity (NI)		Noise disturbance (ND)	
	RC	NRC	RC	NRC	RC	NRC	RC	NRC	RC	NRC	RC	NRC
AQS	1.00	1.00										
SC	0.42	0.56	1.00	1.00								
TVE		−0.32		−0.28	1.00	1.00						
VR	−0.32	−0.48	−0.27	−0.39			1.00	1.00				
NI									1.00	1.00		
ND		−0.41		−0.26		0.25		0.26	0.49	0.36	1.00	1.00
Students talking in the classroom									0.37	0.27	0.47	0.29
Students talking in the neighboring classrooms												0.30
Students moving or shuffling in the neighburing classrooms												0.25

fected by the factors that are not optimized. Speech comprehension is affected by voice reverberation, teachers' vocal effort and noise disturbance in the classrooms with poor acoustics, but only by voice reverberation in the classrooms with better acoustics. Acoustical quality satisfaction is affected by speech comprehension, voice reverberation, teachers' vocal effort and noise disturbance in the classrooms with poor acoustics, but only by speech comprehension and voice reverberation in the classrooms with better acoustics.

Generally, a good correlation exists between noise disturbance and noise intensity, and both of them are well correlated to students talking in the classroom. In the nonrenovated classrooms, poor sound insulation also determines close correlations between noise disturbance and students talking in the neighboring classrooms and students moving or shuffling in the neighboring classrooms. These correlations are only with noise disturbance and not with noise intensity, probably because the noise from these other sources is not very intense but it is very annoying. The above results agree with Kennedy *et al.*,[7] who, in university classrooms, found students talking in the classroom as the factor that is most commonly reported as interfering with the listening environment, followed by intermittent noises in the building but outside the classroom, while constant noise within or outside the building was less likely to be reported as interfering.

3. Consequences caused by poor acoustics

The students were asked to indicate the frequency of a list of perceived consequences caused by poor classroom acoustics on a five-point scale from "never" to "very often." Only the mean values for the students who were not satisfied about the overall classroom acoustics (i.e., the 165 students that marked 1 and 2 on the correspondent satisfaction scale) have been analyzed. The most important consequences of the poor acoustics in the classrooms are "Decrease in concentration" (mean=3.5, s.d=1.2), "Decrease in teacher voice perception" (mean=3.2, s.d=1.1) and "Decrease in students questions perception" (mean=3.1, s.d.=1.2). The most commonly reported adverse consequence of a poor listening environment according to Kennedy *et al.*[7] was failure to hear questions asked by other students in the class followed by concentration broken, which coincides with the present results.

E. Results of the thermal, indoor air and luminous environments

Correlation analyses based on the answers of the full sample, concerning the thermal, indoor air and luminous environments, were performed. These correlations show that, as far as the thermal conditions are concerned, the dissatisfaction is associated with the high temperature and the drafts, that students feel when they open the windows for ventilation and cooling during breaks. External screens on the windows and ventilation systems should be applied. Ventilation is also necessary for indoor air quality since students associate dissatisfaction with the high intensity of odors. As for visual quality, the students associate dissatisfaction with the brightness of the windows and lighting. Blinds or curtains should be mounted on the windows, slightly darker paint should be used on the walls and the lighting system should be correctly designed. Once again, the satisfaction of the thermal and visual conditions depends on the factors for which the students feel discomfort, which are not optimized in the building. Most of the classrooms in fact have windows without screens and are exposed to direct solar radiation, which causes high temperatures inside the classrooms and too much brightness from the windows.

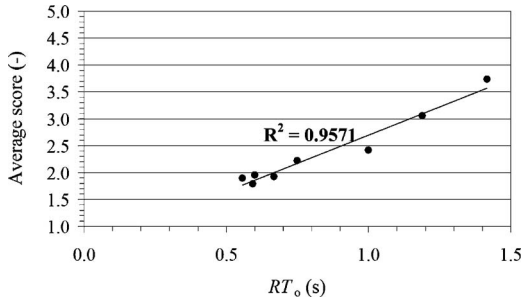


FIG. 4. Average scores for voice reverberation versus measured values of reverberation time (single number frequency averaged between 500 Hz, 1 kHz and 2 kHz) in occupied classrooms, and best-fit regression line. The five-point scale is bounded by the words "very dry" (1) and "very reverberant" (5).

VI. RELATIONSHIPS BETWEEN THE OBJECTIVE AND SUBJECTIVE DATA

A. Voice reverberation

Figure 4 plots the RT_o against the average scores for voice reverberation in the eight classroom types. The good correlation (R^2=0.957) was maintained when the reverberation times were corrected for full occupancy. It seems that students are aware of the different reverberant conditions in the classrooms, and are able to classify the sensations in a judgment scale, even though a larger amount of data would be necessary to confirm this statement.

B. Noise disturbance and intensity

Figure 5 shows the averages of the noise disturbance

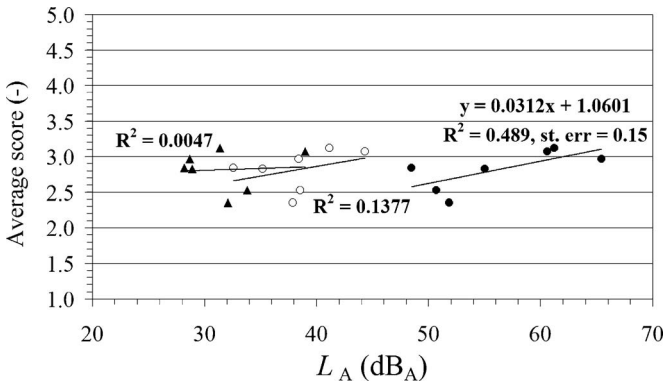


FIG. 5. Average noise disturbance scores versus measured values of L_{Aeq} (white circles), L_{A90} (solid triangles) and $L_{A\,max}$ (solid circles) and best-fit regression lines. The five-point scale is bounded by the words "very low" (1) and "very high" (5).

scores for each classroom type versus the corresponding measured values of L_{Aeq}, L_{A90} and $L_{A\,max}$, and the best-fit lines. Results are given for seven of the eight classroom types, as types M2b and M4 were excluded from the in field measurements (see Sec. IV B 2). A slight correlation exists between the mean subjective scores and the $L_{A\,max}$ (related to single-event noise, measured inside the classrooms), with an R^2 of 0.489 (p value for incorrelation test, $r=0$, is equal to 0.08): noise disturbance scores increase with an increase in the maximum A-weighted sound-pressure levels. Similar results for noise intensity have been observed, where, again, a good correlation is present between the subjective scores and $L_{A\,max}$, with an R^2 coefficient of 0.531 (p value for the incorrelation test equal to 0.06). It should be pointed out that these correlations are only significant when classroom average scores, instead of the answers of the single students, are considered. However, they seem to reveal that a stronger relationship exists between either noise disturbance or intensity and $L_{A\,max}$, more than L_{Aeq} and L_{A90}, so showing that students seem to be more disturbed by intermittent loud noises than by constant noise. This has also been proven in recent research by Dockrell and Shield,[5] who found that for young children (6–11-year olds) external $L_{A\,max}$ levels play a significant role in reported annoyance (caused mainly by trains, motorbikes, lorries and sirens), whereas external L_{A90} and L_{A99} levels play a significant role in determining whether or not children hear sound sources.

C. Speech comprehension

The STI and SNR_A were considered as the predictors of speech comprehension scores. These measures were obtained for six positions in each of the six chosen occupied classroom types. No measurements were carried out in rooms M2b or M4 (see Sec. IV B 2); room L2 was also excluded because of fewer subjective data (only one classroom was surveyed for this type, instead of a minimum of three for the others). The student seating area of each classroom was divided into six approximately equal areas around each measurement point, counting at least four student positions, in order to correlate the measurements to the speech comprehension scores. The average speech comprehension score for each of these groups was obtained by averaging the answers of all the student around the same measurement position for all the classrooms of the same type. Figure 6 shows the average speech comprehension scores versus measured STI. A slight relationship ($R^2=0.342$) can be observed: STI values close to 0.80, which qualifies as excellent intelligibility, can be associated with higher average speech comprehension scores (4.5 on a 1–5 scale), while STI values of about 0.50, corresponding to fair intelligibility, can be associated with the medium score (point 3 of the scale). The same good correlations are maintained when the reverberation times are corrected for full occupancy. A similar analysis with SNR_A showed no correlation with the subjective scores ($R^2=0.072$). Teachers tend to compensate for noise with higher vocal efforts, guaranteeing high values of SNR_A in all the classrooms, but, even with these high SNR_A levels, the students are aware of the detrimental effect of reverberation,

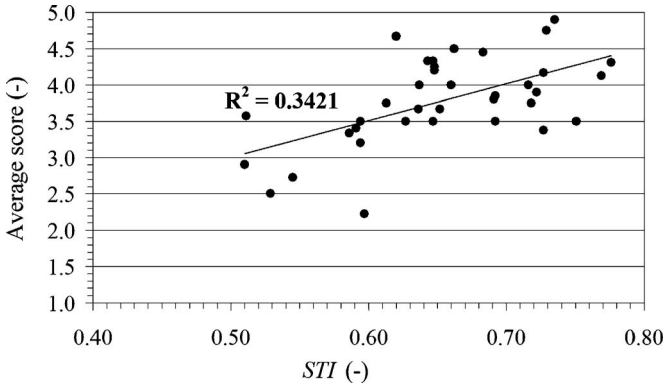


FIG. 6. Average speech comprehension scores for each point in the six chosen classroom types (S1, M1, M2a, M3, L1, EL1) versus STI values and best-fit linear regression function. The five-point scale is bounded by the words "very badly" (1) and "very well" (5).

which is well represented by the better association of the speech comprehension scores with the STI values. These representations are only an attempt to correlate the assessment of speech communication with the measured parameters. A correlation exists between the STI values and the speech comprehension scores in the classrooms, but it is not the same as the correlation between the STI and speech intelligibility, which is obtained with speech intelligibility tests.[2,3,8] In a speech comprehension score there is a speech intelligibility contribution, but also the contribution of other factors that have not been investigated in the survey. Kennedy *et al.*[7] found that other environmental aspects, personal factors, course material and teachers' characteristics were at least as important as STI values in predicting the perception of listening ease (PLE) score in university classrooms. Volberg *et al.*[23] found that, when evaluating the quality of speech communication, the listeners take into account speech intelligibility, but also the effort to understand what the speaker says, how difficult the task is, how annoying the environment and how absorbing other parallel activities are. Hagen *et al.*[6] reported a significant improvement in the subjective evaluation after acoustical interventions were made in classrooms, but not sufficient for successful listening at school. They indicated an improvement of the listening climate due to the correct behavior of the teacher which comprehend loudness of voice, articulation, listening mode, not shouting.

VII. CONCLUSIONS

A subjective survey on perceived environmental quality has been carried out on 51 secondary-school classrooms, some of which have been acoustically renovated, and acoustical measurements were carried out in eight of the 51 classrooms, these eight being representative of the different types of classrooms that are the subject of the survey.

Concerning acoustical measurements, it was confirmed that sound-absorption treatments are necessary also in small occupied classrooms in order to obtain optimal reverberation times, and that the noise levels in the classrooms of the school, far from high traffic arteries with quiet students inside the classrooms, are generally lower than the *acceptable*

target of 40 dBA L_{Aeq}.[14] The noise comes mainly from inside the building and the average free-field vocal effort is between "Normal" and "Raised." It should be pointed out that other studies are necessary to better investigate factors that can influence the teacher's vocal effort.

Concerning the subjective survey, the students awarded a prevalence of influence of visual and acoustical quality on school performances, and with a parity of dissatisfaction in the acoustical, thermal and indoor air quality conditions, it seems that they attributed more relevance in the overall quality judgment, to the acoustical satisfaction. The subjective evaluations of intensity, disturbance and frequency of each noise source are closely correlated, and the highest perceived noisy source are the students talking in the classroom. Acoustical satisfaction was lower in nonrenovated classrooms, and one of the most important consequences of poor acoustics was the decrease in concentration.

From the correlations between objective and subjective data, a stronger relation has been noticed between both noise disturbance and intensity average scores and $L_{A\,max}$ levels, more than L_{Aeq} and L_{A90}, so showing that students seemed to be more disturbed by intermittent loud noises than by constant noise. Teachers compensated for noise guaranteeing SNR_A values higher than the optimal target of 15 dB(A) in all the classrooms while in nonrenovated ones STI values do not meet the minimum criterion of 0.60. Even with these high SNR_A, the students were aware of the detrimental effect of reverberation, which is well represented by the better association of the speech comprehension scores with the STI values. It should be pointed out that in speech comprehension there is a valuable speech-intelligibility contribution but also the contribution of other factors of the listening environment, considered in recent literature, that can strongly improve speech comprehension, and that can be investigated in future studies.

ACKNOWLEDGMENTS

The authors thank the School Building Department of the Province of Turin who funded this work, and the teachers and students of the school, who gave up class time to participate in this study. We are also thankful to the reviewers for the careful reading of the manuscript and for the long lists of detailed corrections and suggestions that have helped greatly to improve the paper.

[1]T. Houtgast, "The effect of ambient noise on speech intelligibility in classrooms," Appl. Acoust. **14**, 15–25 (1981).
[2]J. S. Bradley, "Speech intelligibility studies in classrooms," J. Acoust. Soc. Am. **80**(3), 846–854 (1986).
[3]J. S. Bradley and H. Sato, "Speech intelligibility results for grade 1, 3 and 6 children in real classrooms," in *Proceedings of the 18th International Congress on Acoustics*, Kyoto, Japan, 2004, paper ID: Tu4.B1.2, pp. II-1191–1194.
[4]R. Hétu, C. Truchon-Gagnon, and S. A. Bilodeau, "Problems of noise in school settings: A review of literature and the results of an exploratory study," J. Speech Lang. Path. Audiol. **14**(3), 31–39 (1990).
[5]J. E. Dockrell and B. M. Shield, "Children's perceptions of their acoustic environment at school and at home," J. Acoust. Soc. Am. **115**(6), 2964–2973 (2004).
[6]M. Hagen, J. Kahlert, C. Hemmer-Schanze, L. Huber, and M. Meis, "Developing an Acoustic School design: Steps to improve Hearing and listening at school," Build. Acoust. **11**(4), 293–307 (2004).
[7]S. M. Kennedy, M. Hodgson, L. D. Edgett, N. Lamb, and R. Rempel, "Subjective assessment of listening environments in university classrooms: Perceptions of students," J. Acoust. Soc. Am. **119**(1), 299–309 (2006).
[8]ISO 9921, "Ergonomics—Assessment of speech communication," International Organization for Standardization, Genève (2003).
[9]T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60–72 (1980).
[10]EN 60268-16, "Objective rating of speech intelligibility by speech transmission index," European Committee for Standardization, Brussels, (2003).
[11]C. V. Pavlovic, "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**(2), 413–422 (1987).
[12]D. Byrne *et al.*, "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108–2120 (1994).
[13]H. Sato and J. S. Bradley, "Evaluation of acoustical conditions for speech communication in active elementary school classrooms," in *Proceedings of the 18th International Congress on Acoustics* (Kyoto, Japan, 2004), paper ID: Tu4.B1.1, pp. II-1187–1190.
[14]M. Picard and J. S. Bradley, "Revisiting speech interference in classrooms," Audiology **40**, 221–244 (2001).
[15]B. M. Shield and J. E. Dockrell, "The effects of noise on children at school: A review," Build. Acoust. **10**(2), 97–116 (2003).
[16]ANSI S12.60, "Acoustical performance criteria, design requirements, and guidelines for schools" (American National Standards Institute, New York, 2002).
[17]Department for Education and Skills, "Building Bulletin 93: Acoustic Design of School," The Stationery Office, London, 2003 (www.teachernet.gov.uk, last viewed October 25, 2007).
[18]M. R. Hodgson, "Empirical prediction of speech levels and reverberation in classrooms," Build. Acoust. **8**(1), 1–14 (2001).
[19]M. Barron, and L. J. Lee, "Energy relations in concert auditoriums. I," J. Acoust. Soc. Am. **84**(2), 618–628 (1988).
[20]B. M. Shield and J. E. Dockrell, "External and internal noise surveys of London primary schools," J. Acoust. Soc. Am. **115**(2), 730–738 (2004).
[21]F. Ortalda, *La Survey in Psicologia* ("The Survey in Psychology") (Carocci, Roma, 1998).
[22]EN ISO 10551, "Ergonomics of the thermal environment—Assessment of the influence of the thermal environment using subjective judgement scales," European Committee for Standardization, Brussels, 2001.
[23]L. Volberg, M. Kulka, C. A. Sust, and H. Lazarus, "Speech intelligibility and the subjective assessment of speech quality in near real communication conditions," Acta. Acust. Acust. **92**, 406–416 (2006).

A comparison of filter design structures for multi-channel acoustic communication systems

Pierre M. Dumuid,[a] Ben S. Cazzolato, and Anthony C. Zander
ANVC Group, School of Mechanical Engineering, The University of Adelaide, South Australia 5005

(Received 31 May 2007; revised 9 October 2007; accepted 9 October 2007)

The application of inverse filter designs as a means of providing improved communication performance in acoustic environments is investigated. Tikhonov regularized inverse filters of channel transfer functions calculated in the frequency domain are used as a means of obtaining multi-channel filters. Three classifications of inverse filter structures have been considered using time-domain simulations. The performance of Tikhonov regularized inverse filters designed according to each of these classifications is compared with each other and against a filter design developed by Stojanovic [Stojanovic, M. (2005). "Retrofocusing techniques for high rate acoustic communications," J. Acoust. Soc. Am. **117**, 1173–1185]. It is shown that the filter design developed by Stojanovic requires less regularization and outperforms the Tikhonov regularized inverse filter designs when communicating over a single channel. While the filter developed by Stojanovic is designed to use multiple transmitters to transmit to a single receiver, the filter was implemented in a multi-channel system and proposed to have a focusing similar to that obtained using time-reversal. It was found that for the scenario used in the simulation, the Tikhonov regularized inverse design for full multi-channel inversion achieved better focusing than the design by Stojanovic, where simulation results show 20 dB less cross-talk at the expense of around 2 dB loss in signal strength. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804940]

PACS number(s): 43.60.Ac, 43.60.Dh, 43.60.Fg, 43.60.Gk [EJS] Pages: 174–185

I. INTRODUCTION

Recently there have been numerous papers published on the design of communication systems for shallow underwater acoustic environments. Shallow underwater acoustic environments have been described as extremely difficult media in which to achieve high data rates. The major performance limitations arise from losses due to geometrical spreading and absorption, ambient noise, Doppler spread and reverberation from multi-path, with the latter being the primary limitation.

In the early 1990s, the principal means of combating multi-path in the shallow underwater environment was to use noncoherent modulation techniques (Kilfoyle and Baggeroer, 2000). Coherent techniques were found to be challenging due to the difficulty of obtaining a phase lock and also that the channel was subject to fading. Stojanovic *et al.* (1993) presented a communication design that addressed both of these problems. The design involved incorporating a joint update of the phase lock loop and the taps of the decision feedback filter (DFE). By using multiple receivers (described as spatial diversity), the effects of fading were reduced, and higher data rates were attainable.

More recently, a technique known as time-reversal (TR) has been investigated to achieve a coherent communication link, but with the aim of reducing the computational complexity. Time reversal was first introduced in an experiment by Parvulescu (1995) and Parvulescu and Clay (1965). Dowling (1994) proposed that the use of a passive phase conjugate receiver (an implementation of time-reversal at the receiver) could be a means to achieve coherent communications. The concept of time-reversal is described as follows: If the ocean impulse response between a transmitter and receiver is played temporally in reverse at the transmitter, then the response at the receiver location is found to have temporal and spatial focusing. The temporal focusing can be explained by the fact that the response at the receiver location is the auto-correlation of the channel response. A property of the auto-correlation function is that the signal is symmetric, having a peak value at $t=0$ and decays away from this time. The spatial focusing has been explained by Clay (1966) according to the manner in which the modes of the channel are excited, whereby the superposition of all the modes excited in an appropriate phase results in spatial focus. In many media, reciprocity can be assumed and the signal can also be injected at the original source location resulting in spatial and temporal focusing at the receiver location. Time reversal is generally implemented using an array of transmitters, often called a time-reversal mirror (Jackson and Dowling, 1991).

Although TR was demonstrated in 1961 by Parvulescu and Clay (1965) it was not until 1999 that Edelmann *et al.* (2002) demonstrated a communication system that employed time-reversal. The system developed by Edelmann *et al.* (2002) transmitted a binary phase shift key (PSK) signal using the time-reversal of a signal obtained from the transmission of a 2 ms, 3.5 kHz pure tone pulse and transmitting replicas of this wave form with a positive or negative scaling. Scatter plots showed that TR assists in mitigating the inter-symbol interference (ISI). Several researchers have since implemented time-reversal and also found it to reduce

[a]Author to whom correspondence should be addressed. Electronic mail: pierre.dumuid@adelaide.edu.au

ISI with reduced computational complexity Candy *et al.* (2005); Edelmann (2005); Edelmann *et al.* (2001); Edelmann *et al.* (2005); Flynn *et al.* (2004); Flynn *et al.* (2004); Rouseff (2005); and Yon *et al.* (2003). A number of these investigations have shown that TR without any other processing will always have some ISI present. Stojanovic (2005) in particular showed that TR was severely limited in its performance when compared to a number of other designs. Edelmann *et al.* (2005) and Song *et al.* (2006) have shown that the ISI in TR can be diminished by the use of a decision feedback equalizer (DFE). Song *et al.* (2006) demonstrated that near optimal results can be obtained (with respect to the results obtained by Stojanovic, (2005)), by realizing that a matched filter is not required at the receiver when TR is employed, since TR at the transmitter acts as a matched filter.

Another aspect of TR that is considered advantageous to underwater acoustic communications is the spatial focusing. Spatial focusing has been considered as a possible means to increase the data rate by transmitting different information to each element of the receiver array to achieve a multi-channel communications system (Candy *et al.*, 2005; and Song *et al.*, 2006). Cazzolato *et al.* (2001) showed that using Tikhonov regularized inverse filters achieves greater spatial and temporal focusing than TR for a simulated underwater environment.

The direct inversion of measured impulse responses (IRs) using time-domain techniques (see, for example, Nelson *et al.*, 1995) are particularly complex and require considerable computational effort. To speed up the calculations, Cazzolato *et al.* (2001) employed a method developed by Kirkeby *et al.* (1998) that reduced the computational effort required to design the inverse filter by performing the inversion within the frequency domain. The technique involved the use of a *regularization parameter* to ensure causality such that wraparound did not occur on the conversion back to the time-domain. Dumuid *et al.* (2006) showed that varying the regularization parameter from 0 to ∞ changed the performance from a pure inverse filter to that of a TR filter. An alternative approach to obtaining the inverse filter solution has been developed by Montaldo *et al.* (2004) where an approximation of the inverse filter may be obtained experimentally through iterating a TR based technique. This iterative method was developed for use in an ultrasound application where it was found that iterative TR was faster than performing the direct calculation of the inverse filter. For underwater acoustic communications, however, the transmission times are much longer and the iterative technique becomes impractical due to the long propagation time within the ocean. Higley *et al.* (2006) improved the speed of the iteration by performing the adaption in software, and it was also shown that the technique was found to be mathematically equivalent to the Neuman matrix inverse.

Cazzolato *et al.* (2001) demonstrated that inverse filtering outperformed time-reversal when using a Dirac impulse transmission within a simulation of an underwater environment. The work is extended here, to the transmission of a phase shift key (PSK) communication signal. It is shown that the Tikhonov inverse filter can be implemented in a number of different ways. The performance of these implementations is compared against TR and a design presented in Stojanovic (2005). As the design by Stojanovic was a theoretical filter design, the filter in the current studies was modified to make it practically implementable. Comparisons were made concerning how well the filters reduce the symbol errors for both a single channel and a multi-channel transmission. The filter designs were compared using a simulation obtained from a set of IRs measured from a reverberant open-air experimental configuration.

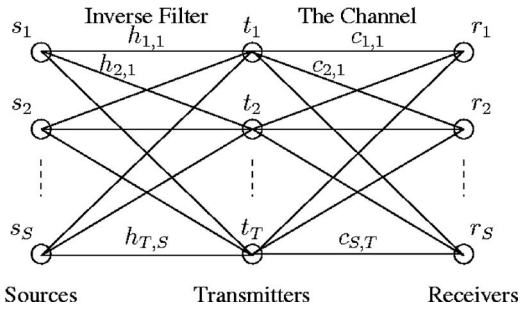


FIG. 1. Schematic of filter connections.

The results obtained from the simulation demonstrate that Tikhonov inverse filtering with appropriate regularization performs better than time-reversal. The filter design presented by Stojanovic (2005) is observed to outperform both time-reversal and Tikhonov inverse filters when used in a single channel scenario. However, the multi-channel implementation of the Tikhonov inverse filter is found to be the only filter able to perform multi-channel communications for the scenario presented.

II. FILTER DESIGNS

A. Design classifications

Three filter classifications are presented as a means by which inverse filters may be designed for multiple input - multiple output (MIMO) systems. They are: *inverse by path* (single input, single output (SISO)), *inverse by channel* (multiple input, single output (MISO)) and *inverse by full MIMO*. These classifications are based on the filter structure shown in Fig. 1. A set of sources, s_i, $i \in [1,S]$, are desired to be replicated at the receivers, r_i, $i \in [1,S]$, with minimal cross-talk through the use of a set of transmitters, t_j, $j \in [1,T]$. The channel through which the signals are transmitted is modeled as a set of IRs, $c_{i,j}(t)$ where $i \in [1,S]$ and $j \in [1,T]$ denote the receiver and transmitter, respectively. In order to achieve the desired response between the source and the receiver, a set of filters, $h_{j,i}(t)$, are generated.

The design of the inverse filters can be classified by the channel responses on which the inverse filter is dependent. Figure 2(a) presents a classification that shall be called *inverse by path*. In this design, each subfilter, $h_{j,i}(t)$ of the inverse filter is only dependent on the single channel response, $c_{i,j}(t)$. This model can be seen as a multi-channel filter developed by using multiple single channel systems at once, where each single channel system consists of a single source, $s_i(t)$, single transmitter, $t_j(t)$, and a single receiver, $r_i(t)$. The general aim of the *inverse by path* design is to find $h_{j,i}$ such that $h_{j,i}(t) * c_{i,j}(t) \simeq \delta(t)$, where $*$ is the convolution

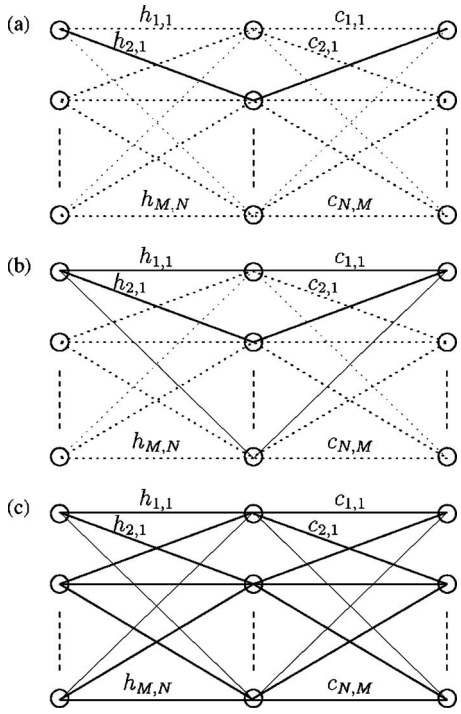


FIG. 2. Schematic of filter design classifications (a) *inverse by path*, (b) *inverse by channel*, and (c) *inverse by full MIMO*.

operator. If only the source, $s_i(t)$, is excited, then the response at receiver i using all transmitters is given by

$$r_i(t) = \left(\sum_{j=1}^{T} h_{j,i}(t) * c_{i,j}(t) \right) * s_i(t) \tag{1}$$

$$\simeq T\delta(t) * s_i(t), \tag{2}$$

where the summation in Eq. (1) is approximately $T\delta(t)$ in a multi-path environment as a result of the correlated peak at $t=0$ for all the terms in the summation and the value of $h_{j,i}(t) * c_{i,j}(t)$ for $t \neq 0$ is often incoherent and averages to zero. In other words, the T transmitters add coherently, increasing the gain T times.

The cross-talk for the *inverse by path* design at receiver i can be determined by measuring the response at this receiver due to a signal that is transmitted to target a different receiver, $r_{i_{ct}}$, where ct stands for "cross-talk." The response is given by

$$r_i(t) = \left(\sum_{j=1}^{T} h_{j,i_{ct}}(t) * c_{i,j}(t) \right) * s_{i_{ct}}(t). \tag{3}$$

The cross-talk is observed to be dependent on the correlations between the filter $h_{j,i_{ct}}(t)$ and the channel path $c_{i,j}(t)$ for $j \in [1,T]$. If the channels are sufficiently uncorrelated then the term in the brackets should tend towards zero as the number of transmitters is increased. The time-reversal mirror is an example of an inverse filter designed according to the *inverse by path* design, and its features of both focusing and temporal compression have been described in the literature extensively.

The second classification, called *inverse by channel*, is shown in Fig. 2(b). This classification encompasses the filter

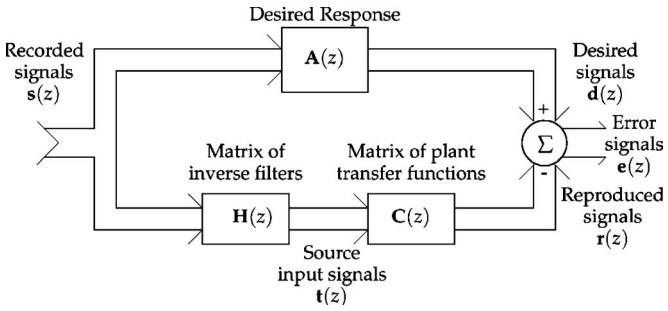


FIG. 3. Generic inverse filter system schematic (Kirkeby, 1998, Fig. 1).

designs where the subfilters, $h_{j,i}(t)$, $j \in [1,T]$ are calculated together to replicate the response $s_i(t)$ at receiver r_i. This design is different from the *inverse by path* design since the filter, $h_{j,i}(t)$ is dependent on multiple responses, $c_{i,j}(t)$, $j \in [1,T]$.

The cross-talk cancellation for the *inverse by channel* design is dependent on $h_{j,i_{ct}}(t)$ and $c_{i,j}(t)$ being uncorrelated. The filters presented in Stojanovic (2005) are examples of inverse filters designed according to the *inverse by channel* classification.

The third classification, called *inverse by full MIMO*, is shown in Fig. 2(c). Filters designed according to this classification require that each subfilter of the inverse filter is calculated based on all the filters in the channel response. Inverse filters designed according to this classification attempt to achieve a MIMO transfer function between sources and receivers that has a desired response between $s_i(t)$ and $r_i(t)$ while also minimizing the cross-talk between channels. Examples of this type of inverse filter design include those by Cazzolato *et al.* (2001), Montaldo *et al.* (2004), and Higley *et al.* (2006).

It should be noted that there are extensions to the design classifications discussed above, an example being the placement of filter at both the transmitter and receiver as represented in the designs by Stojanovic (2005).

B. Tikhonov regularized inverse filter

The design of the Tikhonov regularized inverse filter is based on the system presented in Fig. 3. A set of signals, $\mathbf{s}(z)$, are transformed by the filter, $\mathbf{A}(z)$, to produce a set of signals, $\mathbf{d}(z)$, that are to be replicated by the signals, $\mathbf{r}(z)$, being the output of the electro-acoustic system denoted by $\mathbf{C}(z)$. In order to achieve this, a filter $\mathbf{H}(z)$ is designed given that $\mathbf{C}(z)$ and $\mathbf{A}(z)$ are known. This filter is used on the transmission signals, $\mathbf{s}(z)$, to generate a set of signals $\mathbf{t}(z)$ that when played through the channel $\mathbf{C}(z)$ result in the signals $\mathbf{r}(z)$ at the receivers. Often the transfer matrix, $\mathbf{A}(z)$, is a delay to ensure causality, i.e., $\mathbf{A}(z)=z^{-m}\mathbf{I}$, or in the case of a communication system, the channel spectral shaping filter response, $\mathbf{A}(z)=g(z)\mathbf{I}$. This problem can be expressed as

$$\mathbf{r}(z) = \mathbf{C}(z)\mathbf{t}(z) \tag{4}$$

with the objective that

$$\mathbf{r}(z) = \mathbf{A}(z)\mathbf{s}(z). \tag{5}$$

Given that

FIG. 4. Photograph of the experimental rig used to obtain the impulse response functions showing transmitter and receiver arrays and randomly oriented objects used to emulate a difficult environment through which to transmit.

$$\mathbf{r}(z) = \mathbf{C}(z)\mathbf{H}(z)\mathbf{s}(z), \tag{6}$$

the filter, $\mathbf{H}(z)$ is designed so that $\mathbf{C}(z)\mathbf{H}(z)$ approximates $\mathbf{A}(z)$. Kirkeby *et al.* (1998) proposed a cost function to achieve this, along with a term to regulate the energy of the transmitted signal. The cost function is given by

$$J(z) = \mathbf{e}^{\mathrm{H}}(z^{-1})\mathbf{e}(z) + \beta\mathbf{t}^{\mathrm{H}}(z^{-1})\mathbf{t}(z), \tag{7}$$

where $\mathbf{e}(z) = \mathbf{d}(z) - \mathbf{r}(z)$ is the error signal, and β is a weighting term applied to the energy of the transmitted signal. The solution to this equation is given by (Kirkeby *et al.* (1998), Eq. (8))

$$\mathbf{H}(z) = (\mathbf{C}^{\mathrm{H}}(z^{-1})\mathbf{C}(z) + \beta\mathbf{I})^{-1}\mathbf{C}^{\mathrm{H}}(z^{-1})\mathbf{A}(z) \tag{8}$$

and its frequency domain equivalent,

$$\mathbf{H}(\omega) = (\mathbf{C}^{\mathrm{H}}(\omega)\mathbf{C}(\omega) + \beta\mathbf{I})^{-1}\mathbf{C}^{\mathrm{H}}(\omega)\mathbf{A}(\omega), \tag{9}$$

which was observed by Kirkeby *et al.* (1998) to be the Tikhonov regularized inverse filter design.

Kirkeby *et al.* (1998) observed that if the regularization parameter, β, was large enough then the temporal wrap-around was negligible, allowing a causal filter to be calculated in the frequency domain using fast Fourier transforms. It is also of interest to note that Higley *et al.* (2006) published a method that iteratively approaches Eq. (9).

The Tikhonov inverse filter designs for each classification shall be defined given the inverse filter be expressed as

$$\mathbf{H}(\omega) = \begin{bmatrix} H_{1,1}(\omega) & H_{1,2}(\omega) & \cdots & H_{1,J}(\omega) \\ H_{2,1}(\omega) & H_{2,2}(\omega) & \cdots & H_{2,J}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ H_{I,1}(\omega) & H_{I,2}(\omega) & \cdots & H_{I,J}(\omega) \end{bmatrix}. \tag{10}$$

The filter design according to *inverse by path* design is calculated by

$$H_{i,j}(\omega) = \frac{C^*_{j,i}(\omega)}{C^*_{j,i}(\omega)C_{j,i}(\omega) + \beta} = (|C_{j,i}(\omega)|^2 + \beta)^{-1}C^*_{j,i}(\omega), \tag{11}$$

the filter design according to *inverse by channel* is calculated by

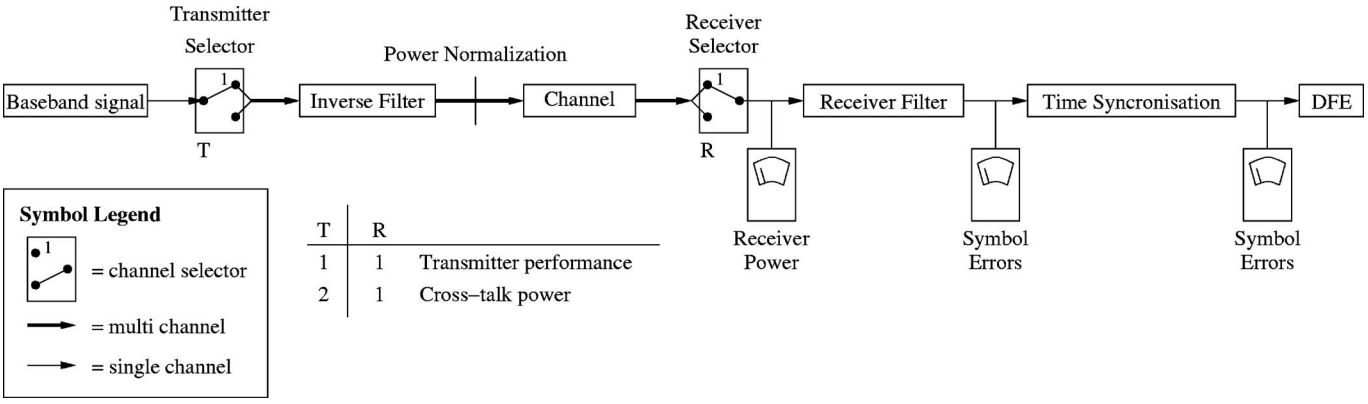


FIG. 5. Schematic of program used to test various filter designs using CONDOR. The "channel selector" block represents which channel of the MIMO stream is selected (when transmitting) or processed (after reception).

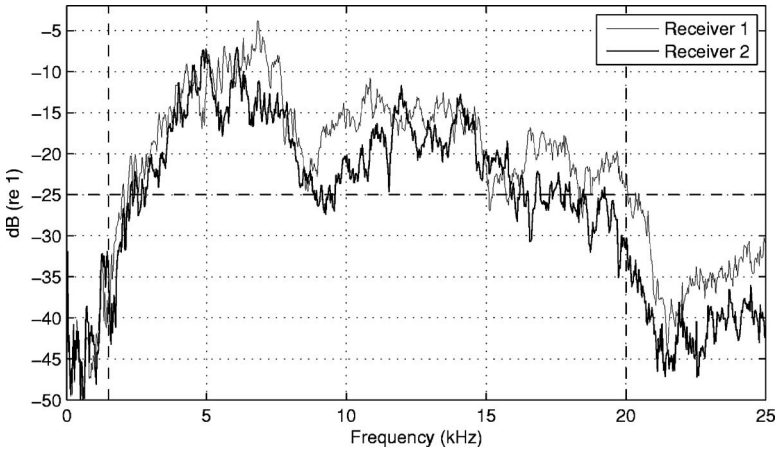


FIG. 6. Average frequency responses between all six transmitters and receivers 1 and 2. The vertical dashed lines show the region in which the simulations occurred, and the horizontal dashed line indicates the chosen operational level.

$$\begin{bmatrix} H_{1,j}(\omega) \\ H_{2,j}(\omega) \\ \vdots \\ H_{N,j}(\omega) \end{bmatrix} = \left(\begin{bmatrix} C^*_{j,1}(\omega) \\ C^*_{j,2}(\omega) \\ \vdots \\ C^*_{j,N}(\omega) \end{bmatrix} [C_{j,1}(\omega) C_{j,2}(\omega) \cdots C_{j,N}(\omega)] + \beta \mathbf{I} \right)^{-1} \times \begin{bmatrix} C^*_{j,1}(\omega) \\ C^*_{j,2}(\omega) \\ \vdots \\ C^*_{j,N}(\omega) \end{bmatrix} \tag{12}$$

and the filter design according to *inverse by full MIMO* is calculated by

$$\mathbf{H}(\omega) = (\mathbf{C}^{\mathrm{H}}(\omega)\mathbf{C}(\omega) + \beta \mathbf{I})^{-1} \mathbf{C}^{\mathrm{H}}(\omega). \tag{13}$$

Dumuid *et al.* (2006) showed that as β tended towards infinity, the filter, $\mathbf{H}(\omega)$ for the *inverse by full MIMO* design tended towards the time-reversal filter. Observing Eqs. (11) to (13), this property holds true for all of the inverse filters considered. Thus, as β approaches infinity, each of these filtering classifications approaches the time-reversal filter design.

C. Regularization of Stojanovic's two-sided filter for no ISI

Stojanovic (2005) compared time-reversal with a set of optimal equalization designs. These designs were developed for systems comprised of either multiple transmitters or multiple receivers, but not both. The filter design by Stojanovic that was used in this simulation is the two-sided filter that utilizes multiple transmitters. Stojanovic (2005) proposed another filter that theoretically performed better by allowing ISI. However, the filter is more difficult to implement and the benefits of using the filter are minimal.

The two-sided filter design developed by Stojanovic is given by

$$H_0(\omega) = K(\omega)\sqrt{X(\omega)}\,\gamma^{-1/4}(\omega), \tag{14}$$

$$H_m(\omega) = K^{-1}(\omega)\sqrt{X(\omega)}\,\gamma^{-3/4}(\omega)C^*_m(\omega), \tag{15}$$

where $H_m(\omega)$ is the filter at the mth transmitting array, $C_m(\omega)$ the channel response between the mth array element, $H_0(\omega)$ the filter at the receiver, $\gamma(\omega) = \Sigma_{m=1}^{M} |C^2_m(\omega)|$ the composite channel power spectral density, $X(\omega)$ the Nyquist transfer function and

$$K(\omega) = \sqrt{\frac{E/\sigma_d^2}{\int_{-\infty}^{+\infty} \sqrt{S_w(\omega)} \dfrac{X(\omega)}{\sqrt{\gamma(\omega)}} d\omega}} S_w^{1/4}(\omega), \tag{16}$$

where E represents the transmission energy, σ_d^2 the average power of the data sequence, and $S_w(\omega)$ the power spectral density of the noise.

This filter design can be separated into four filtering stages:

1. $K(\omega)$ - a filter that is related to the noise spectrum. If the noise if flat then $K(\omega)$ is a constant.
2. $X(\omega)$ - the Nyquist transfer function, being the raised cosine spectrum. The raised cosine spectrum is used as the desired total transfer function that has no ISI, and is commonly observed to be split between the transmitter and receiver (see Proakis, 2001, page 561).
3. $\gamma^{-1/4}(\omega)$ and $\gamma^{-3/4}(\omega)$ - the inverse of the composite channel power spectrum. The composite channel power spectra can be observed to be the total channel response when using time-reversal, and thus these filters combined are an inverse filter.
4. $C^*(\omega)$ - the time-reversal filter.

It can be observed that the two-sided filter given by Eqs. (14) and (15) is simply the time-reversal filter with equalizers fitted at both the transmitter and receiver to compensate for the summation of the time-reversal responses (through the inversion of $\gamma(\omega)$) and a filter to compensate for the noise spectrum (through the use of $K(\omega)$).

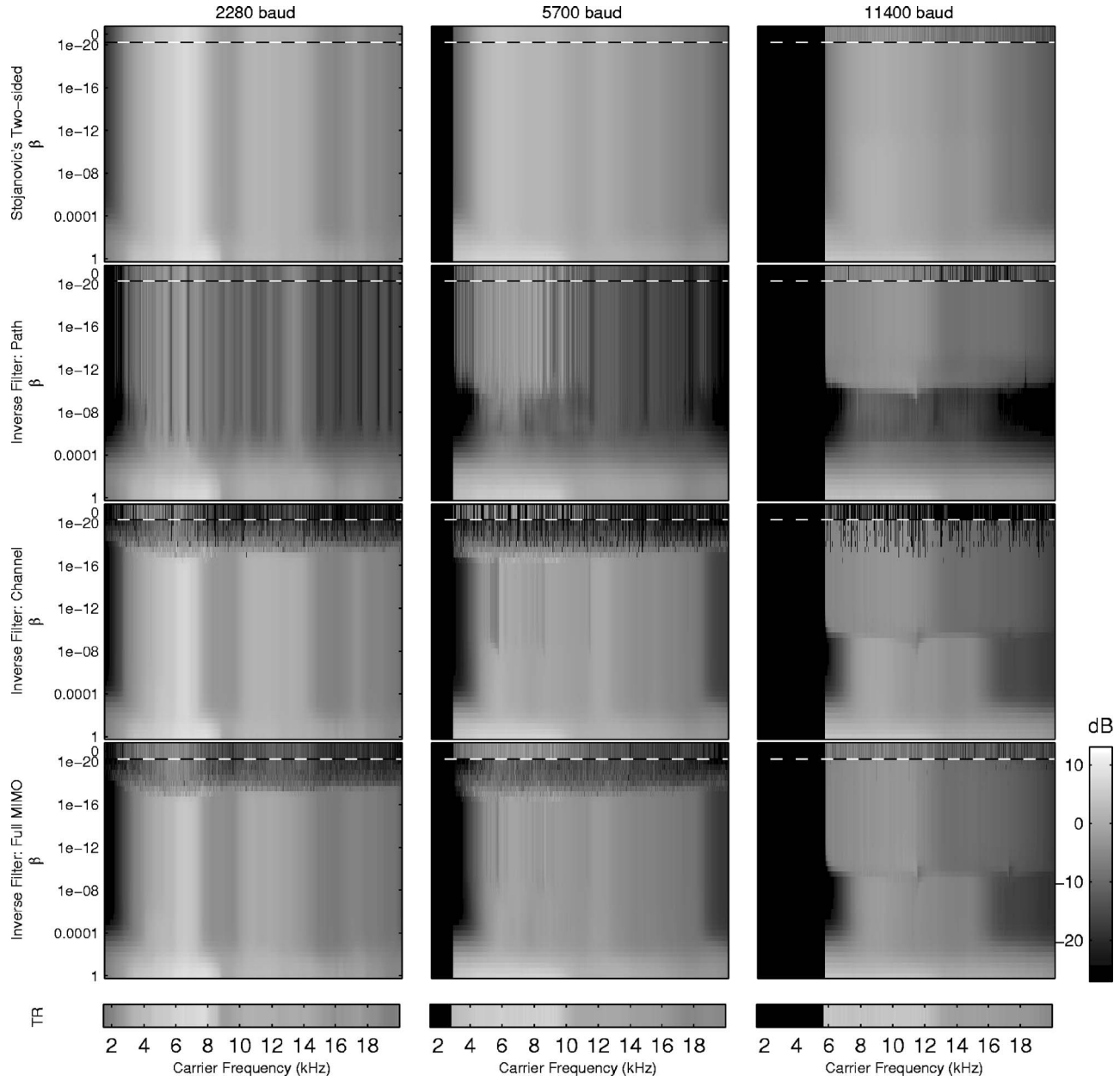


FIG. 7. Receiver power for the four filter designs operating with a carrier frequency ranging from 1.5 to 20 kHz at the symbol rates of 2280, 5700, and 11 400 baud, and with the regularization parameter, β, varying from 10^{-20} to 1. Results for $\beta=0$ are also included in the plots in the top row of pixels.

Stojanovic's work examined the theoretical performance of the two-sided filter. Such an examination does not consider the problems that occur when implementing the filters as time-domain filters. In this work, a small adjustment is made to the design to account for the noncausal wraparound. Wraparound may occur when converting the filters from the frequency domain into time-domain if there are zeros in the composite channel power spectrum. To avoid wraparound, a regularization parameter β can been added to produce a regularized filter design

$$H_0(\omega) = K(\omega)\sqrt{X(\omega)}(\gamma(\omega)+\beta)^{-1/4}e^{-jT\omega/2}$$

$$H_m(\omega) = K^{-1}(\omega)\sqrt{X(\omega)}(\gamma(\omega)+\beta)^{-3/4}e^{-jT\omega/2}C^*(\omega). \quad (17)$$

The term $e^{-jT\omega/2}$ has been added to make the filter casual, T being the duration of the fast Fourier transform window.

In this investigation, the design is also replicated for each channel to create a MIMO filter design as per the *inverse by channel* design of Fig. 2(b).

III. PERFORMANCE COMPARISONS

The following filter designs will be compared:

1. Two-sided filter design by Stojanovic (2005)
2. Time-reversal filter
3. Tikhonov regularized filter formed by
   (a) *inverse by path*
   (b) *inverse by channel*
   (c) *inverse by full MIMO*

The IRs used in the simulation were measured from a laboratory experiment that consisted of two arrays, each having

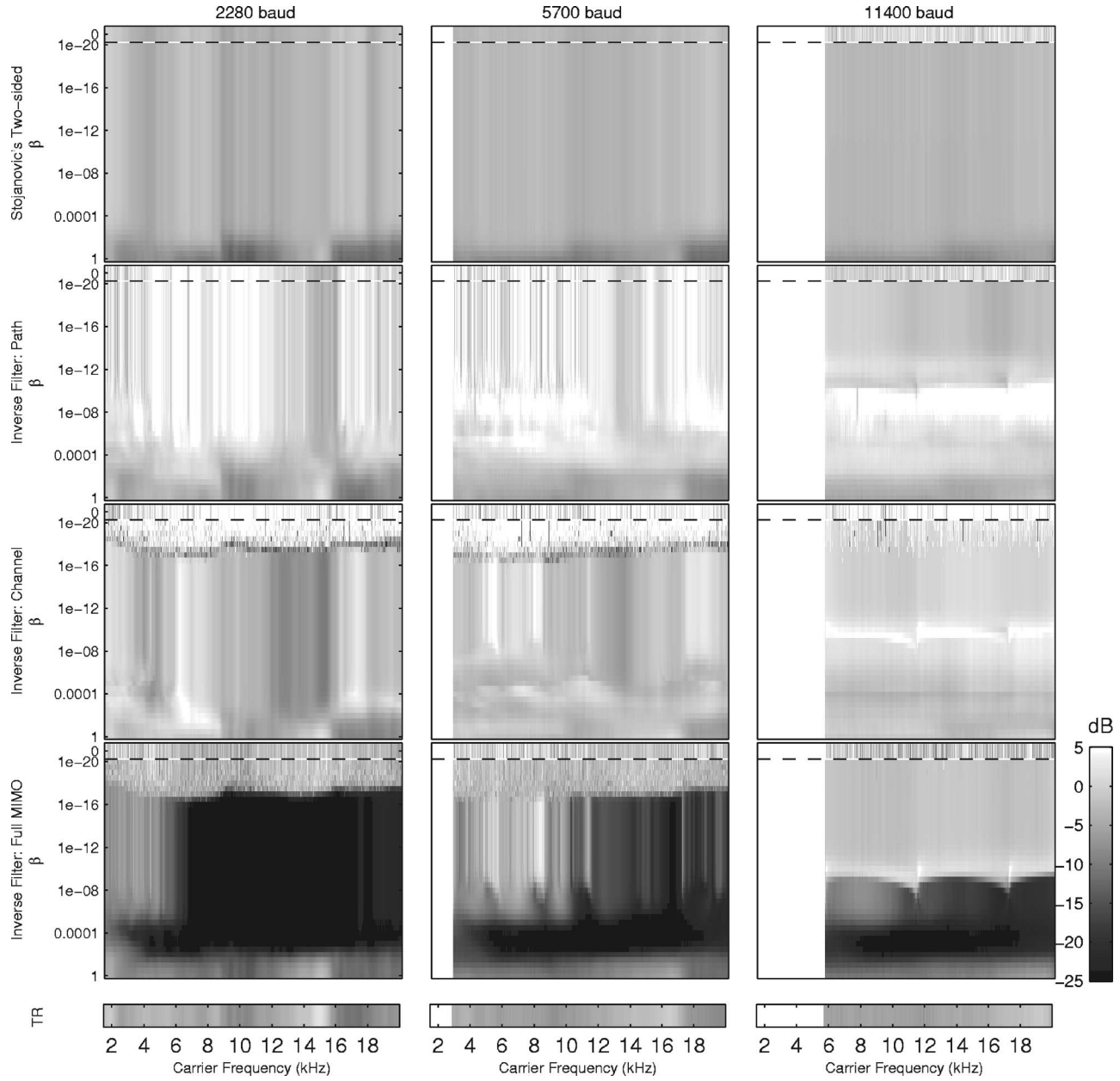


FIG. 8. Amplitude of the cross-talk power relative to the signal power for the four filter designs operating with a carrier frequency ranging from 1.5 to 20 kHz at the symbol rates of 2280, 5700, and 11 400 baud, and with the regularization parameter, β, varying from 10^{-20} to 1. Results for $\beta=0$ are also included in the plots in the top row of pixels.

six co-located speakers and microphones arranged on a 3 $\times 2$ grid with a grid spacing of 55 mm. The two arrays were separated by an approximately 1 meter long open-air channel containing various objects to increase the reverberation. Reverberation was desired to emulate a difficult environment through which to transmit, such as an underwater acoustic environment. Figure 4 shows a photograph of the experimental setup. The quality of the speakers and microphones used in the experiment was rather low to reduce the cost of the experiment, but the poor quality was considered a means of adding an extra degree of difficulty for the system to compensate. The IRs were measured at a sample rate of 55 kHz.

A number of parameters determined the performance of the communication system including the carrier frequency, symbol rate and the regularization parameter. Each filter was examined under a wide range of parameters in order to avoid the possibility that a certain set of conditions would favor one particular filter design over another. To perform the simulation, a number of computers were controlled using CONDOR (being a distributed computing software available from http://www.cs.wisc.edu/condor/). Each computer calculated the system performance for a set of design parameters. A schematic of the simulation performed on each computer is shown in Fig. 5.

The parameters varied for the simulation were the number of transmitter elements used, symbol rate, carrier frequency, regularization parameter and filter type. Each simulation was conducted as follows:

1. A sequence of 800 bits were used to generate 400 sym-

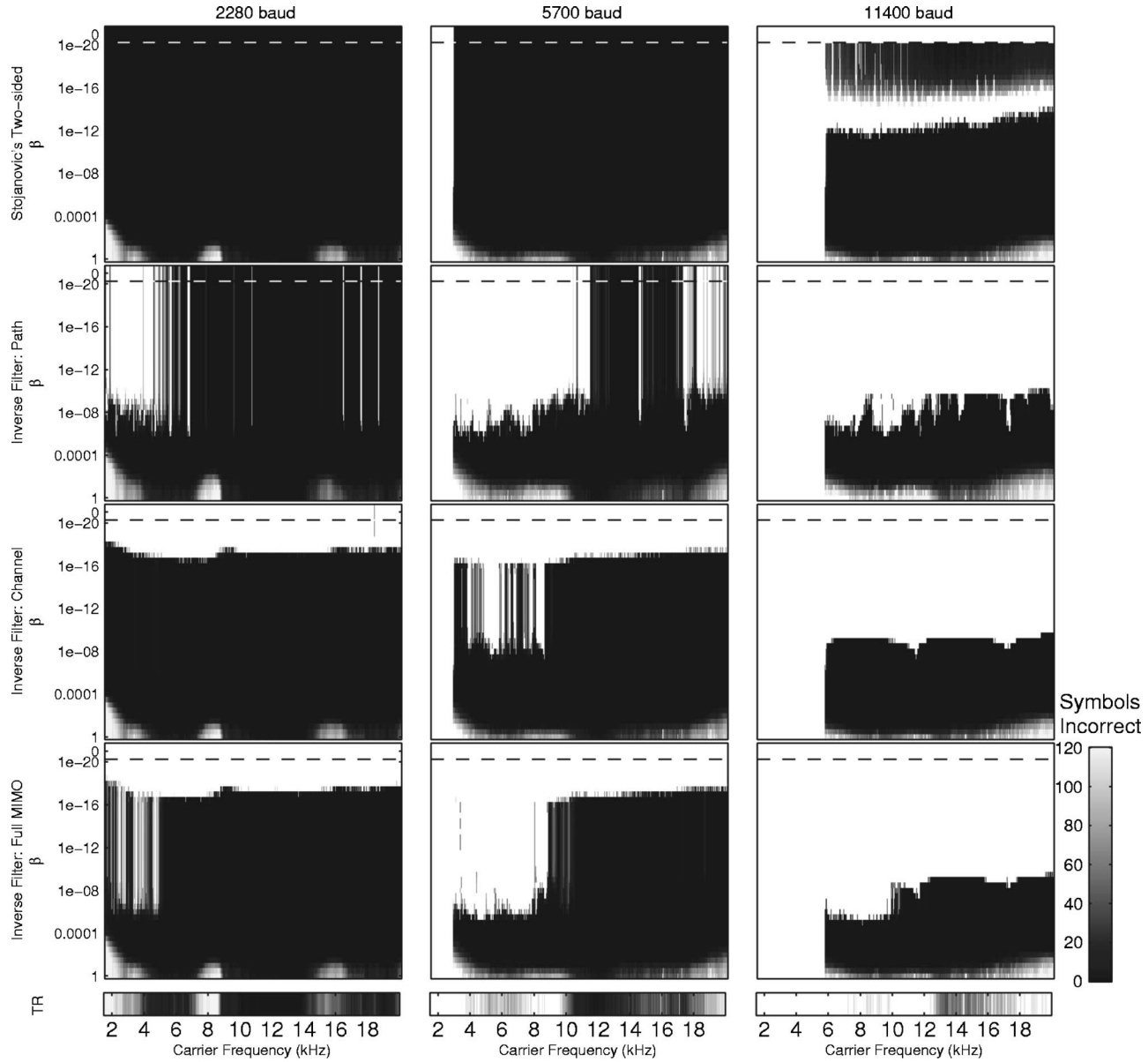


FIG. 9. Symbol error for the four filter designs operating with a carrier frequency ranging from 1.5 to 20 kHz at the symbol rates of 2280, 5700, and 11 400 baud, and with the regularization parameter, β, varying from 10^{-20} to 1. Results for $\beta=0$ are also included in the plots in the top row of pixels. Black indicates zero symbol error.

bols using 4-PSK signal space. These symbols were used to create a base-band signal using a raised cosine spectral filter.

2. The channel IRs were converted to base-band and used to calculate each inverse filter.
3. The base-band signal from step 1 was convolved with each filter to generate the signal to be transmitted. The response at receiver 1 was examined under two conditions: first, for a transmission that is intended to be received at receiver 1, and second for a transmission that is intended to be received at receiver 2. The first condition was used to measure the quality of the transmission, and the second condition to measure the level of cross-talk.
4. The transmission signals were globally normalized to have the same power. In a real-time situation, this would be accomplished by an automatic gain control at the output of the inverse filter.
5. The transmission signals were convolved with the channel responses.
6. The power of each received signal was measured to determine the strength of the signal that would be received for each filter design.
7. The signal was synchronized with the first peak of the training sequence, sampled at the symbol rate and passed into a detector.
8. The signal level and symbol error were measured.

IV. RESULTS

The experiment that was simulated consisted of utilizing six speakers on the transmitter array to transmit to two adjacent microphones at the receiver array. Two adjacent microphones where chosen to increase the cross-talk between the microphones to demonstrate the filter design performance.

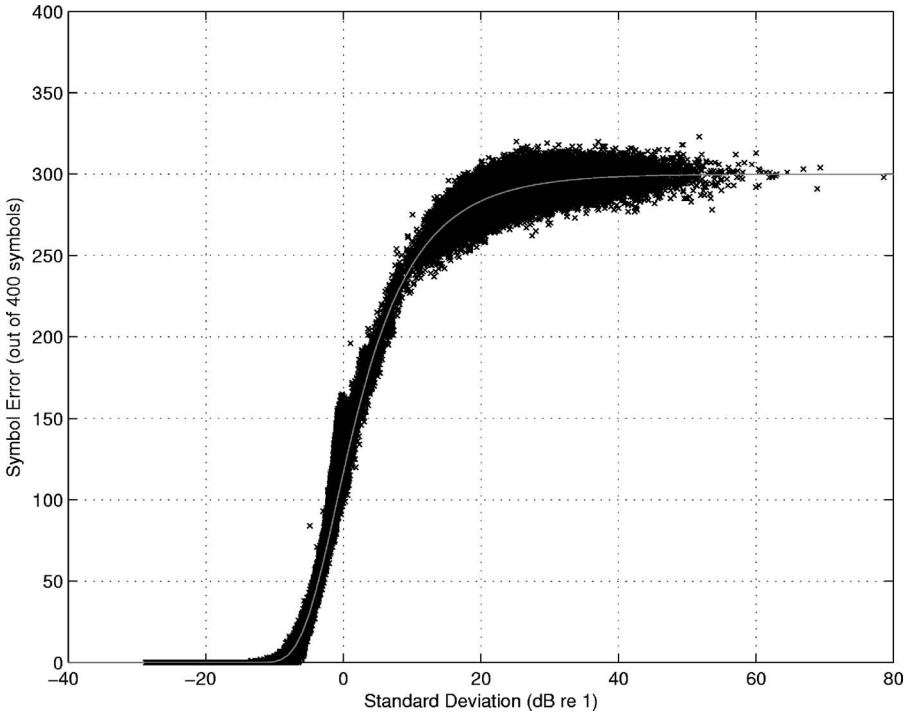


FIG. 10. Scatter of standard deviation versus symbol error for all filter designs. The light curve shows the expected average for a Gaussian distributed symbol spread.

The average frequency responses between the six transmitters and each of the two receivers are shown in Fig. 6. Receivers 1 and 2 are observed to have a fairly similar response. Arbitrarily taking −25 dB as an operational level, the channel was considered operable between approximately 1.5 and 20 kHz, although the response is observed to fluctuate significantly throughout this range.

Figure 7 shows the received power, Fig. 8 the relative cross-talk, and Fig. 9 the symbol error at receiver 1 for different symbol rates, 2280, 5700, and 11 400 baud, (being 25, 10, and 5 samples/symbol, respectively) and each of the filter designs. As time-reversal does not include a regularization parameter, it is included in a thin subplot at the bottom of all three figures. Results are shown for a range of regularization values (ordinate) and carrier frequencies (abscissa). The results for no regularization ($\beta=0$) are shown in the top row of pixels of each plot above the dashed line, while the other pixels range logarithmically from $\beta=10^{-20}$ to 10^{0}. In the simulation, the regularization parameter, β, was normalized against the peak value of the frequency response of the channel. The upper limit of 10^{0} for the regularization parameter, β, was chosen to significantly exceed the largest singular value of the channel responses. This value resulted in the filter effectively being a time-reversal filter. Carrier frequencies below half the symbol rate (Nyquist limit) are meaningless and have been masked out in the figures.

Figure 7 shows that the received power is greatest for the Stojanovic filter design while the *inverse by path* Tikhonov inverse filter is observed to have the least. This can be understood by the way the filter is designed. The filter design by Stojanovic consists of matched filters applied to each receiver followed by a summer that combines the outputs of these filters. The combined signal is passed through an inverse filter that compensates for the sum of these responses. The matched filter can be observed to maximize the signal-to-noise ratio, and the summation results in an averaging that causes the frequency response to be smoother and thus easier to invert. However, the Tikhonov inverse filter consists of an inverse that must compensate for responses that are less smooth, and thus more effort is used to smooth out the response, resulting in less receiver energy. However, as the regularization parameter increases the received power of all the inverse filter designs approach a similar magnitude to that achieved by the design of Stojanovic.

Figure 8 shows the relative cross-talk power (being the difference between the power of the signal for a transmission targeted at the target receiver, and the signal for a transmission signal targeted at the cross-talk receiver.) The filter design developed by Stojanovic generally has less cross-talk compared with the *inverse by path* and *inverse by channel* designs. However, the *inverse by full MIMO* design is observed to have over 20 dB less cross-talk compared with the filter designed by Stojanovic when the regularization is around 10^{-4}. While it is unclear why the design by Stojanovic has less cross-talk than the *inverse by path* and *inverse by channel* designs, the *inverse by full MIMO* can be expected to outperform all the other designs as this is the only design targeted at reducing the cross-talk.

Figure 9 shows that no symbol errors were detected for the Stojanovic filter design having regularization values between 0 to approximately 10^{-5} for 2280 and 5700 baud, and the regularization required for 11 400 baud is much smaller than that required for the Tikhonov inverse filter design. This observation can be attributed to the spectral averaging in the Stojanovic design. In order for a zero (or value close to) to occur within this average, all of the IRs must share a common zero. As the number of transmitters is increased, the probability of a common zero is reduced. At 11 400 baud, regularization was required for Stojanovic's filter design to achieve few symbol errors. Regularization was found to be

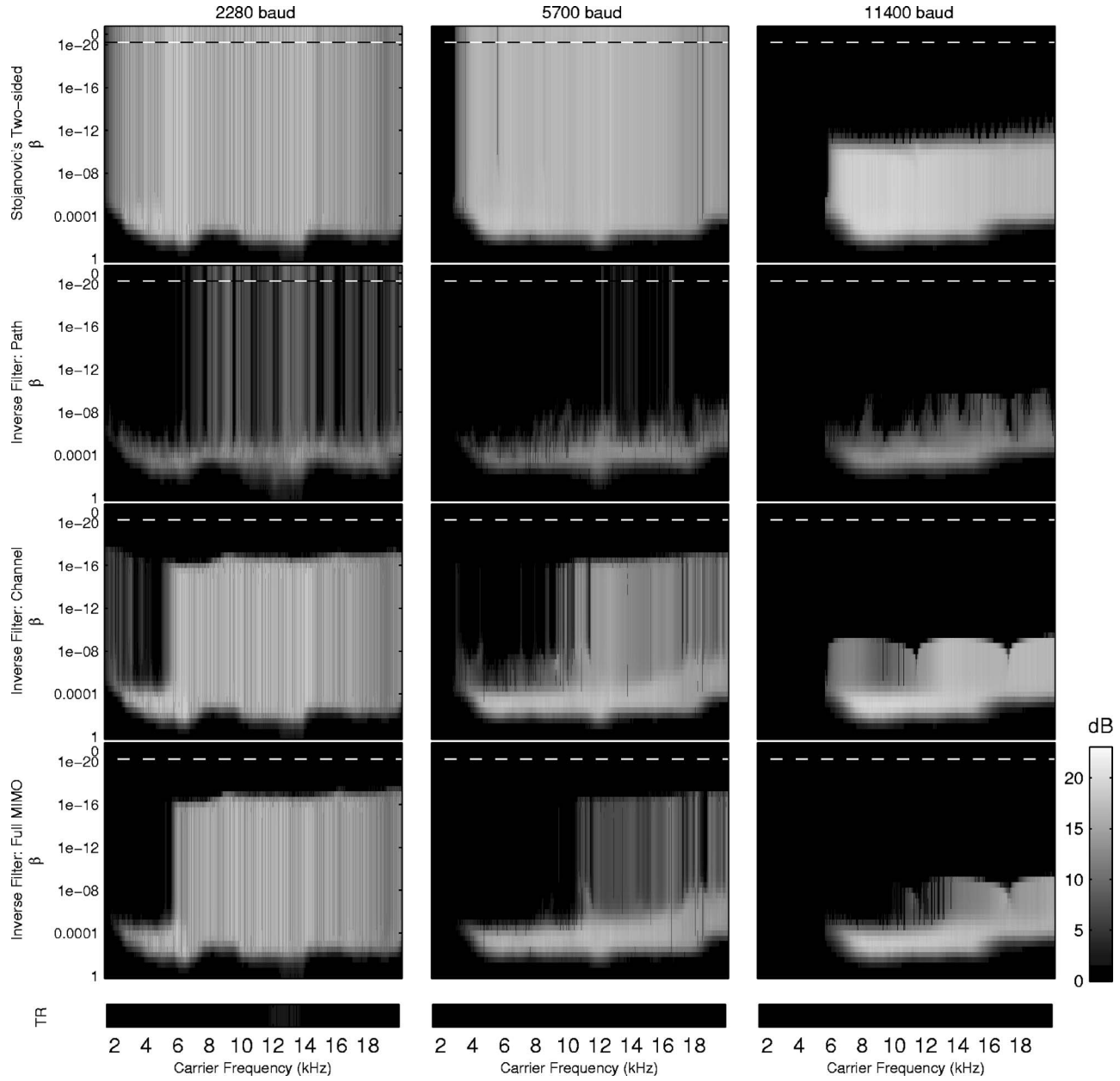


FIG. 11. Margin for increase in standard deviation for 1 in 400 probability of error for the four filter designs operating with a carrier frequency ranging from 1.5 to 20 kHz at the symbol rates of 2280, 5700, and 11 400 baud, and with the regularization parameter, β, varying from 10^{-20} to 1. Results for β=0 are also included in the plots in the top row of pixels.

required for this symbol rate because a common zero existed in the frequency responses of the channels. This common zero was the result of the filtering performed in the band-pass to base-band conversion.

In the simulations the number of symbol errors was determined without the addition of noise in the channel. The omission was so that the symbol errors would be the result of the ISI. If the symbols error resulting from ISI is distributed according to a Gaussian distribution, then the expected number of symbol errors when channel noise is included can be obtained by combining the standard deviation of the noise and the ISI. It should be noted that this approach is only possible because the filters are not adaptive and are calculated from an averaged channel response. When using adaptive filters, the noise causes the taps of the filters to fluctuate. For some adaptive algorithms, this fluctuation is equivalent to adding regularization (see, for example, Proakis, 2001), Eq. 10.2-33).

The ISI cannot generally be assumed Gaussian. However, if the ISI is smaller than the noise, then the error from assuming a Gaussian distribution is small (Shimbo and Celebiler, 1971). To verify that the ISI is Gaussian, a plot of the standard deviation versus the symbol error for all the simulations was plotted against those expected from a Gaussian distribution, and is shown in Fig. 10. The deviation between the Gaussian trend and the results from the simulation is considered small, and thus the Gaussian approximation was considered valid. The relationship between the probability of error for Gaussian interference having a standard deviation of σ for PSK is given by (Proakis, 2001)

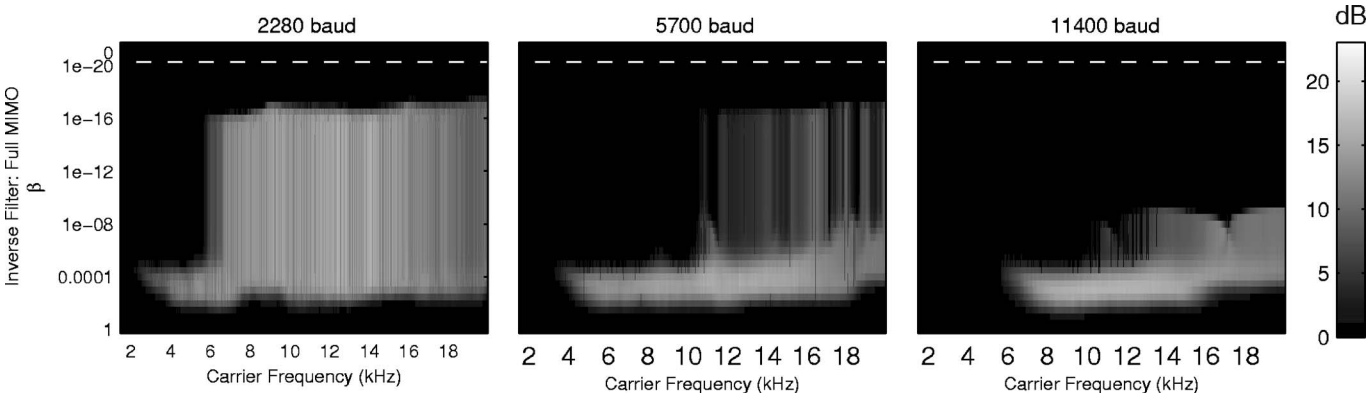


FIG. 12. Margin for increase in standard deviation for 1 in 400 probability of error when cross-talk is included for the filter design *inverse by full MIMO*. All the other filter designs exceeded the permissible level of standard deviation required to attain 1 in 400, and thus are not shown. Operation used a carrier frequency ranging from 1.5 to 20 kHz at the symbol rates of 2280, 5700, and 11 400 symbols/s, and with the regularization parameter, β, varying from 10^{-20} to 1. Results for $\beta=0$ are also included in the plots in the top row of pixels.

$$P_e = 1 - Q^2\left(-\frac{1}{\sigma}\right), \tag{18}$$

where $Q(x)=\frac{1}{\sqrt{2\pi}}\int_{t=x}^{\infty} e^{-\frac{t^2}{2}}dt$ is the MATLAB function, `qfunc`. The correlation between the measured data and the curve demonstrates that the intersymbol interference for these channels can be considered Gaussian. The graph of $400P_e$ approaches a limit of 300 symbol errors. This limit can be explained by considering that as four types of symbols are detectable, each with equal probability, then the probability of a correct symbol is $P_c=1/4$, and the probability of an incorrect symbol is $P_e=(1-P_c)=3/4$, leading to $400P_e=300$.

The margin of noise that may be added before attaining a probability of 1 in 400 chance of error is shown in Fig. 11. The design by Stojanovic is observed to have the greatest margin. However, both the *inverse by channel* and *inverse by full MIMO* Tikhonov inverse filter designs can be seen to have similar margin if an appropriate regularization value is used. Of the Tikhonov designs, the *inverse by path* design is observed to be the least resilient to noise, while the *inverse by channel* filter is the most resilient to noise. This can be explained by the fact that the *inverse by path* filter uses most of the energy to compensate for nulls in each channel rather than for a combination of channel, while the *inverse by full MIMO* filter uses its energy to reduce the cross-talk to the other receivers.

An observation that can be made from the results shown in Figs. 7–9 and 11 is that as β becomes larger (relative to the largest singular value) all the solutions are observed to approach the time-reversal results (shown in the thin lower plots). Time reversal is observed to have the worst performance with respect to both the number of symbol errors and the noise margin.

A number of decision feedback equalizers (DFEs) were implemented at the receiver having both feed-forward and feed-back taps that span 72 ms (twice the length of the channel IRs.) The responses were passed through the DFEs multiple times with a decrease in the step size to obtain the best possible adaption. The DFE were able to reduce the spread of the symbols, but the differences with and without the DFE were not noticeably enough to warrant inclusion of the results in this paper. In particular, the incorporation of the DFE in the TR transmission did not improve the symbol error to the same degree as that of using an inverse filter design. However, increasing the number of transmitted symbols used to train the filters might be found to improve the results.

From the figures, it can be observed that the *inverse by path* does not require any regularization at low symbol rates. This result my be explained by examining an alternative expression for inverse filter given by

$$\mathbf{H}(\omega) = \frac{\mathrm{adj}(\mathbf{C}^{\mathrm{H}}(\omega)\mathbf{C}(\omega)+\beta\mathbf{I})\mathbf{C}^{\mathrm{H}}(\omega)}{\det(\mathbf{C}^{\mathrm{H}}(\omega)\mathbf{C}(\omega)+\beta\mathbf{I})}A(\omega). \tag{19}$$

In the *inverse by path* design, the subfilter for each path is designed independently and the denominator is independent and unique. If there is a single path with a response that is difficult to invert, then only the inverse filter for that path becomes nonfunctional, while the other filters can still operate. However, for the *inverse by channel* and the *inverse by full MIMO* Tikhonov designs, the determinant of the inverse filter is common to a number of the subfilters. When the determinant in Eq. (19) contains a zero, then all the subfilters with this common determinant become unstable.

Cazzolato *et al.* (2001) showed that the focal region of inverse filter was smaller than that for time-reversal. An increase of the distance between the receiver elements could reduce the cross-talk and result in better performance of the time-reversal filter. However, increasing the distance between the receivers would also improve the performance of the *inverse by full MIMO* filter as less effort would be required to eliminate the cross-talk.

Of interest for high speed communications systems is multi-channel transmission. Although the symbol error was not recorded for the simultaneous transmission of two streams, the expected symbol error for dual transmission was considered to be estimated by combining the standard deviation of the received signal with that obtained when transmitting to the cross-talk target. The only filter design that is able to achieve a 1 in 400 probability of error when the standard deviations where combined was the *inverse by full MIMO* Tikhonov filter design. The margin for the standard deviation

for this filter is shown in Fig. 12. The Tikhonov inverse filter is observed to operate best when the regularization is around 10^{-3} for lower carrier and 10^{-4} for higher carrier frequencies. At low symbol rates, the range of reasonable regularization is large, however for faster data rates, the range is much smaller. This can be attributed to the bandwidth that the signal occupies. When the data rate is increased, the bandwidth is increased, and the number of zero's likely to be in the bandwidth increases, resulting in reducing the range of regularization values.

V. CONCLUSION

Three classifications of channel filter design have been discussed, *inverse by path*, *inverse by channel* and *inverse by full MIMO*. Tikhonov regularized inverse filters were implemented according to these classifications and compared with the time-reversal filter and a filter design proposed by Stojanovic (2005). While Stojanovic (2005) presented theoretical results, the filter was shown to be practically implementable by modifying the design to include a regularization parameter. The filter design by Stojanovic outperforms the Tikhonov regularized inverse filter designs when communicating over a single channel. The *inverse by path* and *inverse by channel* performed particularly poorly when compared to the design by Stojanovic, while the *inverse by full MIMO* design was found to have only slightly reduced performance. While the designs by Stojanovic and *inverse by path*, and *inverse by channel* Tikhonov inverse filter designs are not designed for MIMO communications; they were, however, found to reduce the cross-talk. However, the performance of these designs for MIMO communication was found to be poor when compared to the *inverse by MIMO* filter. For the simulation, the *inverse by MIMO* was found to provide 20 dB less cross-talk at the expense of around 2 dB loss in signal strength when compared to the filter design by Stojanovic. In this scenario, the *inverse by MIMO* Tikhonov filter design was the only design that was able to be used to transmit multiple transmission streams.

Candy, J., Poggio, A., Chambers, D., Guidry, B., Robbins, C., and Kent, C. (**2005**). "Multi-channel time-reversal processing for acoustic communications in a highly reverberant environment," J. Acoust. Soc. Am. **118**, 2339–2354.

Cazzolato, B., Nelson, P., Joseph, P., and Brind, R. (**2001**). "Numerical simulation of optimal deconvolution in a shallow-water environment," J. Acoust. Soc. Am. **110**, 170–185.

Clay, C. (**1966**). "Waveguides, arrays, and filters," Geophysics **31**, 501–505.

Dowling, D. (**1994**). "Acoustic pulse compression by passive phase-conjugation," J. Acoust. Soc. Am. **95**, 1450–1458.

Dumuid, P., Cazzolato, B., and Zander, A. (**2006**). "Transducer sensitivity compensation using diagonal preconditioning for time-reversal and Tikhonov inverse filtering in acoustic systems," J. Acoust. Soc. Am. **119**, 372–381.

Edelmann, G. (**2005**). "An overview of time-reversal acoustic communications," *TICA '05*, Istanbul, Turkey.

Edelmann, G., Akal, T., Hodgkiss, W., Kim, S., Kuperman, W., and Song, H. (**2002**). "An initial demonstration of underwater acoustic communication using time-reversal," IEEE J. Ocean. Eng. **27**, 602–609.

Edelmann, G., Hodgkiss, W., Kim, S., Kuperman, W., and Song, H. (**2001**). "Underwater acoustic communication using time-reversal," in *MTS/IEEE OCEANS, 2001 Conference*, Vol. **4**, 2231–2235.

Edelmann, G., Song, H., Kim, S., Hodgkiss, W., Kuperman, W., and Akal, T. (**2005**). "Underwater acoustic communications using time-reversal," IEEE J. Ocean. Eng. **30**, 852–864.

Flynn, J., Ritcey, J., Fox, W., and Rouseff, D. (**2004**). "Performance predictions of acoustic communications by decision-directed passive phase conjugation," Technical Report No. UWEETR-2004-0003, University of Washington, Department of Electrical Engineering, Seattle, Washington.

Flynn, J., Ritcey, J., Rouseff, D., and Fox, W. (**2004**). "Multichannel equalization by decision-directed passive phase conjugation: Experimental results," IEEE J. Ocean. Eng. **29**, 824–836.

Higley, W., Roux, P., and Kuperman, W. (**2006**). "Relationship between time-reversal and linear equalization in digital communications (1)," J. Acoust. Soc. Am. **120**, 35–37.

Jackson, D., and Dowling, D. (**1991**). "Phase conjugation in underwater acoustics," J. Acoust. Soc. Am. **89**, 171–181.

Kilfoyle, D., and Baggeroer, A. (**2000**). "The state of the art in underwater acoustic telemetry," IEEE J. Ocean. Eng. **25**, 4–27.

Kirkeby, O., Nelson, P., Hamada, H., and Orduña Bustamante, F. (**1998**). "Fast deconvolution of multichannel systems using regularization," IEEE Trans. Speech Audio Process. **6**, 189–195.

Montaldo, G., Tanter, M., and Fink, M. (**2004**). "Real time inverse filter focusing through iterative time-reversal," J. Acoust. Soc. Am. **115**, 768–775.

Nelson, P., Orduña Bustamante, F., and Hamada, H. (**1995**). "Inverse filter design and equalization zones in multichannel sound reproduction," IEEE Trans. Speech Audio Process. **3**, 185–192.

Parvulescu, A. (**1995**). "Matched-signal ("MESS") processing by the ocean," J. Acoust. Soc. Am. **98**, 943–960.

Parvulescu, A., and Clay, C. (**1965**). "Reproducibility of signal transmission in the ocean," Radio Electron. Eng. **29**, 223–228.

Proakis, J. (**2001**). *Digital Communications, McGraw-Hill Series in Electrical and Computer Engineering*, 4th ed. (McGraw-Hill, New York).

Rouseff, D. (**2005**). "Intersymbol interference in underwater acoustic communications using time-reversal signal processing," J. Acoust. Soc. Am. **117**, 780–788.

Shimbo, O., and Celebiler, M. (**1971**). "The probability of error due to intersymbol interference and Gaussian noise in digital communication systems," IEEE Trans. Commun. Technol. **19**, 113–119.

Song, H., Hodgkiss, W., Kuperman, W., Stevenson, M., and Akal, T. (**2006**). "Improvement of time-reversal communications using adaptive channel equalizers," IEEE J. Ocean. Eng. **31**, 487–496.

Song, H., Roux, P., Hodgkiss, W., Kuperman, W., Akal, T., and Stevenson, M. (**2006**). "Multiple input/multiple output coherent time-reversal communications in a shallow water acoustic-channel," IEEE J. Ocean. Eng. **31**, 170–178.

Stojanovic, M. (**2005**). "Retrofocusing techniques for high rate acoustic communications," J. Acoust. Soc. Am. **117**, 1173–1185.

Stojanovic, M., Catipovic, J., and Proakis, J. (**1993**). "Adaptive multichannel combining and equalization for underwater acoustic communications," J. Acoust. Soc. Am. **94**, 1621–1631.

Yon, S., Tanter, M., and Fink, M. (**2003**). "Sound focusing in rooms: The time-reversal approach," J. Acoust. Soc. Am. **113**, 1533–1543.

Guided wave arrays for high resolution inspection

Alexander Velichko[a] and Paul D. Wilcox[b]
Department of Mechanical Engineering, University of Bristol, Bristol, BS8 1TR, United Kingdom

(Received 26 April 2007; revised 4 October 2007; accepted 5 October 2007)

The paper describes a general approach for processing data from a guided wave transducer array on a plate-like structure. The raw data set from such an array contains time-domain signals from each transmitter–receiver combination. The technique is based on linear superposition of signals in the frequency domain with some amplitude and phase factors and can be applied to any array geometry and any types of array elements. The problem of finding optimal coefficients, which allow the best resolution to be achieved with the minimum number of array elements, is investigated. It is shown that improvements in resolution are obtained at the expense of sensitivity to noise. A method of quantifying this sensitivity is presented. Results are shown that illustrate the application of the technique to a linear array and an array of circular geometry (containing a single ring of elements). Experimental data obtained from a guided wave array containing electromagnetic acoustic transducer elements for exciting and detecting the *S*0 Lamb wave mode in a 5-mm-thick aluminum plate are processed with different algorithms and the results are discussed. Generalization of the technique for the case of multimode media is suggested.
[DOI: 10.1121/1.2804699]

PACS number(s): 43.60.Fg, 43.20.Bi, 43.35.Cg, 43.35.Zc [TDM] Pages: 186–196

I. INTRODUCTION

Ultrasonic guided waves are widely used in many areas of nondestructive evaluation.[1–4] In this paper the problem of synthesizing a high resolution image of reflectors in a plate-like structure with a guided wave array of arbitrary shape is considered. The complete raw data set of signals from every transmitter–receiver combination is collected and then post-processed. The general approach is to multiply the transmitted and received signals by suitable amplitude and phase factors and add them together in order to focus the beam on every point within the test structure. The algorithms differ in the way in which these phase coefficients are calculated.

One technique is based on applying such phase shifts that all signals have equal phases at the focusing point. For an ultrasonic linear array this method is known as the total focusing method.[5] In the current paper this method is referred to as the basic-phased addition algorithm.[4]

In principle the basic-phased addition algorithm can be applied to any array geometry, but its performance is good only for a limited number of cases. For example, it produces acceptable results for a linear array,[5] but in the case of a circular array a large number of elements is required to obtain acceptable resolution.[4]

This paper is concerned with finding optimal phase coefficients, which allow the best resolution to be achieved for any array geometry with the minimum number of array elements. There is no requirement for the array elements to be omnidirectional, i.e., to have equal transmission and reception sensitivity in all directions.

In order to obtain better resolution and higher sensitivity it is necessary to use higher order guided wave modes at higher frequencies, where multiple modes exist. Generalization of the above-mentioned algorithms for the case of multimode media is suggested.

II. PRELIMINARIES

A. Array and data acquisition

An ultrasonic guided wave array system on an isotropic plate structure is considered. The array elements behave as point sources, some or all of them may be transmitters and some or all of them may be receivers. The time traces from each transmitter–receiver combination are collected and converted to the frequency, ω, domain. Let n_T elements located at the points $\mathbf{r}_{(T)i}$, $i=1,\ldots,n_T$, function as transmitters and n_R elements located at the points $\mathbf{r}_{(R)i}$, $i=1,\ldots,n_R$, function as receivers. The complex spectrum of the nth signal for the ith transmitter and jth receiver is denoted by $s_n(\omega)$, $n=1,\ldots,N$, where $N=n_T n_R$ is the total number of transmitter receiver pairs in the array.

The wave field from any transmitter propagates into the plate, interacts with the scatterers, and generates a scattered wave field. It is initially assumed that there is only one wave mode. The wave field in the far zone from a point source located at the point $\mathbf{r}_0$ has the form:

$$u = b(\mathbf{n},\omega)\frac{e^{ikR}}{\sqrt{R}},$$

$$\mathbf{R} = \mathbf{r} - \mathbf{r}_0, \quad R = |\mathbf{R}|, \quad \mathbf{n} = \frac{\mathbf{R}}{R}, \tag{1}$$

where $k=2\pi/\lambda$ is the wave number, λ is the wavelength, and function $b(\mathbf{n},\omega)$ describes the angular directivity of the source.

To model the scattered wave field the Born approximation is used. This means that the reflectors are assumed to act

[a]Electronic mail: a.velichko@bristol.ac.uk
[b]Electronic mail: p.wilcox@bristol.ac.uk

as point scatterers to the incident wave field and multiple scattering between them is ignored. Therefore, the scattered signals $s_n(\omega)$ can be written in the following form:

$$s_n(\omega) = \int B(\mathbf{r}, \mathbf{n}_{(T)i}, \mathbf{n}_{(R)j}, \omega) G_n(\mathbf{r}, \omega) d\mathbf{r}. \tag{2}$$

The function B in this expression is the density of distribution of scatterer amplitude at the point $\mathbf{r}$ and in general case depends on the direction $\mathbf{n}_{(T)i}$ of the incoming wave and direction $\mathbf{n}_{(R)j}$ of the reflected wave. The function $G_n(\mathbf{r}, \omega)$ is the Green's function (the scattered signal from a point reflector) for the nth transmitter–receiver pair and has the form:

$$G_n(\mathbf{r}, \omega) = u_{(T)i} u_{(R)j},$$

$$u_{(T)i} = b_{(T)i}(\mathbf{n}_{(T)i}, \omega) \frac{e^{ikR_{(T)i}}}{\sqrt{R_{(T)i}}},$$

$$u_{(R)j} = b_{(R)j}(\mathbf{n}_{(R)j}, \omega) \frac{e^{ikR_{(R)j}}}{\sqrt{R_{(R)j}}}, \tag{3}$$

where $R_{(T)i} = |\mathbf{r} - \mathbf{r}_{(T)i}|$, $R_{(R)j} = |\mathbf{r} - \mathbf{r}_{(R)j}|$.

The objective of processing the array data $s_n(\omega)$ is to obtain an estimation of function B from the system (2).

B. Far field approximation

The polar coordinate system r, φ, where r and φ represent, respectively, radial and angular position, is defined with its origin at the nominal center of the array. If all reflectors are situated in the far field of the whole array, $|\mathbf{r}_{(T)i}| \ll r$, $|\mathbf{r}_{(R)i}| \ll r$, then

$$R_{(T)i} \approx r - \mathbf{r}_{(T)i}\mathbf{n}, \quad R_{(R)j} \approx r - \mathbf{r}_{(R)j}\mathbf{n}. \tag{4}$$

Here $\mathbf{n}$ is the unity vector $\mathbf{n} = \{\cos\varphi, \sin\varphi\}$. In the far field of the array the directivity functions of the array elements, $b_{(T)i}(\mathbf{n}_{(T)i}, \omega)$ and $b_{(R)j}(\mathbf{n}_{(R)j}, \omega)$, can be replaced by $b_{(T)i}(\mathbf{n}, \omega) \equiv b_{(T)i}(\varphi, \omega)$ and $b_{(R)j}(\mathbf{n}, \omega) \equiv b_{(R)j}(\varphi, \omega)$. Under such assumptions

$$u_{(T)i} = K_{(T)i} \frac{e^{ikr}}{\sqrt{r}}, \quad u_{(R)j} = K_{(R)j} \frac{e^{ikr}}{\sqrt{r}},$$

$$K_{(T)i} = b_{(T)i}(\varphi, \omega) e^{-ik\mathbf{r}_{(T)i}\mathbf{n}},$$

$$K_{(R)j} = b_{(R)j}(\varphi, \omega) e^{-ik\mathbf{r}_{(R)j}\mathbf{n}}. \tag{5}$$

In the far field of the array the function B depends only on the location of the reflector r, φ, and frequency ω, $B \equiv B(r, \varphi, \omega)$. Then integral (2) can be written as

$$s_n(\omega) = \int \int B(r, \varphi, \omega) G_n(r, \varphi, \omega) r dr d\varphi. \tag{6}$$

Substitution of expressions (5) into formula (3) gives the following far field approximation for the Green's function G_n:

$$G_n(r, \varphi, \omega) = K_n(\varphi, \omega) \frac{e^{i2kr}}{r}, \quad K_n = K_{(T)i} K_{(R)j}, \tag{7}$$

and the function K_n can be written as

$$K_n(\varphi, \omega) = b_{(T)i}(\varphi, \omega) b_{(R)j}(\varphi, \omega) e^{-ik\mathbf{r}_n\mathbf{n}},$$

$$\mathbf{r}_n = \mathbf{r}_{(T)i} + \mathbf{r}_{(R)j}. \tag{8}$$

Thus in far field of the array the Green's function G_n is the multiplication of two terms. The first term, $K_n(\varphi, \omega)$, depends on index n of transmitter–receiver combination and direction φ. The second term, e^{i2kr}/r, depends only on radial distance r. Therefore, Eq. (6) can be written as two equations:

$$s_n(\omega) = \int_0^{2\pi} C(\varphi, \omega) K_n(\varphi, \omega) d\varphi, \tag{9}$$

$$C(\varphi, \omega) = \int B(r, \varphi, \omega) e^{i2kr} dr. \tag{10}$$

The processing algorithm can be divided into two parts. The first part is the determination of the function $C(\varphi, \omega)$ from the system of equations (9) and associated with angular resolution. From Eq. (10) it is seen that the function $C(\varphi, \omega)$ represents the reflected signal from the φ direction and the total signal s_n is the sum of these signals for all possible directions. The function K_n in this case is an angular Green's function for nth transmitter–receiver pair.

The second part is the determination of the function $B(r, \varphi)$ as a function of the radial distance r for each direction φ from Eq. (10). This step is associated with radial resolution.

Therefore, the angular resolution is the extraction of a reflected signal from the data set $s_n(\omega)$ for each direction φ and the radial resolution is the mapping of this signal to propagation distance r. Such division of the whole problem (2) into the angular resolution (9) and the radial resolution (10) is possible only in the far field to the array. In the near field to the array the approximations (4) are not valid and the Green's function G_n cannot be split into angular and radial parts.

C. Phased methods

The general idea of angular resolution is to multiply the signals $s_n(\omega)$ by suitable phase factors $t_n(\varphi_0, \omega)$ in order to focus the beam in each direction φ_0:

$$C^{(1)}(\varphi_0, \omega) = \sum_n t_n(\varphi_0, \omega) s_n(\omega), \tag{11}$$

here $C^{(1)}(\varphi_0, \omega)$ is the approximation to the function $C(\varphi_0, \omega)$. Using Eq. (9), the functions $C^{(1)}(\varphi_0, \omega)$ and $C(\varphi, \omega)$ can be related by

$$C^{(1)}(\varphi_0, \omega) = \int_0^{2\pi} P(\varphi_0, \varphi, \omega) C(\varphi, \omega) d\varphi, \tag{12}$$

$$P(\varphi_0, \varphi, \omega) = \sum_n t_n(\varphi_0, \omega) K_n(\varphi, \omega). \tag{13}$$

The function $P(\varphi_0,\varphi,\omega)$ is called the point spread function (PSF) and is the result for a point reflector located at the direction φ. For ideal resolution (i.e., $C^{(1)}=C$) the PSF is a delta function $\delta(\varphi-\varphi_0)$, which has infinitesimal width and no sidelobes. In practice the phase coefficients $\mathbf{t}(\varphi_0,\omega)=\{t_1,\ldots,t_N\}^T$ need to be chosen to make PSF "close" to the delta function. The processing algorithms differ in the way in which the phase coefficients are calculated.

For radial resolution the dispersion compensation algorithm is used. The resulting image, $I(r_0,\varphi_0)$, is given by

$$I(r_0,\varphi_0)=\int_{-\infty}^{+\infty} C^{(1)}(\varphi_0,\omega)e^{-i2k(\omega)r_0}d\omega. \tag{14}$$

The details of dispersion compensation method and its numerical implementation can be found in Ref. 6.

III. ANGULAR RESOLUTION

From comparison of expressions (8) and (5) it can be seen that the angular Green's function K_n for nth transmitter–receiver pair has the same form as the angular Green's function for the point transmitter located at the point $\mathbf{r}_n$. Consider the field radiated by an array of N transmitting elements. If each of the transmitters in such an array has an amplitude t_n then the total transmitted wave field will be

$$u(\varphi,r)=P(\varphi)\frac{e^{ikr}}{\sqrt{r}}, \tag{15}$$

where the function $P(\varphi)$ characterizes the angular pattern of the wave field and is defined by the formula (13) for the point spread function.

Therefore the angular resolution is equivalent to the problem of focusing of signal by a system of transmitters located at the points $\mathbf{r}_n=\mathbf{r}_{(T)i}+\mathbf{r}_{(R)j}$ and different methods of pattern synthesis for phased arrays can be applied.[7] For example, for a linear array of elements, where each element acts as transmitter and receiver, the equivalent transmitter array is a linear array of a twice bigger aperture. For a circular array with transducers around the perimeter only, the equivalent transmitter array consists of a circular area of twice the diameter densely populated with elements. The array layouts for this case are shown in Fig. 1.

A. Basic-phased addition method

The most physically obvious approach to computing the phase coefficients, t_n, is to add such phase shifts to each signal K_n that all signals in the steering direction φ_0 have equal phases. It is analogous to the delaying or advancing of t domain signals that is used in conventional synthetic aperture focusing technique, but with the added effect of taking the dispersive nature of guided waves into account.[3,4] From Eq. (8) it is seen that these phase shifts are $\Phi_n=k\mathbf{r}_n\mathbf{n}_0$, $\mathbf{n}_0=\{\cos\varphi_0,\sin\varphi_0\}$.

This method can also be formulated in a different way. Namely, the coefficients $t_n(\varphi_0)$ are chosen to maximize the PSF at the steering direction φ_0:

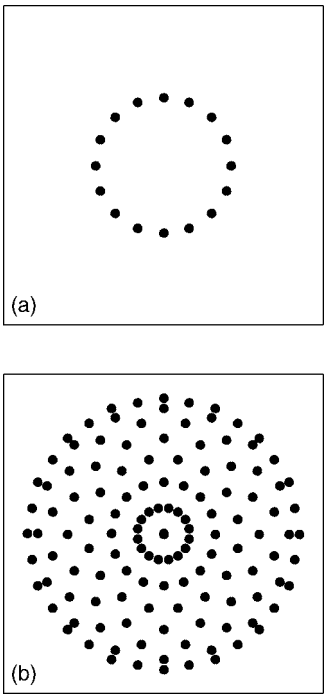


FIG. 1. Layout of circular array with elements acting as transmitters and receivers (a) and equivalent array of transmitters (b).

$$\frac{|P(\varphi_0,\varphi_0)|^2}{|\mathbf{t}(\varphi_0)|^2}\to\max. \tag{16}$$

The term $|\mathbf{t}(\varphi_0)|^2$ in the denominator of this expression represents the restriction on the variation range of the coefficients $\mathbf{t}$. Otherwise the condition $|P(\varphi_0,\varphi_0)|\to\max$ does not give the coefficients $t_n(\varphi_0)$ because $|P(\varphi_0,\varphi_0)|\to\infty$ as $|\mathbf{t}|\to\infty$.

B. Maximization of contrast method

An alternative way of calculating the coefficients $\mathbf{t}(\varphi_0)$ is to maximize the amplitude of the PSF at the focusing direction φ_0 relative to the amplitude of the PSF at all other directions:

$$\frac{|P(\varphi_0,\varphi_0)|^2}{\int_0^T |P(\varphi_0,\varphi)|^2 d\varphi}\to\max. \tag{17}$$

Here it is assumed that the range of azimuthal angle φ is $0\leqslant\varphi\leqslant T$.

In the theory of antenna arrays this method is also known as gain optimization.[7]

The problems (16) and (17) can be written in general form:

$$\mu(\mathbf{t})=\frac{|P(\varphi_0,\varphi_0)|^2}{\alpha|\mathbf{t}(\varphi_0)|^2+\beta\int_0^T |P(\varphi_0,\varphi)|^2 d\varphi}\to\max, \tag{18}$$

where $\alpha,\beta\geqslant 0$ are some parameters.

Using expression (13) the function $\mu(\mathbf{t})$ can be written as

$$\mu(\mathbf{t}) = \frac{\mathbf{t}^{*T}\mathbf{A}\mathbf{t}}{\mathbf{t}^{*T}\mathbf{B}\mathbf{t}}, \tag{19}$$

where an asterisk denotes complex conjugation, and elements of matrices $\mathbf{A}$ and $\mathbf{B}$ have the form

$$A_{ij} = K_i^*(\varphi_0)K_j(\varphi_0),$$

$$B_{ij} = \alpha\delta_{ij} + \beta\int_0^T K_i^*(\varphi)K_j(\varphi)d\varphi. \tag{20}$$

Therefore the numerator and denominator of the function $\mu(\mathbf{t})$ are quadratic forms of the vector $\mathbf{t}$ and the matrix $\mathbf{B}$ is positive definite. From the theory of quadratic form it follows that the extrema vector $\mathbf{t}$ is defined by the eigenvalue problem

$$\mathbf{A}\mathbf{t} = \mu\mathbf{B}\mathbf{t}, \tag{21}$$

and corresponds to the maximum eigenvalue μ_0. Substitution of expression (20) for the matrix $\mathbf{A}$ into Eq. (21) gives

$$\mathbf{K}_0^* P(\varphi_0, \varphi_0) = \mu_0\mathbf{B}\mathbf{t}(\varphi_0),$$

where vector $\mathbf{K}_0$ is $\mathbf{K}_0 = \{K_1(\varphi_0), \ldots, K_N(\varphi_0)\}^T$. Therefore, the optimal coefficients are

$$\mathbf{t}(\varphi_0) = c_0\mathbf{B}^{-1}\mathbf{K}_0^*. \tag{22}$$

It is seen that optimal coefficients are scaled by arbitrary constant c_0. From Eq. (11) the approximate solution $C^{(1)}\times(\varphi_0)$ for the method of maximization of contrast can be written in matrix form as

$$C^{(1)}(\varphi_0) = c_0\mathbf{K}_0^{*T}(\mathbf{B}^{-1})^T\mathbf{s}, \quad \mathbf{s} = \{s_1, \ldots, s_N\}^T. \tag{23}$$

If parameter β in expression (20) for the matrix $\mathbf{B}$ is zero, then matrix $\mathbf{B}$ is the identity matrix and formula (22) gives coefficients for the basic phased addition method, i.e., $\mathbf{t}(\varphi_0) = c_0\mathbf{K}_0^*$.

If the function $B(r, \varphi)$ is bounded then the normalization factor c_0 is defined from the condition

$$\int_0^T P(\varphi_0, \varphi)d\varphi = 1. \tag{24}$$

If distribution of reflectors is discrete then function $B(r, \varphi)$ is the sum of delta functions $a\delta(\varphi-\varphi_0)\delta(r-r_0)r^{-1}$, where a is the amplitude of the reflector. In this case an amplitude a instead of density B needs to be estimated and the normalization factor c_0 is defined from the condition

$$P(\varphi_0, \varphi_0) = 1. \tag{25}$$

Furthermore, the resulting image $I(r_0, \varphi_0)$ obtained by Eq. (14) must be normalized by the factor r_0, i.e., the resulting image is $r_0 I(r_0, \varphi_0)$.

C. Direct solution to the angular resolution problem

An alternative method of solving the angular resolution problem given by Eq. (9) is to solve it directly by using regularization methods. By replacing the integral in Eq. (9) by a sum at discrete points $\varphi_m = (m-1)\Delta\varphi, m = 1, \ldots, M$, an $N\times M$ system of linear equations is defined:

$$\Delta\varphi\mathbf{K}\mathbf{C} = \mathbf{s}, \tag{26}$$

where elements of the matrix $\mathbf{K}$ are $K_{nm} = K_n(\varphi_m)$ and $\mathbf{C} = \{C(\varphi_1), \ldots, C(\varphi_M)\}^T$.

To solve the system (26) the Tikhonov regularization method[8] can be used and the approximation, $\mathbf{C}^{(1)}$, to the solution $\mathbf{C}$ satisfies the minimization problem

$$|\Delta\varphi\mathbf{K}\mathbf{C}^{(1)} - \mathbf{s}|^2 + \gamma|\mathbf{C}^{(1)}|^2 \to \min, \tag{27}$$

where γ is a regularization parameter.

Condition (27) gives the following expression for the approximate solution:

$$\mathbf{C}^{(1)} = (\gamma\mathbf{E} + \Delta\varphi\mathbf{K}^+\mathbf{K})^{-1}\mathbf{K}^+\mathbf{s}, \tag{28}$$

where $\mathbf{K}^+ = \mathbf{K}^{*T}$. Performing direct matrix multiplication it can be seen that

$$(\gamma\mathbf{E} + \Delta\varphi\mathbf{K}^+\mathbf{K})^{-1}\mathbf{K}^+ = \mathbf{K}^+(\gamma\mathbf{E} + \Delta\varphi\mathbf{K}\mathbf{K}^+)^{-1}. \tag{29}$$

Therefore the mth component $C_m^{(1)} \equiv C^{(1)}(\varphi_m)$ of vector $\mathbf{C}^{(1)}$ is given by

$$C^{(1)}(\varphi_m) = \mathbf{K}_m^{*T}(\gamma\mathbf{E} + \Delta\varphi\mathbf{K}\mathbf{K}^+)^{-1}\mathbf{s}. \tag{30}$$

Matrix $\Delta\varphi\mathbf{K}\mathbf{K}^+$ is the approximation to the integral $\int_0^T K_i(\varphi)K_j^*(\varphi)d\varphi$. From Eq. (20) it follows that formula (30) gives the same solution as the maximization of contrast method (23) with $\alpha = \gamma$, $\beta = 1$, and $c_0 = 1$. In other words, direct solution by regularization is exactly equivalent to the maximization of contrast method.

D. Stability of angular resolution methods

In practice it is important for the resolution methods to be stable with respect to errors in initial data. Stability means that small errors in the initial data, s_n, cause small errors in the solution $C^{(1)}$ given by Eq. (11).

For a point reflector the model (7) is used and the angular resolution in this case is given by the point spread function (13). If all array elements are omnidirectional then the angular Green's function $K_n = e^{-ik\mathbf{r}_n\mathbf{n}}$. Consider the angular Green's function with amplitude and phase errors q_n and ϵ_n:

$$\tilde{K}_n = (1 + q_n)e^{i\epsilon_n}K_n. \tag{31}$$

The errors q_n and ϵ_n are all assumed to be random, independent, and equally distributed within intervals $\pm\Delta q$ and $\pm\Delta\epsilon$ with probability densities $1/(2\Delta q)$ and $1/(2\Delta\epsilon)$, respectively. The PSF with the errors is given by

$$\tilde{P} = \sum_n t_n K_n(1 + q_n)e^{i\epsilon_n}. \tag{32}$$

As the average of q_n is zero, the mean value of P is

$$M\{\tilde{P}\} = \sum_n t_n K_n \frac{1}{2\Delta\epsilon}\int_{-\Delta\epsilon}^{\Delta\epsilon} e^{i\epsilon_n}d\epsilon_n = P\frac{\sin\Delta\epsilon}{\Delta\epsilon}. \tag{33}$$

Formula (33) shows that mean value of the PSF is always stable. To characterize the stability of the angular resolution the variance of the PSF can be taken:

$$\sigma_P^2 = M\{|\tilde{P} - M\{\tilde{P}\}|^2\} = M\{|\tilde{P}|^2\} - |M\{\tilde{P}\}|^2. \tag{34}$$

The calculation of σ_P^2 is the same as in Ref. 9 and the result is given by

$$\sigma_P^2 = \left(\frac{(\Delta q)^2}{3} + 1 - \left(\frac{\sin \Delta\epsilon}{\Delta\epsilon} \right)^2 \right) |\mathbf{t}|^2, \tag{35}$$

where $|\mathbf{t}|^2 = \Sigma_n |t_n|^2$. Let the PSF be normalized by the condition $P(\varphi_0, \varphi_0) = 1$. Then

$$1 = \left| \sum_n t_n K_n \right|^2 \leqslant \sum_n |t_n|^2 \sum_n |K_n|^2, \tag{36}$$

and, as $|K_n| = 1$,

$$|\mathbf{t}| \geqslant \frac{1}{\sqrt{N}}. \tag{37}$$

For the basic-phased addition method from Eq. (22) follows $t_n = c_0 K_n^*$. The normalization constant c_0 is defined from the condition $P = 1$ and is equal to N^{-1}. Therefore, for the basic-phased addition method $|\mathbf{t}| = 1/\sqrt{N}$. So, for the given errors Δq and $\Delta\epsilon$ the basic-phased addition method has the minimum possible standard deviation $\sigma_{P_{\min}}$ of the PSF. If this minimum standard deviation is taken as a reference value then

$$\tilde{\sigma}_P \equiv |\mathbf{t}| \sqrt{N} \tag{38}$$

is the standard deviation of PSF with any other set of coefficients, $\mathbf{t}$, relative to the minimum possible standard deviation $\sigma_{P_{\min}}$ that can be obtained.

If parameter β in formula (20) for the matrix $\mathbf{B}$ is nonzero then it is enough to consider only the case $\beta = 1$. In this case the coefficients t_n given by Eq. (22) depend only on parameter α. The dependence $\tilde{\sigma}_P(\alpha)$ for difference types of arrays is discussed in the next section.

E. Examples

Two types of array layouts are considered to demonstrate the results of angular resolution algorithms. The system under consideration is 1-mm-thick aluminum plate, and the guided wave mode of interest is the S_0 Lamb wave mode at a center frequency of 1 MHz. At this frequency, its wavelength, λ_0, is 5.35 mm.

Note that there is a maximum interelement spacing requirement in an array that must be adhered to in order to prevent the appearance of grating lobes. This maximum spacing does not depend on the method of array processing and it is approximately equal to $\lambda_{\min}/2$, where $\lambda_{\min}$ is the shortest wavelength of guided waves within the frequency range of the transmitted signal.

The first array layout is a linear array of 16 omnidirectional elements. The aperture of array is $8\lambda_0$ (42.8 mm), so the interelement spacing is $0.5\lambda_0$. The angle φ varies from 0 to 180°. The array is illustrated schematically in Fig. 2(a). The second array layout consists of a single circular ring of 16 omnidirectional elements, the diameter of the ring is $8\lambda_0/\pi$ (13.6 mm). The angle φ in this case varies from 0 to 360°. The array layout is shown in Fig. 2(b). Each element of both arrays acts as transmitter and receiver, so the total number of transmitter–receiver combinations in each array is $N = 256$.

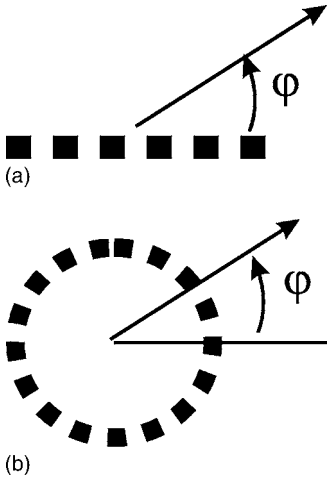


FIG. 2. Layout of (a) linear array and (b) circular array.

If the size of array (the aperture for linear array or the radius for circular array) increases then the angular resolution improves. However, at the same time the maximum interelement spacing requirement leads to a linear increase in the number of elements.[4]

Figure 3(a) shows the standard deviation $\tilde{\sigma}_P$ for the linear array as a function of parameter α for different focusing directions φ_0. As α is increased then the coefficients $\mathbf{t}$ tend to the coefficients given by the basic-phased addition method

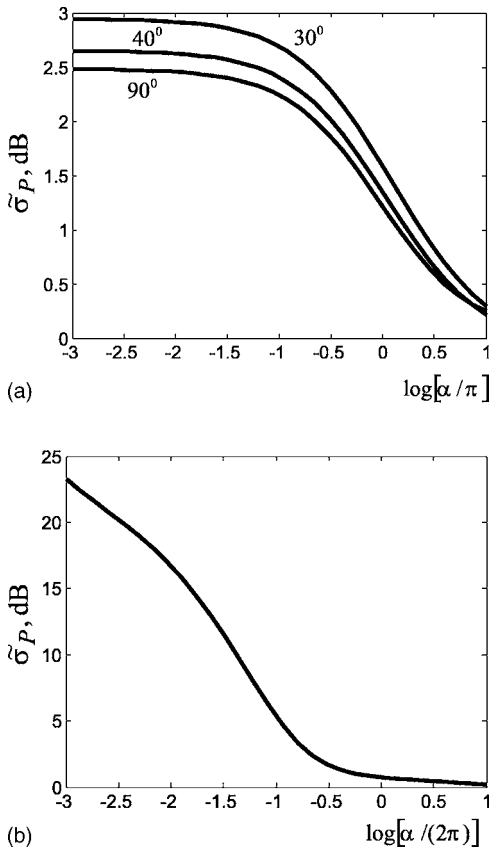
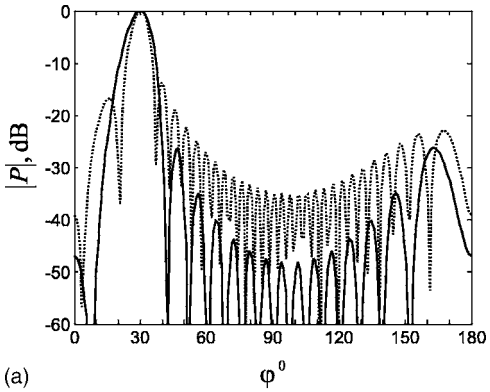
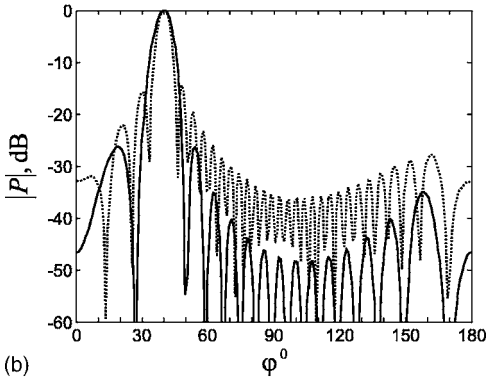


FIG. 3. Standard deviation of the PSF obtained by the maximization of contrast method (dB scale, reference value is the standard deviation of the PSF for the basic-phased addition method) for (a) linear array, focusing directions are 30°, 40°, 90° and (b) circular array, focusing direction is 180°.





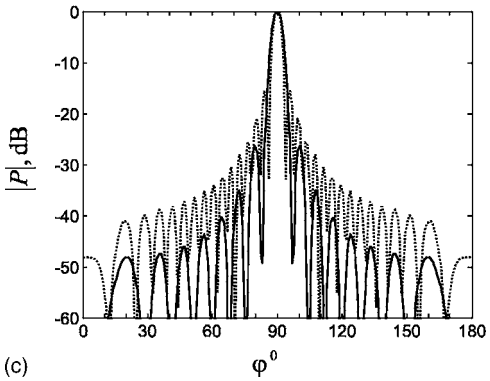


FIG. 4. Point spread functions obtained by the basic-phased addition method (solid line) and the maximization of contrast method (dotted line) for linear array and focusing direction 30° (a), 40° (b), and 90° (c).

and $\tilde{\sigma}_P$ tends to unity. For each α the standard deviation is lowest for the focusing direction $\varphi_0=90°$. In other words the linear array becomes slightly more sensitive to noise when steered at angles away from the normal direction.

The dependence $\tilde{\sigma}_P(\alpha)$ for the circular array is shown in Fig. 3(b). As the array is approximately symmetrical relative to the focusing directions φ_0, it is enough to consider only one angle $\varphi_0=180°$. It can be seen that for values of $\alpha/(2\pi)<0.1$ the maximization of contrast method for the given configuration of a circular array is very unstable.

Figure 4 shows the PSF $P(\varphi_0,\varphi)$ as a function of the angle φ for the linear array and different focusing directions φ_0. The regularization parameter α was taken equal to π. The maximization of contrast method gives a PSF with a more narrow main lobe than the basic-phased addition method. However, the sidelobe amplitude for the maximization of contrast algorithm is much higher.

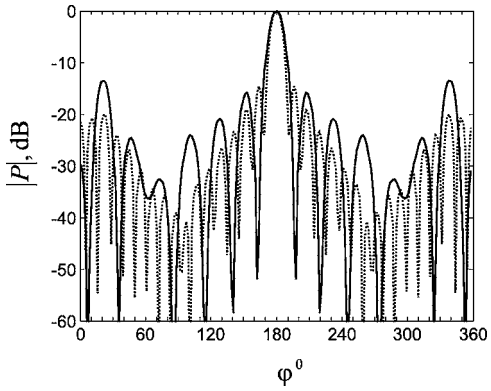


FIG. 5. Point spread functions obtained by the basic-phased addition method (solid line) and the maximization of contrast method (dotted line) for circular array and focusing direction 180°.

Figure 5 shows the PSF for the circular array and $\varphi_0=180°$. Parameter α was set equal to 2π. As for the linear array, the maximization of contrast method gives a more narrow main lobe than the basic-phased addition method. The sidelobes near the focusing direction have a similar level (−15 dB) for both methods. However, for $|\varphi-\varphi_0|>20°$ the sidelobe amplitude for the maximization of contrast algorithm is less than −19 dB while over the same range of angles the sidelobe amplitude for the basic-phased addition algorithm is −13 dB. Hence in this case the maximization of contrast method could be regarded as an improvement over the basic-phased addition method.

IV. MAXIMIZATION OF CONTRAST OVER AN INTERVAL

In the maximization of contrast method (17) the PSF is maximized at one focusing direction with respect to the PSF amplitude in other directions. The resulting PSF has a main lobe of finite width which characterizes the minimum resolvable angular interval. This leads to the idea of modifying the maximization of contrast method to include this interval. The idea of the method of maximization of contrast over an interval is to maximize the PSF in some interval over the focusing direction rather than in a single direction. Such a method was proposed in Ref. 10 for the problem of focusing of bulk waves by a group of surface sources. It can be written in the form:

$$\frac{\int_{\varphi_0-\Delta\varphi_1}^{\varphi_0+\Delta\varphi_2}|P(\varphi_0,\varphi)|^2d\varphi}{\alpha|\mathbf{t}(\varphi_0)|^2+\int_0^T|P(\varphi_0,\varphi)|^2d\varphi}\rightarrow\max. \tag{39}$$

The total maximization angular interval is $\Delta\varphi=\Delta\varphi_1+\Delta\varphi_2$, here $\Delta\varphi_1$, $\Delta\varphi_2\geqslant 0$ and in general case $\Delta\varphi_1\neq\Delta\varphi_2$. The term $\alpha|\mathbf{t}(\varphi_0)|^2$ in the denominator, where $\alpha\geqslant 0$ is a regularization parameter, is put in to make the PSF stable [as in Eq. (18)].

In the same way as the problem of maximization of contrast in one direction (18), problem (39) can be written as eigenvalue problem (21) and extrema coefficient vector $\mathbf{t}$ corresponds to the maximum eigenvalue. Matrix $\mathbf{B}$ remains the same and matrix $\mathbf{A}$ becomes

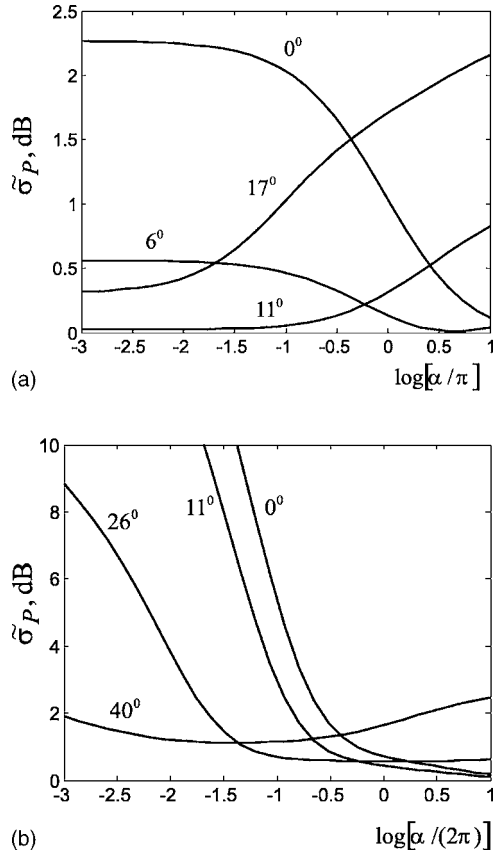


FIG. 6. Standard deviation of the PSF obtained by the maximization of contrast method over the interval (dB scale, reference value is the standard deviation of the PSF for the basic-phased addition method) for (a) linear array, focusing direction is 90°, $\Delta\varphi=0°, 6°, 11°, 17°$ and (b) circular array, focusing direction is 180°, $\Delta\varphi=0°, 11°, 26°, 40°$.

$$A_{ij} = \int_{\varphi_0-\Delta\varphi_1}^{\varphi_0+\Delta\varphi_2} K_i^*(\varphi) K_j(\varphi) d\varphi. \tag{40}$$

In this case the solution cannot be written in an analytical form as solution (22) and the eigenvalue problem (21) must be solved numerically. The numerical procedure has been implemented using functions written in the MATLAB (The Mathworks Inc., Natick, MA) modeling environment.

Figure 6(a) shows the plots of the function $\tilde{\sigma}_P(\alpha)$ for the different maximization intervals $\Delta\varphi$ in the case of the linear array. The focusing direction is 90° and subintervals $\Delta\varphi_1$ and $\Delta\varphi_2$ are equal to each other. Figure 6(b) shows the plots of the $\tilde{\sigma}_P(\alpha)$ for the case of the circular array. It can be seen that behavior of the $\tilde{\sigma}_P(\alpha)$ strongly depends on the width of the angular interval $\Delta\varphi$ and, therefore, for different intervals $\Delta\varphi$ different values of parameter α should be taken in order to keep an acceptably low sensitivity to noise.

Figure 7 shows the PSF $P(\varphi_0, \varphi)$ for the different maximization intervals $\Delta\varphi$ for the linear type of array and the focusing direction $\varphi_0=90°$. The zero angular interval corresponds to the method of maximization of contrast in one direction (18). For the angular intervals 0° and 6° the regularization parameter $\alpha=\pi$ was taken, and for the intervals 11° and 17° parameter α was chosen equal to 0.001π. From Fig. 6(a) it can be seen that such choice of α gives the standard deviation, $\tilde{\sigma}_P$, of less than 0.6 dB.

Figure 8 shows the PSF $P(\varphi_0, \varphi)$ for the circular array and the focusing direction $\varphi_0=180°$. The regularization parameter α was taken to make the standard deviation equal to 1.5 dB for all angular intervals. This gives values of α of

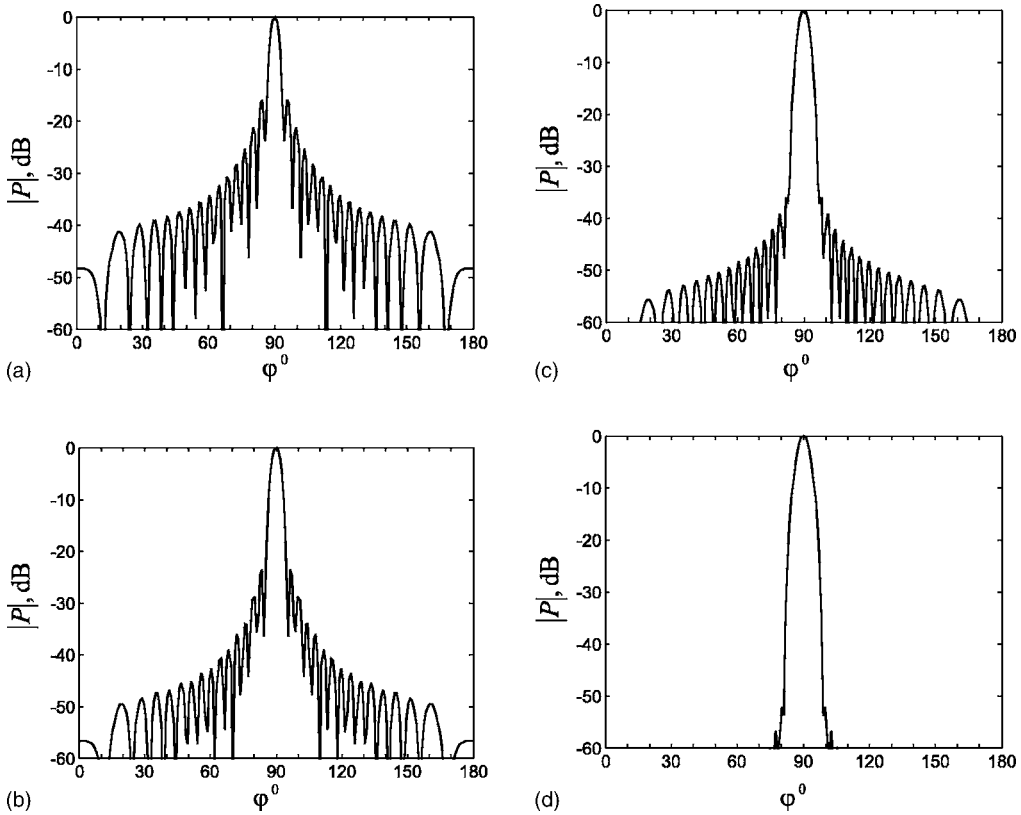


FIG. 7. Point spread functions obtained by the method of maximization of contrast over an interval for the linear array, focusing direction 90°, and angular intervals 0° (a), 6° (b), 11° (c), and 17° (d).

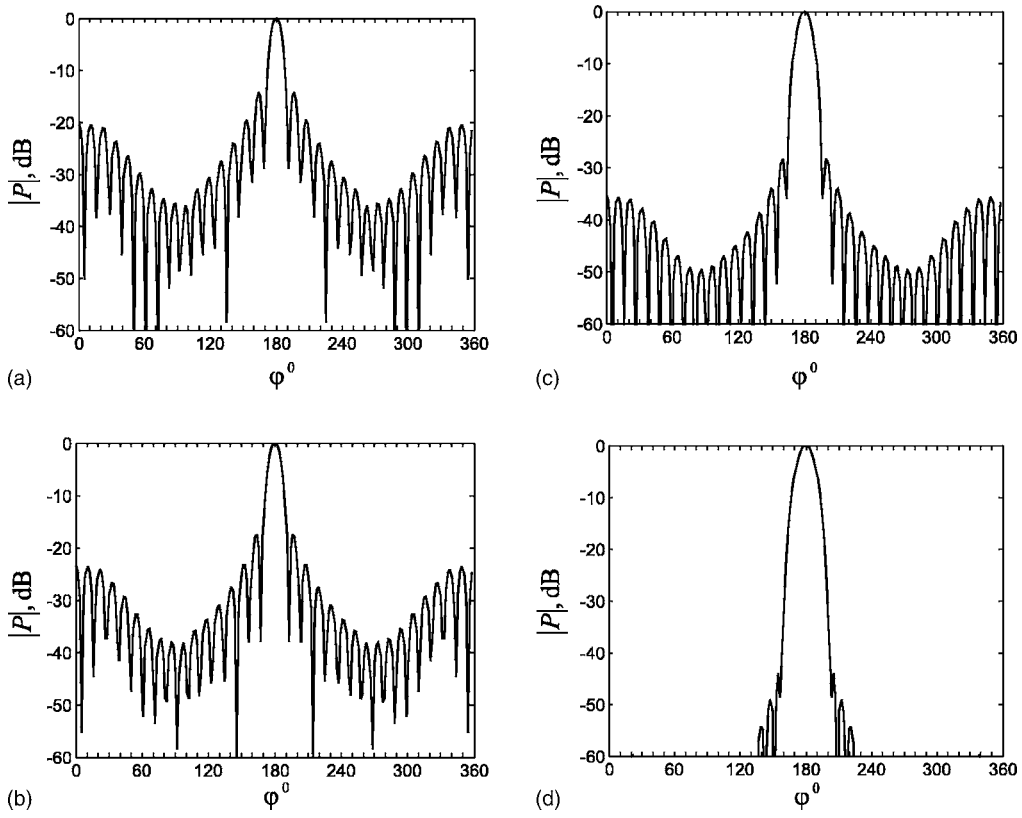


FIG. 8. Point spread functions obtained by the method of maximization of contrast over an interval for the circular array, focusing direction 180°, and angular intervals 0° (a), 11° (b), 26° (c), and 40° (d).

0.6π, 0.32π, 0.048π, and 0.0028π for the intervals, $\Delta\varphi$, of 0°, 11°, 26°, and 40°, respectively. It is seen that with increasing of $\Delta\varphi$ the width of the main lobe increases and the sidelobe amplitude decreases.

Figure 9 shows the plots of the function $\tilde{\sigma}_P(\alpha)$ for the linear array and the focusing direction 40°. Four different combinations of angular subintervals $\Delta\varphi_1$, $\Delta\varphi_2$ are considered: $\Delta\varphi_1=\Delta\varphi_2=0$; $\Delta\varphi_1=\Delta\varphi_2=8°$; $\Delta\varphi_1=10°$, $\Delta\varphi_2=13°$, and $\Delta\varphi_1=\Delta\varphi_2=16°$. The corresponding PSFs are shown in Fig. 10. The parameter α was taken equal to π for the first combination and $\alpha=0.001\pi$ for the other ones.

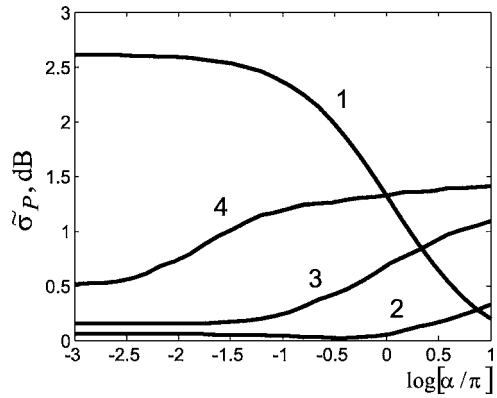


FIG. 9. Standard deviation of the PSF obtained by the maximization of contrast method over the interval (dB scale, reference value is the standard deviation of the PSF for the basic-phased addition method) for the linear array, focusing direction 40°, and combinations of the angular subintervals $\Delta\varphi_1=\Delta\varphi_2=0$ (1), $\Delta\varphi_1=\Delta\varphi_2=8°$ (2), $\Delta\varphi_1=10°$, $\Delta\varphi_2=13°$ (3), and $\Delta\varphi_1=\Delta\varphi_2=16°$ (4).

V. EXPERIMENTAL EXAMPLE

The experimental example considered is an array containing 16 transmitter elements and 32 receiver elements arranged in concentric rings with pitch circle diameters of 52 and 136 mm, as shown in Fig. 11. The design of this array was described in Ref. 4. This paper also presented results obtained using basic-phased addition method and deconvolution algorithm.

The array elements are electromagnetic acoustic transducers (EMATs) designed to excite and detect the S_0 Lamb wave mode and have equal transmission and reception sensitivity in all directions. The array was used on a 1.05 m by 1.25 m by 5-mm-thick aluminum plate specimen with square cut edges and 2-mm-thick steel disk was bonded to the surface of the plate at the location shown to simulate a defect. The transmitted signal used is a five cycle Hanning windowed toneburst with a center frequency of 200 kHz. A more detailed description of the experimental setup and array elements can be found in Ref. 4.

The results of data processing are shown in Fig. 12. From Figs. 12(a) and 12(b) it is seen that the basic-phased addition method and the method of maximization of contrast in one direction give many angular sidelobes. Figure 12(c) shows the result of processing the data by the method of maximization of contrast in the interval. The maximization interval was taken to suppress the sidelobe level to below −40 dB and is given by

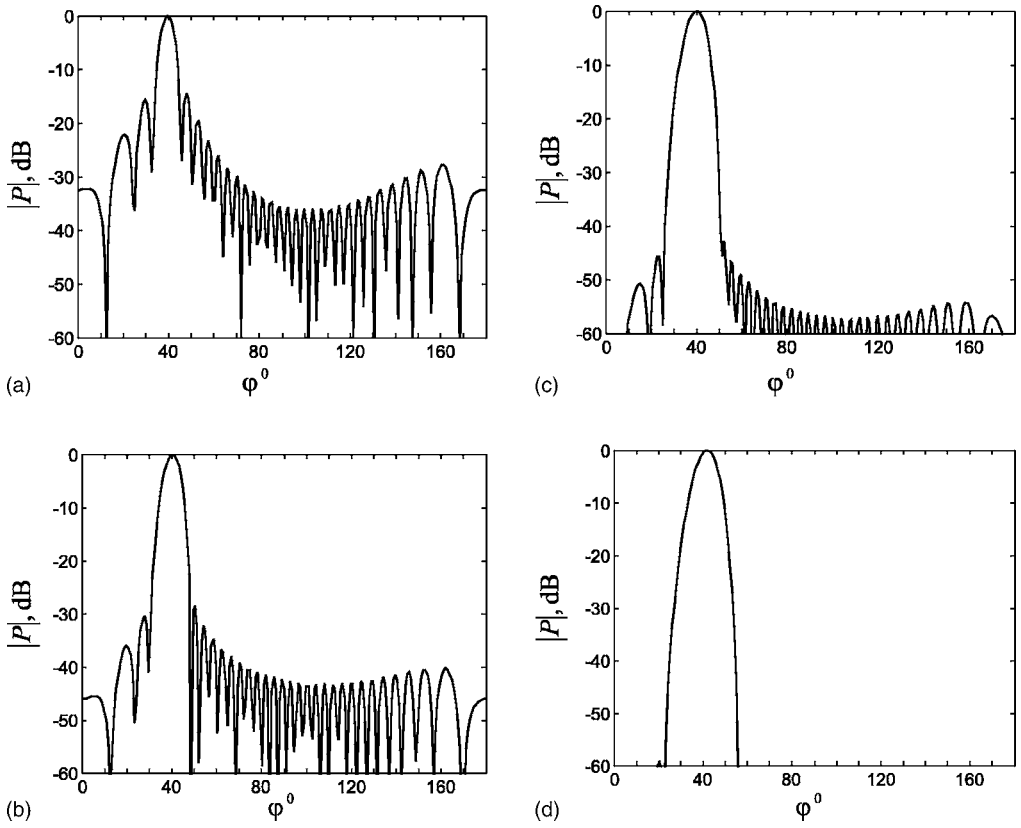


FIG. 10. Point spread functions obtained by the method of maximization of contrast over an interval for the linear array, focusing direction 40°, and angular intervals $\Delta\varphi_1=\Delta\varphi_2=0$ (a), $\Delta\varphi_1=\Delta\varphi_2=8°$ (b), $\Delta\varphi_1=10°$, $\Delta\varphi_2=13°$ (c), and $\Delta\varphi_1=\Delta\varphi_2=16°$ (d).

$$\Delta\varphi = \frac{\Delta\varphi_0}{\lambda_0}\lambda, \tag{41}$$

where λ is wavelength, $\Delta\varphi_0=23°$ is maximization interval for the center frequency, and $\lambda_0=26.7$ mm is wavelength at the center frequency. The regularization parameter α was chosen to make the standard deviation $\tilde{\sigma}_P$ less than 1.5 dB for all angular intervals.

It can be seen that all sidelobes are successfully suppressed, but only to a level of around −30 dB. This is consistent with the expected sensitivity of the technique to the noise according to the analysis of the stability of the angular resolution methods given in Sec. III D. It is also instructive to estimate the errors in the angular Green's function caused by far field approximation (4). The next term in the expansion (4) has the form:

$$R_{(T)i} \approx r - \mathbf{r}_{(T)i}\mathbf{n} + \frac{1}{2r}(\mathbf{r}_{(T)i}\mathbf{l})^2, \tag{42}$$

where $\mathbf{l}=\{-\sin\varphi, \cos\varphi\}$. From Eq. (31) it follows that the phase error ϵ_n is given by

$$\epsilon_n = \frac{\pi[(\mathbf{r}_{(T)i}+\mathbf{r}_{(R)j})\mathbf{l}]^2}{r\lambda_0}. \tag{43}$$

From Fig. 11 it is seen that the nearest reflector to the array is the lower edge of the plate at the distance $r=0.38$ m from the center of the array. For the given geometry of the array $|(\mathbf{r}_{(T)i}+\mathbf{r}_{(R)j})| \leqslant 0.094$ m, and, hence, $0 \leqslant \epsilon_n \leqslant 2\Delta\epsilon$, where $\Delta\epsilon=1.37$. Finally, expression (35) (assuming that $\Delta q=0$) gives for sidelobe contribution caused by phase errors a value of about −30 dB.

This demonstrates the practical application of the technique to real experimental data.

VI. MULTIMODE RESOLUTION

A. Algorithm

The above-described processing algorithms can be generalized for structures in which more than one guided wave mode may exist. If there are several modes the problem of maximization (39) can be written for every transmitted–received mode combination:

$$\frac{\int_{\varphi_0-\Delta\varphi_m/2}^{\varphi_0+\Delta\varphi_m/2} |P_{mm}(\varphi_0,\varphi)|^2 d\varphi}{\alpha|\mathbf{t}(\varphi_0)|^2 + \sum_n \int_0^{2\pi} |P_{mn}(\varphi_0,\varphi)|^2 d\varphi} \to \max. \tag{44}$$

Here index m refers to the mode combination of interest and index n refers to all possible mode combinations. Each function $P_{mn}(\varphi_0,\varphi)$ describes the response from a point reflector for the nth mode combination if the extracted mode combination is the mth. For the case of ideal resolution $P_{mn}(\varphi_0,\varphi)=\delta_{mn}\delta(\varphi-\varphi_0)$.

B. Example

The same configuration of array as in Sec. V on 5-mm-thick aluminum plate is considered. The inner ring has

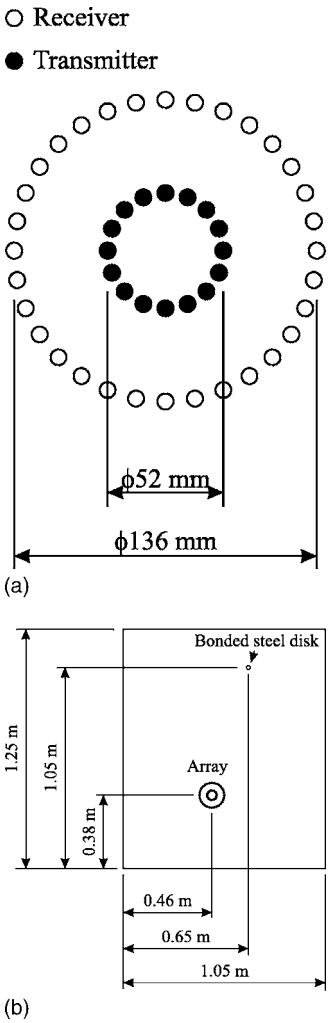


FIG. 11. (a) Geometry of EMAT array and (b) experimental arrangement on 5-mm-thick aluminum plate specimen.

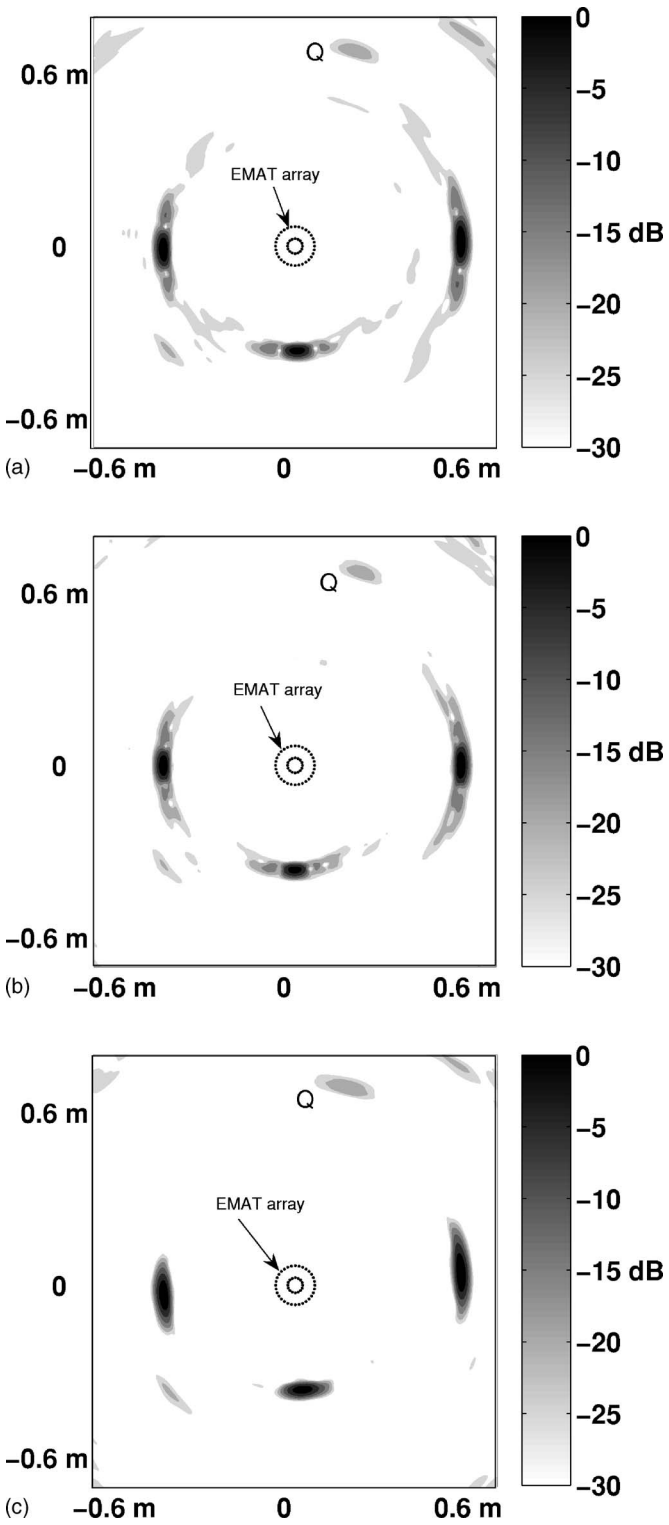


FIG. 12. Experimental results obtained from EMAT array on a 5-mm-thick aluminum plate processed with (a) basic-phased addition algorithm, (b) method of maximization of contrast, and (c) method of maximization of contrast in the interval. The signals labeled Q are from a 20-mm-diameter, 2-mm-thick steel disk bonded to the surface of the plate.

diameter 40 mm and contains 16 transducers, the outer ring has diameter 80 mm and contains 32 transducers. Each transducer operates as omnidirectional transmitter and receiver. Note that the general method does not require omnidirectionality of array elements and this assumption is taken for simplicity of the analysis only. It is supposed that there are three modes in the system, S_0, A_0, and SH_0 and transducers are equally sensitive to all three modes. The frequency is 200 kHz and wavelengths are 26.7, 11.7, and 15.6 mm for S_0, A_0, and SH_0 modes, respectively.

Figure 13 shows the set of PSFs for the case of $SH_0 - SH_0$ mode combination extraction. The focusing direction is 180° and the maximization interval is 17°. The parameter α was taken equal to 18π and gives a standard deviation $\tilde{\sigma}_P$ of 1 dB. As all array elements act as transmitters and receivers, then $P_{mn}=P_{nm}$, and in Fig. 13 results for only six mode combination of SH_0-SH_0, S_0-S_0, A_0-A_0, S_0-A_0, S_0-SH_0, and A_0-SH_0 are shown. It can be seen that all mode combinations have sidelobe amplitude below −30 dB. This level of modal selectivity is obtained entirely by the data processing method. However, some additional suppression of unwanted mode combinations may be achieved at the array element level as well.

VII. CONCLUSION

A general procedure for array processing has been presented. The approach is based on multiplying the transmitted

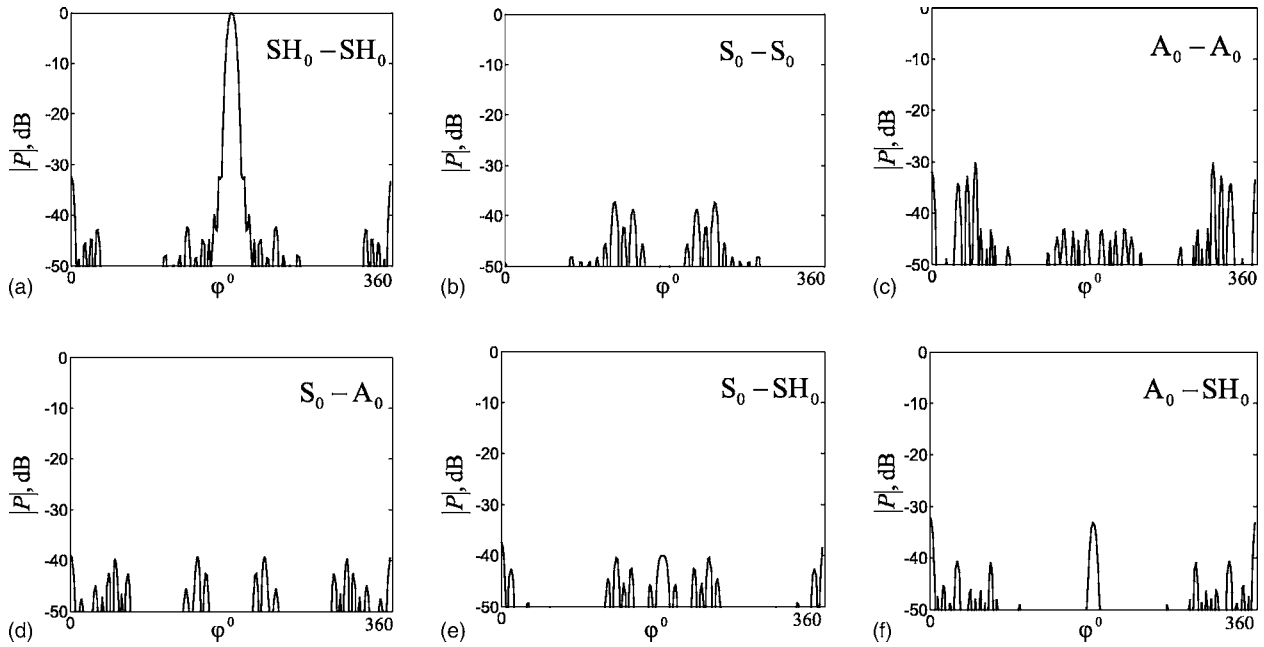


FIG. 13. Amplitude of point spread functions in the case of SH_0-SH_0 mode extraction for different mode combinations: (a) SH_0-SH_0, (b) S_0-S_0, (c) A_0-A_0, (d) S_0-A_0, (e) S_0-SH_0, and (f) A_0-SH_0.

and received signals by suitable phase factors in order to focus the beam on every point within a test structure. The technique is applicable to any geometry of array and any type of array elements. Only the far field case has been considered as this enables angular and radial resolution to be separated. It is possible to extend the technique to near field resolution, but analysis of the results and computations in this case become more complicated.

Different methods of calculating the phase coefficients have been considered. The performance of different algorithms has been tested on modeling and experimental data. Generalization of the above-presented algorithms for the case of multimode media has been suggested.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the UK Research Centre in NDE and by BNFL, Nexia Solutions, and DSTL.

[1]D. E. Chimenti, "Guided waves in plates and their use in materials characterization," Appl. Mech. Rev. **50**, 247–284 (1997).
[2]D. N. Alleyne, B. Pavlakovic, M. J. S. Lowe, and P. Cawley, "Rapid long-range inspection of chemical plant pipework using guided waves," Insight **43**, 93–96, 101 (2001).
[3]R. Sicard, J. Goyette, and D. Zellouf, "A saft algorithm for Lamb wave imaging of isotropic plate-like structures," Ultrasonics **39**, 487–494 (2002).
[4]P. D. Wilcox, "Omni-directional guided wave transducer arrays for the rapid inspection of large areas of plate structures," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 699–709 (2003).
[5]C. Holmes, B. Drinkwater, and P. Wilcox, "Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation," NDT & E Int. **38**, 701–711 (2005).
[6]P. D. Wilcox, "A rapid signal processing technique to remove the effect of dispersion from guided wave signals," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 419–427 (2003).
[7]R. J. Mailloux, *Phased Array Antenna Handbook* (Artech House, Boston, 1994).
[8]A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems* (Winston, Washington, DC, 1977).
[9]W. Wirth, *Radar Techniques Using Array Antennas* (IEE, London, 2001).
[10]V. A. Babeshko, E. V. Glushkov, and J. F. Zinchenko, *The Dynamics of Inhomogeneous Linearly-Elastic Bodies* (Nauka, Moscow, 1989) (in Russian).

Middle-ear circuit model parameters based on a population of human ears

Kevin N. O'Connor and Sunil Puria[a)]
Department of Mechanical Engineering, Stanford University, 496 Lomita Mall, Durand Building, Room 206, Stanford, California 94305, Department of Otolaryngology/Head and Neck Surgery, 801 Welch Road, Stanford, California 94305, and Palo Alto Veterans Affairs, 3801 Miranda Ave., Palo Alto, California 94304

(Received 28 June 2007; revised 1 November 2007; accepted 2 November 2007)

Middle-ear circuit model parameters are selected to produce overall magnitude and phase agreement with pressure to stapes velocity transfer function measurements made on 16 human temporal bones, up to approximately 12 kHz. The circuit model, which was previously used for the cat, represents the tympanic membrane (TM) as a distributed parameter acoustic transmission line, and ossicular chain and cochlea as a network of lumped circuit elements. For some ears the TM transmission line primarily affects the magnitude of the response, while for others it primarily affects the phase. Model responses also compare favorably with velocity ratio data between the umbo and stapes footplate as well as between the umbo and incus, and exhibit similar characteristics to three previous input impedance measurements, including two from living ears. Similarities are also shown between the model magnitude and adjusted pressure to stapes velocity measurements from living ears, suggesting that the model may suitably approximate the behavior of living ears. In addition to fitting individual measurements, a set of parameters is selected to produce agreement with the *mean* of the 16 measurements up to 10 kHz, to allow the main features of the ensemble to be reproduced from a single parameter set. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2817358]

PACS number(s): 43.64.Bt, 43.64.Ha [BLM] Pages: 197–211

I. INTRODUCTION

It is well established that the function of the middle ear in terrestrial mammals is to efficiently transfer, for a broad range of frequencies relevant to the animal, acoustic energy between the low-density air of the ear canal and the higher-density fluid of the cochlea (Békésy, 1960; Wever and Lawrence, 1950). A critical first step in the transduction process takes place at the tympanic membrane (TM), which converts ear canal pressure into vibrations of the malleus-incus complex. Vibrations of the stapes, coupled through incus vibrations, result in fluid pressure inside the cochlea.

There exist a significant number of physiological measurements that characterize the human middle ear. Mathematical models attempt to encapsulate the physiology, and analog circuit models have been popular. Models that represent the TM motion as a single piston are limited in their range of validity to frequencies below a few kHz (Onchi, 1949; Zwislocki, 1962; Kringlebotn, 1988; Shera and Zweig, 1992). Attempts have been made to extend the frequency range with a two-piston model of the TM (Shaw and Stinson, 1983; Goode *et al.*, 1994), however even the two-piston descriptions are limited in their region of validity to below 4–6 kHz (Puria *et al.*, 1997). To overcome these limitations, a new model was introduced in which the eardrum is represented as a distributed parameter transmission line (Puria and Allen, 1998). This model was shown to match measurements made on the cat middle ear up to 20 kHz.

The purpose of this paper is to obtain parameters for the model topology representing the eardrum as a transmission line, to produce model responses that resemble measurements made on the human middle ear. Measurements made on 16 human cadaver ears of stapes velocity normalized by the pressure difference between the ear canal and middle-ear cavity are used as starting points, with model parameters selected to capture the essential magnitude and phase behavior of each measurement. The resulting model transfer functions are compared to past measurements made on living ears (Huber *et al.*, 2001), in which adjustments have been made for frequencies below 2 kHz in an attempt to account for possible methodological differences between live and cadaver measurements (Chien *et al.*, 2006). Mean velocity transfer functions from two other studies are used to help with parameter selection (Aibara *et al.*, 2001; Willi *et al.*, 2002), and impedance measurements from three previous studies (Hudde, 1983; Voss *et al.*, 2000; Farmer-Fedor and Rabbitt, 2002), two of which are based on living ears, are used as an independent check of the resulting model parameters. An additional parameter set is reported which provides agreement with the mean of the 16 measured transfer functions up to 10 kHz, thus providing a convenient way to use the model to approximately reproduce the mean behavior.

The present model for the human middle ear is shown to be in closer agreement with physiological measurements than three previous lumped-element models (Kringlebotn, 1988; Goode *et al.*, 1994; Feng and Gan, 2004), and there is some correspondence to measurements made in living ears.

[a)]Author to whom correspondence should be addressed. Electronic mail: puria@stanford.edu

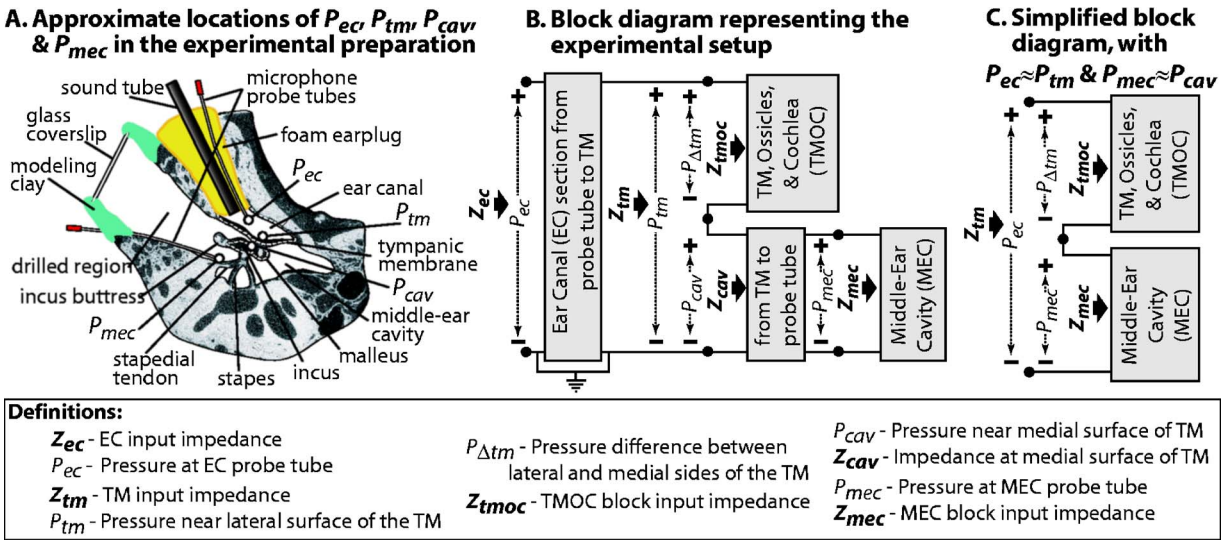


FIG. 1. (Color online) Experimental setup diagram (A) with a corresponding block diagram representation (B), and the simplified version of the block diagram used for model parameter selection (C). Section A depicts a portion of the experimental setup used to measure ear canal pressure (P_{ec}), middle-ear cavity pressure (P_{mec}), and stapes velocity (not shown). Pressures at the lateral and medial surfaces of the tympanic membrane (TM), near the umbo, are also shown as P_{tm} and P_{cav}, respectively. Section B depicts the assumed relationships between the four pressures in section A in terms of a block diagram. P_{ec} and P_{tm} are related by a two-port block representing the pressure and volume velocity transformations between those two points in the ear canal. P_{cav} and P_{mec} are similarly related by a two-port block representing the pressure and volume velocity transformations between those two points in the middle-ear cavity. $P_{\Delta tm}$ represents the pressure difference between the two sides of the TM, and is equal to P_{tm}-P_{cav} in this diagram. The tympanic membrane, ossicular chain, and cochlea (TMOC) are represented as a block with input impedance Z_{tmoc}, where Z_{tmoc} is equal to $P_{\Delta tm}$ divided by the volume velocity absorbed by the TM. Other input impedances are shown as Z_{ec} at the P_{ec} measurement location, Z_{tm} at the P_{tm} location, Z_{cav} at the P_{cav} location, and at the P_{mec} location the input impedance of the middle-ear cavity is shown as Z_{mec}. Section C contains a simplified version of the block diagram in section B, in which it is assumed that P_{ec} is approximately equal to P_{tm}, and P_{mec} is approximately equal to P_{cav}. These assumptions allow $P_{\Delta tm}$ to be determined directly from the measurements as P_{ec}-P_{mec}, which is why this version of the diagram is used for model fitting. While from a physical standpoint these approximations are mostly applicable at low frequencies, it is assumed that they hold for the full frequency range due to their benefits in reducing the complexity of the model.

II. MODEL OVERVIEW

The focus of this study is on modeling the behavior of the tympanic membrane, ossicular chain, and cochlea (TMOC). The ear canal (EC) and middle-ear cavity (MEC) are not modeled in this study. Figure 1(a) illustrates a portion of the experimental setup for obtaining the EC and MEC pressure and stapes velocity measurements that are used to find TMOC model parameters. The EC pressure, P_{ec}, was measured in the small air space within the ear canal between the foam plug and tympanic membrane (TM), at a distance from the lateral TM surface estimated to be on the order of 2–3 mm. The MEC pressure, P_{mec}, was measured within the drilled middle-ear cavity at a distance from the anterior TM surface estimated to be on the order of 4–5 mm. Because the measured pressures were spatially separated from the surfaces of the TM, pressures are also labeled in the diagram on the lateral and medial surfaces of the TM, as P_{tm} and P_{cav}, respectively.

Figure 1(b) contains a block diagram illustrating the relationships between the four pressures labeled in Fig. 1(a), as well as impedance blocks representing the TMOC (with input impedance Z_{tmoc}) and MEC (with input impedance Z_{mec}). The spatial transformation between P_{ec} and P_{tm} is accomplished with a two-port block representing the portion of the ear canal between those two points, with an input impedance Z_{ec}. The spatial separation between P_{mec} and P_{cav} is similarly modeled using a two-port block, with input impedance Z_{cav}.

The measured P_{ec} and P_{mec} pressures should differ somewhat from pressures measured at the surface of the TM, especially at high frequencies. In the interest of simplicity, however, it was decided to assume that P_{ec} is approximately equal to P_{tm} and P_{mec} is approximately equal to P_{cav}, for the purpose of fitting the model to measurements. These simplifications are shown in Fig. 1(c), in which the EC two port is removed with P_{ec} replacing P_{tm}, and the MEC two port is removed with P_{mec} replacing P_{cav}. The input impedance seen at the lateral TM surface, Z_{tm}, is then defined as the sum of the TMOC input impedance, Z_{tmoc}, and the MEC input impedance, Z_{mec}.

The pressure difference between the two sides of the TM, $P_{\Delta tm}$, is equal to $P_{tm}-P_{cav}$ in Fig. 1(b), and is approximated to be equal to $P_{ec}-P_{mec}$ in Fig. 1(c), which is the form used for model fitting. The assumption that the TMOC block is driven exclusively by the pressure difference across the TM (O'Connor and Puria, 2006), and that the MEC can be treated as a one-port block in series with the TMOC block [Fig. 1(c)] follows from the Zwislocki (1962) model formulation, and specifically ignores the effects of middle-ear cavity pressure on the oval and round windows, as explored in other formulations such as that of Shera and Zweig (1992). By defining the model according to Fig. 1(c), it is possible to approximately isolate the TMOC block from the EC and MEC blocks by using $P_{ec}-P_{mec}$ as the input to the block, thus simplifying the model fitting procedure. Once the TMOC block is characterized, it is possible to produce an approximate full middle-ear model by later adding suitable MEC and EC models in a manner similar to that shown in Fig. 1(b).

A. TMOC block

Figure 2 illustrates the TMOC model block in two equivalent forms: Fig. 2(a) shows the original form of the model, based on Fig. 5(b) of Puria and Allen (1998), in which transformers implement the effects of the TM area, malleus-incus lever ratio, and stapes footplate area; and Fig. 2(b) shows an equivalent version of the model in which many of the model parameters have been redefined, and relabeled with a "T" attached to their subscripts, such that the effects of the transformers are taken into account by their values. Equations at the bottom of Fig. 2(b) indicate how to convert between the untransformed parameters found in Fig. 2(a) and the transformed parameters found in Fig. 2(b). While the two versions of the model are interchangeable, the transformed version [Fig. 2(b)] is somewhat easier to think about and has the benefit of allowing the magnitudes of model parameters to be compared directly even though in [Fig. 2(a)] they lie on opposite sides of one or more transformers.

The model is displayed in the form of an electrical circuit, with pressure and force both analogous to voltage, and volume velocity and velocity both analogous to current. Consistent with these analogies, acoustic and mechanical stiffness elements are represented as electrical capacitors, acoustic and mechanical masses are represented as electrical inductors, and acoustic and mechanical resistances are represented as electrical resistors (Beranek, 1954; Pierce, 1989). All model variables and parameters are defined within Fig. 2.

B. TM model

The two-port block representing the tympanic membrane is implemented as a one-dimensional cylindrical lossless acoustic transmission line with characteristic impedance Z_{0tm} and a wave propagation delay from the input to the output of T_{tm}. While this distributed parameter representation is a simplification, it captures the essential concepts of there being a propagation lag for sound transferring through the TM, plus the idea that there can be reflections at the umbo depending on the impedance attached to it, Z_{ocT}. With the presence of reflections at the umbo, it becomes possible for constructive or destructive interference to occur between the incident and reflected pressure waves at each point along the transmission line such that the total pressure applied at the umbo, P_{uT}, can either be more or less than the pressure applied to the entrance of the transmission line, $P_{\Delta tm}$, for a given frequency.

The relationship between the characteristic impedance of the transmission line, Z_{0tm}, and the load impedance of the transmission line, Z_{ocT}, determines when and how reflections occur. If the load impedance has the same value as the characteristic impedance, for example, then there are no reflections since the load behaves in the same manner as the rest of the line, and therefore does not act as a boundary. When the load impedance has a different magnitude from the characteristic impedance, and/or has a nonzero imaginary part, then reflections do occur, whose magnitude and phase vary with respect to the incident wave depending on the relationship between the characteristic and load impedances. A simple way of representing the relationship between the reflected and incident waves is by using the "reflection coefficient," denoted by Γ, which is defined here as the reflected pressure wave's complex magnitude, P_-, divided by the incident pressure wave's complex magnitude, P_+. The reflection coefficient is also expressed, as a function of angular frequency ω, in terms of Z_{ocT} and Z_{0tm} as follows:

$$\Gamma(\omega) = \frac{P_-(\omega)}{P_+(\omega)} = \frac{\frac{Z_{ocT}(\omega)}{Z_{0tm}} - 1}{\frac{Z_{ocT}(\omega)}{Z_{0tm}} + 1}. \quad (1)$$

The ratio between the pressure at the output of the transmission line, P_{uT}, and the pressure at the input of the transmission line, $P_{\Delta tm}$, which is referred to as P_{gain}, can be derived from the standard equations governing lossless transmission line behavior, in which the complex pressure amplitude at position x and frequency ω, $P(x,\omega)$, is expressed as $P_+(\omega)e^{-i\omega x/c} + P_-(\omega)e^{i\omega x/c}$ (where c is the speed of sound in air and i is $\sqrt{-1}$). By defining $P_{uT}(\omega)$ as $P(0,\omega)$ and $P_{\Delta tm}(\omega)$ as $P(-\ell,\omega)$ (where ℓ is the length of the transmission line), substituting Eq. (1), and substituting T_{tm} for ℓ/c, it follows that $P_{gain}(\omega)$ can be expressed as

$$P_{gain}(\omega) = \frac{P_{uT}(\omega)}{P_{\Delta tm}(\omega)} = \frac{1+\Gamma(\omega)}{e^{i\omega T_{tm}} + \Gamma(\omega)e^{-i\omega T_{tm}}}. \quad (2)$$

P_{gain} is used to compute the pressure driving the ossicular chain from the pressure at the input to the transmission line.

Finally, the input impedance of the transmission line, Z_{tmoc}, which is equal to $P_{\Delta tm}$ divided by the corresponding volume velocity entering the transmission line, U_{tm}, can be derived given that for a lossless transmission line the volume velocity at position x and frequency ω, $U(x,\omega)$, can be expressed as $[P_+(\omega)/Z_0]e^{-i\omega x/c} - [P_-(\omega)/Z_0]e^{i\omega x/c}$ (in which Z_0 is the characteristic impedance of the line). Evaluating at $x = -\ell$, substituting T_{tm} for ℓ/c, replacing Z_0 with Z_{0tm}, and substituting Eq. (1) where appropriate, it follows that $Z_{tmoc}(\omega)$ can be expressed as

$$Z_{tmoc}(\omega) = \frac{P_{\Delta tm}(\omega)}{U_{tm}(\omega)} = Z_{0tm}\left(\frac{1+\Gamma(\omega)e^{-i2\omega T_{tm}}}{1-\Gamma(\omega)e^{-i2\omega T_{tm}}}\right). \quad (3)$$

III. METHODS

A. Measurements

1. $V_{st}/P_{\Delta tm}$

The primary curves used for fitting the model are 16 measurements, from two studies referred to as set A and set B, of stapes velocity (V_{st}) normalized by $P_{\Delta tm}$, where $P_{\Delta tm}$ is defined as $P_{ec} - P_{mec}$ [see Fig. 1(c)]. The measurements in set A have been previously published in O'Connor and Puria (2006) for four previously frozen human temporal bone cores. The measurements in set B are from O'Connor *et al.* (2008) for 12 human temporal bone cores, of which 4 were previously frozen and 8 were fresh. The preparation and measurement procedures were similar between the two studies, although for the 2007 study the sound source was capable of producing larger pressures at high frequencies than

the sound source used in the 2006 study. For the mean $V_{st}/P_{\Delta tm}$ magnitudes, the signal to noise ratio (SNR) for set A is above 6 dB for all frequencies below 11.5 kHz, whereas for set B the SNR of the mean curve is above 6 dB for the full reported frequency range [see the noise floor curves in Figs. 4(a) and 4(b)].

2. V_{st}/P_{ec} from living ears

As an independent check of the model fits, but not to assist in the fitting procedure, comparisons to the model are made using V_{st}/P_{ec} measurements adapted from the stapes displacement measurements reported by Huber *et al.* (2001) on seven living ears. The original measurements exhibit lower magnitudes than typical temporal bone measurements below 2 kHz, by around a factor of 3. In Chien *et al.* (2006) it was argued that this observed magnitude difference below 2 kHz may be the result of systematic V_{st} measurement angle differences between studies on living ears versus temporal bone ears. The curve used for comparison with the model, then, is the adjusted Huber *et al.* (2001) mean from Fig. 14 of Chien *et al.* (2006), in which the assumed effects of the measurement angle difference are accounted for below around 2 kHz. The data above 2 kHz were not adjusted.

3. V_{st}/V_u

An umbo velocity to stapes velocity transfer function mean curve (V_{st}/V_u) constructed using published stapes velocity and previously unpublished umbo velocity data measured in the Aibara *et al.* (2001) study, based on 11 fresh bones, is another piece of data used for comparison with the model. Because it is a mean curve (rather than a measurement from an individual ear), it is based on a different set of ears from those in sets A and B, and because of some differences in how the bones were set up in the Aibara *et al.* (2001) study versus the studies for sets A and B, it is used more as an approximate guide than as something to be fit as closely as possible.

4. V_i/V_u

The umbo velocity to incus velocity (V_i/V_u) transfer function mean published as the "TF-ui" curve in Fig. 7 of Willi *et al.* (2002), based on nine fresh bones, is another curve used for rough comparison with the model.

5. Input impedance

The mean of middle-ear input impedance measurements from Voss *et al.* (2000), based on 13 fresh temporal bones and reported up to 4 kHz, is used for rough comparison with the model, but not to assist in the fitting procedure. Adjustments were made in the study to remove the effects of the ear canal such that their reported measurements should roughly correspond to Z_{tm}, or $Z_{tmoc}+Z_{mec}$, in Fig. 1(c). Since the focus of this study is on determining the Z_{tmoc} model by itself, without attempting to model the middle-ear cavity, the middle-ear cavity model provided in the Voss *et al.* (2000) study (from Fig. 12 and Table I for bone "24L" of that study) is subtracted from their input impedance measurements to produce approximate Z_{tmoc} measurements that can be compared directly to the present model.

The mean power reflectance and pressure reflectance phase curves published in Fig. 16 of Farmer-Fedor and Rabbitt (2002) between 1 and 15 kHz and based on measurements made approximately 2 mm from the medial end of the ear canal in living ears, were converted to an impedance to serve as a further basis for rough comparison with the model, but not to assist in parameter selection. The square root was taken of the power reflectance to obtain the pressure reflectance magnitude, and this was combined with the pressure reflectance phase to obtain the complex pressure reflectance. A characteristic impedance of 3.28×10^7 Kg/(s m^4) at the 2 mm measurement location was estimated by multiplying the density of air (1.18 Kg/m^3) by the speed of sound in air (347 m/s) and dividing by the ear canal cross-sectional area at 2 mm as plotted in Fig. 4(a) of the Farmer-Fedor and Rabbitt (2002) paper (originally from Stinson and Lawton, 1989; yielding an area of approximately 0.125×10^{-4} m^2). The reflectance was then converted to an approximate corresponding impedance by using the formula in Eq. (1) for plane waves, substituting the appropriate reflectance, impedance, and characteristic impedance variables, and solving for impedance. Because both a portion of the ear canal and the middle-ear cavity are present for the reflectance measurements, the resulting impedance is labeled "Z_{ec}" in reference to Fig. 1(b).

The real and imaginary parts of the normalized measured impedances published in Fig. 4 of Hudde (1983) between 1 and 19 kHz for six living ears were converted to magnitude and phase form, using the same characteristic impedance that was estimated above, and averaged to produce another "Z_{ec}" impedance curve for rough comparison with the model, but also not to assist in parameter selection.

B. Fitting procedure

The model parameters in Fig. 2(b) were selected for each of the 16 $V_{st}/P_{\Delta tm}$ measurements from sets A and B using a systematic manual fitting procedure. The procedure involved starting with a simplified version of the circuit model with only one element and one branch, and then sequentially restoring elements and branches to the model and assigning element values at each stage using specific criteria, until values were assigned to all elements in the model. This approach was chosen over automated approaches involving the minimization of an error function because it affords more insight into the behavior of each model element and allows varying emphasis to be placed on fitting particular features of the data over other features.

The fitting procedure was performed with the help of MATLAB (The MathWorks, Natick, MA) scripts that were written to display the data curves and various aspects of the model on the same plot, and to make it easy to regenerate the plots as parameters were changed. The steps in the fitting procedure are briefly described below.

1. Stapes-cochlea resistance: $R_{scT} = R_{alT} + R_{cT}$

In the first step, the model contained only the transformed stapes-cochlea resistance, R_{scT}, which is the sum of the R_{alT} and R_{cT} parameters shown in Fig. 2(b). The two resistances in the stapes-cochlea branch were combined into one variable to simplify the fitting procedure. This parameter was set to the average value of the cochlear input impedance magnitude measured in Aibara *et al.* (2001) from 0.1 to 5 kHz, converted to root-mean-square (rms) form, and then transformed [see the equations in Fig. 2(b)] to yield a value of 4.84×10^7 Kg/(s m^4).

2. Series stiffness elements: K_{mT} and $K_{scT} = K_{alT} + K_{rwT}$

In the second step, K_{mT} and K_{scT}, which is equal to $K_{alT} + K_{rwT}$, were introduced into the model in series with R_{scT}. The two stiffness elements were then assigned equal values such that the model and measured $U_{scT}/P_{\Delta tm}$ transfer functions (converted from the $V_{st}/P_{\Delta tm}$ measurements) agreed at low frequencies.

3. Ossicular joint stiffness elements: K_{imjT} and K_{isjT}

In the third step, the transformed IMJ and ISJ branch stiffnesses, K_{imjT} and K_{isjT} respectively, were introduced within two parallel shunt branches between K_{mT} and K_{scT}. The measurements of V_{st}/V_u and V_i/V_u from two other studies were then compared against the corresponding model curves. The V_{st}/V_i model transfer function was also plotted for reference, and values of K_{imjT}, K_{isjT}, and K_{scT} were set to produce approximate agreement between the data and model among these velocity transfer functions at low frequencies.

Because three stiffness parameters were being adjusted, but only two reference measurements were available, it was not possible to uniquely determine all three from the data. It was assumed that the IS joint is typically stiffer than the IM joint, so a relatively high value was assigned to K_{isjT} compared to K_{imjT} and K_{scT}. By changing these stiffnesses, the low frequency $U_{scT}/P_{\Delta tm}$ agreement was typically altered, so K_{mT} was then adjusted to restore this low frequency agreement without affecting the velocity transfer functions.

4. Transmission line parameters: T_{tm} and Z_{0tm}

In the fourth step, the transmission line was introduced into the model. The delay, T_{tm}, was set such that good overall phase agreement was seen for the $U_{scT}/P_{\Delta tm}$ transfer function, and Z_{0tm} was set to match the Z_{ocT} load impedance magnitude at 10 kHz such that it updated whenever Z_{ocT} subsequently changed. In the event that Z_{0tm} needed to be adjusted further, it was scaled by an additional factor. By matching the characteristic impedance to the load impedance at 10 kHz, the goal was to reduce reflections in the transmission line at high frequencies, and to consequently reduce the size of peaks that appeared in the transfer function magnitude as a result of these reflections. Reducing the reflections in the transmission line becomes more important once masses are introduced into the circuit, since their purely imaginary impedances become larger at higher frequencies, which tends to increase reflections in the transmission line.

5. Ossicle masses: M_{mT}, M_{iT}, and M_{sT}

The fifth step involved introducing the three transformed masses into the model. The stapes mass, M_{sT}, was kept fixed in this study to reduce the complexity of the fitting procedure, and was assigned a value based on the 3.5 mg stapes mass reported in Beer *et al.* (1999), which was divided by A_{fp}^2 to yield 3.53×10^5 Kg/m^4 for M_s (since it is treated as an acoustic mass in this model), and then transformed to produce a value of 575 Kg/m^4 for M_{sT}.

The incus mass, M_{iT}, was introduced next and set to improve the high frequency magnitude roll-off of the $U_{scT}/P_{\Delta tm}$ model transfer function.

The malleus mass, M_{mT}, was introduced last, and set in such a way as to tailor the effects of the transmission line to the measurements. Because M_{mT} is responsible for the largest portion of the imaginary part of Z_{ocT} at high frequencies, its value can have a large effect on the reflections occurring within the transmission line and hence the magnitude peaks produced by the transmission line. For measurements with relatively small phase group delays, and consequently having relatively small T_{tm} values, such as those in set B, M_{mT} was sometimes set such that a peak due to the transmission line corresponded to a peak in the measurements in the vicinity above 10 kHz. For measurements with relatively large phase group delays, and hence relatively large T_{tm} values, such as those in set A, however, the peaks due to the transmission line typically began to occur at frequencies that were too low to match features in the data. In these cases it was typically necessary to assign a relatively small value to M_{mT} so as to reduce the size of these peaks, though an effort was also made to keep the value of M_{mT} as high as possible in such cases rather than allowing it to approach zero.

Once all of the masses were introduced, T_{tm} was adjusted to restore $U_{scT}/P_{\Delta tm}$ phase agreement.

6. Malleus and ossicular joint damping: R_{mT}, R_{imjT}, and R_{isjT}

For the sixth step, R_{mT}, R_{imjT}, and R_{isjT} were introduced. R_{mT} was set first, to bring the $U_{scT}/P_{\Delta tm}$ transfer function magnitude into better mid-frequency agreement with the data. R_{mT} also has a significant effect on the real part of Z_{ocT}, and so was used to reduce the size of magnitude peaks due to the transmission line.

R_{isjT} was set next, to adjust the slope of the magnitude roll-off above the Z_{iT} and K_{isjT} impedance breakpoint. A larger R_{isjT} both reduces the magnitude peak that sometimes occurs around that breakpoint, and makes the subsequent roll-off less steep. In some cases R_{isjT} was kept small so as to minimize its effect on the peak and the roll-off when they already agreed well with the data.

R_{imjT} was set last, and was used to reduce the volume velocity lost to the IMJ branch at higher frequencies, by causing the IMJ impedance magnitude to level off at high frequencies rather than continue to decrease.

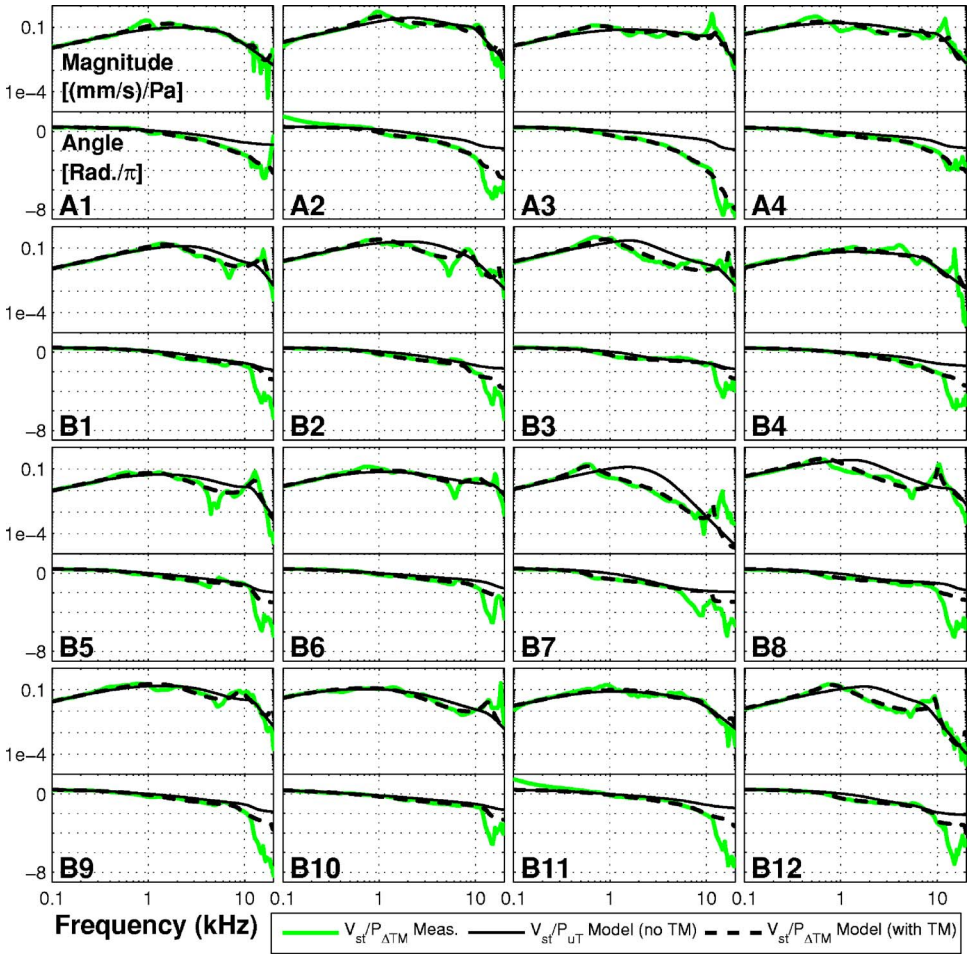


FIG. 3. (Color online) $V_{st}/P_{\Delta tm}$ measured data and model fits with and without the tympanic membrane (TM) for all 16 ears. The ear name is shown in the lower-left of each panel, with the upper half of each panel containing the magnitude curves in units of (mm/s)/Pa, and the lower half containing the phase curves in units of radians normalized by π. The frequency range for each panel is 0.1–20 kHz. The thick solid lines represent the $V_{st}/(P_{ec}\text{-}P_{mec})$ transfer functions measured for each ear, which are referred to throughout this paper as "$V_{st}/P_{\Delta tm}$" in accordance with the simplifying assumptions in section C of Fig. 1. The thick dashed lines represent the $V_{st}/P_{\Delta tm}$ model curves fit to the measurements, and the thin solid lines represent the models curves *without* the TM, which are equivalent to V_{st}/P_{uT} (see section A of Fig. 2). The effects of the transmission line representation of the TM on the magnitude and phase can be seen for each ear by comparing the versions of the model with and without the TM.

7. Additional tweaking

After setting all of the parameters, it was often useful to readjust some of the values in order to improve certain aspects of the fit using the various strategies described above.

C. Comparisons to other models

A version of the present model, fit to the $V_{st}/P_{\Delta tm}$ mean of the 16 ears in sets A and B and referred to as "AB," is compared to three previously published lumped-element models from the Kringlebotn (1988), Goode *et al.* (1994), and Feng and Gan (2004) studies. For the Kringlebotn (1988) and Goode *et al.* (1994) models, only the TM, ossicular chain, and cochlea portions of the models were included. The ear canal and middle-ear cavity portions of those models were excluded so as to enable direct comparisons to the present AB model. While the Feng and Gan (2004) model did not contain a middle-ear cavity model, its ear canal representation was removed (consisting of the M1, K2, and C2 elements in their model) to allow it to be directly compared to the present model.

All three previous models are displayed for the $V_{st}/P_{\Delta tm}$ transfer function and the Z_{tmoc} input impedance, but only the Feng and Gan (2004) model is shown for V_i/V_u and V_{st}/V_i ratios since the Kringlebotn (1988) and Goode *et al.* (1994) models do not explicitly differentiate the incus and umbo velocities.

IV. RESULTS

A. $V_{st}/P_{\Delta tm}$ model fits

1. Individual fits

Figure 3 displays, for all 16 ears, the $V_{st}/P_{\Delta tm}$ measurements (thick solid lines) and model fits without and with the transmission line (thin solid and thick dashed lines respectively). The ear name (i.e., A1–A4 and B1–B12) is labeled within each subplot, with magnitudes occupying the upper half [in units of (mm/s)/Pa], and phases occupying the lower half (in units of radians normalized by π) of each subplot. The model curves without the transmission line are equivalent to V_{st}/P_{uT}, and are included to show how the transmission line affects the magnitude and phase of the model for each ear.

The individual model fits reveal how well the model is able to match the magnitude and phase features of each $V_{st}/P_{\Delta tm}$ measurement. While the overall features of the measured magnitude and phase curves are well matched by the model, it is typical for the model curves to partially skip over some of the localized peaks and dips present in the measurements, e.g., the magnitude dip in B2 above 5 kHz, or the magnitude peak in A1 around 1 kHz.

For some ears (e.g., A1–A4, B4, B6, and B11), the transmission line mostly affects the phase of the model, by providing additional group delay. For many of the other ears, however, the effects of the transmission line are more pronounced in how it affects the shape of the magnitude curve.

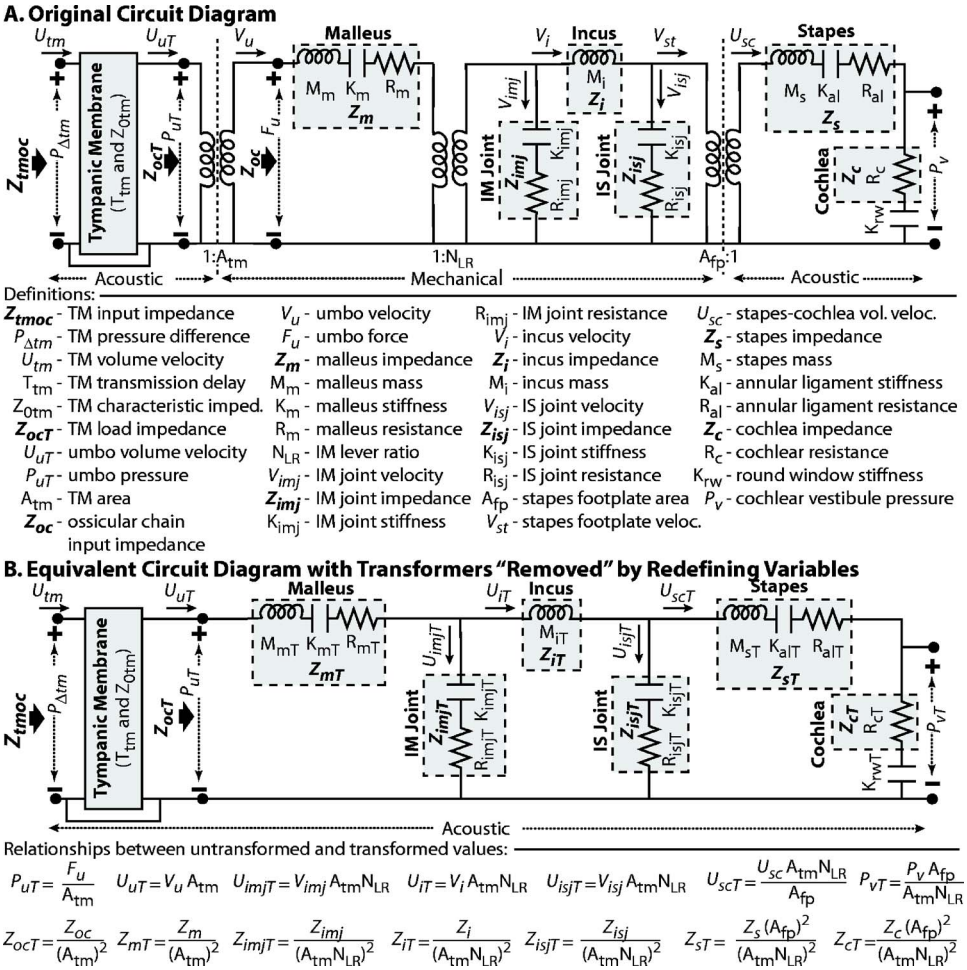


FIG. 2. (Color online) Model of the TMOC block depicted in sections B and C of Fig. 1, in its original form (A) and an alternate form in which the effects of the three transformers have been absorbed into the model parameters (B). The tympanic membrane is represented as a one-dimensional acoustic transmission line with associated delay T_{tm} and characteristic impedance Z_{0tm}. The ossicular chain and cochlea are represented as a network of electrical circuit elements with acoustic or mechanical interpretations (see definitions within the figure). Section A depicts the TMOC model adapted from Puria and Allen (1998), with three transformers representing the effects of the TM area (A_{tm}), the effective lever ratio of the malleus-incus complex (N_{LR}), and the area of the stapes footplate (A_{fp}). Section B shows an alternate version of the model in section A with the transformers "removed" such that all variables are represented by their acoustic equivalents as seen from the left of all three transformers. Variables redefined in this way have a "T" appended to their subscripts to indicate that they have been "Transformed," and equations are shown under the panel for converting between transformed and untransformed versions of the variables. In the case of impedance blocks, such as Z_{mT} for the malleus, the same conversion equation applies for all elements within the block (i.e., M_{mT}, K_{mT}, and R_{mT} in this case). By redefining the variables in this manner, it becomes possible to make direct quantitative comparisons between variables that were previously located on opposite sides of one or more transformer. For this reason the transformed version of the circuit (section B) was used for much of the model-fitting procedure. To convert the transformed parameter values to their untransformed equivalents, values of 6×10^{-5} m^2 for A_{tm}, 1.3 for N_{LR}, and 3.14×10^{-6} m^2 for A_{fp} are used.

2. Comparisons of model and data ensembles

Figure 4 shows mean and mean±standard error of the mean (SEM) curves of the $V_{st}/P_{\Delta tm}$ measurements (circular markers) and curves of the model with the TM (up-pointing triangular markers) for set A [Fig. 4(a)], and set B [Fig. 4(b)], along with the corresponding V_{st} noise floor measurements normalized by $P_{\Delta tm}$. In [Fig. 4(c)], the model fit to the mean of all 16 ears (AB) is shown, with and without the TM (up-pointing and down-pointing triangular markers, respectively), along with the mean and mean±SEM curves for the 16 ears in sets A and B (circular markers). For comparison, a V_{st}/P_{ec} mean magnitude curve, based on measurements made by Huber *et al.* (2001) on seven living ears, is also shown in Figs. 4(a)–4(c) (square-shaped markers), for which magnitude adjustments were made by Chien *et al.* (2006) on frequencies below approximately 2 kHz to account for possible methodological differences between living ear and temporal bone ear studies. Also for purposes of comparison, $V_{st}/P_{\Delta tm}$ model curves from three other studies are shown in Fig. 4(c): Kringlebotn, 1988 (x-shaped markers); Goode *et al.*, 1994 (no markers); and Feng and Gan, 2004 (+-shaped markers). For set A, the mean model and measurement magnitudes (upper half of Fig. 4(a)) show generally good agreement over the full frequency range, with the exception of the peak in the data near 12 kHz (due to the inability to match the peaks seen in A3 and A4: see Fig. 3). In the measurements for set A the signal to noise ratio for the mean magnitude falls below 6 dB above 11.5 kHz, however. Some smaller differences between the means can be attributed to the inability to fully match the peaks and dips in the 0.8–2 kHz vicinity, as well as some other cases where peaks and dips in the model do not correspond to features in the

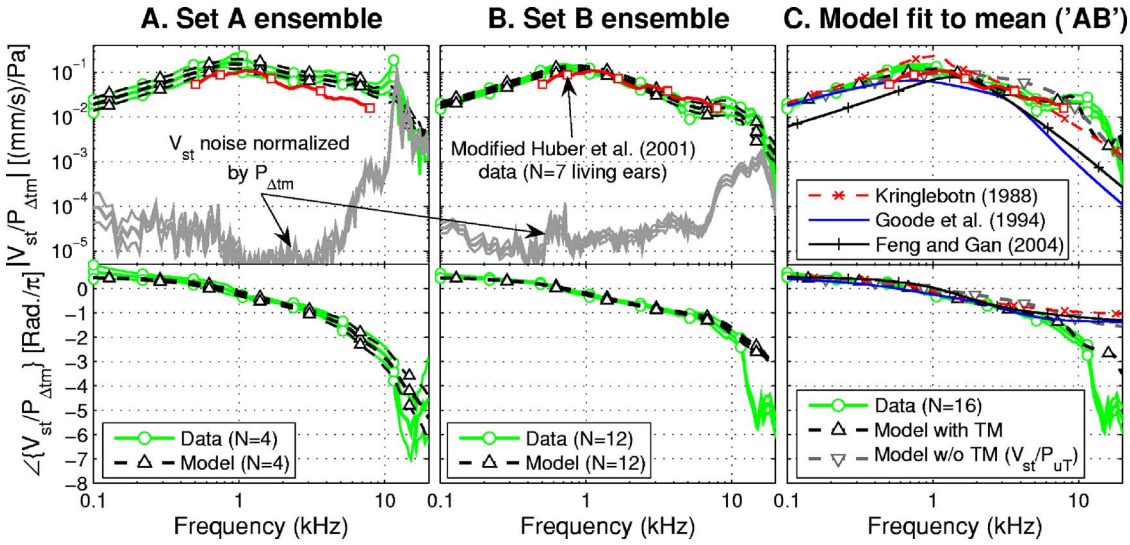


FIG. 4. (Color online) Model $V_{st}/P_{\Delta tm}$ curves compared to the corresponding measured curves for the set A and set B ensembles (columns A and B, respectively), as well as comparisons between the ensemble of the 16 measurements in sets A and B to a model fit to the mean of that ensemble (column C). The $V_{st}/P_{\Delta tm}$ measurements in columns A, B, and C are shown with circular markers in terms of the mean and mean±SEM (standard error of the mean) of each ensemble, based on 4, 12, and 16 ears, respectively. The upper half of each panel contains magnitude curves in units of (mm/s)/Pa, while the lower half contains phase curves in units of radians/π. The corresponding ensembles of models are shown in columns A and B in terms of the mean and mean±SEM with up-pointing triangular markers. A V_{st}/P_{ec} magnitude mean based on measurements from Huber *et al.* (2001) on seven living subjects, and modified below approximately 2 kHz by Chien *et al.* (2006) in an attempt to account for methodological differences, is shown in columns A, B, and C with square-shaped markers. Columns A and B also contain the V_{st} noise floor measurements normalized by $P_{\Delta tm}$ (also in terms of the mean and mean±SEM) for the set A and set B studies, respectively. Column C contains plots from a single model, referred to as AB, that was fit to the mean of the 16 $V_{st}/P_{\Delta tm}$ measurements in sets A and B. The AB $V_{st}/P_{\Delta tm}$ model curve is shown with up-pointing triangular markers, and a version of this model without the TM is also shown (equivalent to V_{st}/P_{uT}) with down-pointing triangular markers. Additional $V_{st}/P_{\Delta tm}$ equivalent curves from three previous models are also shown in column C for comparison with the AB model: Kringlebotn (1988) with *x*-shaped markers, Goode *et al.* (1994) with no markers, and Feng and Gan (2004) with +-shaped markers.

data. The modified Huber *et al.* (2001) magnitude curve (square markers) appears a little below the set A model and data, with a steeper roll-off above 2 kHz. The phases of set A (lower half of Fig. 4(a)) show good agreement between the model and data up to around 12 kHz, with the exception of below 0.2 kHz, where the phase mean rises up due to a feature in the data for A2 (Fig. 3). The phase difference above 12 kHz can be attributed in part to the steep drop in phase seen in A2 and A3 which could not be fully matched by the model.

For set B, the mean magnitudes (upper half of Fig. 4(b)) also show generally good agreement between the model and measurements, with the main exception being that the model does not fully capture some of the peaks (e.g., B3 and B4) and dips (e.g., B1, B2, B4, B5, B6, B8, and B9) found in the 4–10 kHz vicinity for many of the ears in set B (Fig. 3). Despite this, due to the generally lower T_{tm} values in set B, it was often possible to utilize the TM to convincingly model some of the high frequency magnitude peaks where it was more challenging in set A. Due to the use of an improved sound source, the mean magnitude for set B stays at least 6 dB above the noise for the entire frequency range. The modified Huber *et al.* (2001) magnitude curve (square markers) bears a close resemblance to the set B data and model, especially above 1 kHz. Phase agreement between the means (lower half of Fig. 4(b)) is good up to around 12 kHz, with some small differences appearing above 6 kHz. Above 12 kHz the measured phase exhibits a steep drop that the model has trouble representing. In the case of B7, which features an unusually early and steep magnitude roll-off, the phase drops significantly below the model as early as 5 kHz. The magnitude dips found in the 4–10 kHz vicinity for several ears, which are difficult to model fully, typically correspond in frequency with shifts in the phase that are similarly difficult to model. B11 exhibits a similar low frequency phase rise to that of A2, but it has a smaller effect on the mean since all of the 11 other ears lack such a feature.

The magnitude of the model fit to the mean of all 16 ears (AB) exhibits reasonably good agreement with the mean below 10 kHz (upper half of Fig. 4(c)). Above 10 kHz the model magnitude drops off sooner than the mean. For this model fit, the transmission line has a noticeable effect on both the magnitude and the phase, which can be seen by comparing the model curves with the TM (up-pointing triangular markers) and without the TM (down-pointing triangular markers) in the upper and lower halves of Fig. 4(c). The modified magnitude curve of Huber *et al.* (2001) (square markers) bears an even closer resemblance to the 16 ear data mean and the AB model magnitude than the set B data and model. The parameters for the AB fit are based on those for ear B11, with different values for Z_{0tm} and M_{mT}, and a slightly different value for R_{mT}, but all other parameters the same (see Fig. 7 and Table I). The model curves from other studies are described in the Discussion section.

B. Velocity ratios

Figure 5 shows the measured V_{st}/V_u (from Aibara *et al.*, 2001) and V_i/V_u (from Willi *et al.*, 2002) mean curves (solid lines with down-pointing and right-pointing triangular mark-

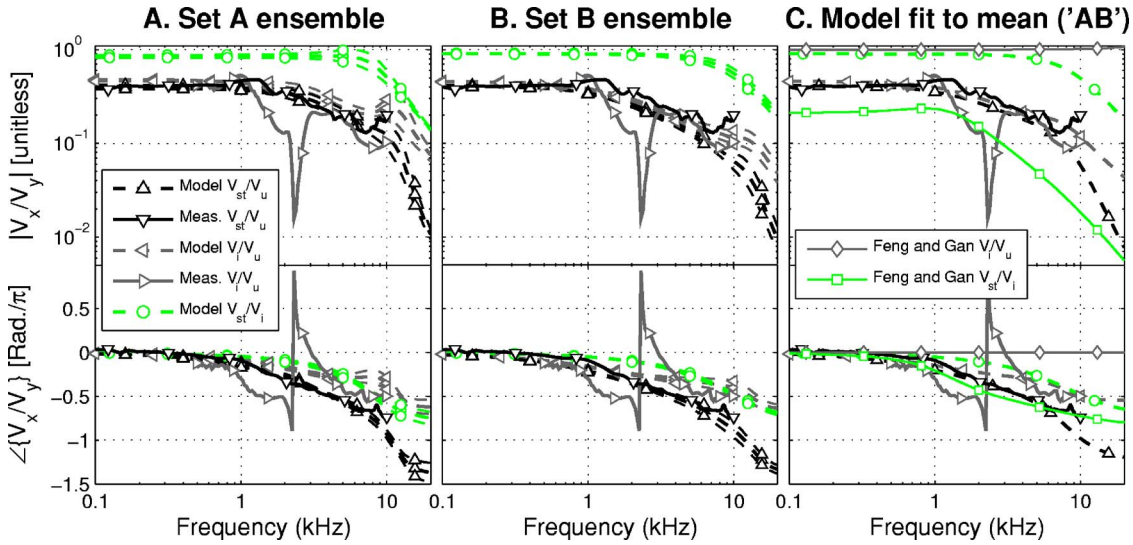


FIG. 5. (Color online) Model and measured V_x/V_y velocity ratios for the set A ensemble (A), the set B ensemble (B), and the AB model based on the 16 ear mean (C), where V_x and V_y, respectively, refer to V_{st} and V_u, V_{st} and V_i, or V_i and V_u [refer to Fig. 2(a) for definitions]. The upper half of each column contains unitless velocity ratio magnitudes, while the lower half contains phase curves in units of radians/π. Measured V_{st}/V_u data from Aibara *et al.* (2001) are shown in all three columns with solid lines and down-pointing triangular markers, while the corresponding V_{st}/V_u model curves are shown with dashed lines and up-pointing triangular markers. Similarly, the measured V_i/V_u data from Willi *et al.* (2002) are shown with solid lines and right-pointing triangular markers, while the corresponding V_i/V_u model curves are shown with dashed lines and left-pointed triangular markers. Model V_{st}/V_i curves are also shown with dashed lines and circular markers. Column C also contains equivalent V_i/V_u and V_{st}/V_i model curves from the Feng and Gan (2004) model with thin lines and diamond and square-shaped markers, respectively.

ers, respectively) along with the corresponding mean and mean±SEM curves from the models (dashed lines with up-pointing and left-pointing triangular markers, respectively) for set A [Fig. 5(a)] and from the models for set B [Fig. 5(b)], as well as curves for the AB model fit to the mean of all 16 ears [Fig. 5(c)]. Model curves for V_{st}/V_i are also shown (dashed lines with circular markers). In [Fig. 5(c)], curves of V_i/V_u (thin solid lines with diamond-shaped markers) and V_{st}/V_i (thin solid lines with square-shaped markers) are also shown for the Feng and Gan (2004) model for comparison. Note that if V_{isj} [see Fig. 2(a)] were zero in the present model, then V_{st}/V_i would be equal to 1, whereas (because of the transformer before the IMJ branch) if V_{imj} were zero, then V_i/V_u would be equal to $1/N_{LR}$. V_i/V_u and V_{st}/V_i are also sometimes referred to as the IMJ and ISJ "slippage," respectively. For the V_i/V_u data of Willi *et al.* (2002), which exhibits a large magnitude dip and phase jump between 1 and 3 kHz for unknown reasons, the model comparison focused on the regions below 1 kHz and above 3 kHz.

For set A, the model V_{st}/V_u magnitude [upper half of Fig. 5(a)] lies close to the Aibara *et al.* (2001) curve for most of its range. Because the model V_{st}/V_i mean is less than 1 for all frequencies and V_{st}/V_u is equal to the product of V_i/V_u and V_{st}/V_i, the model V_i/V_u mean lies above the model V_{st}/V_u mean for all frequencies. This causes some small disagreements with the data of Willi *et al.* (2002) (not considering the 1–3 kHz region) when it dips below the V_{st}/V_u data. The model V_{st}/V_u phase [lower half of Fig. 5(a)] also closely matches the Aibara *et al.* (2001) curve, while the model V_i/V_u phase matches the data of Willi *et al.* (2002) more approximately.

For set B [Fig. 5(b)], the model curves are similar to those for set A, but shifted down slightly such that the V_i/V_u curve most closely matches the data rather than the V_{st}/V_u curve. For the AB model fit to the 16 ear mean [Fig. 5(c)], all three model velocity ratios bear a close resemblance to those from set B. The Feng and Gan (2004) V_i/V_u and V_{st}/V_i model curves, generally inconsistent with physiological measurements, are described in the Discussion section.

C. Input impedance

Figure 6 contains Z_{tmoc} curves from the present model (circular markers with dashed lines) for the set A ensemble of models [Fig. 6(a)], the set B ensemble of models [Fig. 6(b)], and the AB model [Fig. 6(c)]. In [Fig. 6(c)], the AB model is also shown without the TM, which corresponds to Z_{ocT} in Fig. 2 (circular markers with a dotted line). In Figs. 6(a) and 6(b), converted impedance measurements based on previously published results in other studies are also shown. An approximation of Z_{tmoc} based on data in the Voss *et al.* (2000) study (square-shaped markers), only available up to 4 kHz, was calculated by subtracting their reported middle-ear cavity impedance model from the mean of their reported Z_{tm} measurements. Two Z_{ec} measurements from living ears, available above 1 kHz, based on reflectance data reported in Farmer-Fedor and Rabbitt (2002; up-pointing triangular markers), and impedance data reported in Hudde (1983; diamond-shaped markers) are also shown. When computing both of these Z_{ec} magnitude curves from the published results, it was necessary to multiply by a characteristic impedance value, which was somewhat arbitrarily set to 3.28×10^7 Kg/(s m^4) (see Sec. A 4 in the Methods section), so the exact magnitudes of these curves is not certain. The two Z_{ec} measurements are affected by middle-ear cavities, though these tend to have relatively small effects on the total input impedance above 1 kHz (Voss *et al.* 2000), and by a (pre-

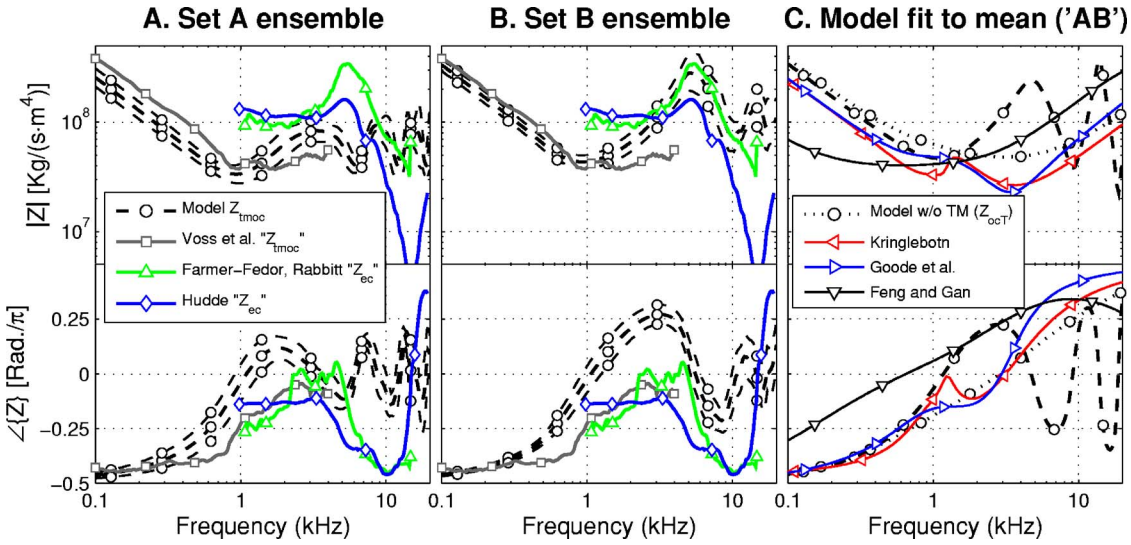


FIG. 6. (Color online) Model Z_{tmoc} curves for the set A ensemble (A), the set B ensemble (B), and the AB model based on the 16 ear mean (C), with converted Z_{tmoc} and Z_{ec} curves from other studies and models shown for comparison. The upper half of each column contains acoustic impedance magnitudes in units of Kg/(s m^4), while the lower half of each column contains phase curves in units of radians/π. The Z_{tmoc} model curves from the present study are shown with dashed lines and circular markers in all three columns. Column C also contains the input impedance of the present model without the TM, equivalent to Z_{ocT} in Fig. 2, with dotted lines and circular markers. Columns A and B also contain: 1. a mean measurement from Voss *et al.* (2000) of their equivalent Z_{tm} mean [see Fig. 1(b)] with their model of Z_{cav} subtracted to yield an approximation of Z_{tmoc} (solid lines and square-shaped markers); 2. a mean curve corresponding approximately to Z_{ec} [see Fig. 1(b)] based on reflectance measurements from living ears in Farmer-Fedor and Rabbitt (2002) (solid lines and up-pointing triangular markers); and 3. a mean curve also corresponding approximately to Z_{ec} based on the real and imaginary parts of impedance curves from living ears published in Hudde (1983) (solid lines and diamond-shaped markers). The impedance curves based on other studies were not used to assist with the model fitting procedure and are provided for purposes of comparison only. In addition to the AB Z_{tmoc} and Z_{ocT} curves, column C also contains three Z_{tmoc} curves from the Kringlebotn (1988; left-pointing triangular markers), Goode *et al.* (1994; right-pointing triangular markers), and Feng and Gan (2004; down-pointing triangular markers) models.

sumably 2–3 mm long) section of the ear canal, which should be considered when comparing with the other curves. Model input impedance curves from other studies, corresponding to Z_{tmoc}, are also shown in Fig. 6(c) for the Kringlebotn (1988) model, the Goode *et al.* (1994) model, and the Feng and Gan (2004) model, for comparison (see the Discussion section).

The Voss *et al.* (2000) "Z_{tmoc}" input impedance magnitude typically lies within 2 dB of the present set B and AB Z_{tmoc} model curves below 1.5 kHz, while the set A model typically lies within 4 dB of the modified Voss *et al.* curve within that range. Above 1.5 kHz the present models rise above the Voss *et al.* curve up to the 4 kHz limit for those measurements. The magnitudes of the present model curves rise and fall periodically above 1 kHz due to the transmission line representation of the TM, with set A's [Fig. 6(a)] more closely spaced peaks beginning at a lower frequency and not rising as high as those of set B [Fig. 6(b)] and AB [Fig. 6(c)]. The largest peak-to-dip magnitude variation in the AB Z_{tmoc} curve [upper half of Fig. 6(c)] due to these oscillations is 24 dB, which occurs between 14 and 19 kHz, and the largest magnitude variation below 10 kHz is 16 dB (between 4.8 and 9.2 kHz). The Farmer-Fedor and Rabbitt (2002), and Hudde (1983) magnitude curves also exhibit peaking and dipping effects, at similar frequencies to the set B mean [upper half of Fig. 6(b)].

The phase of the adjusted Voss *et al.* (2000) measurements [lower half of Figs. 6(a) and 6(b)] begins near −0.5 radians/π, implying that the impedance behaves mostly like a stiffness element, then slowly moves closer to 0, implying that the impedance behaves more resistively, before beginning to drop again around 2.5 kHz. For the three Z_{tmoc} model curves from the present study [lower half of Figs. 6(a)–6(c)], the phases also begin near −0.5 radians/π, then increase to values above 0, which implies some mass-like behavior, and continue with oscillations around 0 radians/π. As was the case for the magnitudes, the oscillations in the phase for set A start at a lower frequency, are more closely spaced, and do not peak as high as those for the other two curves. The Farmer-Fedor and Rabbitt (2002), and Hudde (1983) phase curves are fairly similar to the Voss *et al.* (2000) phase over their 1–4 kHz overlapping frequency range, and both decrease again toward −0.5 radians/π by 10 kHz. The Hudde (1983) curve then proceeds to rise up to around 0.4 radians/π just before 20 kHz.

D. Model parameters

1. Transformed parameters

Figure 7 graphically displays the transformed parameters [see Fig. 2(b)] used for each of the individual models shown in Fig. 3, as well as transformed parameters corresponding to the AB model fit to the mean of all 16 ears [see Figs. 4(c), 5(c), and 6(c)]. Transformed stiffness elements occupy the top panel, with transformed resistances (including Z_{0tm}, since it also has units of resistance), transformed masses, and the TM delay occupying each of the lower panels, respectively. Parameter values are plotted from left to right for ears A1–B12, followed by values for the AB model, with a unique symbol used to distinguish the parameters be-

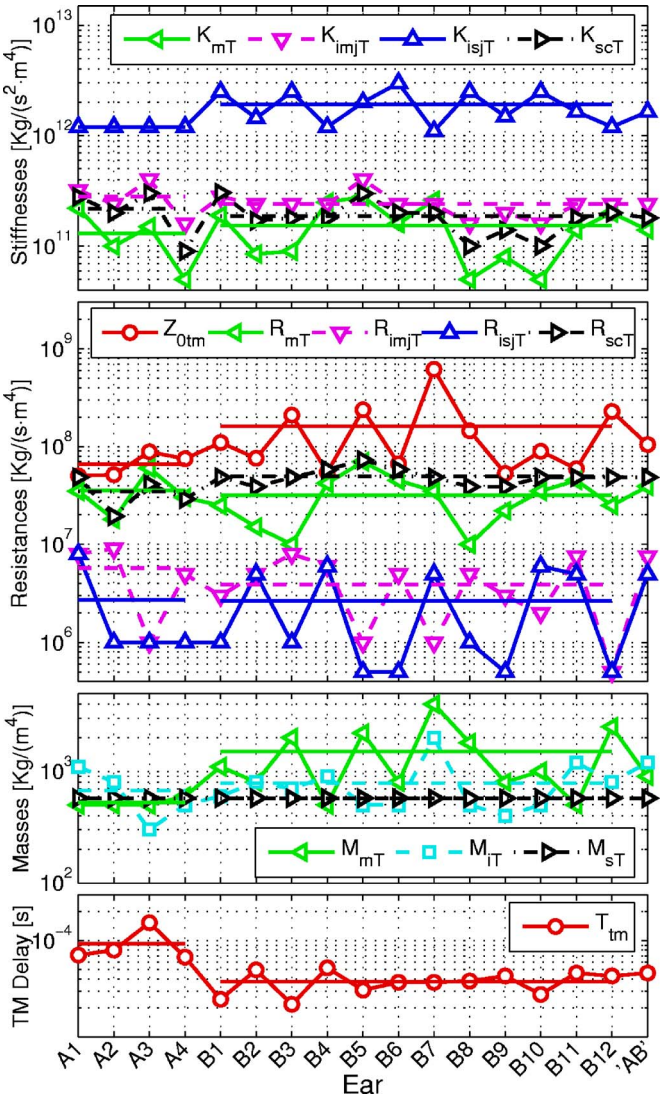


FIG. 7. (Color online) Transformed versions of the model parameters [see Fig. 2(b)] plotted for each ear (labeled A1 through B12 from left to right) plus those for the AB model, which was fit to the $V_{st}/P_{\Delta tm}$ mean of all 16 ears. The top panel shows transformed stiffness values in units of Kg/(s^2 m^4), the next panel shows transformed resistance values in units of Kg/(s m^4), followed by transformed mass values in units of Kg/m^4, and finally TM delay values in units of seconds. Each parameter is plotted using a distinct marker within a given panel, and the same marker is used in different panels for elements belonging to the same impedance block (e.g., K_{mT}, R_{mT}, and M_{mT} are all plotted with a left-facing triangle). Parameter means are also plotted for ears A1–A4 and B1–B12 as shown by horizontal lines extending over the relevant ear ranges with the same line styles used for plotting the individual parameter values. K_{scT} and R_{scT} are the total stiffness and resistance, respectively, for the stapes-cochlea branch, and are equal to $K_{alT}+K_{rwT}$ and $R_{alT}+R_{cT}$, respectively. To convert the displayed transformed parameter values to their untransformed equivalents in the context of Fig. 2(a), refer to the equations at the bottom of the Fig. 2(b), and use values of 6×10^{-5} m^2 for A_{tm}, 1.3 for N_{LR}, and 3.14×10^{-6} m^2 for A_{fp}.

longing to different sections of the model. For example, left-pointing triangles are used to represent the transformed stiffness, resistance, and mass of the malleus, while down-pointing triangles are used to represent the transformed stiffness and resistance of the IMJ. For a given parameter, lines connect the plotted values for the different ears, with the line style varied in some cases to make the lines easier to tell apart. Horizontal lines are plotted for each parameter, with a matching line style, to indicate the mean value of that parameter for set A (the line extending from A1–A4), and for set B (the line extending from B1–B12). To convert these transformed values into their untransformed versions [Fig. 2(a)], one can use the equations listed at the bottom of Fig. 2(b).

2. Untransformed parameters

While Fig. 7 graphically shows the transformed parameters of Fig. 2(b) for the individual ears and AB model, Table I shows the average value (and SEM) of each untransformed parameter in Fig. 2(a) from the four models in set A, the 12 models in set B, the 16 models in sets A and B together, and the 1 model based on the mean of the 16 ears, thus allowing each parameter to be compared numerically between different model ensembles. The AB model parameters can be used to reconstruct the AB model curves in Figs. 4(c), 5(c), and 6(c). Recall that K_{sc} is equal to $K_{al}+K_{rw}$, and R_{sc} is equal to $R_{al}+R_c$. To calculate the cochlear vestibule pressure (P_v) from the present model, it is necessary to decide how to split K_{sc} and R_{sc} into their separate parts, as described in Secs. B 1 and B 3 of the Discussion section below.

V. DISCUSSION

A. Comparisons to previous models

The $V_{st}/P_{\Delta tm}$ model curves from other studies, based on versions of the models without included middle-ear cavity or ear canal models, are shown in Fig. 4(c). With the exception of the peak in the 1 kHz vicinity and the roll-off above 6 kHz, the Kringlebotn (1988) model magnitude (thin solid lines with *x*-shaped markers) compares favorably to the present AB model with the TM (dashed lines and up-pointing triangular markers) and data (solid lines and circular markers). The Goode *et al.* (1994) model magnitude (thin solid lines with no markers) lies below the present model and data with the exception of very low frequencies, but it nonetheless bears a reasonable resemblance to the present data and AB model with the TM below 3 kHz. The Feng and Gan (2004) model magnitude (thin solid lines with +-shaped markers) lies well below the present model and data for most of the frequency range, with the exception of approximately 1–3 kHz. The phase for the present AB model with the TM resembles the data up to around 12 kHz, whereas the Kringlebotn (1988), Feng and Gan (2004), and Goode *et al.* (1994) models only follow the data phase up to 2, 4, and 6 kHz, respectively, above which the data phase decreases below the models. Overall, the AB $V_{st}/P_{\Delta tm}$ model curve performs better than the other three models in terms of matching the data magnitude and phase.

In Fig. 5(c), model curves are plotted of V_i/V_u (thin solid lines with diamond-shaped markers) and V_{st}/V_i (thin solid lines with square-shaped markers) from the Feng and Gan model (2004). Analogous curves from the Goode *et al.* (1994) and Kringlebotn (1988) models could not be created since those models do not explicitly distinguish between incus and umbo velocities. The Feng and Gan (2004) V_i/V_u curve has a magnitude very close to 1, implying that the malleus-incus lever ratio is not accounted for in the model

TABLE I. Untransformed parameter value means with SEM (standard error of the mean) for the models in set A, set B, the combination of sets A and B, and the AB model based on fitting the 16 ear $V_{st}/P_{\Delta tm}$ mean from sets A and B. Note that K_{sc} equals $K_{al}+K_{rw}$, and R_{sc} equals $R_{al}+R_c$.

Parameter		Set A (4 ears)		Set B (12 ears)		Sets A and B (16 ears)		Fit to 16 ear
Name	Units	Mean	SEM	Mean	SEM	Mean	SEM	Mean (AB)
T_{tm}	s	9.31×10^{-5}	2.04×10^{-5}	3.76×10^{-5}	0.28×10^{-5}	5.15×10^{-5}	0.80×10^{-5}	4.61×10^{-5}
Z_{0tm}	Kg/(s m^4)	6.62×10^{7}	0.92×10^{7}	16.2×10^{7}	4.6×10^{7}	13.8×10^{7}	3.6×10^{7}	10.5×10^{7}
A_{tm}	m^2	6.0×10^{-5}	0	6.0×10^{-5}	0	6.0×10^{-5}	0	6.0×10^{-5}
M_m	Kg	1.89×10^{-6}	0.09×10^{-6}	5.40×10^{-6}	1.08×10^{-6}	4.52×10^{-6}	0.89×10^{-6}	3.24×10^{-6}
K_m	Kg/s^2	468	131	551	86	530	71	504
R_m	Kg/s	0.129	0.032	0.114	0.018	0.118	0.015	0.140
N_{LR}	Unitless	1.3	0	1.3	0	1.3	0	1.3
K_{imj}	Kg/s^2	1.70×10^{3}	0.31×10^{3}	1.46×10^{3}	0.11×10^{3}	1.52×10^{3}	0.11×10^{3}	1.46×10^{3}
R_{imj}	Kg/s	3.50×10^{-2}	1.09×10^{-2}	2.38×10^{-2}	0.45×10^{-2}	2.66×10^{-2}	0.43×10^{-2}	4.56×10^{-2}
M_i	Kg	4.11×10^{-6}	1.07×10^{-6}	4.77×10^{-6}	0.78×10^{-6}	4.60×10^{-6}	0.63×10^{-7}	7.30×10^{-6}
K_{isj}	Kg/s^2	7.30×10^{3}	0	11.7×10^{3}	1.15×10^{3}	10.6×10^{3}	1.0×10^{3}	10.0×10^{3}
R_{isj}	Kg/s	1.67×10^{-2}	1.07×10^{-2}	1.62×10^{-2}	0.43×10^{-2}	1.64×10^{-2}	0.40×10^{-2}	3.04×10^{-2}
A_{fp}	m^2	3.14×10^{-6}	0	3.14×10^{-6}	0	3.14×10^{-6}	0	3.14×10^{-6}
M_s	Kg/m^4	3.55×10^{5}	0	3.55×10^{5}	0	3.55×10^{5}	0	3.55×10^{5}
K_{sc}	Kg/(s^2 m^4)	1.34×10^{14}	0.29×10^{14}	1.16×10^{14}	0.14×10^{14}	1.21×10^{14}	0.11×10^{14}	1.11×10^{14}
R_{sc}	Kg/(s m^4)	2.13×10^{10}	0.40×10^{10}	3.06×10^{10}	0.17×10^{10}	2.83×10^{10}	0.19×10^{10}	2.99×10^{10}

and that there is very little velocity lost in that joint. The Feng and Gan (2004) V_{st}/V_i magnitude lies much lower, on the other hand, which indicates that practically all of the velocity is lost in the IS joint rather than the IM joint. The roles are reversed for the present model such that the majority of the velocity is lost in the IM joint and much less is lost in the IS joint, which is consistent with measurements.

Figure 6(c) shows the equivalent Z_{tmoc} model curves for the Kringlebotn (1988; thin solid lines with left-pointing triangular markers), Goode *et al.* (1994; thin solid lines with right-pointing triangular markers), and Feng and Gan (2004; thin solid lines with down-pointing triangular markers) models. The present AB Z_{tmoc} model with the TM (dashed line with circular markers) is similar to the set B model mean [Fig. 6(b)], and therefore also shares a close resemblance with the adjusted Voss *et al.* (2000) magnitude curve below 1.5 kHz. The other model magnitudes all lie below the present Z_{tmoc} model magnitude over nearly the entire range below 7 kHz, above which their magnitudes periodically rise above the oscillations in the AB Z_{tmoc} model with the TM. The AB Z_{ocT} model magnitude without the TM (dotted line and circular markers), decreases smoothly until 3–4 kHz, then proceeds to smoothly rise for the rest of the frequency range, which contrasts significantly from the oscillatory behavior of the AB Z_{tmoc} magnitude which includes the effects of the TM model. The AB Z_{tmoc} phase resembles the Goode *et al.* (1994) and Kringlebotn (1988) phases up to 0.8 and 1.3 kHz, respectively, above which the AB Z_{tmoc} model proceeds to oscillate above and below 0 radians/π while the other models rise and stay above 0 radians/π above 3 kHz (0.6 kHz for the Feng and Gan phase).

The AB Z_{tmoc} model with the TM exhibits low frequency agreement with the adjusted Voss *et al.* (2000) magnitude below 1.5 kHz, and phase below 0.3 kHz that largely surpasses the other models, and exhibits some qualitatively similar magnitude features above 3 kHz and phase features above 4.5 kHz to the Farmer-Fedor and Rabbitt (2002) and Hudde (1983) measurements that are not present in the other models, including the AB model without the TM (Z_{ocT}). The phases of the models without the transmission line representation of the TM all rise and stay well above 0 radians/π for frequencies higher than 0.6–3 kHz, which indicates mass-like behavior inconsistent with the Z_{ec} measurements, while the AB Z_{tmoc} model phase oscillates between mass and stiffness-like behavior for high frequencies.

B. Model parameters

1. Stiffness elements

K_{isjT} is by far the highest stiffness value of those assigned (Fig. 7), mainly to make the V_i/V_u and V_{st}/V_u transfer functions similar in magnitude at low frequencies, as implied by the Willi *et al.* (2002) and Aibara *et al.* (2001) data, by bringing the V_{st}/V_i transfer function close to 1. K_{mT} and K_{scT} are varied together initially to set the low frequency volume velocity, which typically places them in the same vicinity, with K_{imjT} varied next to get the velocity transfer functions to agree with the data. To restore the low frequency volume velocity agreement with the data after fitting the velocity ratios, K_{mT} is then typically lowered, since K_{mT} has no effect on the velocity transfer functions, which is why in Fig. 7 the K_{mT} mean is less than that for K_{scT}.

The K_{scT} variable, equal to the sum of K_{alT} and K_{rwT}, is used in the fitting procedure instead of the separate stiffnesses since both stiffnesses are located on the same circuit branch and cannot be distinguished by curve-fitting alone. In Puria (2003), the reverse middle ear stiffness, K_{mer}, which is measured from the cochlea outward to the stapes and is an approximate measure of the untransformed K_{al} parameter, is reported as 0.81×10^{14} Kg/(s m^4). If K_{al} is equated to K_{mer} from Puria (2003) for all models, then the average value for K_{rw} becomes 0.53×10^{14} for set A, 0.35×10^{14} for set B, 0.4×10^{14} for the mean of sets A and B, and 0.3×10^{14} for the AB model based on the 16 ear mean. In contrast, the

value of K_{rw} reported for the Puria and Allen (1998) model for the cat is 0.12×10^{14}, which is slightly higher than the 0.1×10^{14} value reported by Lynch *et al.* (1982) for the cat.

2. Characteristic impedance Z_{0tm}

The transmission line characteristic impedance values tend on average to be higher for set B than for set A (see Fig. 7 and Table I). For the most part Z_{0tm} was set automatically to match the magnitude of the ossicular chain input impedance, Z_{ocT}, at 10 kHz. Beyond that, Z_{0tm} was sometimes increased further so as to produce desired effects in the P_{gain} function [Eq. (2)] to shape the magnitude of the $V_{st}/P_{\Delta tm}$ transfer function. Because Z_{ocT} at high frequencies is significantly affected by the size of the malleus mass, and Z_{0tm} is directly tied to Z_{ocT}, Z_{0tm} for the most part tends to rise and fall with M_{mT}.

3. Resistance elements

R_{scT}, equal to the sum of R_{alT} and R_{cT}, is initially set to match the (transformed and rms-adjusted) measured cochlear resistance from Aibara *et al.* (2001), and for most ears it is not varied significantly from that initial value. The other resistances in the circuit can be compared to this reference point (in transformed units) in Fig. 7. R_{mT} tends to be varied somewhat more than R_{scT} to adjust the amount of velocity entering the circuit to match the data, and on the transformed scale it averages to a value not far below that of R_{scT}. R_{imjT} and R_{isjT} both end up being considerably lower on average than the other two resistances, thus allowing the shunt branches to pass more volume velocity at high frequencies, which is often necessary to produce the magnitude roll-off that is often observed in the $V_{st}/P_{\Delta tm}$ transfer functions.

In Puria (2003) the reverse middle-ear resistance, R_{mer}, which is measured from the cochlea outward to the stapes and is an approximate measure of the untransformed R_{al} parameter, is reported as 1×10^{10} Kg/(s m^4). If R_{al} is equated to this R_{mer} value for all models, then the average value for the untransformed R_c parameter becomes 1.13×10^{10} for set A, 2.06×10^{10} for set B, 1.83×10^{10} for the mean of sets A and B, and 1.99×10^{10} for the AB model based on the 16 ear mean. The measured cochlear input impedance from Aibara *et al.* (2001) (with an rms adjustment), which was used as the initial value for R_{sc}, is 2.9×10^{10} by comparison.

4. Mass elements

Both the malleus and incus mass parameters in the model end up being significantly smaller than typical physical measurements of these masses. From Table 1 of Beer *et al.* (1999), malleus, incus, and stapes masses were reported as 25.6, 27.6, and 3.5 mg, respectively. Malleus and incus masses were reported from Sim *et al.* (2007) as 30.3 and 32.0 mg, respectively, by comparison. The corresponding 16 ear parameter averages for malleus and incus mass in the present study, however, are 4.52 and 4.6 mg, respectively, which are between a factor of 5 and 6 smaller than the physical values reported by Beer *et al.* (1999), and closer to a factor of 7 smaller than the physical values reported in Sim *et al.* (2007). Had a deliberate effort not been made to keep the malleus masses as high as possible without seriously compromising the fits, the resulting values for the malleus mass could have easily been set much lower. The stapes mass in the model is equal to the Beer *et al.* (1999) value for all ears, however, since it was deliberately assigned that way. Puria (2003) reports a reverse middle ear mass, M_{mer}, which is measured from the cochlea to the stapes and should be approximately equal to M_s (plus some additional mass due to the connection of the rest of the ossicular chain to the stapes), with a value of 4.34×10^{-6} Kg (or 4.34 mg) when converted from an acoustic mass to a mechanical mass by multiplying by A_{fp}^2, by comparison.

In Puria and Allen (1998), both the incus and the malleus mass parameters that resulted from the model-fitting procedure for the cat were considerably smaller than the typical physical masses for the cat, by a factor of 30 for the malleus, and a factor of 39 for the incus.

The lower mass values in the model may be consistent with a hypothesized physical process by which the malleus and incus switch their dominant rotational axes at high frequencies to minimize inertia. At low frequencies the classical hinging motion of the malleus-incus complex around the anterior-posterior axis is thought to be dominant, but at high frequencies the ossicular inertia may cause reduced ossicular motion around that axis. At high frequencies, then, the malleus handle is thought to switch from a hinging motion to a twisting motion around the inferior-superior axis of the malleus handle such that inertia is reduced and the system behaves as though it has lower effective mass at high frequencies, which is where the effects of mass are highest (see Puria *et al.*, 2006, and Puria *et al.*, 2007, for a discussion of this hypothesis).

5. Group delay in the TM and ossicular chain

The choice to model the TM as a transmission line stems mostly from the need to account for the group delay that is evident in the V_{st}/P_{ec} and $V_{st}/P_{\Delta tm}$ measurements. In this section, group delays are determined for the measurements as well as different portions of the model, to estimate how the TM and ossicular chain each contribute to the overall delay (Table II). Group delays are determined by plotting the phase on a linear frequency axis in units of radians/s, measuring the slope of the phase over a region where it approximates a straight line (by finding the best straight line fit to the phase), and multiplying the measured slope by −1. For the measurements and the model with the TM, the group delay is measured over the 2–11.1 kHz frequency range, over which the phase is approximately linear. For the model without the TM, however, the group delay is measured from 6 to 14 kHz since the phase is not typically linear for lower frequencies.

In Table II, the group delays of the measured V_{st}/P_{ec} and $V_{st}/P_{\Delta tm}$ means (for set A, set B, and the combination of sets A and B) are shown in rows 1 and 2, respectively. By comparison, the V_{st}/P_{ec} group delay from the Aibara *et al.* (2001) measurements was found to be 62 μs, which is closer to the value measured for set B (row 1), but still around 14 μs lower. From the standpoint of group delay, the V_{st}/P_{ec} and $V_{st}/P_{\Delta tm}$ measurements are essentially interchangeable since

TABLE II. Summary of measured and estimated group delays for different segments of the middle ear sound conduction path, based on the data and mean model phases for set A, set B, and the combination of sets A and B.

		Measured or Estimated Delay (μs)		
	Description of Data and Model Delays	Set A	Set B	A and B
Data	1. V_{st}/P_{ec}	134.4	76.1	90.6
	2. $V_{st}/P_{\Delta tm}$	133.8	75.0	89.7
	3. EC probe tube distance to TM ($\approx$2.5 mm)	7.2	7.2	7.2
Model	4. $V_{st}/P_{\Delta tm}$	134.3	62.1	80.2
	5. $V_{st}/P_{\Delta tm}$ with EC probe tube delay subtracted (4 minus 3)	127.1	54.9	73.0
	6. V_{st}/P_{uT} (umbo to stapes delay from model with TM removed)	51.6	37.3	40.9
	7. Estimated TM delay (5 minus 6)	75.5	17.6	32.1
	8. Average T_{tm} values with EC probe tube delay subtracted	85.9	30.1	44.3

they are within 1.1 μs of each other. This implies that the middle-ear cavity has a negligible effect on group delay for these measurements. Though it is not considered directly in the model, the separation between the ear canal probe tube and the TM has a delay associated with it that is implicit in the measurements and model fits to those measurements. Since the separation is estimated to be between 2 and 3 mm, this associated delay is estimated to be 7.2 μs (the time for a wave moving at 347 m/s to travel 2.5 mm), and is listed in row 3.

Rows 4–8 pertain to group delay measurements made on the mean model phase curves. Row 4 lists the group delay of the $V_{st}/P_{\Delta tm}$ model measurements, which for set A is quite close to the corresponding delays for the V_{st}/P_{ec} and $V_{st}/P_{\Delta tm}$ data (rows 1 and 2). For set B and the combined A and B sets, however, the model delay is lower by as much as 14 μs. Because the V_{st}/P_{ec} and $V_{st}/P_{\Delta tm}$ group delay measurements for the data are essentially interchangeable, the model $V_{st}/P_{\Delta tm}$ group delays are assumed to be directly comparable to the data V_{st}/P_{ec} group delays. For this reason, the effects of the ear canal probe tube separation are accounted for in row 5 by subtracting the ear canal delay in row 3 from the model $V_{st}/P_{\Delta tm}$ group delay in row 4. The group delay of the ossicular chain portion of the model is estimated by measuring the group delay of the model without the TM, and is listed in row 6. Row 7 lists the estimated portion of the delay associated with the TM, which is found by subtracting the estimated ossicular chain delay in row 6 from the overall delay (with EC probe tube delay subtracted) in row 5. Finally, row 8 lists the average T_{tm} parameter values representing the TM delay in the model, with the EC probe tube delay subtracted, for comparison with the delays calculated in row 7. The adjusted T_{tm} parameter values in row 8 are all higher than the estimated model TM delays in row 7, by as much as 12.5 μs. This disparity may in part be a result of how the group delays for the model were only measured in frequency ranges for which the phase appears linear, such that the range for the full model (2–11.1 kHz) was different from the range for the model without the TM (6–14 kHz).

The estimated human middle-ear delay range of 75–134.4 μs (rows 1 and 2) is generally higher than the estimate of 80 μs for the cat (Puria and Allen, 1998), and more than a factor of 2 higher than the estimated range of 25–30 μs for gerbil (Olson, 1998; Dong and Olson, 2006). For the cat, Puria and Allen (1998) estimated the TM component to be approximately 40 μs, compared to the present estimated range of 17.6–75.5 μs (row 7) for human. For the gerbil, about 17 μs can be accounted for as being due to the ear canal and the ossicular chain, leaving a delay of at most 13 μs for the gerbil eardrum (de La Rochefoucauld and Olson, 2007).

C. Comparisons to data from living ears

The Huber *et al.* (2001) V_{st}/P_{ec} magnitude mean in Fig. 4, based on seven living ears and modified by Chien *et al.* (2006) below 2 kHz in an attempt to account for methodological differences between living ear and cadaver ear studies, shows close resemblance to the set B data and model, and an even closer resemblance to the 16 ear mean and AB model. While the pressure in the middle-ear cavity was not subtracted from the drive pressure for these measurements, the effects of this difference are expected to be minimal since the middle-ear cavities were not sealed during these measurements, such that P_{mec} should be small. Chien *et al.* (2006) applied a frequency-dependent magnitude scale factor roughly equal to a factor of 3 below 1 kHz and a factor of 1 above 2 kHz, with a linear transition between 1 and 2 kHz, which brings the Huber *et al.* (2001) data into closer agreement with temporal bone measurements below 2 kHz, but leaves the data unchanged above 2 kHz. The close resemblance between the adjusted Huber *et al.* (2001) curve and the AB model, especially above 2 kHz, suggests that the present model may be applicable to living ears as well as cadaver temporal bone ears.

The impedance curves in Fig. 6 that were adapted from Hudde (1983) and Farmer-Fedor and Rabbitt (2002), are also based on measurements made on living ears. Both of these curves were scaled by a characteristic impedance that was chosen in the absence of specific information from these studies, so the most appropriate scaling of these magnitude curves is not known. Additionally, these measurements are affected by the ear canal and middle-ear cavities of the subjects. While these curves exhibit several differences between the present model and data in the magnitude and phase, they do exhibit some qualitative similarities with the set B and AB models above 2 kHz that support the idea that the present model may approximate the behavior of living ears as well as temporal bone ears.

VI. CONCLUSIONS

A one-dimensional model of the tympanic membrane (TM), ossicular chain, and cochlea consisting of all lumped-parameter elements except for the TM, which is represented as a distributed-parameter transmission line, was used to produce overall magnitude and phase agreement with 16 pressure to stapes velocity transfer function measurements from human temporal bones, from two different studies, up to 12 kHz. Some of the local peaks and dips within the data could only be matched approximately, but the model's overall magnitude and phase similarity to the data surpasses that of three previous models consisting exclusively of lumped-parameter elements (Kringlebotn, 1988; Goode *et al.*, 1994; and Feng and Gan, 2004), particularly above 6 kHz. The present set B and AB models also closely resemble a pressure to stapes velocity mean adapted from measurements on living ears (Huber *et al.*, 2001; Chien *et al.*, 2006). Model agreement with available velocity ratio data between the umbo and stapes and the umbo and incus was also achieved, in a more physically realistic manner than the one previous model (Feng and Gan, 2004) for which comparisons were possible. The model input impedance exhibits a number of similar features with measured impedance curves from three previous studies (Voss *et al.*, 2000; Farmer-Fedor and Rabbitt, 2002; and Hudde, 1983), two of which involved measurements on living ears, and exhibits closer similarity overall to the measurements than the lumped-parameter element models used for comparison.

Larger values for the tympanic membrane transmission line delay parameter typically required smaller values to be assigned to the malleus mass in order to avoid producing undesired magnitude peaks and "stair-stepping" effects in the phase. The effects of the transmission line varied for individual ears between primarily affecting the magnitude, primarily affecting the phase, and affecting both more or less equally. The model calculations are consistent with the idea that there is significant transmission delay associated with the human TM, and that the ossicular chain also contributes to overall observed middle-ear delay.

In addition to selecting model parameters to fit the individual measurements, a parameter set was selected to fit the *mean* of the 16 measurements (AB), which performs well up to 10 kHz, to make it possible for the overall features of the ensemble of models to be captured using a single set of parameters.

The similarities between the present models and measurements made on living ears suggest that the model may be appropriate for representing the behavior of living ears in addition to the behavior of cadaveric ears.

ACKNOWLEDGMENT

This work was supported by Grant No. DC005969 from the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health.

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (**2001**). "Human middle-ear sound transfer function and cochlear input impedance," Hear. Res. **152**, 100–109.

Beer, H. J., Bornitz, M., Hardtke, H. J., Schmidt, R., Hofmann, G., Vogel, U., Zahnert, T., and Huttenbrink, K. B. (**1999**). "Modeling of components of the human middle ear and simulation of their dynamic behavior," Audiol. Neuro-Otol. **4**, 156–162.

Békésy, G. (**1960**). *Experiments in Hearing* (McGraw–Hill, New York).

Beranek, L. L. (**1954**). *Acoustics* (McGraw–Hill, New York).

Chien, W., Ravicz, M. E., Merchant, S. N., and Rosowski, J. J. (**2006**). "The effect of methodological differences in the measurement of stapes motion in live and cadaver ears," Audiol. Neuro-Otol. **11**, 183–197.

de La Rochefoucauld, O., and Olson, E. (**2007**). "Exploring sound transmission through the middle ear by tracing middle ear delay," *Association for Research in Oto-Laryngology Mid-Winter Meeting*, Denver, CO.

Dong, W., and Olson, E. S. (**2006**). "Middle ear forward and reverse transmission in gerbil," J. Neurophysiol. **95**, 2951–2961.

Farmer-Fedor, B. L., and Rabbitt, R. D. (**2002**). "Acoustic intensity, impedance and reflection coefficient in the human ear canal," J. Acoust. Soc. Am. **112**, 600–620.

Feng, B., and Gan, R. Z. (**2004**). "Lumped parametric model of the human ear for sound transmission," Biomechan. Model Mechanobiol. **3**, 33–47.

Goode, R. L., Killion, M., Nakamura, K., and Nishihara, S. (**1994**). "New knowledge about the function of the human middle ear: Development of an improved analog model," Am. J. Otol. **15**, 145–154.

Huber, A., Linder, T., Ferrazzini, M., Schmid, S., Dillier, N., Stoeckli, S., and Fisch, U. (**2001**). "Intraoperative assessment of stapes movement," Ann. Otol. Rhinol. Laryngol. **110**, 31–35.

Hudde, H. (**1983**). "Measurement of the eardrum impedance of human ears," J. Acoust. Soc. Am. **73**, 242–247.

Kringlebotn, M. (**1988**). "Network model for the human middle ear," Scand. Audiol. **17**, 75–85.

Lynch, T. J., III, Nedzelnitsky, V., and Peake, W. T. (**1982**). "Input impedance of the cochlea in cat," J. Acoust. Soc. Am. **72**, 108–130.

O'Connor, K. N., and Puria, S. (**2006**). "Middle ear cavity and ear canal pressure-driven stapes velocity responses in human cadaveric temporal bones," J. Acoust. Soc. Am. **120**, 1517–1528.

O'Connor, K. N., Tam, M., Blevins, N. H., and Puria, S. (**2008**). "Tympanic membrane collagen fibers: A key to high frequency sound conduction," Laryngoscope (in press).

Olson, E. S. (**1998**). "Observing middle and inner ear mechanics with novel intracochlear pressure sensors," J. Acoust. Soc. Am. **103**, 3445–3463.

Onchi, Y. (**1949**). "A study of the mechanism of the middle ear," J. Acoust. Soc. Am. **21**, 404–410.

Pierce, A. D. (**1989**). *Acoustics: An Introduction to Its Physical Principles and Applications* (Acoustical Society of America, Woodbury, NY).

Puria, S., Peake, W. T., and Rosowski, J. J. (**1997**). "Sound-pressure measurements in the cochlear vestibule of human-cadaver ears," J. Acoust. Soc. Am. **101**, 2754–2770.

Puria, S., and Allen, J. B. (**1998**). "Measurements and model of the cat middle ear: Evidence of tympanic membrane acoustic delay," J. Acoust. Soc. Am. **104**, 3463–3481.

Puria, S. (**2003**). "Measurements of human middle ear forward and reverse acoustics: Implications for otoacoustic emissions," J. Acoust. Soc. Am. **113**, 2773–2789.

Puria, S., Sim, J. H., Shin, M., Tuck-Lee, J., and Steele, C. R. (**2006**). "Middle ear morphometry from cadaveric temporal bone microCT imaging," in *Middle Ear Mechanics in Research and Otology*, edited by A. Eiber and A. Huber (World Scientific, Zurich, Switzerland).

Puria, S., Sim, J. H., Shin, M., and Steele, C. R. (**2007**). "A gear in the middle ear," *Association for Research in Oto-Laryngology Mid-Winter Meeting*, Denver, CO.

Shaw, E. A. G., and Stinson, M. R. (**1983**). *The Human External and Middle Ear: Models and Concepts* (Delft U. P., Delft, The Netherlands).

Shera, C. A., and Zweig, G. (**1992**). "Middle-ear phenomenology: The view from the three windows," J. Acoust. Soc. Am. **92**, 1356–1370.

Sim, J. H., Puria, S., and Steele, C. R. (**2007**). "Calculation of inertial properties of the malleus-incus complex from micro-CT imaging," J. Mech. Mater. Struct. **2**, 1515–1524.

Stinson, M. R., and Lawton, B. W. (**1989**). "Specification of the geometry of the human ear canal for the prediction of sound-pressure level distribution," J. Acoust. Soc. Am. **85**, 2492–2503.

Voss, S. E., Rosowski, J. J., Merchant, S. N., and Peake, W. T. (**2000**). "Acoustic responses of the human middle ear," Hear. Res. **150**, 43–69.

Wever, E. G., and Lawrence, M. (**1950**). "The acoustic pathways to the cochlea," J. Acoust. Soc. Am. **22**, 460–467.

Willi, U. B., Ferrazzini, M. A., and Huber, A. M. (**2002**). "The incudomalleolar joint and sound transmission losses," Hear. Res. **174**, 32–44.

Zwislocki, J. J. (**1962**). "Analysis of middle ear function: Part I: Input Impedance," J. Acoust. Soc. Am. **34**, 1514–1523.

Reconciling the origin of the transient evoked ototacoustic emission in humans

Robert H. Withnell,[a] Chantel Hazlewood, and Amber Knowlton
Department of Speech and Hearing Sciences, Indiana University, Bloomington, Indiana 47405, USA

(Received 29 June 2007; revised 4 October 2007; accepted 4 October 2007)

A pervasive theme in the literature for the transient evoked otoacoustic emission (TEOAE) measured from the human ear canal has been one of the emission arising solely (or largely) from a single, place-fixed mechanism. Here TEOAEs are reported measured in the absence of significant stimulus contamination at stimulus onset, providing for the identification of a TEOAE response beginning within the time window that is typically removed by windowing. Contrary to previous studies, it was found that in humans, as has previously been found in guinea pig, the TEOAE appears to arise from two generation mechanisms, the relative contributions of these two mechanisms being time and stimulus-level dependent. The method of windowing the earliest part of the ear canal measurement to remove stimulus artifact removes part of the TEOAE i.e., much of the component arising from a nonlinear generation mechanism. This reconciliation of TEOAE origin is consistent with all OAEs in mammals arising in a stimulus-level dependent manner from two mechanisms of generation, one linear, one nonlinear, as suggested by Shera and Guinan [J. Acoust. Soc. Am. **105**, 782–798 (1999)]. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804635]

PACS number(s): 43.64.Jb, 43.64.Kc [BLM] Pages: 212–221

I. INTRODUCTION

The relationship between the stimulus spectrum and the otoacoustic emission response spectrum for transient-evoked otoacoustic emissions (TEOAEs) has repeatedly been found to have a one-to-one correspondence, consistent with a linear generation mechanism (e.g., Kemp and Chum, 1980; Kemp, 1986; Probst *et al.*, 1986; Hauser *et al.*, 1991, Xu *et al.*, 1994; Prieve *et al.*, 1996; Killan and Kapadia, 2006; Kalluri and Shera, 2007). Discordant with these studies using exclusively human subjects have been studies using guinea pigs, which have suggested a contribution from a nonlinear mechanism of generation (Avan *et al.*, 1995; Withnell and Yates, 1998; Yates and Withnell, 1999). And there have been a few studies with human subjects that suggest other than a solely linear mechanism of generation for the TEOAE (Avan *et al.*, 1993, 1997; Carvalho *et al.*, 2003). Insight into this disparity was provided by Withnell and McKinley (2005), who reported that, in guinea pigs, the dominant mechanism of generation of the TEOAE appeared to be time dependent, shifting from a wave-fixed generation mechanism to a place-fixed generation mechanism over the time course of the response.

Wave and place fixed, first described by Kemp (e.g., see Kemp, 1986), argues for two mechanisms of OAE generation, distinguishable by their phase characteristic, with the region of generation shifting with the traveling wave for the wave-fixed and being spatially fixed for the place-fixed mechanism. For an exactly scaling-symmetric cochlea, OAEs would be produced solely by a wave-fixed mechanism. A place-fixed mechanism involves random variations in the impedance of the basilar membrane versus length (Zweig and Shera, 1995), producing, it seems, a breaking of scaling symmetry near threshold. Acting through the cochlear amplifier feedback loop, these small variations in basilar membrane impedance act as a reflection source from which intracochlear standing waves can develop between the reflected wave and the incident wave. At or near threshold, a variation in hearing threshold versus stimulus frequency on the order of 20 dB can be observed (e.g., Long, 1984), presumably associated with these intracochlear standing waves (Talmadge *et al.*, 1998).

Studies of the origin of the TEOAE in humans have windowed or filtered the first 2.5–5 ms of the ear canal recording or nonlinear derived response (e.g., Kemp *et al.*, 1990) to remove the ringing of the speaker associated with delivering a click stimulus subsequent to loading it with the ear canal. This windowing removes any part of the TEOAE that is present in the early part of the response. As a result, examinations of TEOAE origin (e.g., Probst *et al.*, 1986; Xu *et al.*, 1994; Killan and Kapadia, 2006) that have found the TEOAE to arise from a linear generation mechanism require that the windowed early part of the response contain no TEOAE or that this early part of the TEOAE have the same generation mechanism. A time dependence to the origin of the TEOAE in guinea pig raises the possibility that a similar relationship may exist in humans and that if the early part of the response can be preserved and not removed with windowing, then this early part of the TEOAE may have a nonlinear generation mechanism. To explore this possibility, a similar approach to measuring the TEOAE was completed to that performed previously in guinea pig (see Withnell *et al.*, 1998), with the earphone not physically coupled to the ear canal and sound delivered to the ear with the recording microphone positioned in the ear canal.

[a]Electronic mail: rwithnel@indiana.edu

 0001-4966/2008/123(1)/212/10/$23.00

II. METHOD

A. Subject

Seven adult females, less than 30 years of age, served as subjects for this study. Data for this study were collected from one ear of each subject, hearing being within clinically normal limits for the ear tested.

B. Signal generation and data acquisition

Signal generation and data acquisition were computer-controlled using custom software and a Card Deluxe soundcard with 96 kHz sampling rate. Sound stimuli were delivered by a Beyer DT48 loudspeaker positioned approximately 4 cm from the entrance to the ear canal. A Sennheiser M series electrostatic microphone (6 mm diameter) was placed in the ear canal to measure ear canal sound pressures, the depth of insertion defined by the length of the microphone capsule (<1 cm). The frequency response of the loudspeaker at the position of the microphone in the ear canal was determined by delivering pseudorandom electrical noise (a sum of sine waves, spectrally flat with random phase) to the loudspeaker; the impulse response of the loudspeaker was then calculated from the frequency response measured at the microphone. The stimulus wave form for evoking a TEOAE was generated using a sinc function $[\sin(\omega t)/(\omega t)]$ deconvolved with the impulse response of the loudspeaker (see Yates and Withnell, 1999, for further details). This provided for a stimulus with a flat amplitude spectrum and linear phase delay at the measurement microphone in the ear canal. TEOAEs were obtained using the nonlinear derived extraction technique (Kemp *et al.*, 1990) with a 21.3 ms time base, a 9 dB stimulus level ratio, and 1000–4000 averages [see Withnell and McKinley (2005) for more details regarding extraction of the TEOAE from ear canal sound pressure recordings]. For the data reported here, TEOAEs were obtained over a range of stimulus levels and bandwidths.

C. Data analysis

Time domain windowing to separate TEOAE components with different phase characteristics was achieved using a recursive exponential filter (see Shera and Zweig, 1993; Kalluri and Shera, 2001)

TEOAE early latency component=TEOAE$\cdot F(t)$,

TEOAE late component=TEOAE$\cdot F(t)$,

where $F(t)=1/\Gamma_n(\tau)$, $\Gamma_n(\tau)$ is defined recursively as $\Gamma_{n+1}(\tau)=e^{\Gamma_n(\tau)-1}$, with $\Gamma_1(\tau)=e^{\tau 2}$,

$\tau=t/\tau_{\text{cut}}$, where t is time and τ_{cut} is the length of the window, filter order $(n)=16$.

The value of τ_{cut} was chosen such that it:

i. Separates the TEOAE into two components with different phase characteristics, one with a shallow slope, the other a steep slope.

ii Matches the amplitudes of the two components with the TEOAE phase. The slope of the TEOAE phase identifies which of the early and late components dominates the TEOAE and the amplitudes of each of the components should be in agreement with this, e.g., when the slope of the TEOAE phase is steep, the late component should be larger in amplitude than the early component.

This study was completed with the approval of the Human Ethics Committee, Indiana University.

III. RESULTS

Figure 1 shows examples of time-averaged, nonlinear derived, ear canal sound pressure recordings from six of the subjects, the acoustic stimulus varying in bandwidth but having a similar peak sound pressure level (~71 dB pSPL) at the measurement microphone. Each of the panels includes the stimulus wave form, scaled to the amplitude of the time-averaged, nonlinear derived, ear canal sound pressure recordings. Inset in each panel is the earliest part of the recording. Each of the first four panels show the earliest part of the response in time to be higher in frequency than the later part of the response. Figure 1(e) shows the earliest part of the response to have a higher frequency than the later part but the response as a whole is delayed relative to the responses in Figs. 1(a)–1(d). Stimulus contamination of the response coincident in time with the stimulus is notable in Fig. 1(a) and may be present in other panels.

A. Quantifying stimulus contamination of the nonlinear derived response

When a short duration electrical pulse is delivered to an earphone/loudspeaker, it rings at its resonant frequency (where the electrical pulse has a frequency spectrum that includes one or more of the resonant frequencies of the loudspeaker). If the loudspeaker is driven with a sufficiently large current that the diaphragm response is no longer linear, then the speaker ringing at the resonant frequency will no longer be linear. This speaker-generated nonlinearity will contaminate the nonlinear derived response recorded from a human ear with normal hearing. Figure 2 shows the nonlinear derived sound pressure level recorded in response to a short duration stimulus[1] over a range of stimulus levels with the Senheisser microphone placed in the ear canal of KEMAR, with the Beyer DT48 loudspeaker used to generate the acoustic stimulus positioned approximately 4 cm from the ear canal entrance. KEMAR provides a passive cavity with an input impedance that is intended to match the input impedance of the average human ear (at higher stimulus levels where the cochlear input impedance is resistive). Each of the panels includes the stimulus wave form, scaled to the amplitude of the time-averaged, nonlinear derived, ear canal sound pressure recordings. Each panel shows the stimulus amplitude at the microphone of the stimulus i.e., 57–77 dB pSPL, but this is not the determinant of stimulus contamination; rather, it is the magnitude of the input voltage to the speaker coupled with the acoustic load. The uppermost panel shows little or no stimulus contamination of the nonlinear derived response. The bottom panel shows stimulus contamination with the earphone ringing. The amplitude spectrum of this ringing shows two resonant modes, one centered on ~3.2 kHz with another mode centered on ~4.5 kHz. Note that the stimulus contamination is largest coincident in time

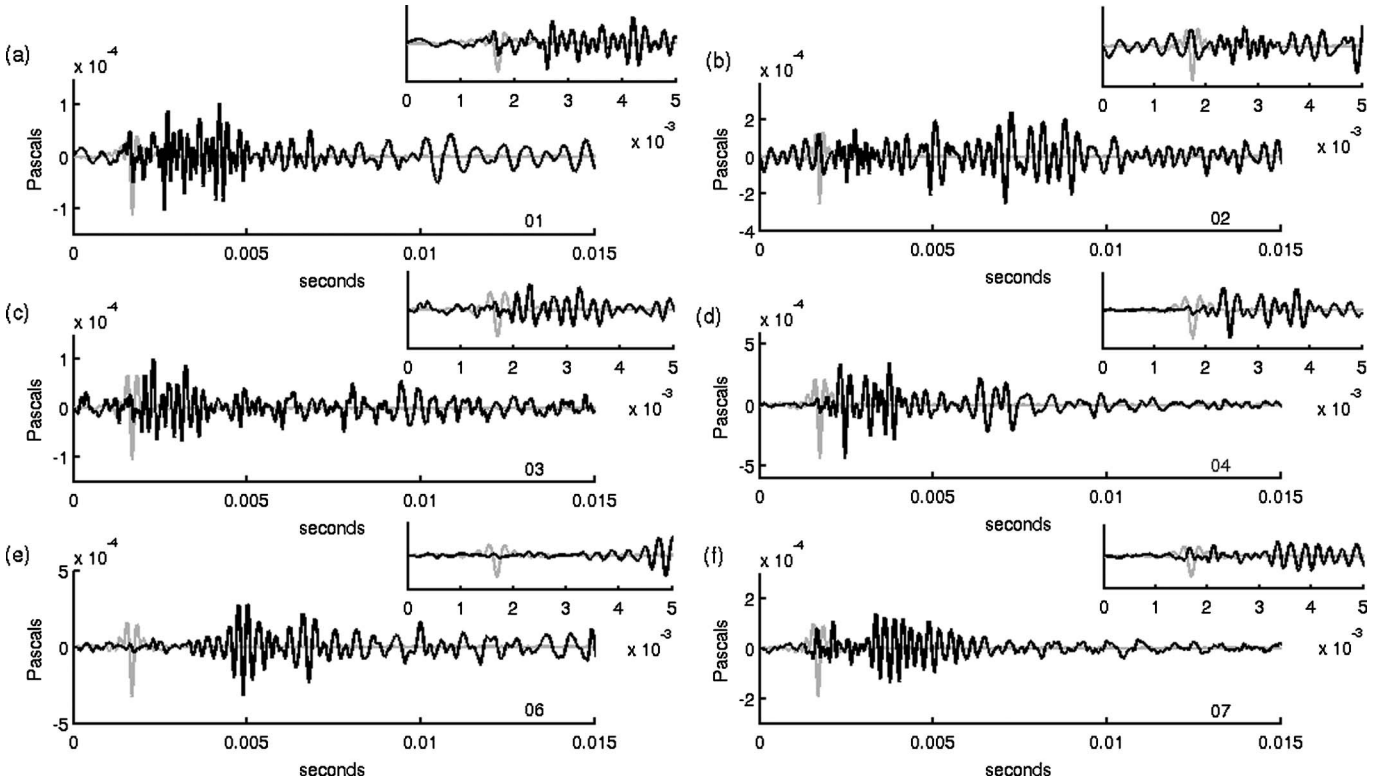


FIG. 1. TEOAE responses from six subjects, the acoustic stimulus varying in bandwidth but having a similar peak sound pressure level (~71 dB pSPL) at the measurement microphone. Each of the panels includes the stimulus wave form, scaled to the amplitude of the time-averaged, nonlinear derived, ear canal sound pressure recordings. Inset in each panel is the earliest part of the recording. Stimulus level ratio in each case was 9 dB. Stimulus bandwidths were (a) 0.5–5.5 kHz, (b) 1.5–7 kHz, (c) 1.5–5 kHz, (d) 1–5 kHz, (e) 1–5 kHz, and (f) 1–5 kHz.

with the stimulus, decaying to the noise floor within ~1.5 ms of the stimulus peak amplitude. The stimulus generated in the bottom panel was for the maximum input voltage provided by the software. At five stimulus levels (three of which are shown), stimulus contamination, if present, was largest coincident in time with the stimulus. It would appear that the upper limit to the magnitude of the stimulus contamination of the nonlinear derived response is defined at stimulus onset, prior to the onset of a physiological response, and that stimulus contamination has a very short duration.

It is apparent in Fig. 1 that the magnitude of the speaker-generated nonlinearity at stimulus onset is insufficient in amplitude for speaker-generated nonlinearity to contaminate significantly these TEOAE responses recorded from six human ears.

B. TEOAE versus stimulus level

Figure 3 shows the nonlinear derived response recorded over a range of stimulus levels from one subject. Figure panels are arranged in descending stimulus level (top to bottom) from 78 to 63 dB pSPL in 5 dB steps. Each panel shows the nonlinear residual and the click stimulus wave form (lighter lines), the click stimulus peak occurring at 1.7 ms. In all four recordings, a robust TEOAE is observed; stimulus contamination coincident in time with the stimulus is small (or not present) relative to the response that follows. As in Fig. 1, the earliest part of the response is higher in frequency than the response later in time. Inspection of the four panels reveals that the early part of the TEOAE (<3.7 ms or <2 ms

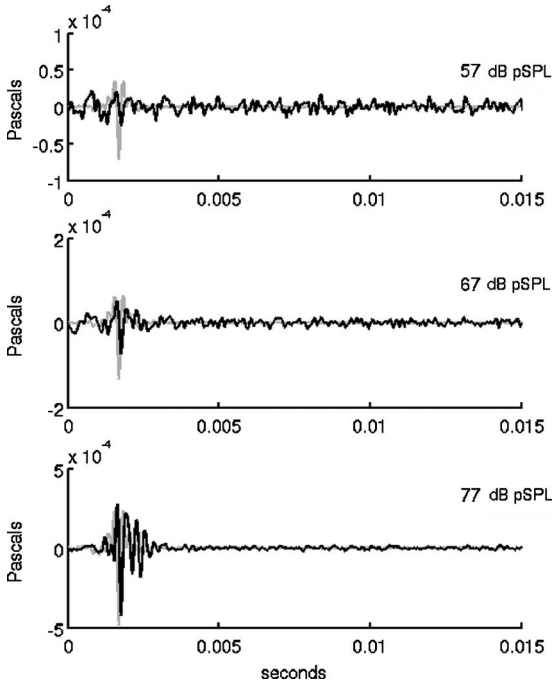


FIG. 2. The nonlinear derived sound pressure level recorded in response to a short duration stimulus over a range of stimulus levels in KEMAR. Each of the panels includes the stimulus wave form, scaled to the amplitude of the time-averaged, nonlinear derived, ear canal sound pressure recordings. Stimulus bandwidth was 1–5 kHz. The uppermost panel shows little or no stimulus contamination of the nonlinear derived response. The bottom panel shows stimulus contamination with the earphone ringing. Stimulus contamination is largest coincident in time with the stimulus, decaying to the noise floor within ~1.5 ms of the stimulus peak amplitude.

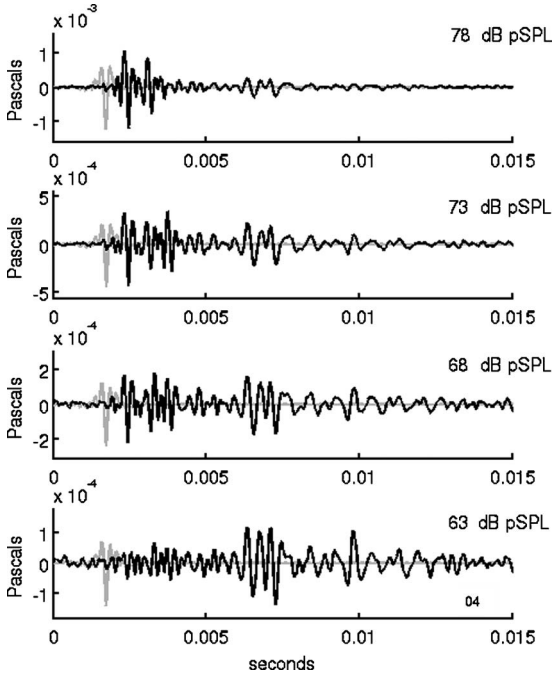


FIG. 3. The TEOAE recorded over a range of stimulus levels from one subject. Panels are arranged in descending stimulus level (top to bottom) from 78 to 63 dB pSPL in 5 dB steps. Each panel shows the nonlinear residual and the click stimulus wave form (lighter lines), the click stimulus peak occurring at 1.7 ms. In all four recordings, a robust TEOAE is observed; stimulus contamination coincident in time with the stimulus is small (or not present) relative to the response that follows. Stimulus bandwidth was 1–5 kHz.

poststimulus peak) decreases in amplitude relative to the later part (>4.7 ms) with decreasing stimulus level. The method of TEOAE extraction used in this study (nonlinear derived extraction technique) however, increasingly underestimates the TEOAE as stimulus level is reduced and the stimulus level ratio is maintained constant due to OAE growth not being saturated (Lonsbury-Martin *et al.*, 1988) and so the relative amplitude relationship of the early and late parts of the TEOAE may be affected.

Examples of TEOAEs obtained across a range of stimulus levels from two other subjects are shown in Fig. 4. As observed in Fig. 3, the early part of the TEOAE decreases in amplitude relative to the later part as stimulus level decreases. In contrast to Fig. 3, stimulus contamination of the response is more evident, particularly for the responses to the highest stimulus levels.

C. Time-domain windowing of the TEOAE

Here we examine the proposal that the TEOAE in humans can be separated into two components, the early part of the TEOAE having a shallow phase slope consistent with a wave-fixed mechanism and the late part having a steep phase slope consistent with a place-fixed mechanism. Equipment used to record TEOAEs in humans typically window some part of the averaged ear canal sound pressure recording poststimulus onset (e.g., 2.5 ms, Kemp *et al.*, 1990) to reduce stimulus contamination of the recording. This windowing would remove some or all of the early part of the TEOAE response in Figs. 3 and 4. Figure 5 shows amplitude and phase spectra corresponding to the TEOAE in the top panel of Fig. 3 and the spectra resulting from time-domain windowing this TEOAE. Also shown are the original TEOAE time wave form, and the early and late TEOAE wave forms extracted by time domain windowing. The black line is the original TEOAE. The phase spectrum of the TEOAE shows the slope of the phase to be rotating rapidly up to about 2.2 kHz with a shallow slope from 2.2 to 5.5 kHz. This indicates that the early component should be larger in amplitude than the late component over the 2.2–5.5 kHz frequency range and vice-versa below 2.2 kHz. For a cut-off value of 3.7 ms (see Sec. III D for how this value was ar-

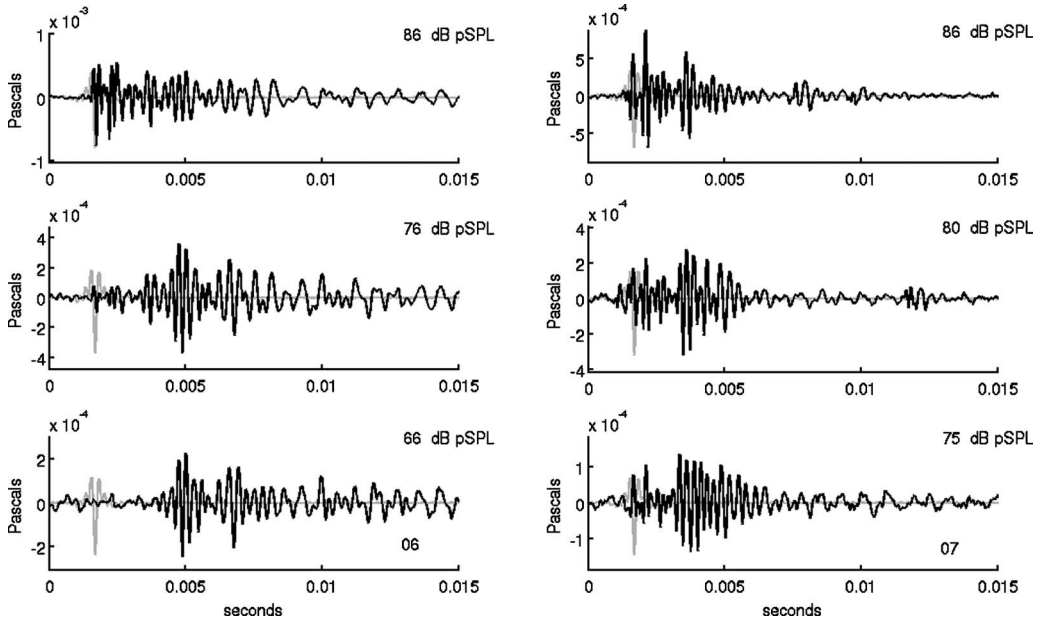


FIG. 4. Examples of TEOAEs obtained across a range of stimulus levels from two other subjects. As observed in Fig. 3, the early part of the TEOAE decreases in amplitude relative to the later part as stimulus level decreases. In contrast to Fig. 3, stimulus contamination of the response is more evident, particularly for the responses to the highest stimulus levels. Stimulus bandwidth was 1–5 kHz.

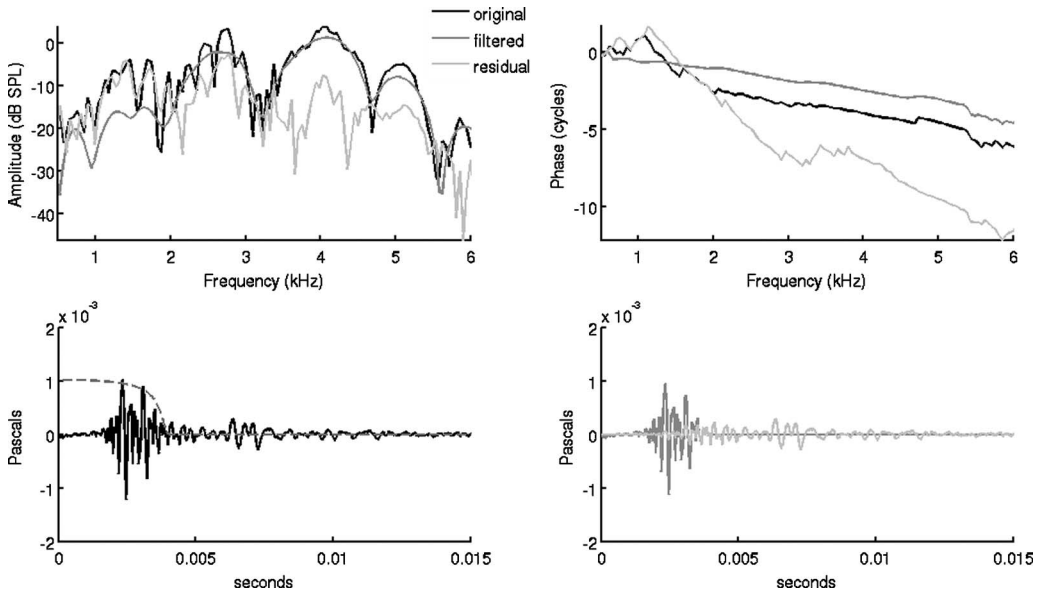


FIG. 5. (a) The amplitude spectra corresponding to the TEOAE in (c) and the spectra resulting from time-domain windowing this TEOAE. (b) The phase spectra corresponding to the TEOAE in (c) and the spectra resulting from windowing this TEOAE. (c) TheTEOAE time wave form (the top panel of Fig. 3) and the windowing function, $F(t)$. (d) TEOAE component wave forms extracted by time domain windowing.

rived at), the early part of the TEOAE rotates ~1.5 cycles over a 2 kHz range, corresponding to a group delay of ~0.8 ms, a value that suggests either a wave-fixed mechanism of generation or the emission arising from the basal region of the cochlea. In contrast, the late component has a group delay of ~6.9 ms centered on 2 kHz and ~3.8 ms centered on 4 kHz, values that suggest forward travel times of >3.5 and 1.9 ms [see Shera *et al.* (2005), Eq. (59)].

Figure 6 is as for Fig. 5, but from another subject corresponding to the TEOAE obtained to a stimulus level of 86 dB pSPL from Fig. 4 (left top panel). Unlike the TEOAE analyzed in Fig. 5 (top panel of Fig. 3), there is evidence of stimulus contamination of the TEOAE. To reduce this stimulus contamination, the response was time-domain windowed with a time cut of 2.38 ms (0.68 ms poststimulus peak) before analysis, the wave form obtained after this windowing then being time-domain windowed to separate the early and late components of the TEOAE response. In contrast to the TEOAE in Fig. 5, this TEOAE has a phase response that rotates rapidly with frequency over most of the frequency range, indicating that the late part of the TEOAE should be larger than the early component at most frequencies. For a cut-off value of 3.6 ms, the early part of the TEOAE rotates ~1.8 cycles over a 2 kHz range, corresponding to a nonphysical group delay for the 2.5–4.5 kHz region of the cochlea of ~0.9 ms. In contrast, the late component has a

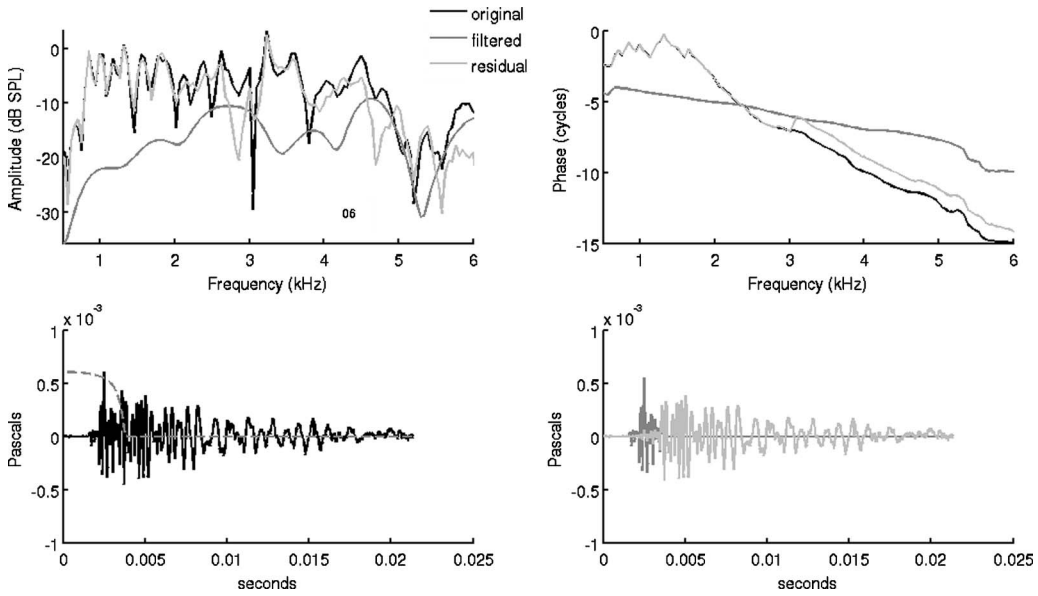


FIG. 6. As for Fig. 5, but from another subject corresponding to the TEOAE obtained to a stimulus level of 86 dB pSPL from Fig. 4 (left top panel). To reduce stimulus contamination, the response was time-domain windowed with a time cut of 2.38 ms (0.68 ms poststimulus peak) before analysis, the wave form obtained after this windowing then being time-domain windowed to separate the early and late components of the TEOAE response.

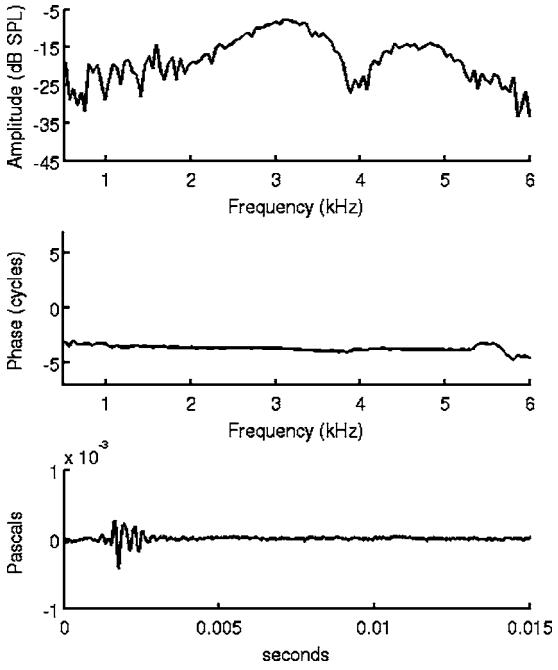


FIG. 7. The stimulus artifact recorded with KEMAR. It is notable that the amplitude and phase spectra for the artifact differs significantly from the TEOAE recorded from a human subject (see Figs. 5 and 6). The amplitude spectrum shows two resonant modes, one centered on ~3.2 kHz, the other centered on ~4.5 kHz. The phase of the artifact has a zero slope, consistent with no delay, the onset of the artifact coincident in time with the stimulus.

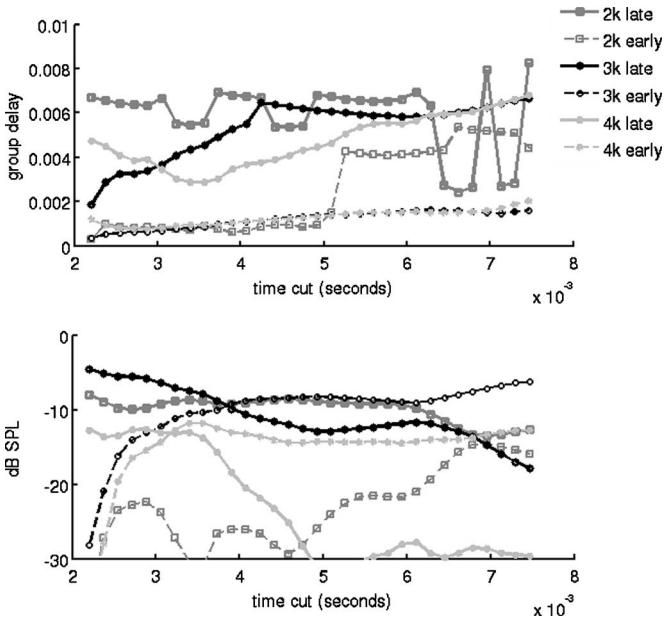


FIG. 8. The slope of the phase and the amplitude for the early and late components of the TEOAE in Fig. 5 at 2, 3, and 4 kHz over a range of values of τ_{cut}.

group delay of ~6.4 ms centered on 2 kHz and ~4.2 ms centered on 4 kHz, values that suggest forward travel times of >3.2 and 2.1 ms. These values are similar to that found for the TEOAE from the subject in Fig. 5.

Figure 7 shows the equivalent representation of Figs. 5 and 6 for the stimulus artifact recorded with KEMAR. It is notable that the amplitude spectra for the artifact differs significantly from the TEOAE recorded from a human subject (see Figs. 5 and 6), and the phase of the artifact has a zero slope consistent with no delay while the early component of the TEOAE has a negative phase slope. The negative slope for the early component of the TEOAE in Figs. 5 and 6 is thought to be due to a breaking of scaling and the slope of the phase of the early component of the TEOAEs clearly distinguish them from stimulus artifact.

D. Determination of τ_{cut}

Determination of the appropriate time cut, τ_{cut}, for the analysis presented in Figs. 5 and 6 was made by evaluating the rate of change of phase versus frequency ($d\Phi/d\omega$) at 2, 2.5, 3, 3.5, and 4 kHz for the early and late components for a range of τ_{cut}. For τ_{cut}'s shorter than the optimum value, some of the component arising from a nonlinear mechanism remains in the late part of the response, reducing the rate of change of phase versus frequency for this component, or the "group delay." Figure 8 shows the slope of the phase and the amplitude for the early and late components of the TEOAE in Fig. 5 at 2, 3, and 4 kHz over a range of values of τ_{cut} The optimal τ_{cut} is between 3.4 and 3.75 ms; for cut-off values beyond 3.75 ms the value of $d\Phi/d\omega$ [Fig. 8(a)] for the 4 kHz late component of the TEOAE starts to increase by virtue of the main energy in this part of the response being removed by windowing. Figure 8(b) shows the amplitudes of the 3 and 4 kHz late components to, in general, be a decreasing function of τ_{cut} these component include a significant early component at low τ_{cut}'s. As τ_{cut} increases, the late component should plateau in amplitude until a cut-off value is reached where some of the late component is removed with time domain windowing; for TEOAE arising from a linear coherent reflection mechanism this cut-off value would be a function of frequency, higher frequency components being removed first, i.e., with the lowest cut-off value. Concomitantly, the early component of the TEOAE will increase in amplitude versus τ_{cut} until it plateaus, any subsequent alteration in amplitude as τ_{cut} increases dependent on the magnitude and phase of the late component versus the early component. The 2 and 3 kHz early component frequencies seem to reach a plateau in amplitude in the 3.4 and 3.75 ms regions. This emphasizes an observation made by Withnell and McKinley (2005) that there is no one cut-off value that optimally separates both components. A cut-off value for the recursive exponential filter (defined as $1/e$) that separates the early and late components of 3.7 ms is within the optimal τ_{cut} range of 3.4–3.75 ms.

E. TEOAE early and late components versus stimulus level

The stimulus-level dependence of the contributions of the early versus late components is illustrated in Fig. 9 with the amplitude and phase spectra of the TEOAE and the early and late components corresponding to the four TEOAEs obtained over a stimulus range of 78 to 63 dB pSPL from Fig. 3. Values of τ_{cut} at each stimulus level with justification for the selection of τ_{cut} are given in Table I. Panel A, Fig. 9 provides the amplitude and phase spectra of the TEOAE and early and late components for a stimulus level of 78 dB

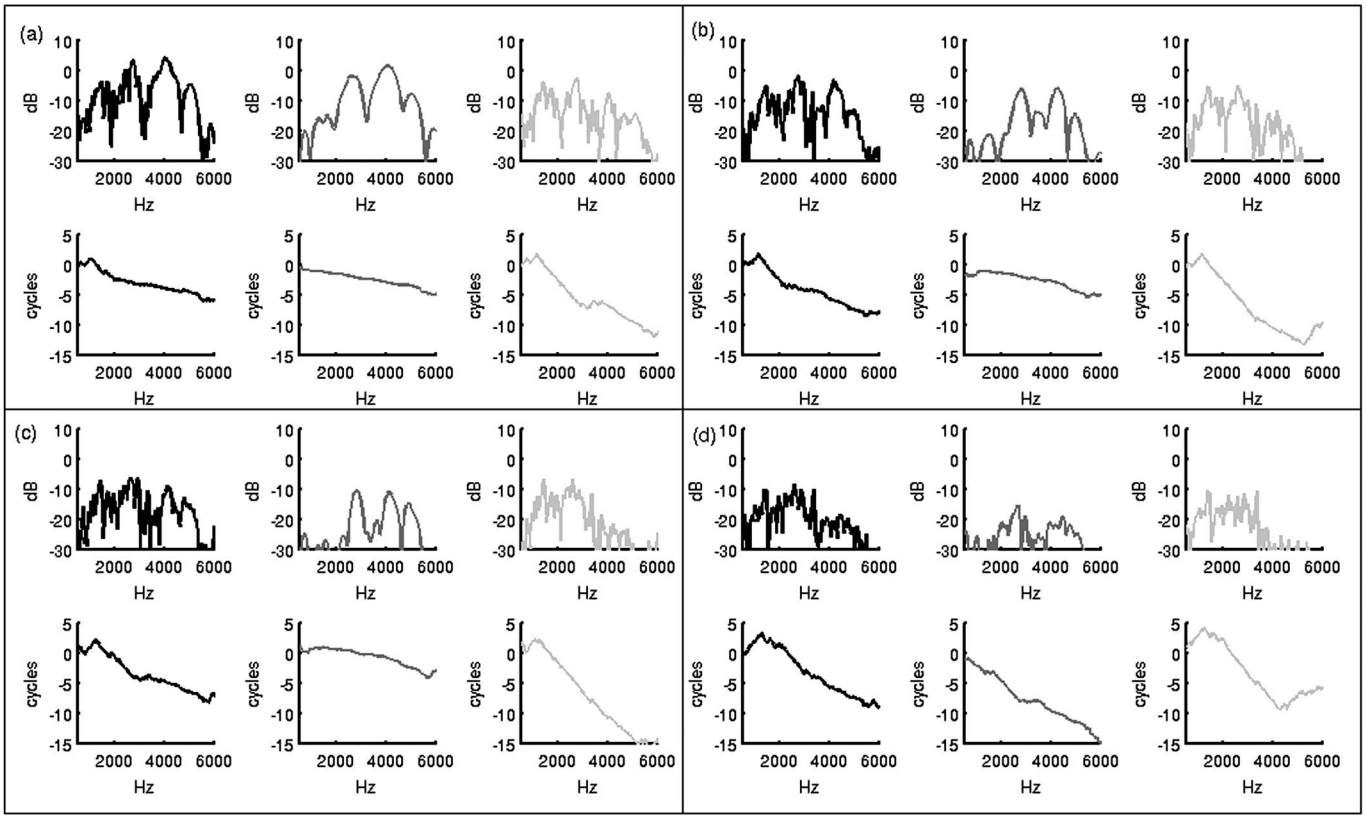


FIG. 9. The stimulus-level dependence of the contributions of the early vs late components is illustrated over a stimulus range of 78 to 63 dB pSPL (TEOAEs from Fig. 3). See the text for further details.

pSPL (see Fig. 5), panels B to D for stimulus levels of 73–63 dB pSPL. TEOAE amplitude spectra decrease in magnitude as stimulus level decreases. Inspection of the phase spectra for the TEOAE reveals that the slope of the phase increases with decreasing stimulus level (steep phase component extends to higher frequencies), arguing for a decrease in the relative contribution of the early component and an increase in the relative contribution of the late component as stimulus level decreases. Time-domain windowing of the TEOAE can isolate an early and a late component with signature phases that argue for distinct mechanisms of generation. The early component amplitude spectrum in Fig. 9 exhibits a microstructure that argues for a more complex origin than can be ascribed to a simple reflection from a nonlinear variation in basilar membrane impedance (Talmadge *et al.*, 2000) where the gain of the cochlear amplifier feedback loop is constant or a slowly varying function of frequency. The early component phase spectra in panels A to C have a slope that does not alter significantly between 1.5 and 4 kHz with decreasing stimulus level; the lack of change in slope of the phase of this component as stimulus level decreases argues against the delay being a construct of the slope of the phase being nonzero may represent a breaking of scaling symmetry (cochlear Q is not constant as a function of frequency). Human cochlear Q's are not thought to be constant as a function of frequency (Shera and Guinan, 2003), although a recent report of cochlear Q's inferred from stimulus frequency otoacoustic emission measurements suggested cochlear Q's to be constant from 1 to 4 kHz (Schairer *et al.*, 2006). The bandwidth of the amplitude spectra of the late component of the TEOAE becomes smaller with decreasing stimulus level, the basal region of the cochlea contributing less to this component. No stimulus-level dependent variation in the slope of the late component phase is evident up to 3 kHz.

TABLE I. Values of τ_{cut} at each stimulus level with justification for the selection of τ_{cut} corresponding to Fig. 9.

Stimulus level	τ_{cut} (ms)	Justification
78	3.73	See the text
73	4.24	TEOAE phase has a steep slope below 2.3 kHz arguing for the late component to be larger than the early component over this frequency range. A saw-tooth pattern is observed from 2.3 to 3 kHz, indicating that the early and late components have similar amplitudes over this frequency range. Above 3 kHz, TEOAE phase has a shallow slope and so the early component must be larger than the late component.
68	4.92	TEOAE phase has a steep slope below 2.7 kHz, a saw-tooth pattern from 2.7 to 3.8 kHz, and a shallow slope above 3.8 kHz
63	6.79	TEOAE phase has a steep slope below 3.0 kHz, a saw-tooth pattern from 3.0 to 3.7 kHz, and a shallow slope above 3.7 kHz

IV. DISCUSSION

The TEOAE in humans has consistently been found to arise predominantly from one mechanism of generation, a

place-fixed mechanism (e.g., Kemp, 1986; Probst *et al.*, 1986; Prieve *et al.*, 1996; Killan and Kapadia, 2006; Kalluri and Shera, 2007). The TEOAE phase spectrum in this case shows a steep slope. In all previous studies that have examined the origin of the TEOAE in humans, the early part of the TEOAE recording was contaminated by stimulus artifact and so was removed prior to analysis, typically the first 2.5 ms (e.g., Kemp *et al.*, 1990), but 6 ms (Prieve *et al.*, 1996) and more (e.g., Killan and Kapadia, 2006) have been extracted from the recording prior to analysis. In all of these studies, stimulus contamination of the TEOAE response was produced by ringing of the earphone centered on the major resonant frequencies of the earphone with the earphone physically coupled to the ear canal.

Here, TEOAEs were recorded with the earphone not physically coupled to the ear canal. Measurements in KEMAR with the earphone not physically coupled to the ear canal suggested that the upper limit to the magnitude of the stimulus contamination of the TEOAE is defined at stimulus onset, prior to the onset of a physiological response. In the absence of significant stimulus contamination at stimulus onset, TEOAEs were recorded with a response beginning within the time window that typically is removed by windowing (see Fig. 1). Time-domain windowing of the TEOAE using a recursive exponential filter suggests that the TEOAE can be separated into two components, one with a shallow phase slope that gives a short group delay, the other having a steep phase slope that gives a group delay suggestive of a round-trip cochlear delay. The shallow phase slope for the early component of the TEOAE differs from stimulus artifact that has zero phase slope and so identifies this early component as a physiological response. Examination of the TEOAE versus stimulus level (Fig. 9) reveals a stimulus-level dependence for the two components identified as contributing to the total TEOAE. As stimulus level increases, the absolute and relative contribution of the early component reduces as stimulus increases while the relative contribution of the late component decreases. It seems then that the TEOAE in humans, in accord with other types of OAE recorded in rodents (TEOAEs: Yates and Withnell, 1999; SFOAEs: Goodman *et al.*, 2003; DPOAEs: Withnell *et al.*, 2003; Schneider *et al.*, 2003) and humans (DPOAEs: Talmadge *et al.*, 1999; Kalluri and Shera, 2001; Knight and Kemp, 2001; SFOAEs: Schairer and Keefe, 2005), has two components with phase spectra that suggest that the TEOAE arises from two distinct mechanisms of generation.

It could be argued that time-domain windowing the TEOAE will produce two components, one with a shallow phase, the other with a steeper phase, simply by virtue of the windowing process. The group delay for the early component (e.g., the early component of the TEOAE in Figs. 5 and 6 had a "group delay" of ~0.8 ms) it could be argued represents contributions from the tails of the traveling waves generated by the stimulus, the response having a short delay by virtue of it arising basal to the active region of the traveling wave for each frequency component. Two testable predictions arise from this explanation: (i) The early component of the TEOAE should have a group delay that is an increasing function of τ_{cut} and (ii) the group delay for the early component of the TEOAE should increase with decreasing stimulus level.

With respect to the first prediction, inspection of the group delay for the early component as a function of τ_{cut} in Fig. 8 at 3 kHz reveals that up to a $\tau_{cut} \sim 7.5$ or 5.8 ms poststimulus peak the group delay is between 0.6 and 1.6 ms. The delay increases from 0.6 to 1.6 ms over the range $2.5 < \tau_{cut} < 6.4$ ms, and then stays relatively constant at a delay of ~1.6 ms from $6.4 < \tau_{cut} < 7.5$ ms. Note that we are considering here the proposition that the TEOAE arises from one generation mechanism so that increasing τ_{cut} increases the OAE within the time window and so eventually, as τ_{cut} increases, it should encompass the whole TEOAE. A $\tau_{cut} \sim 7.5$ or 5.8 ms poststimulus peak provides for a TEOAE that should receive significant contribution from the cochlear region near the characteristic place for 3 kHz (mean SFOAE group delay at 3 kHz to 5.5 ms; Shera and Guinan, 2003). A time dependence for the 3 kHz component represents an increasing contribution from OAE later in time but the group delay peaks at 1.6 ms, a value not in accord with one mechanism of generation, i.e., the TEOAE at 3 kHz encompassed within a time window of 5.8 ms poststimulus peak should have a much longer delay if the OAE arises from only mechanism and the amplitude of the emission is dependent on the displacement amplitude of the basilar membrane. Alternatively, an argument based on one mechanism where the emission is generated in the basal region of the cochlea would not provide for a late component with an OAE group delay commensurate with a round-trip cochlear travel time.

The second prediction can be tested by examining Fig. 9. The phase spectra for the early component of the TEOAE has a slope that does not alter significantly between 1.5 and 4 kHz with stimulus level from 78 to 68 dB pSPL, i.e., the group delay for the early component of the TEOAE does not increase with decreasing stimulus level. It would appear, then, that the group delay for the early component extracted by time-domain windowing is consistent with the emission arising from a wave-fixed mechanism and inconsistent with the delay being a construct of the windowing process.

Zweig and Shera (1995) provided a theoretical description for the generation of OAEs exhibiting a steep phase slope, the coherent reflection filter (CRF) theory. The CRF theory is expected to apply only in the low-level, linear region of operation of the cochlea; with increasing stimulus level, the traveling wave peak broadens, presumably reducing the phase coherence of reflections across this region. In addition to reflection from randomly distributed cochlear irregularities (place-fixed OAE), emissions also are thought to arise as a consequence of cochlear nonlinearity acting through the cochlear amplifier feedback loop (wave-fixed OAE). To date, no theory of cochlear mechanical function provides an adequate description of OAE generation incorporating a stimulus-level dependent cochlear nonlinearity.

Windowing or removing the early part of the TEOAE recording so as to isolate TEOAE that is generated predominantly by one mechanism, a linear, place-fixed mechanism, serendipitously has clinical virtue. This TEOAE component will have a one-to-one correspondence with the stimulus fre-

quencies that generated it subject to variation in cochlear reflectance. The amplitude spectrum microstructure, according to the CRF theory, is produced by random variation in cochlear reflectance (Zweig and Shera, 1995). The SFOAE, according to the CRF theory, arises from reflections from cochlear irregularity with the signal source being the amplitude of displacement of the basilar membrane traveling wave, a source that is largest at the peak of the traveling wave subject to phase coherence across this peak. A number of studies have established that SFOAEs arise predominantly from this tip region by virtue of their group delay (e.g., Shera and Guinan, 2003; Goodman *et al.*, 2004) with TEOAEs being effectively a composite of SFOAEs (Kalluri and Shera, 2007).

In guinea pigs, it has been suggested that the TEOAE arises predominantly from intermodulation distortion energy generated by the cochlear nonlinear response to the stimulus component frequencies (Yates and Withnell, 1999). A nonlinear generation mechanism will be both within channel and between channel, the extent of the between channel contribution (intermodulation distortion) being presumably a consequence of the amount of overlap of the cochlear filters (Withnell and McKinley, 2005). In humans it has been found that cochlear filters are sharper than rodents (Shera *et al.*, 2002) and so the contribution of intermodulation distortion products to the TEOAE in humans is presumably less than in rodents. The within channel nonlinear contribution presumably arises from the outer hair cell nonlinearity/ies acting through the cochlear amplifier feedback loop generating a periodic basilar membrane response to a sinusoidal input that is a sum of the fundamental plus higher order harmonic distortion. Higher order harmonics do not couple well into the basilar membrane and the resultant fundamental response is not linearly related to the change in input stimulus level.

Previous studies suggesting that the TEOAE recorded from a human ear arises solely from one mechanism is presumably a result of windowing the earliest part of the TEOAE response, removing much of the component arising from a nonlinear generation mechanism. The TEOAE, as has been found in guinea pig, appears to arise from two distinct mechanisms, the relative contributions of these two mechanisms being time dependent and stimulus level dependent. It seems, then, that all OAEs in mammals arise in a stimulus level dependent manner from two mechanisms of generation, one linear, one nonlinear, as suggested by Shera and Guinan (1999).

ACKNOWLEDGMENT

The authors would like to thank Chris Shera and two anonymous reviewers for valuable feedback on this manuscript.

[1]The stimulus wave form was generated using a sinc function $[\sin(\omega t)/(\omega t)]$ deconvolved with the impulse response of the loudspeaker, providing a stimulus with a flat amplitude spectrum and linear phase delay at the measurement microphone in the ear canal.

Avan, P., Bonfils, P., Loth, D., Elbez, M., and Erminy, M. (**1995**). "Transient-evoked otoacoustic emissions and high-frequency acoustic trauma in the guinea pig," J. Acoust. Soc. Am. **97**, 3012–3020.

Avan, P., Bonfils, P., Loth, D., and Wit, H. P. (**1993**). "Temporal patterns of transient-evoked otoacoustic emissions in normal and impaired cochleae," Hear. Res. **70**, 109–120.

Avan, P., Elbez, M., and Bonfils, P. (**1997**). "Click-evoked otoacoustic emissions and the influence of high-frequency hearing losses in humans," J. Acoust. Soc. Am. **101**, 2771–2777.

Carvalho, S., Buki, B., Bonfils, P., and Avan, P. (**2003**). "Effect of click intensity on click-evoked otoacoustic emission wave forms: Implications for the origin of emissions," Hear. Res. **175**, 215–225.

Goodman, S. S., Withnell, R. H., De Boer, E., Lilly, D. J., and Nuttall, A. L. (**2004**). "Cochlear delays measured with amplitude-modulated tone-burst evoked OAEs," Hear. Res. **188**, 57–69.

Goodman, S. S., Withnell, R. H., and Shera, C. A. (**2003**). "The origin of SFOAE microstructure in the guinea pig," Hear. Res. **183**, 7–17.

Hauser, R., Probst, R., and Lohle, E. (**1991**). "Click- and tone-burst-evoked otoacoustic emissions in normally hearing ears and in ears with high-frequency sensorineural hearing loss," Eur. Arch. Otorhinolaryngol. **248**, 345–352.

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.

Kalluri, R., and Shera, C. A. (**2007**) "Near equivalence of human click-evoked and stimulus-frequency otoacoustic emissions," J. Acoust. Soc. Am. **121**, 2097–2110.

Kemp, D. T. (**1986**). "Otoacoustic emissions, travelling waves and cochlear mechanisms," Hear. Res. **22**, 95–104.

Kemp, D. T., and Chum, R. (**1980**). "Properties of the generator of stimulated acoustic emissions," Hear. Res. **2**, 213–232.

Kemp, D. T., Ryan, S., and Bray, P. (**1990**). "A guide to the effective use of otoacoustic emissions," Ear Hear. **11**, 93–105.

Killan, E. C., and Kapadia, S. (**2006**). "Simultaneous suppression of tone burst-evoked otoacoustic emissions—Effect of level and presentation paradigm," Hear. Res. **212**, 65–73.

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.

Long, G. R. (**1984**). "The microstructure of quiet and masked thresholds," Hear. Res. **15**, 73–87.

Lonsbury-Martin, B. L., Martin, G. K., Probst, R., and Coats, A. C. (**1988**). "Spontaneous otoacoustic emissions in a nonhuman primate. II. Cochlear anatomy," Hear. Res. **33**, 69–93.

Prieve, B. A., Gorga, M. P., and Neely, S. T. (**1996**). "Click- and tone-burst-evoked otoacoustic emissions in normal-hearing and hearing-impaired ears," J. Acoust. Soc. Am. **99**, 3077–3086.

Probst, R., Coats, A. C., Martin, G. K., and Lonsbury-Martin, B. L. (**1986**). "Spontaneous, click-, and toneburst-evoked otoacoustic emissions from normal ears," Hear. Res. **21**, 261–275.

Schairer, K. S., Ellison, J. C., Fitzpatrick, D., and Keefe, D. H. (**2006**). "Use of stimulus-frequency otoacoustic emission latency and level to investigate cochlear mechanics in human ears," J. Acoust. Soc. Am. **120**, 901–914.

Schairer, K. S., and Keefe, D. H. (**2005**). "Simultaneous recording of stimulus-frequency and distortion-product otoacoustic emission input-output functions in human ears," J. Acoust. Soc. Am. **117**, 818–832.

Schneider, S., Prijs, V. F., and Schoonhoven, R. (**2003**). "Amplitude and phase of distortion product otoacoustic emissions in the guinea pig in an (f1,f2) area study," J. Acoust. Soc. Am. **113**, 3285–3296.

Shera, C. A., and Guinan, J. J., Jr. (**1999**). "Evoked otoacoustic emissions arise by two fundamentally different mechanisms: A taxonomy for mammalian OAEs," J. Acoust. Soc. Am. **105**, 782–798.

Shera, C. A., and Guinan, J. J., Jr. (**2003**). "Stimulus-frequency-emission group delay: A test of coherent reflection filtering and a window on cochlear tuning," J. Acoust. Soc. Am. **113**, 2762–2772.

Shera, C. A., Guinan, J. J., and Oxenham, A. J. (**2002**). "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," Proc. Natl. Acad. Sci. U.S.A. **99**, 3318–3323.

Shera, C. A., Tubis, A., and Talmadge, C. L. (**2005**). "Coherent reflection in a two-dimensional cochlea: Short-wave versus long-wave scattering in the generation of reflection-source otoacoustic emissions," J. Acoust. Soc. Am. **118**, 287–313.

Shera, C. A., and Zweig, G. (**1993**). "Noninvasive measurement of the cochlear traveling-wave ratio," J. Acoust. Soc. Am. **93**, 3333–3352.

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.
Talmadge, C. L., Tubis, A., Long, G. R., and Piskorski, P. (**1998**). "Modeling otoacoustic emission and hearing threshold fine structures," J. Acoust. Soc. Am. **104**, 1517–1543.
Talmadge, C. L., Tubis, A., Long, G. R., and Tong, C. (**2000**). "Modeling the combined effects of basilar membrane nonlinearity and roughness on stimulus frequency otoacoustic emission fine structure," J. Acoust. Soc. Am. **108**, 2911–2932.
Withnell, R. H., Kirk, D. L., and Yates, G. K. (**1998**). "Otoacoustic emissions measured with a physically open recording system," J. Acoust. Soc. Am. **104**, 350–355.
Withnell, R. H., and McKinley, S. (**2005**). "Delay dependence for the origin of the nonlinear derived transient evoked otoacoustic emission," J. Acoust. Soc. Am. **117**, 281–291.
Withnell, R. H., Shaffer, L. A., and Talmadge, C. L. (**2003**). "Generation of DPOAEs in the guinea pig," Hear. Res. **178**, 106–117.
Withnell, R. H., and Yates, G. K. (**1998**). "Enhancement of the transient-evoked otoacoustic emission produced by the addition of a pure tone in the guinea pig," J. Acoust. Soc. Am. **104**, 344–349.
Xu, L., Probst, R., Harris, F. P., and Roede, J. (**1994**). "Peripheral analysis of frequency in human ears revealed by tone burst evoked otoacoustic emissions," Hear. Res. **74**, 173–180.
Yates, G. K., and Withnell, R. H. (**1999**). "The role of intermodulation distortion in transient-evoked otoacoustic emissions," Hear. Res. **136**, 49–64.
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.

Supporting evidence for reverse cochlear traveling waves

W. Dong[a] and E. S. Olson
Department of Otolaryngology, Head and Neck Surgery, Columbia University, P & S 11-452, 630 West 168th Street, New York, New York 10032, USA

(Received 6 September 2007; revised 25 October 2007; accepted 27 October 2007; corrected 5 March 2007)

As a result of the cochlea's nonlinear mechanics, stimulation by two tones results in the generation of distortion products (DPs) at frequencies flanking the primary tones. DPs are measurable in the ear canal as oto-acoustic emissions, and are used to noninvasively explore cochlear mechanics and diagnose hearing loss. Theories of DP emissions generally include both forward and reverse cochlear traveling waves. However, a recent experiment failed to detect the reverse-traveling wave and concluded that the dominant emission path was directly through the fluid as a compression pressure [Ren, 2004, Nat. Neurosc. **7**, 333–334]. To explore this further, we measured intracochlear DPs simultaneously with emissions over a wide frequency range, both close to and remote from the basilar membrane. Our results support the existence of the reverse-traveling wave: (1) They show spatial variation in DPs that is at odds with a compression pressure. (2) Although they confirm a forward-traveling character of intraocochlear DPs in a broad frequency region of the best frequency, this behavior does not refute the existence of reverse-traveling waves. (3) Finally, the results show that, in cases in which it can be expected, the DP emission is delayed relative to the DP in a way that supports reverse-traveling-wave theory. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2816566]

PACS number(s): 43.64.Kc, 43.64.Jb, 43.64.Bt [BLM] Pages: 222–240

I. INTRODUCTION

The discovery of oto-acoustic emissions was exciting because emissions are generated within the cochlea and detected within the ear canal (EC), thus providing a noninvasive view into cochlear mechanics (Kemp, 1978). However, the mechanism by which the emissions are transported out through the cochlea is not certain. The usual view was that a reverse-traveling wave was the dominant pressure mode, but recent results were interpreted as favoring a compression mode (Ren, 2004; He *et al.*, 2007). The present paper presents a comprehensive investigation, from cochlear mechanics to emissions, bearing on this question. The introduction below provides a background to cochlear pressure modes and reviews aspects of emission theory that are useful for the interpretation of our results.

Sound at the eardrum is transmitted via the middle ear to the auditory inner ear, the cochlea. Within the cochlea, the sound is carried along the cochlea's sensory tissue (organ of Corti) by a wave, known as the cochlear traveling wave (von Bekesy, 1960). The wave is supported by sensory tissue flexibility and fluid inertia and travels much slower than sound travels in water. It peaks at frequency-specific locations: low frequencies in the apex, high frequencies in the base. By measuring the mechanical response at one longitudinal location [Fig. 1(A)] and varying the frequency, the location's peak or "best" frequency (BF) is identified [Fig. 1(B)]. In a healthy cochlea at frequencies close to the BF the amplitude of the cochlear traveling wave is enhanced by active outer hair cell forces, which show saturating nonlinearity [Fig. 1(B)] (Robles and Ruggero, 2001; Cooper, 2003). The signature of the wave is found in the phase of the response to tonal stimuli, which can be delayed by several cycles relative to the input in the ear canal [Fig. 1(C)]. The phase versus frequency slope is termed the "group delay." In the intracochlear response, the group delay is large at frequencies close to the BF, reflecting the slow speed of the wave there. In passive cochlear models the slowing of the wave is related to its peaking, via conservation of energy (Lighthill, 1981). Species differences in the size of the group delay have been correlated with sharpness of frequency tuning (Shera *et al.*, 2002; Shera and Guinan, 2003; Ruggero and Temchin, 2007). Thus, the group delay is an interesting auditory metric.

In addition to launching the cochlear traveling wave along the sensory tissue, the vibration of the stapes compresses the cochlear fluid, causing a compression (soundwave) pressure that exists along with the traveling-wave pressure. The compression pressure increases linearly with the intensity of stapes motion. The compression pressure mode was predicted theoretically by Peterson and Bogert (1950). It fills the cochlea approximately instantaneously (travels at the speed of sound in water) and can be roughly thought of as a background pressure that varies in time with the plunging motion of the stapes. (At high frequencies, standing-wave resonances are expected for this mode. Although interesting in its own right, this aspect of the compression pressure is not important for the present analysis.) The compression pressure is often called the cochlear "fast wave," as opposed to the slow cochlear traveling wave. The experimentally measured spatial variation of intracochlear pressure supports the idea that the pressure is composed of

[a]Electronic mail: wd2015@columbia.edu

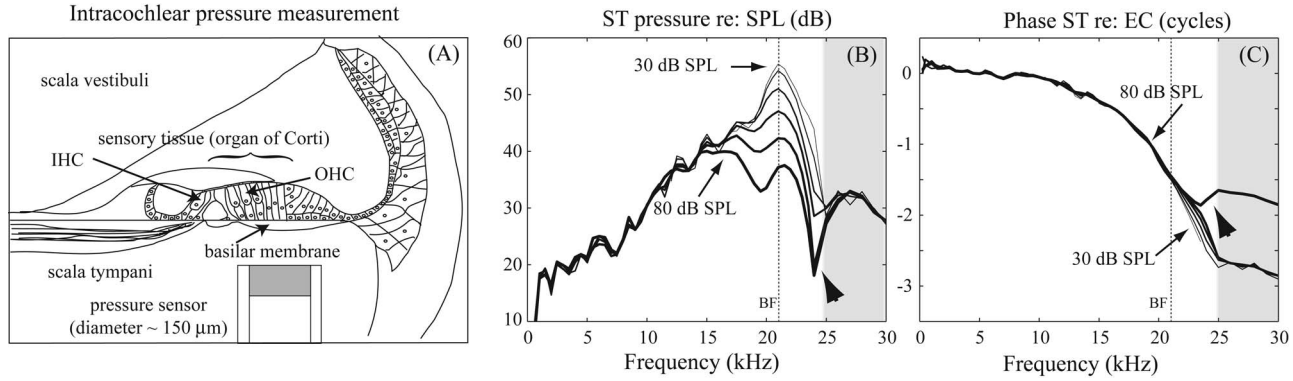


FIG. 1. Illustration of intracochlear pressure measurements and single tone response characteristics close to the BM. (A) Intracochlear pressure was measured in scala tympani (ST), close to or far from the sensory tissue's BM in the basal turn of the gerbil cochlea. The pressure sensor is shown to scale. Also indicated are the cochlear inner and outer hair cells, the mechanotransducers of audition. (B) and (C) Single-tone ST pressure responses measured close to the BM in turn one of the gerbil cochlea (wg81, 20 μm from BM, BF=21 kHz). (B) Amplitude normalized by the stimulus level in the EC. (C) Phase relative to the pressure in the EC. Significant aspects of cochlear mechanics are: (1) Cochlear nonlinearity: the normalized response curves fan out in the frequency region of the peak (15–25 kHz). 30 dB SPL responses showed a sharp peak at the BF of 21 kHz (dotted line), the 80 dB responses showed broader tuning with the peak shifted toward lower frequency. (2) Traveling wave: the phase of the response accumulated three complete cycles of delay relative to the stimulus in the ear canal. The phase changed most rapidly at frequencies around the BF. (3) Compression wave: at frequencies above 24 kHz, the phase changed little with frequency, which indicates that the fast wave dominated the responses. The amplitude in this frequency region was well above the noise floor and scaled linearly with SPL (gray band). (4) Interference between traveling and compression waves: the deep notch at 24 kHz with stimuli of 70 and 80 dB SPL corresponds to a region in which the phase jumps with level (arrowheads).

the sum of a traveling-wave mode and a compression mode (Olson, 1998, 1999, 2001, Dong and Olson, 2005b). Close to the BM and at frequencies near the local BF, rapid spatial pressure variations are present, evincing the fluid accelerations of the cochlear traveling wave. Such pressure variations were predicted in the 1970s by traveling-wave models (Steele and Taber, 1979a). At frequencies well above BF (where BM motion is very small), or far from the BM, the pressure is still large but is nearly spatially invariant, reflecting the lack of fluid acceleration and the dominance of the compression mode in this frequency region. (This behavior is shown, for example, in Fig. 4 of Dong and Olson (2005b); Fig. 10 of Olson (1998); Fig. 2 of Olson (1999); and Figs. 7 and 9 of Olson (2001)). In Figs. 1(B) and 1(C) of the present paper, the compression pressure region is identified by its linearity and relatively flat phase response (gray band). The destructive summation of the compression and traveling-wave modes produces notches in the pressure [arrowheads in Figs. 1(B) and 1(C)]. Because it can be mathematically uncoupled from the traveling-wave pressure and is not expected to lead to hair cell excitation (due to the very small motions it gives rise to), the compression pressure has been neglected in most cochlear models.

The cochlea's nonlinear mechanics produce distortion. For example, when two pure tones are used as stimulus, the response of a healthy cochlea contains a family of tones at frequencies corresponding to algebraic combinations of the input frequencies, termed "distortion products" (DPs). After being generated, a DP might travel apically to its own BF place, as demonstrated in both perceptual studies and intracochlear measurements (Goldstein, 1967; Robles *et al.*, 1997). The DPs also emerge from the cochlea, travel through the middle ear, and are detected in the EC as a DP otoacoustic emission (DPOAE) (Kemp, 1978). DPOAEs exhibit characteristic behavior: the amplitude contains frequency- and level-dependent notches ("fine structure") and the DPOAE phase versus frequency can be either nearly flat or rapidly varying as the frequency of the primaries is swept at a constant ratio.

Based on these and other observations, the analysis of cochlear emissions has relied on a "dual-component" theory. The conceptual framework for the theory and the quantitative description that builds upon it are based on cochlear wave mechanics, in which both forward and reverse cochlear traveling waves are present (Kemp, 1986; Zweig and Shera, 1995; Talmadge *et al.*, 1998; Shera and Guinan, 1999). The theory predicts that after being generated by nonlinear outer hair cell forces, a DP will travel as a cochlear traveling wave in reverse to emerge as a DPOAE (as a "generator" or "wave-fixed" component). In a scaling-symmetric cochlea the amount of phase accumulation to the generation place (where f_2 and f_1 have substantial overlap) is approximately unchanged as the primary frequencies are swept (at a fixed ratio). Therefore the phase of the generator component of the DPOAE is not expected to change with frequency when f_1 and f_2 are swept at fixed ratio (Shera *et al.*, 2000). In addition, irregularities in the cochlea's sensory tissue will give rise to reflections of the DP. The largest reflector contributions will come from close to the DP's BF place, where the DP will be large due to cochlear filtering and amplification. The reflected wave will contribute to the DPOAE (as a "reflector" or "place-fixed" component). In the dual-component theory, amplitude fine structure is due to interference between the generator and reflector components (e.g., Stover *et al.*, 1996; Talmadge *et al.*, 1999; Kalluri and Shera, 2001). Because the phase of a single tone changes rapidly in the region of its own BF place, the phase of the reflector component of the DPOAE changes rapidly with frequency. In this description, the slope of the phase-frequency response (group delay) of the reflector component at a given frequency is expected to be roughly twice the group delay of a single tone at the intracochlear BF region corresponding to

that frequency. The dual-component theory, and the separation of the observed emissions into reflector and generator emission types, was found to provide a useful framework for the interpretation of some of our results. This separation is aided by the fixed-ratio stimulus paradigm (as opposed to fixed f_1 or fixed f_2) (Knight and Kemp, 2000; Shera *et al.*, 2000; Kalluri and Shera, 2001; Knight and Kemp, 2001) and therefore we use this paradigm most extensively.

In order to use emissions to study the cochlea we must first understand if and how they are shaped as they travel out through the middle ear. The study of Avan *et al.* (1998) related DPs measured close to the stapes to DPOAEs to quantify middle-ear reverse transmission in the guinea pig. Similarly, Dong and Olson (2006) compared intracochlear DPs close to the stapes to DPOAEs to quantify middle-ear reverse transmission in the gerbil. The main result of that study, which will be required for the present study (also in the gerbil), was that the middle ear did not introduce very much structure into the emissions but, like forward transmission (Olson, 1998), had a fairly flat transfer function at a level of $\sim$−40 dB (compared to the $\sim$+25 dB forward gain of the middle ear) and a phase that was delaylike with a delay of $\sim$38 μs (slightly longer than the $\sim$32 μs forward delay). The delaylike character of middle-ear transmission both in the forward and reverse directions is a robust finding; inter-animal variability in delay time was small and could be attributed to the location of the microphone probe tube within the EC. Superimposed on this simple description, fine structure in the reverse transfer function was present at the $\pm$5–10 dB level in the amplitude, and at the $\pm$20°–30° level in phase and appeared to be linked to the acoustic load of the speaker and microphone system. Scala Vestibuli (SV) pressure was not measured in the present study. Therefore, the average middle ear forward and reverse phase responses from Dong and Olson (2006) are presented in many of the figures in order to illustrate the contribution of the middle ear to the observed phase behavior.

There have been just a handful of joint DP and DPOAE measurements that were designed to probe the intracochlear travel of DPs. Cooper and Shera (2004) compared DPOAE and DP phase-frequency responses measured in the same guinea pig (sequentially in time), and showed that the results were consistent with the reverse-traveling wave: the phase-frequency slope (group delay) of the DPOAE at frequencies close to the BF of the intracochlear measurement location was approximately twice the group delay of a BF tone, measured within the cochlea. This result indicated that the travel time out of the cochlea was similar to the travel time in. Another emission type (stimulus frequency emission) was also measured and led to the same conclusion. In addition, the study showed that the amplitude of the DP response in BM motion was tuned in frequency but that at frequencies below the BF it contained amplitude notches not present in the tuning of the primaries. This behavior was also observed in our previous DP studies (Dong and Olson, 2005b, a; Olson and Dong, 2006). The behavior is predicted due to destructive interference between locally or basally generated distortion and apically generated distortion, traveling out of the cochlea through the base. These results supported the existence of reverse-traveling waves.

However, the results of some intracochlear measurements call into question the existence of the reverse-traveling wave. Measurements of BM motion that extended for $\sim$1 mm in the longitudinal direction were used to look for the reverse-going wave (Ren, 2004; He *et al.*, 2007). DP cochlear traveling waves were detected, but their spatial–temporal character indicated that they were traveling in the forward direction. Moreover, the delay calculated with the phase-frequency slope of the DPs measured at the stapes (a proxy for the DPOAEs) was shorter than that of the DP on the BM. The interpretation of those results was that the DP traveled out of the cochlea to the stapes directly through the cochlear fluid as a (fast) compression pressure, and stapes excitation by the DP launched a forward-traveling wave. The possibility for a role of compression pressure in emissions is supported by the presence of emissions in animals such as frogs that lack a clear traveling wave (Wilson, 1980; van Dijk and Manley, 2001; Ruggero, 2004). However, in the mammalian cochlea the basic mechanics for reverse cochlear traveling waves is the same as for forward waves (flexible sensory tissue coupled to fluid inertia) and it is reasonable to expect that in mammals DPs would primarily travel out as a reverse wave—unless something about the sensory tissue's cellular architecture suppresses reverse waves. For sure, the fact that an experiment designed to directly observe a reverse cochlear traveling wave (Ren, 2004; He *et al.*, 2007) failed to do so has reinvigorated the compression DP theory.

The manner in which distortion products, and emissions in general, emerge from the cochlea is important for two prominent reasons. First, basic models of cochlear mechanics allow reverse waves, and their absence (or very small size) would point to a directional advantage for forward-going waves. This restriction would constrain cochlear models. Second, as noted above, cochlear group delays have been correlated with tuning. If emission phases are used to gauge intracochlear group delays it is important to know if the phase accumulation that is used in the group delay calculation reflects forward and reverse travel, or just forward travel. In this contribution we weigh in on the reverse wave/compression wave question. We simultaneously measured intracochlear pressure DPs both close to and far from the BM and DPOAEs. We first describe relevant aspects of the DPOAE and DP frequency responses individually. Then we relate the DPOAEs to the DPs to explore their relationship to one another. In particular, we explore their relative phase behaviors as this can be examined for a reverse-traveling-wave character.

II. METHODS

A. Animal preparation

The gerbil audiogram extends to $\sim$50 kHz (Ryan, 1976) and the BF of our intracochlear location was $\sim$20 kHz. Animal procedures were approved by the Institutional Animal Care and Use Committee of Columbia University. The experimental animals were young adults, 50–70 g. The gerbils

were deeply anesthetized throughout the experiment and euthanized at the end. (Forty milligrams per kilogram ketamine was administered first to sedate the animal, then an initial dose of 60 mg/kg sodium pentobarbital, with supplemental doses of 10 mg/kg for maintenance of deep anesthesia. Twenty mg/kg of the analgesic buprenorphine was administered every 6 h. At the end of the experiment the animal was overdosed with sodium pentobarbital.) A tracheotomy was performed to maintain a patent airway. The animal core temperature and head holder were maintained at ~37 °C with a heating blanket and a power resistor on the head holder. The left bulla was widely opened with great care to access the cochlea. A small hole (diameter ~200 μm) was hand drilled through the bony wall in the basal-turn scala tympani (ST). To gauge cochlear health, the compound action potential (CAP) thresholds and DPOAE were recorded with a silver electrode at the round window and EC several times during the experiment, in particular before and after making the ST hole.

B. Sound stimulation and data acquisition

The ear was acoustically stimulated via a single Radio Shack 40-1377 tweeter coupled to the ear canal in a closed-system configuration. Acoustic stimuli were generated and collected digitally using Tucker Davis Technologies System 3. One- and two-tone stimuli were used in the experiments. The single tone was swept from 0.2 to 50 kHz with frequency spacing of 500–1000 Hz, at stimulus levels from 40 to 90 dB SPL. For two-tone stimuli, two equal-intensity primary tones were delivered at a fixed f_2:f_1 ratio, 1.05 or 1.25, and f_1 and f_2 were swept in frequency with an f_2 frequency spacing of 100–400 Hz or f_2 was fixed at BF and the f_2:f_1 ratio was varied by sweeping f_1. Frequency sweeps were useful because the responses could be examined for tuning and for a traveling-wave-like phase. For the constant f_2:f_1 ratio paradigm, the overlapping pattern of the responses to the primaries was maintained approximately constant and simply moved longitudinally along the organ of Corti as the stimulus frequencies were swept, as described and motivated in Sec. I.

The sound pressure level was calibrated in the EC within 3 mm of the tympanic membrane using a Bruel & Kjaer probe-tube microphone system (Sokolich, 1977). The frequency-dependent transfer function of the probe tube was accounted for when setting the sound pressure level and analyzing the data. This microphone also served as the receiver of otoacoustic emission pressure, as described previously (Dong and Olson, 2006). Data acquisition times were either 0.4 or 1 s for individual stimulus sets. The longer time was used with some of the low level stimuli to increase the signal-to-noise ratio. The noise level indicated in the figures was determined by taking the average of the microphone voltage spectra recorded during data collection, after removing the prominent peaks at primary and $2f_1-f_2$ and $2f_2-f_1$ DP frequencies. The source of noise is photon shot noise in the case of the sensor, and electronic noise in the case of the probe microphone system, and both are flat in frequency. The voltage noise was converted to equivalent pressure noise using sensor and microphone voltage-to-pressure conversions. The microphone conversion is not flat with frequency, and so the noise level of the microphone expressed in pressure exhibits mild frequency structure. With 1 s data acquisition time its noise level was ~5–10 dB SPL up to 30 kHz, and slightly higher at higher frequencies. In order to measure the DPOAE over a wide frequency range with equal-intensity primary tones, the stimulus level needed to be above 60 dB SPL. The noise floor of our intracochlear pressure measurement is 50–60 dB SPL, and the DP was typically above this level only with primary levels at least 60 dB SPL. Therefore, for the study of distortion products, both our intracochlear and ear canal pressure measurement systems limit us to stimulus levels above ~60 dB SPL. Measurements made at lower levels would extend these studies and offer further information about some of the key issues. However, distortion generated with relatively high sound stimuli likely has the same basic origin (cochlear outer hair cells) as low level distortion (Kim *et al.*, 1980; Lukashkin *et al.*, 2002; Mills, 2002; Avan *et al.*, 2003), and we will show below that its tuning and physiological fragility are consistent with this view. Acoustic distortion was checked in a cavity and was at least 70 dB beneath the level of the primaries when the primary level was 100 dB SPL, and smaller for lower primary levels (Dong and Olson, 2005b), so artifactual distortion did not influence our results.

The design, construction, and calibration of the fiber-optic pressure sensor has been described (Olson, 1998). The sensors used in the present study were 150 μm in outer diameter, slightly smaller than the 167 μm outer diameter of the original design. The sensors operate with wide band sensitivity (Pa/V) that is flat to within a few decibels to at least 50 kHz. The sensitivity can change due to small adjustments of the sensitive membrane. Therefore, the absolute sensitivity is known to ~±10 dB. An individual sensor tip might be used in several experiments or just one if the sensitive membrane was damaged. They were positioned close to the BM in the basal-turn ST of the cochlea (BF~20 kHz) by advancing in micrometer steps with a stepper motor. The distance to the BM was judged by tapping it, whereupon a noisy low-frequency signal was observed on the monitor oscilloscope. Most of the measurements were made close to the sensory tissue, but sometimes additional measurements were made at various distances from the BM. The question of whether the sensor perturbed cochlear mechanics is salient. It has been explored previously (Olson, 2001) and below we explore it further.

C. Phase calculation

When considering ST primary or single-tone pressure responses to the EC stimulus, the phase is simply referenced as

$$\phi_{\text{ST-re-EC}} = \phi_{\text{ST}} - \phi_{\text{EC}}$$

(at primary or single-tone frequencies).

When considering the DPOAE in the EC as a response to the DP in ST (when looking for a reverse wave), the phase is simply referenced as

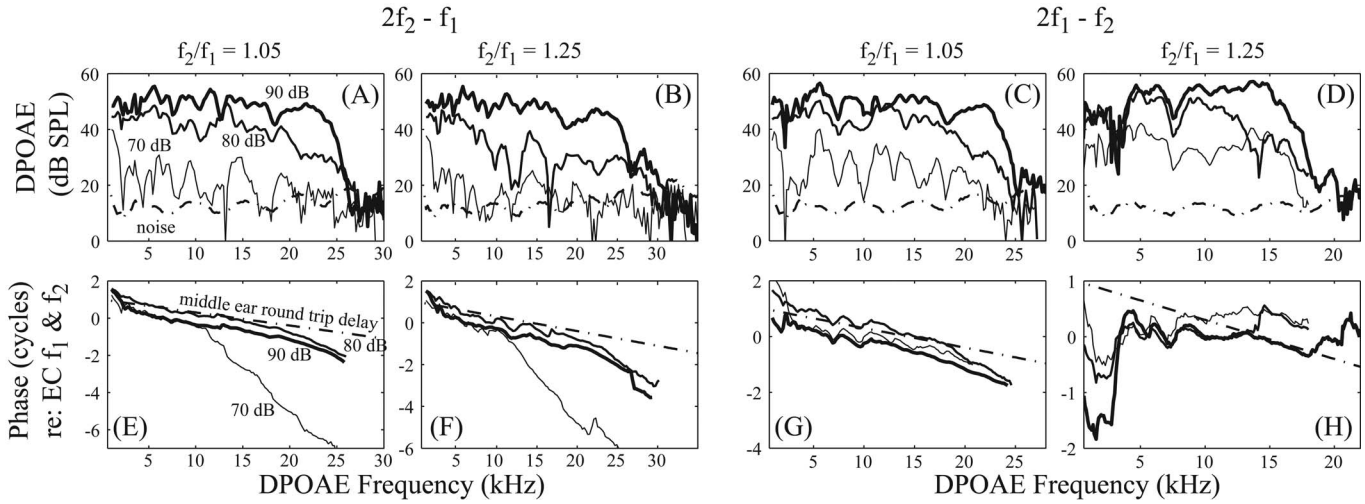


FIG. 2. The $2f_1-f_2$ and $2f_2-f_1$ components measured from an ear with intact cochlea showing the complex nature of the DPOAE. The bulla was open for measuring CAP thresholds at the round window and the cochlea was intact. (A) and (B) The $2f_2-f_1$ amplitude with $f_2:f_1$ ratio of 1.05 and 1.25. (C) and (D) The $2f_1-f_2$ amplitude with $f_2:f_1$ ratio of 1.05 and 1.25. Dotted–dashed line shows the noise floor of the B & K probe-tube microphone. (E)–(H) DPOAE phase referenced to f_1 and f_2 phases in the EC. Dotted–dashed line shows the average middle-ear round-trip delay from Dong and Olson (2006). The two primaries were equal intensity of 70 (thin line), 80 (medium), and 90 dB SPL (thick). $f_2:f_1$ ratio was fixed either at 1.05 or 1.25. f_2 swept from 1000 to 30 000 Hz in 200 Hz steps (animal wg92).

$$\phi_{\text{EC-re-ST}} = \phi_{\text{EC}} - \phi_{\text{ST}} \quad \text{(at DP, DPOAE frequencies)}.$$

Because the DP frequencies are not contained in the stimulus, to reference the DP or DPOAE to the stimulus in the EC requires the following (Dong and Olson, 2005b, 2006), where ϕ_{ECf1} and ϕ_{ECf2} are the phases of the EC primaries:

$$\phi_{2f1-f2\text{-re-EC}} = \phi_{2f1-f2} - (2\phi_{\text{ECf1}} - \phi_{\text{ECf2}}) \quad (2f_1 - f_2 \text{ DP or DPOAE}),$$

$$\phi_{2f2-f1\text{-re-EC}} = \phi_{2f2-f1} - (2\phi_{\text{ECf2}} - \phi_{\text{ECf1}}) \quad (2f_2 - f_1 \text{ DP or DPOAE}).$$

In order to account for the phase behavior that is attributable to middle-ear transmission we use the average middle-ear transmission delays observed in Dong and Olson (2006). The phase versus frequency due to middle-ear forward transmission is found as $(-1)\times$frequency $\times$the forward delay of 32 μs. Similarly, the phase versus frequency due to middle-ear reverse transmission and round-trip transmission are found, respectively, as $(-1)\times f \times 38$ μs (the reverse delay) and $(-1)\times f\times 70$ μs (sum of forward and reverse delays). The forward and reverse middle-ear delay times are similar and in some cases the small difference between them is neglected.

III. RESULTS

Seventeen animals contributed to this study. All these animals were preparations with robust nonlinear tuning, nonlinear responses, and starting healthy CAP thresholds. Results from individual animals are shown to illustrate particular points, and grouped data show the generality of the results. Animal wg96 is an important case because some of the DP and DPOAE data were collected with the sensor close to the BM but before tapping it, so these results inform questions about sensor perturbation. Because tapping did not affect the results in wg96, the data from this experiment are also representative of other animals in which tapping occurred prior to most data collection.

A. Basic characteristics of DPOAEs and DPs

1. DPOAEs are complex

In Fig. 2 we show DPOAEs produced with equal-intensity primary levels of 70, 80, and 90 dB SPL in a gerbil with an intact, unopened cochlea. The amplitudes of the $2f_1-f_2$ and $2f_2-f_1$ DPOAE components are in panels 2(A)–2(D), their phases relative to the primaries (f_1,f_2) in the EC are in panels 2(E)–2(H). The DPOAE was level, frequency, and ratio dependent. The DPOAE was observable up to 20–30 kHz, and then dropped to the noise floor. The $2f_1-f_2$ and $2f_2-f_1$ amplitudes were similar with a $f_2:f_1$ ratio of 1.05 [Figs. 2(A) and 2(C)], while the $2f_1-f_2$ was bigger than $2f_2-f_1$ with a $f_2:f_1$ ratio of 1.25 [Figs. 2(B) and 2(D)], especially at the lower primary levels. The DPOAE generally grew with increasing primary level. In general, the DPOAEs elicited with moderate level primaries [e.g., 70 dB data in Figs. 2(A)–2(D)] displayed more fine structure than DPOAEs elicited with high level primaries. The phase of $2f_2-f_1$ changed rapidly with frequency above 12 kHz with the 70 dB SPL primary level [Figs. 2(E) and 2(F)]. In the region of the rapidly varying phase, this DPOAE would be interpreted as a place-fixed, reflector emission. The phase-frequency response of the $2f_1-f_2$ component at a wide $f_2:f_1$ ratio of 1.25 was flat through a fairly broad frequency range [Fig. 2(H)]. The flat phase is associated with the wave-fixed emission type. Because the middle ear round-trip delay is unavoidable, within the cochlea the phase-frequency response must actually slope upward (by simple subtraction of phase slopes), and we will see in what follows that this is in fact the case. The phase of the DPOAE was in other cases very much like the round-trip middle ear delay [Figs.

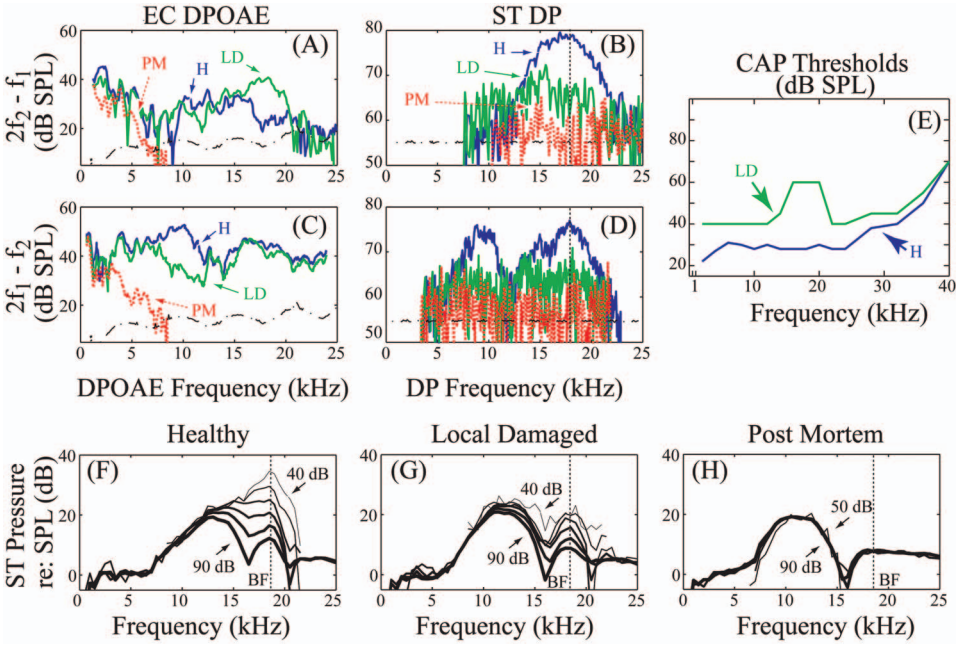


FIG. 3. Simultaneous recording of DPOAE in the EC and DP in the ST close to the BM with different cochlear conditions: healthy (H) (blue), locally damaged (LD) (green), and postmortem (PM) (red). The pressure sensor was positioned ~10 μm from the BM at the basal turn of the cochlea with BF ~18 kHz. (A) and (C) DPOAE $2f_2-f_1$ and $2f_1-f_2$. (B) and (D) DP $2f_2-f_1$ and $2f_1-f_2$. ($L_1=L_2=80$ dB SPL, $f_2:f_1=1.25$, f_2 swept from 1000 to 40 000 in 100 Hz steps.) (E) CAP thresholds under healthy and locally damaged conditions. (F)–(H) Single-tone amplitude responses (normalized to SPL) from the same position/conditions with stimulation levels of 40–90 dB SPL in 10 dB steps. Vertical dotted line indicates the BF position. Single-tone nonlinearity and local DPs were severely reduced with damage and the CAP thresholds were elevated, especially in the frequency region of damage (B), (D), (E), and (G). Postmortem, single-tone responses became linear and the DPs were reduced almost to the noise level (B), (D), and (H). DPOAEs changed over a limited frequency region due to local damage, and were greatly reduced postmortem (A), (C) (animal wg95).

2(E)–2(G)], although often with a superposed jagged structure that lined up with ripples in the magnitude. After accounting for the middle ear round-trip phase, the cochlear phase-frequency response would be flat, and in the dual-component emissions theory, this emission type would also be interpreted as a type of wave-fixed emission.

2. Local damage reduces DPs and changes DPOAEs, death reduces both dramatically

Because we worked at relatively high stimulus levels, it is important to demonstrate the physiological basis of the DPs and DPOAEs we measure. Figure 3 shows how the simultaneously recorded DPOAE [Figs. 3(A) and 3(C)] and DP [Figs. 3(B) and 3(D)] changed due to local damage and death. The cochlear condition is illustrated by CAP thresholds [Fig. 3(E)] and the ST pressure responses measured close to the BM with single-tone stimuli [Figs. 3(F)–3(H)]. The $2f_1-f_2$ and $2f_2-f_1$ are plotted versus their own frequencies, under intact cochlea (blue), local cochlear damage (green), and postmortem (red) conditions.

Before damage, the CAP measurements showed a typical threshold curve, with thresholds about 30 dB SPL up to 24 kHz, and then growing with frequency [Fig. 3(E)]. The single-tone responses showed saturating nonlinearity [Fig. 3(F)]. The BF of this position (peak frequency at low stimulus level) was ~18 kHz. The $2f_2-f_1$ DP was tuned approximately to the BF [Fig. 3(B)] and the $2f_1-f_2$ DP had two peaks [Fig. 3(D)], which will be discussed further in regard to Fig. 5. The DPOAE was broadband, with notches [Figs. 3(A) and 3(C)]. After using the sensor to cause local permanent damage to the organ of Corti, CAP thresholds were elevated at all frequencies except for the highest tested with a pronounced elevation in the 16–21 kHz region [Fig. 3(E)]. (In this experiment the sensor was inadvertently bumped and the local damage technique was not well controlled. We have reproduced elements of this response to localized damage in other experiments by distending the BM with the sensor by ~10 μm.) In addition, the single-tone response gain and degree of nonlinearity was reduced (but not eliminated), indicating that the cochlea's active nonlinear mechanics were damaged [Fig. 3(G)]. At frequencies far from the BF the changes were small and the similarity with predamage values shows that the sensor itself was not damaged when it was used to damage the organ of Corti. The DPs [Figs. 3(B) and 3(D)] were markedly reduced at peak frequencies, and less reduced at frequencies on the low side of the peaks. This applies to both peaks in Fig. 3(D). The DPOAEs [Figs. 3(A) and 3(C)] were only substantially affected in the frequency region corresponding approximately to the BF of the damaged region [slightly below the BF region in Fig. 3(C)]. The $2f_1-f_2$ DPOAE [Fig. 3(C)] decreased; the $2f_2-f_1$ DPOAE [Fig. 3(A)] actually *increased*. This might be because the local damage caused the sensory tissue to be less smooth, enhancing the reflection of DPs. In the postmortem condition, the single-tone responses became linear [Fig. 3(H)], the DPs went into the noise [Figs. 3(B) and 3(D)], and the DPOAEs went into the noise at frequencies above ~7 kHz [Figs. 3(A) and 3(C)]. At frequencies below 5 kHz the DPOAEs remained robust in these data, taken ~30 min postmortem. In other animals, more than 1 hour postmortem the DPOAE disappeared beneath the noise floor even at a stimulus level of $L_1=L_2=90$ dB SPL. In sum, this figure illustrates that the DPOAEs and DPs are both sensitive to localized damage and death and that the DPOAE is sourced by intracochlear distortion even at high stimulus levels.

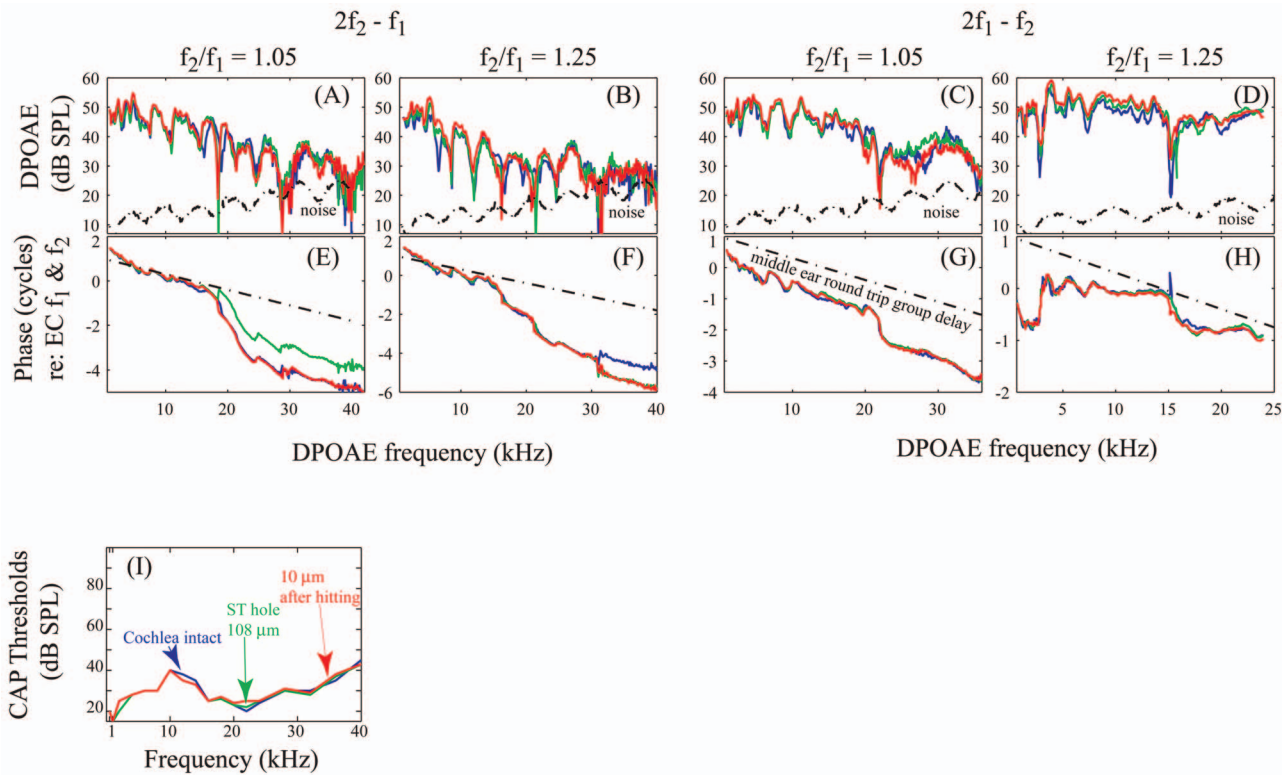


FIG. 4. Lack of sensor perturbation to DPOAE and CAP thresholds. (A)–(D) $2f_2-f_1$ and $2f_1-f_2$ DPOAE amplitude. Dotted–dashed lines indicate the B & K probe-tube microphone noise floor in the EC. (E)–(H) $2f_2-f_1$ and $2f_1-f_2$ DPOAE phase relative to EC f_1 and f_2 phases. Dotted–dashed lines represent the middle-ear round-trip delay. (I) CAP thresholds. Blue lines show data collected with the cochlea intact. Green lines indicate responses with the ST hole and sensor at 108 μm from BM and red lines represent data with the sensor 10 μm from BM after tapping it. The primary stimuli were equal-intensity tones of 80 dB SPL with $f_2:f_1$ ratio fixed at either 1.05 or 1.25. f_2 frequency was swept from 1000 to 40 000 Hz in steps of 100 Hz. The DPOAE and CAP thresholds were nearly unchanged after making the ST hole, introducing the sensor into ST, and after tapping the BM (animal wg96).

3. Introducing the sensor into the cochlea does not perturb cochlear mechanics

The micropressure sensors have been used for nearly a decade to explore cochlear mechanics (Olson, 1998, 1999, 2001, 2004; Dong and Olson, 2005b, a; Olson and Dong, 2006; Decraemer *et al.*, 2007). In previous studies, the presence of the sensor close to the BM sometimes caused the CAP threshold to be elevated a few decibels and in other cases had no effect on CAP (Olson, 2001, Fig. 6). To further explore the possibility that the presence of the pressure sensor perturbs cochlear mechanics, DPOAEs and CAP thresholds are plotted in Fig. 4 under three different conditions: prior to opening the cochlea (blue), after making the sensor hole and positioning the sensor within the cochlea before using it to tap the BM (green), and after hitting the BM and withdrawing 10 μm (red). Changes to the DPOAE [panels 4(A)–4(H) and CAP [panel 4(I)] were barely discernible, and could be due to slight repositioning of the acoustic system. The most pronounced change, in the phase of the low ratio 1.05 $2f_2-f_1$, is tied to the deep notch at ~20 kHz and phase unwrapping, as after they have diverged, the green, and red and blue curves differ by a complete cycle. In summary, Fig. 4 illustrates lack of sensor perturbation. Therefore, unless noted, the data presented here are expected to reflect normal cochlear mechanics.

B. Further characteristics of the DP

The characteristics of locally measured DPs help us to understand their relationship to DPOAEs. In Dong and Olson (2005b, a) and Olson and Dong (2006), locally measured intracochlear DPs at a low $f_2:f_1$ ratio (1.05) were explored in some detail. At frequencies close to the local BF the DPs were tuned and showed similar group delays to the primaries, suggesting they were locally or basally generated. However, at frequencies somewhat below the BF, the group delays were longer than those of the primaries, which indicated nonlocal DP generation. Here, we also show DP results taken with the paradigm with a relatively high $f_2:f_1$ ratio (1.25) to further our understanding of intracochlear DPs.

1. Variations with frequency indicate that DPs are composed of a combination of locally and distantly generated components

Figure 5 shows the $2f_1-f_2$ DP in ST pressure measured close to the BM from wg93. The stimuli were equal-intensity tones of 70, 80, and 90 dB SPL (blue, green and red). The high level results are the most extensive (out of the noise), and 70 dB data extend the results to moderate stimulus levels. The responses are plotted versus the DP frequency in panels 5(A) and 5(B) (amplitude) and panels 5(C) and 5(D) (phase relative to EC f_1 and f_2 primary phases). The amplitude data are replotted versus f_2 in panels 5(E) and 5(F). Single-tone tuning curves and phase are plotted as dotted lines, for comparison.

The low ratio 1.05 data [Fig. 5(A)] show tuning that was much like single-tone tuning in the region of the peak. With primary levels of 70 dB the response was tuned much like

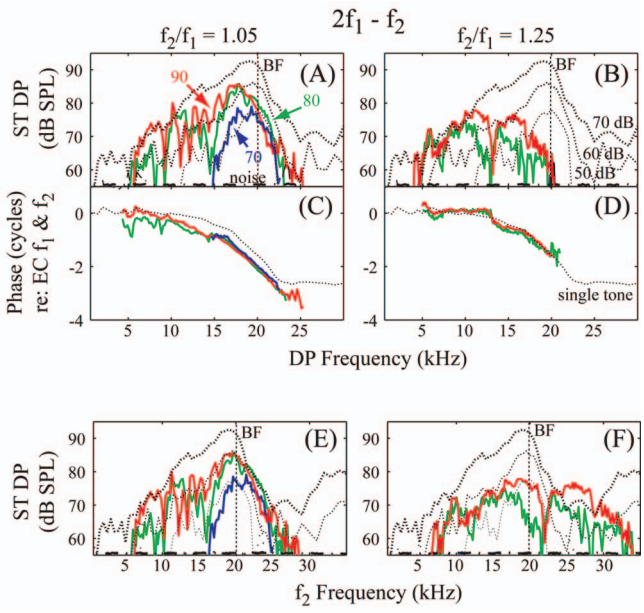


FIG. 5. Scala tympani $2f_1-f_2$ DP amplitude and phase show both locally/basally and remotely generated components. Sensor was positioned 10 μm from the BM, $L_1=L_2=70$, 80, and 90 dB SPL. $f_2:f_1$ was fixed at either 1.05 or 1.25. f_2 was swept from 1000 to 35 000 in 200 Hz steps. (A) and (B) Amplitude of the $2f_1-f_2$ DP with $f_2:f_1$ ratio of 1.05 and 1.25, respectively. (C) and (D) Corresponding phase, referenced to the EC primaries. In (A)–(D) DPs are plotted vs their own frequency. (E) and (F) DP amplitude is plotted vs the f_2 frequency, to better understand the basis for tuning. Solid colored lines show the DP responses; dotted lines show single-tone pressure responses at 50, 60, and 70 dB SPL, measured at the same position, for comparison. Dotted–dashed lines show the noise floor of the sensor and vertical dotted line indicates the BF position (animal wg93).

the 50 dB SPL single-tone response. With primary levels of 80 and 90 dB the peak was shifted to slightly lower frequencies. This shift suggests that the tuning of the DPs resulted from both primary tuning (pregeneration) and DP tuning postgeneration. We pursue this by plotting the DP versus f_2 frequency in Fig. 5(E). The single-tone curve was slightly better fit in Fig. 5(E) than in Fig. 5(A) (at the higher sound levels), supporting the idea that the DP tuning was substantially influenced by primary tuning. In the region of the peak the phase versus DP frequency was much like the single-tone response [Fig. 5(C)], indicating a forward-traveling DP. Below 15 kHz, the DP amplitude showed features that are reminiscent of DPOAE fine structure [Figs. 5(A) and 5(E)]. In this sub-BF frequency region the phase [Fig. 5(C)] departed from the single-tone behavior; it was steeper, and more structured. Thus, the amplitude and phase of the low-frequency DP make it clear that it is not purely locally/basally generated and it is likely that apically produced distortion is contributing.

The high ratio 1.25 $2f_1-f_2$ DPs (right panels) were not tuned like the single tone when plotted versus the DP frequency. Because at high ratio (1.25) the primary and DP frequencies are much more separated than with the low ratio (1.05) condition, it is to be expected that the DP tuning would be more complicated. In Fig. 5(B) two well defined peaks were apparent in the amplitude and correspond to flat (lower-frequency peak) and single-tonelike (higher-frequency peak) phase responses, respectively [Fig. 5(D)]. The dual-peak structure was also observed in other animals [e.g., Fig. 3(D)]. The higher-frequency peak likely corresponds to basally produced distortion traveling forward, but it is curious that it cuts off at frequencies below the BF. When plotted versus f_2 (Fig. 5(F)) it is seen that the second peak extends to f_2 frequencies above 30 kHz. Frequencies of 30 kHz and above are represented in the basal region of the gerbil cochlea that is exposed by the large round window opening. It is possible that in this open cochlea the basal region is beginning to deteriorate (Overstreet *et al.*, 2003). This could account for the second peak's premature high-frequency dropoff compared to the single-tone response in Fig. 5(B). With this understanding, the second peak is convincingly a forward-traveling DP. The lower-frequency peak had a flat phase [Fig. 5(D)], which was similar to the DPOAE response typically observed for the $2f_1-f_2$ component at a high ratio (1.25) [Figs. 2(H) and 4(H)]. In fact, the sub-BF frequency phase was not flat, but sloped slightly upward. When describing the flat $2f_1-f_2$ phase with the $f_2:f_1$ ratio of 1.25 in Fig. 2 we predicted this type of upward sloping DP phase in order to account for both the flat DPOAE phase and the middle-ear delay. The lower-frequency peak at the 90 dB stimulus level in Fig. 5(F) was tuned somewhat like the single-tone f_2 response, but the correspondence is less clear at the 80 dB stimulus level. The sharp notch separating the low- and high-frequency peaks suggests interference, such as would occur if locally and remotely generated components were summing destructively. In sum, the lower-frequency peak in Figs. 5(B) and 5(F) appears to be a combination of locally generated distortion (or basally generated distortion that traveled forward) and distantly generated distortion, traveling out of the cochlea from more apical locations.

2. Spatial variations show that the DPs drop off steeply with distance from the BM

Intracochlear pressure is composed of compression and traveling-wave modes. In Sec. I we described how these appear in measurements; a review point is that the traveling-wave mode is dominant close to the BM at frequencies close to BF and diminishes rapidly with distance from the BM, whereas the compression mode is dominant far from the BM or at frequencies higher than the BF, and is not spatially varying. In Fig. 7 of Dong and Olson (2005b) we showed that DP pressures decreased with distance from the BM. Here we confirm this finding. Figures 6(A) and 6(A′) show the $2f_2-f_1$ DP when the sensor was far from the BM (thin lines) and close to it (thick lines). For comparison, the f_2 primary response at these locations is also plotted [Figs. 6(B) and 6(B′)] The response phases (both f_2 and DP) are shown in Figs. 6(C) and 6(C′). We also show the concurrently measured DPOAE in Figs. 6(D) and 6(D′). The DPOAE changes were small, suggesting that the sensor's location did not influence cochlear mechanics, and confirming the results from Fig. 4. The f_2 amplitude was bigger at the closer than it was at the more distant location, with the largest changes in the region of the BF peak. (Based on low intensity single-tone tuning curves the BF was ~18 kHz for wg95 and 21 kHz for wg92.) The notches in the primary amplitude [arrowheads in

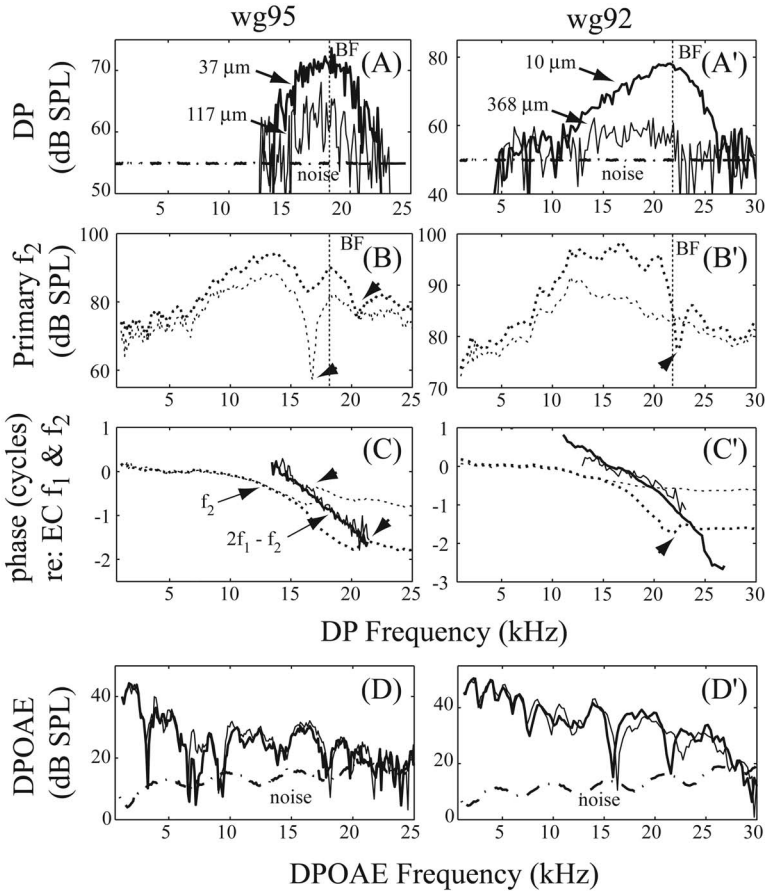


FIG. 6. DP spatial variation shows that the DP drops off with distance from the BM. (A) and (A′) $2f_2-f_1$ DP amplitude measured at distances far (thin) and close (thick) to the BM. Dotted–dashed lines indicate the sensor noise floor. Vertical dotted lines represent the BF position. (B) and (B′) Amplitude of primary f_2 measured close to (thick dotted) and far from (thin dotted) the BM. (C) and (C′) Phase referenced to EC f_1 and f_2. DP and primary f_2 responses are plotted in solid and dotted lines, respectively. Arrowheads indicate frequency/position combinations for which destructive interference between compression and traveling modes is apparent. (D) and (D′) Concurrently measured DPOAE is stable, indicating a lack of significant sensor perturbation. ($L_1=L_2=80$ dB SPL, $f_2:f_1=1.25$, f_2 swept in 100 Hz (animal wg95) or 200 Hz (animal wg92) steps). The DP decreased as the distance from the BM increased, similar to the primary response in the BF frequency region.

Figs. 6(B) and 6(B′)] coincide with phase jumps [arrowheads in Figs. 6(C) and 6(C′)], and are due to destructive interference between traveling- and compression wave pressure modes. At frequencies well above the BF, the compression wave dominated the f_2 responses, which led to the phase and amplitude plateaus. When the sensor was positioned far from the BM the region of compression wave dominance started at a lower frequency [phase plateau region thin-dotted versus thick-dotted lines in Figs. 6(C) and 6(C′)]. The phase plateaus at the different locations are separated by a full cycle. Over much of the frequency range, the DP phase lined up with the forward-traveling f_2 phase (so the DP is also forward traveling, or locally generated). However, in Fig. 6(C) and 6(C′) at frequencies less than 15 kHz, the DP phase was steeper than the primary phase, suggesting that in this region, apically generated distortion traveling backward was contributing to the measured DP.

Most importantly, the DP amplitude decreased substantially with distance from the BM [Figs. 6(A) and 6(A′)]. The rapid spatial variations in the DP prove that it is not a compression pressure, since the observed pressure differences must drive bulk acceleration of the cochlear fluid. The spatial variations observed in the DPs are similar to those of the traveling-wave pressure mode of the primary response, which argues for a traveling-wave mode of DP travel. In general, the DP steadily dropped to the noise level as the measurement position was moved further from the BM. Therefore, the noise level (~55–60 dB SPL) gives the upper limit for the size of a compression-type (approximately space-filling) DP pressure.

C. Relating DPOAEs to DPs

A robust feature of cochlear mechanics is the phase-frequency relationship of the forward-traveling wave, illustrated in Fig. 1(C). Therefore, a very compelling demonstration of a reverse-traveling wave would be the observation of the same phase-frequency relationship, in reverse. The forward-traveling-wave phase is found by referencing the intracochlear response to the EC stimulus and the reverse-traveling phase can be sought by referencing the EC emission to the intracochlear DP. If this DPOAE–DP phase overlies the forward-traveling-wave phase, the reverse-traveling wave will be supported. In the following figures, we show the result of this referencing. We will find that the DPOAE–DP phase results cannot stand on their own definitively, but that coupled with the individual DP and DPOAE results, compelling conclusions can be drawn.

Figures 7–10 show the simultaneous measurements of DPOAE and DP in animals wg81 and wg96. The amplitude and phase of $2f_2-f_1$ DPOAE and DP are in panels (A)–(D); amplitude and phase of $2f_1-f_2$ DPOAE and DP are in panels (F)–(I); the DPOAE–DP phase is in panels (E) and (J)—it is this plot that will be explored for a reverse-traveling-wave character. DPs in these figures were measured close to the BM (within 30 μm). For comparison/reference, average middle-ear round-trip delay is plotted in panels (B) and (G) (dotted–dashed lines) and middle-ear reverse transmission delay is plotted in panels (E) and (J) (dotted–dashed line). The single-tone phase response to a 50 dB SPL stimulus tone from the same experimental animal and conditions as the DP,

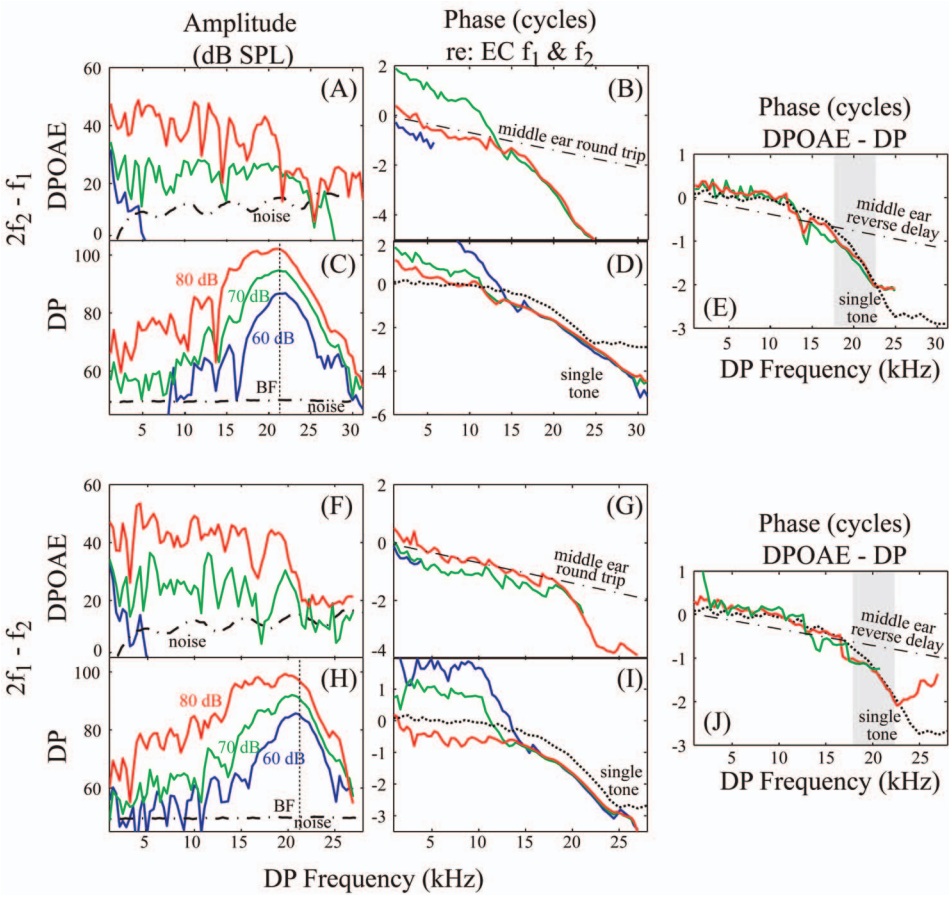


FIG. 7. Simultaneous recording of the DP and DPOAE with $f_2:f_1=1.05$. $L_1=L_2=60$ (blue), 70 (green), and 80 dB SPL (red), the sensor was positioned 20 μm from BM. f_2 was swept from 1000 to 40 000 in 400 Hz steps. (A), (C), (F) and (H) The $2f_2-f_1$ and $2f_1-f_2$ DP and DPOAE amplitudes. Dotted–dashed lines indicate the noise floor of the sensor and B and K probe-tube microphone. Vertical dotted lines represent the BF position. (B), (D), (G) and (I) Phases of DP and DPOAE referenced to EC f_1 and f_2 phases. Dotted–dashed lines show the middle-ear round-trip delay and dotted lines show the single-tone forward-traveling-wave phase, for comparison. (E) and (J) Phase of the DPOAE referenced to the DP phase. When this phase overlaid the forward-traveling wave of the single tone, the reverse wave is considered to be detected. Within the BF frequency region (gray bar), the observation of the reverse wave depends upon the DPOAE phase being rapidly varying, as then the phase contains information about the traveling-wave delay. In the sub-BF region, the reverse wave can be expected as long as the DP amplitude and phase behavior [in panels (C) and (D), (H) and (I)] indicates that the DP is not dominated by local distortion. Except for in these two regions, a reverse wave cannot be expected to be detected (animal wg81).

DPOAE data are plotted as a dotted line in panels (D), (E), (I), and (J). This curve is included to show the forward-traveling-wave phase.

1. Sub-BF region

The BF of both animals in Figs. 7–10 was 21 kHz. At frequencies below ~15 kHz amplitude and phase fine structure was apparent in the DP as well as the DPOAE. The DP phase was not similar to the single-tone response, and the tuning of the DP was not like the single tone (confirming results from Fig. 5) so the DP did not appear to be forward traveling. In some cases details of fine structure in the DP were precisely echoed in the DPOAE (prominent cases are marked with double arrows in Fig. 9) but the general finding is that the sub-BF region of the DP possessed fine structure. Middle ear transmission imposes fine structure on the DPOAE (Dong and Olson, 2006) so we cannot expect every detail in the sub-BF DP to be evident in the DPOAE. The DP fine structure is likely produced by summing interference, either between locally or basally generated and apically generated distortion, or simply between wave-fixed and place-fixed distortion components, both arriving at the basal measurement location from more apical locations. The DPOAEs and DPs exhibited similar level dependence; in Fig. 8, when the stimulus level increased from 60 to 70 dB there was less than a 10 dB change in the response, but the increase to 80 dB caused a greater than 10 dB change in both the DP and DPOAE at most frequencies below 15 kHz [Fig. 8(F) and 8(H)]. The phase of both the DPOAE and DP also changed in character with this level change [Figs. 8(G) and 8(I)]. In the DP the phase slope went from slightly upward sloping to flat and in the DPOAE from flat to slightly downward sloping. In other cases the low-frequency phase of the DP and DPOAE were both relatively rapidly varying, as in panels (B) and (D) of Figs. 7–10 (the $2f_2-f_1$ DP). The rapidly varying DP phase is not similar to the single-tone response and thus does not appear to be locally/basally generated. On the other hand, the large group delay the steep phase represents can be explained as being a reflector emission from more apical locations.

When we plotted the phase DPOAE–DP, it was very similar to the forward-traveling-wave phase in the sub-BF region panels (E) and (J) of Figs. 7–10, which is as expected for a reverse-traveling wave. However, the phase variation was small in this region and not significantly different from what middle-ear transmission delay would produce. Nevertheless, the strong similarity between the forward and reverse phases measured in the same experimental preparation demonstrates a compelling similarity between sound traveling in and sound traveling out of the cochlea.

2. BF region

As we and others have noted (Robles *et al.*, 1997; Dong and Olson, 2005b, a; Olson and Dong, 2006), in a broad region of the BF, the intracochlear DP is usually tuned like a single tone, with a forward-traveling-wave phase. This characteristic is expected for locally generated distortion or a basally generated DP traveling forward. It was also seen in

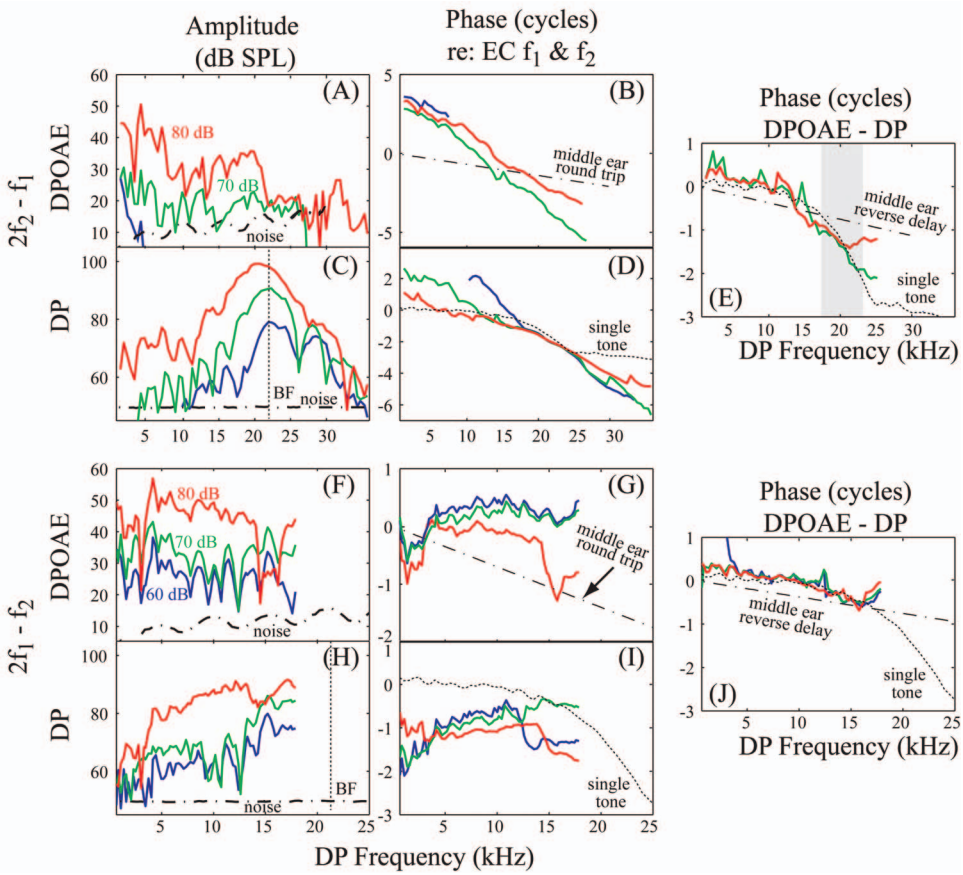


FIG. 8. Simultaneous recording of the DP and DPOAE with $f_2:f_1=1.25$ with the same format, measurement position, and animal as in Fig. 7. In the relative phase data of panel (E), in the BF region [gray bar in panel (E)] the reverse wave was apparent in the 70 dB data (~18–23 kHz region) but less so at the higher, 80 dB SPL level. At the higher level, the DPOAE phase was slightly less steep than at the 70 dB level in the 20–25 kHz region. The $2f_1-f_2$ DP was not recorded in the BF region for this ratio=1.25 case.

the data of Figs. 7–10. Therefore, unlike the sub-BF region, we have no reason to expect that in general the locally measured DP at frequencies around the BF will be responsible for the DPOAE. When we reference the DPOAE phase to the DP, the results take several different shapes.

Starting with Fig. 7(E), the DPOAE–DP phase, after a wiggle at 13–15 kHz, was similar to the forward single-tone phase through and slightly above 20 kHz. The basis for the reverse-wave phase behavior in the BF region can be found in the DPOAE phase, which underwent a change in slope from shallow to steep at ~15 kHz. Comparing the DPOAE and DP phases in the BF region in Fig. 7(E), the DPOAE phase slope was approximately two times the DP phase slope, and the DP phase was forward traveling in character. Therefore it is little wonder that when comparing these phases directly (DPOAE–DP), a reverse wave is indicated. Based on this phase behavior, the DP does appear to be responsible for the DPOAE.

In Fig. 7(J), the $2f_1-f_2$ DPOAE–DP phase was like the single-tone phase in the 20 kHz region. As above, the reason for the similarity is in the DPOAE phase of 7(G), which became steep in the ~18–22 kHz region, with a slope that was approximately twice the forward-traveling DP phase of Fig. 7(I). In Fig. 8(E), the DPOAE–DP phase was similar to the single-tone phase, but only at frequencies very close to the BF. The DPOAE at the 80 dB SPL stimulus level was offset by a full cycle from the 70 dB SPL DPOAE over most of the frequency range, so these two phases are actually approximately equal, although they appear different in the plot. Above 20 kHz, they do diverge slightly, with the 80 dB phase data flattening out somewhat. The 70 dB DPOAE–DP phase was similar to the single-tone data up to ~22 kHz, but with the 80 dB stimulus level the correspondence was more limited.

Data from another animal, in Figs. 9 and 10, showed similar trends. In Figs. 9(E) and 9(J) (60 and 70 dB data), and 10(E) (70 dB data) in the BF region the DPOAE–DP phase was similar to the single-tone phase. In all these cases, the DP was forward traveling, and the DPOAE was steep, with a slope in the BF region approximately equal to twice the forward-traveling phase. However, we see something unexpected in Figs. 9(G), 9(I), and 9(J). The $2f_1-f_2$ DPOAE–DP phase, 80 dB data, in Fig. 9(J) sloped upward with frequency: the DPOAE phase led the DP phase, *as if the DPOAE preceded the DP*. This seemingly paradoxical result is similar to the previously reported finding in which, based on their relative phases, the DP in stapes vibration appeared to precede the intracochlear DP in BM motion (Ren, 2004). The explanation offered in that study was that the DP exited the cochlea essentially instantaneously as a compression (sound) wave, and set the stapes in vibration, which then launched a forward-going traveling wave. On the other hand, the paradox can be adequately explained without invoking an instantaneous compression pressure by revisiting the DPOAE and DP results. At frequencies greater than 15 kHz, the DP phase was that of a forward-going traveling wave [Fig. 9(I)]. This is readily understood as locally generated distortion by the forward-traveling primaries, or basally generated distortion traveling forward. Within the BF frequency region the DPOAE phase had a nearly flat phase [Fig. 9(G)].

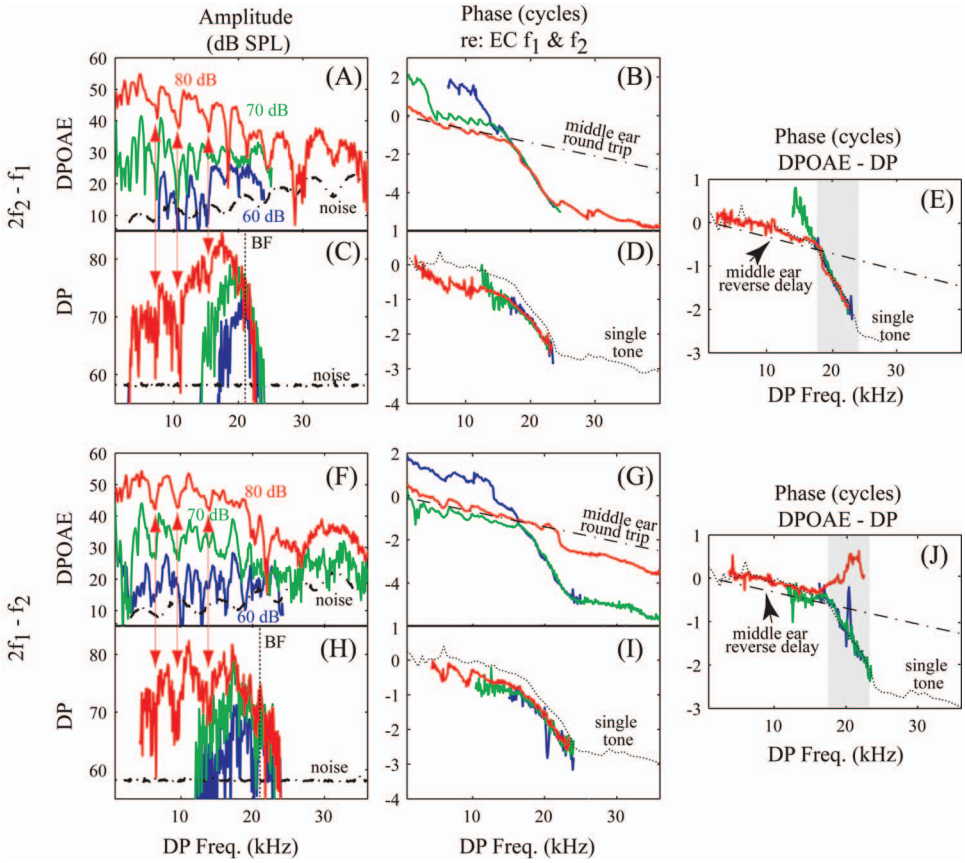


FIG. 9. Simultaneous recording of the DP and DPOAE with f_2:f_1=1.05. L_1 =L_2=60 (blue), 70 (green), and 80 dB SPL (red), sensor positioned 10 μm from the BM, f_2 swept from 1000 to 40 000 in 100 Hz steps. 60 dB SPL data were averaged 50 times rather than the usual 20 times. In the sub-BF region (less than ~14 kHz), the double arrows indicate that the fine structure in DP amplitude was echoed in the DPOAE. The DPOAE–DP phase supported the reverse wave in all cases in which data were available. In the BF region (gray bars) the reverse-traveling wave was only apparent when the DPOAE phase was rapidly varying. When the DPOAE phase was not rapidly varying [panel (G)] (80 dB) relating the DPOAE and DP phases caused a paradoxical result [panel (J)] in which the DPOAE appeared to lead the DP at frequencies close to BF (animal wg96).

This DPOAE phase can also be interpreted within the theory of cochlear emissions, as a wave-fixed/generator emission (Kemp, 1986). However, in a wave-fixed type emission, the slope of the phase-frequency curve is not related to delay. Since the DPOAE and DP phases can be understood individually within the framework of forward and reverse cochlear traveling waves, their relative phase, the DPOAE–DP phase, is therefore also in keeping with that framework, and *is not* at odds with it.

We performed similar measurements with a fixed f_2 stimulus paradigm, and Fig. 11 shows the simultaneous recordings of the $2f_1-f_2$ DP and DPOAE from animal wg96 (the same as in Figs. 9 and 10). With this paradigm the position of the f_2 response pattern was fixed and the overlapping region of f_1 and f_2 response patterns decreased as the f_1 frequency swept from high to low (increasing the f_2:f_1 ratio). Both the DPOAE and the DP were ratio dependent. The amplitude of the DP was ~80 dB from 11 to 19 kHz (11 kHz corresponding to f_2:f_1=1.29), and then decreased quickly with increasing f_2:f_1 ratio. On the other hand, the DPOAE increased slightly as the ratio increased from 1.05 and peaked at 14 kHz [Fig. 11(A)], corresponding to an f_2:f_1 ratio of 1.18. Similar to the DP, the DPOAE also decreased quickly below 11 kHz, corresponding to a f_2:f_1 ratio greater than 1.29 [Fig. 11(A)]. Above 15 kHz, the DP phase-frequency slope was similar to that of the single-tone forward-traveling-wave phase, but below 15 kHz the phase slope was greater than that of the forward-traveling-wave phase. This suggests that at frequencies around the BF, the DP was locally/basally generated, while at sub-BF frequencies it was not. The DPOAE phase was steeper than the DP at frequencies below 15 kHz, but then became shallower. [Unlike the fixed f_2:f_1 ratio paradigm, with the fixed f_2 paradigm the steeply varying DPOAE phase is not a clear indicator of place-fixed emission type, as even a wave-fixed emission is expected to have substantial phase-frequency slope (Shera *et al.*, 2000), so this useful analysis cue is missing in these data.] When we compared DPOAE to DP phases [Fig. 11(C)], the results were similar to those with a fixed ratio. In the sub-BF region, the DPOAE–DP phase overlaid the forward-traveling-wave phase, supporting the reverse-traveling-wave hypothesis. At 15–17 kHz (the region in which the DPOAE phase flattened somewhat) the DPOAE–DP phase sloped upward, as though the DPOAE led the DP, which seems counter to the reverse wave hypothesis (Ren, 2004). In summary, the fixed f_2 DPOAE–DP phase results were similar to the fixed f_2:f_1 ratio results, and showed both reverse-wave and paradoxical behavior. The interpretation of the fixed-ratio-result of Figs. 7–10 benefitted from the ease of separation into wave-fixed/generator and place-fixed/reflector type emissions. These interpretative cues are missing with the fixed f_2 paradigm, but the data are consistent with the same interpretation as in the fixed-ratio case.

3. Grouped data: DPOAE and DP

The exploration for evidence of a reverse wave relied on analysis of the phase responses of the DPOAE and DP. Above, these two responses were compared in individual animals and certain patterns emerged. In particular, in the BF region the reverse wave was only detectable when the

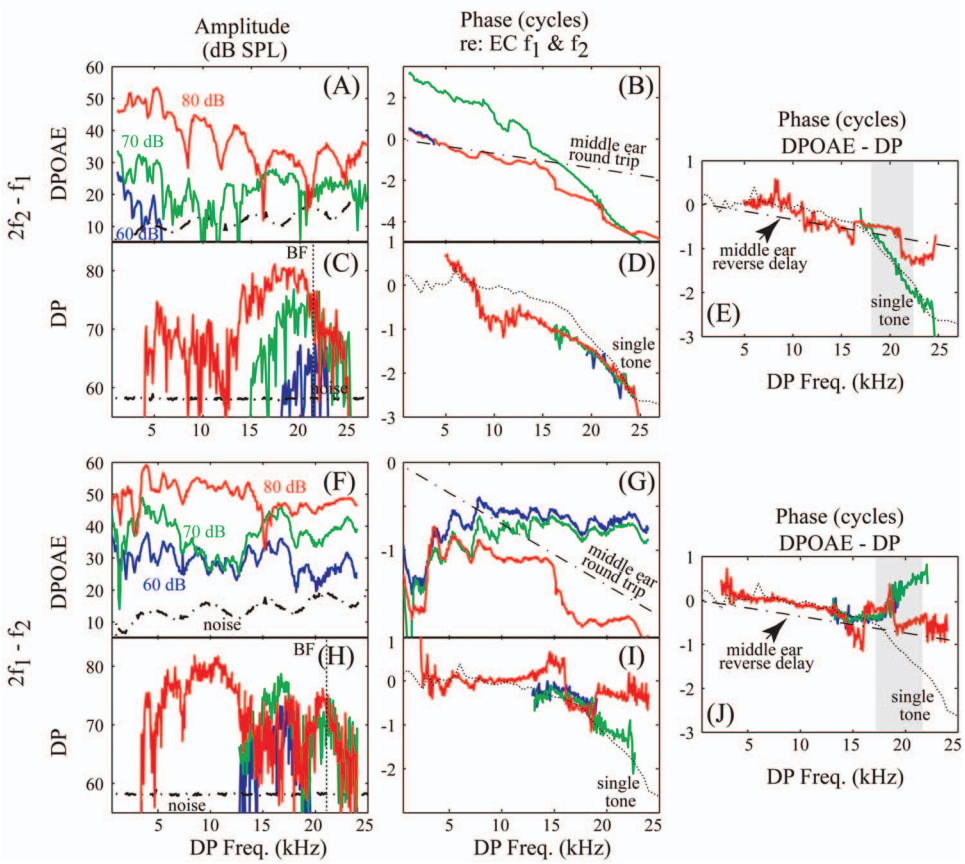


FIG. 10. Simultaneous recording of the DP and DPOAE with $f2:f1=1.25$ with the same format, measurement position, and animal as in Fig. 9. In the BF region (gray bars), the detection of the reverse-traveling wave depends upon the DPOAE phase, which is level dependent. In the BF region (gray bars) the reverse-traveling wave was only apparent when the DPOAE phase was rapidly varying [70 dB stimulus level of panel (B)]. When the DPOAE phase was not rapidly varying [panels (G) and (J)] relating the DPOAE and DP phases caused a paradoxical result in which the DPOAE appeared to lead the DP at frequencies close to BF. In the BF region, the 80 dB stimulus level data in panel (B) are subject to a stair–steplike behavior (coupled to sharp fine structure in the amplitude) that gives rise to a messy DPOAE–DP result.

DPOAE had a rapidly varying phase-frequency response. In order to convey the generality of the DPOAE phase results, in Fig. 12 we show grouped DPOAE data from the eight animals in this study for which the 70 dB stimulus gave emission responses above the noise floor over a relatively wide frequency region. These are more interesting than the 80 dB DPOAEs since they are more apt to show the steep phase-frequency slope of the reflector/place-fixed emission type, whereas the 80 dB DPOAEs tend to show the flat phase-frequency slope of the generator/wave-fixed emission (see DPOAE phase level dependence in Fig. 2). Recall that the group delay of the place-fixed/reflector emission type is related to the group delay of the forward-traveling wave at its BF. If the reverse wave exists, the group delay is expected to be approximately two times the forward-traveling wave group delay at its BF, and if it does not exist, it is expected to be approximately equal to the forward group delay. In order to aid in the comparison, we include a single-tone 50 dB phase-frequency curve from one animal (wg81), multiplied by 2. (The intracochlear location of measurement of all the animals in this study was ~20 kHz, so it is sufficient to include the comparison curve from a single animal.) In contrast to the rapidly varying place-fixed emission phase, the flat phase of the wave-fixed/generator emission, while powerful in its relation to scaling symmetry, does not contain information about forward and reverse group delays (Kemp, 1986; Zweig and Shera, 1995; Shera and Guinan, 1999). Theoretically, the wave-fixed/generator emission is flat

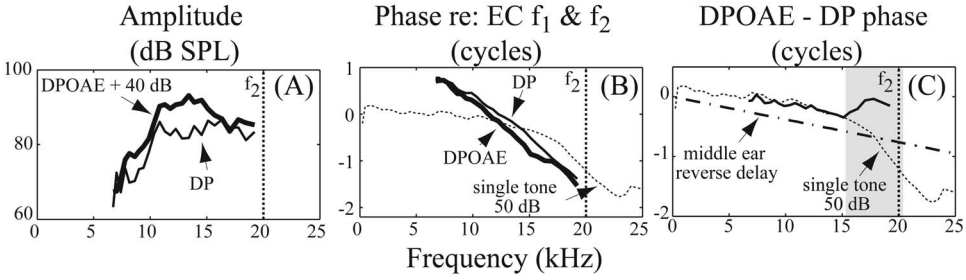


FIG. 11. Simultaneous recordings of $2f_1-f_2$ DP and DPOAE with f_2 fixed at BF. $L_1=L_2=80$ dB SPL, $f_2:f_1$ changed from 1.02 to 1.5 in steps of 0.02. The sensor was positioned 10 μm from the BM. (A) The $2f_1-f_2$ DPOAE (thick line) and DP (thin line) amplitude. The DPOAE amplitude was moved up 40 dB to facilitate the comparison. The vertical line indicates the BF=f_2 frequency. (B) The $2f_1-f_2$ DPOAE (thick) and DP (thin) phase, referenced to EC f_1 and f_2. (C) Phase of DPOAE–DP. The dotted lines show single-tone 50 dB stimulus level response phase relative to the pressure stimulus in the EC. Dotted–dashed lines show average middle-ear reverse delay. The reverse wave was detected in the sub-BF region below 15 kHz (C). Above 15 kHz the DPOAE seemed to lead the DP. Because the stimulus paradigm was not a fixed ratio, the DPOAE phase does not allow for identification of wave-fixed and place-fixed emission type, making interpretation of these data less straightforward than with the fixed-ratio paradigm (animal wg96).

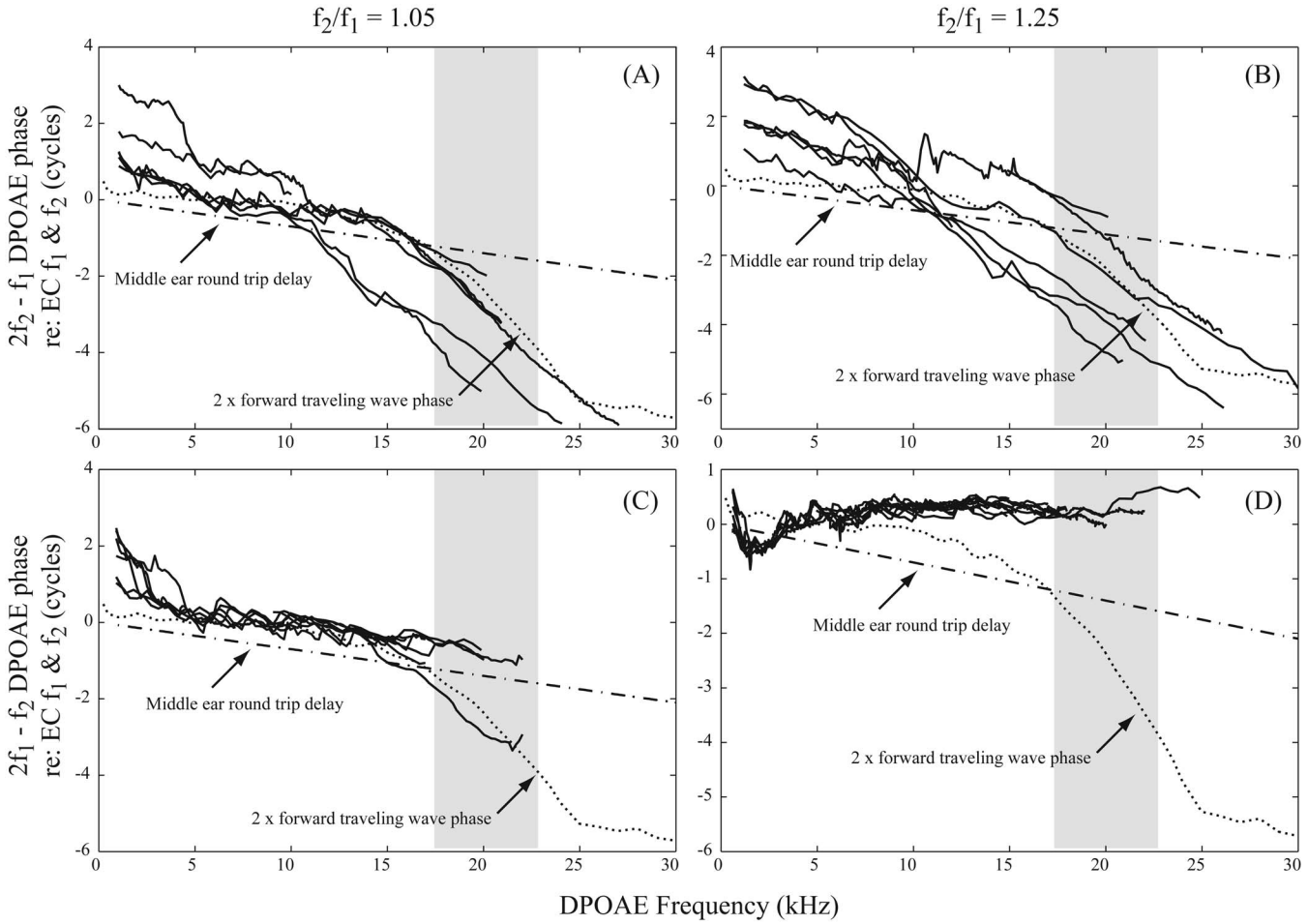


FIG. 12. The grouped phases of the $2f_2-f_1$ and $2f_1-f_2$ DPOAEs from eight animals illustrate the generality of the responses. These preparations had robust DPOAE responses over a wide frequency range for 70 dB primaries. (When the primaries are at a level of 80 dB SPL, the phases tend to flatten, and are less interesting.) All phases were referenced to EC f_1 and f_2 phases. (A) and (B) The $2f_2-f_1$ phase with $f_2:f_1$ ratios of 1.05 and 1.25, respectively. (C) and (D) The $2f_1-f_2$ phase with $f_2:f_1$ ratios of 1.05 and 1.25, respectively. The dotted line is twice the single-tone pressure response phase measured close to the BM from a representative case (wg81, 50 dB SPL), shown for comparison. The dotted–dashed line also shows the average middle-ear round-trip delay for comparison. Gray bars identify the 20 kHz region that can be compared to the DP data gathered during this study, for which the BF was ~20 kHz (animals wg81, wg89, wg90, wg92, wg93, wg94, wg95 and wg96).

within the cochlea, and when measured in the EC would evince middle-ear forward and reverse delays. Therefore, in order to aid the comparison, we also include the average middle-ear round-trip delay phase.

Immediately apparent in Fig. 12 is that the character of each DPOAE type is conserved among animals. The $2f_1-f_2$, 1.05 $f_2:f_1$ ratio DPOAE phase [Fig. 12(C)] was slightly downward sloping at frequencies below 15 kHz, with a slope almost identical to the middle-ear round-trip delay [see also Figs. 2(G), 4(G), 7(G), and 9(G)]. Therefore, within the cochlea this phase would be flat, consistent with a wave-fixed emission. Wiggles appear in the phase data of panel 12(C), suggesting the presence of a lesser contribution from a place-fixed/reflector emission component. The 1.25 $f_2:f_1$ ratio $2f_1-f_2$ phase in panel 12(D) did not show this middle-ear delay, rather the phase slope tended to be nearly flat, or even upward sloping. This was also seen in Figs. 2(H), 4(H), 8(G), and 10(G). Therefore, within the cochlea the DP phase-frequency curve will slope clearly upward. The $2f_2-f_1$ phase often had a gradual slope that was approximately equal to the round-trip middle-ear delay up to a certain frequency, and then abruptly changed to a steeper slope [Figs. 12(A) and 12(B), also in Figs. 2(E) and 2(F), 4(E) and 4(F), 7(B), 8(B), 9(B), and 10(B)]. The underlying cochlear mechanism for the shift point is unknown. We observed in previous figures that the shift generally occurred at lower frequencies for lower stimulus levels [Figs. 2(E) and 2(F), 7(B) and 9(B) and 9(G)]. Theoretically, the steep slope phase corresponds to the place-fixed/reflector emission type. In the ~20 kHz region, the steep slope is similar to twice the forward-traveling-wave delay found in our intracochlear measurements, with BF of ~20 kHz. This is consistent with the presence of a reverse-traveling wave from the 20 kHz region giving rise to the 20 kHz DPOAE. To quantify this correspondence, we found the group delay (negative slope) of the phase-frequency curves at 20 kHz for all the curves in Fig. 12 whose slope was significantly steeper than the middle-ear round-trip delay in the 20 kHz region. This was all of the curves in panels 12(A) and 12(B). The group delays averaged 0.41 ms, with a standard deviation of 0.11 ms. This can be compared with the forward-traveling-wave data (dotted line), whose group delay at 20 kHz was 0.23 ms. When multiplied by 2, as in

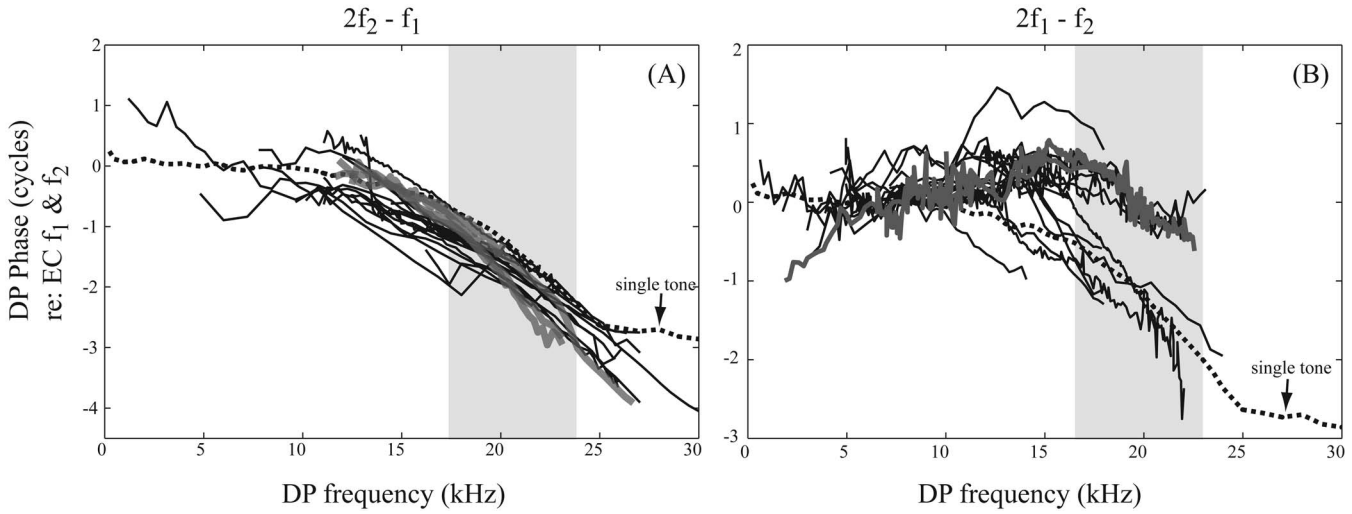


FIG. 13. The $2f_2-f_1$ and $2f_1-f_2$ DP phases from all 17 animals that contributed to this study. The sensor was positioned within 30 μm of the BM, $L_1=L_2=70$ (gray) and 80 dB SPL (black), $f_2:f_1=1.25$. (A) The $2f_2-f_1$ phase; (B) $2f_1-f_2$ phase. All phases referenced to EC f_1 and f_2 phases. The dotted line shows the single-tone response phase measured close to the BM from a representative case (wg81, 50 dB SPL), shown for comparison. The $2f_2-f_1$ phases overlaid the single-tone (forward-traveling wave) phase at frequencies in a broad BF region, indicating that local distortion dominated the DP there. The $2f_1-f_2$ DP phase sloped up at low frequencies, and was not similar to the single-tone response there. In the region of the BF, the phase typically sloped downward similarly to the single-tone phase, suggesting a forward-traveling-wave DP in this frequency region. The gray bar identifies the ~20 kHz BF region probed in this study.

the figure, this is 0.46 ms. Scatter is expected; the reflector emission is not expected to emerge from a point in the cochlea, but rather from the region within the cochlea where the DP is relatively large (Shera and Guinan, 2003).

Figure 13 shows grouped DP phase data from all the animals used in this study. The stimulus level of the plotted curves was in general 80 dB SPL (black curves), but several results found with a stimulus level of 70 dB SPL curves are included (gray curves) and show that those results are similar. (Also see Figs. 5 and 7–10, which show that level does not change the DP phase significantly in the BF region.) The 80 dB data are emphasized in Fig. 13 in order to include all the study animals in the plot with results above the noise through a wide frequency region, and the similarity with 70 dB data when they are available justifies this emphasis. With the 1.25 $f_2:f_1$ ratio, the $2f_1-f_2$ and $2f_2-f_1$ results were quite different from one another. [The 1.05 $f_2:f_1$ ratio DPs both look similar to each other, and were similar to the $2f_2-f_1$ results at the 1.25 $f_2:f_1$ ratio, and are not shown. Extensive observations on the 1.05 ratio DP were presented previously (Dong and Olson, 2005b, a; Olson and Dong, 2006).] Like the DPOAE, the character of the different DPs is conserved among animals. Throughout the broad BF region, the $2f_2-f_1$ DP phase has the forward-traveling-wave character that is easily attributable to local or basally generated distortion. At frequencies somewhat below the BF (below ~15 kHz), this similarity breaks down, signaling a lack of dominantly local distortion. These characteristics confirm the generality of the $2f_2-f_1$ results highlighted in Figs. 7 and 9. The $2f_1-f_2$ DP sloped mildly upward up to ~15 kHz. Close to the 20 kHz BF, the slope was usually downward, with a slope similar to the single-tone phase, which would occur if the DP had traveled to the BF location postgeneration. These characteristics of the grouped data confirm the generality of the $2f_1-f_2$ results in Figs. 5, 8, and 10.

IV. DISCUSSION

DPOAEs are a noninvasive measure of cochlear mechanics. The single place measurement of the DPOAE reflects processes that occurred within the cochlea's complex three-dimensional (3D) structure. The use of the DPOAE to probe these processes has proceeded with certain simplifying assumptions. A robust feature of cochlear mechanics is a forward-traveling wave, and one of the assumptions built into theories of emissions was that they proceed out of the cochlea as a reverse-traveling wave. Many characteristics of DPOAEs do support the reverse-wave picture. However, a recent study aimed at detecting the reverse wave within the cochlea did not detect it (Ren, 2004; He *et al.*, 2007). Those experiments worked as follows: a 1 mm extent of BM was examined. BM velocity was measured. Two-tone stimulation was applied, and the response measured and analyzed to reveal the spatiotemporal pattern of f_1, f_2 and $2f_1-f_2$. Responses at f_1 and f_2 frequencies were observed to travel forward: their phase lags accumulated with distance from the stapes. A wave traveling in the other direction—out of the cochlea, would accumulate phase lag as it got closer to the stapes. When the $2f_1-f_2$ intracochlear DP was examined, it also accumulated phase lag with distance from stapes, indicating a forward-traveling wave. Following up, intracochlear DPs in BM motion were compared with stapes DPs (a proxy for the DPOAE). A fixed f_2, swept f_1 paradigm was used, with f_2 either close to or ~1/2 octave below the local BF. The phase-frequency slope of the BM DP was compared to the DP on the stapes. The BM DP had the steeper slope (indicating greater delay), suggesting that the stapes DP preceded the BM DP. Faced with these observations, an alternative process to the reverse wave was hypothesized, in which

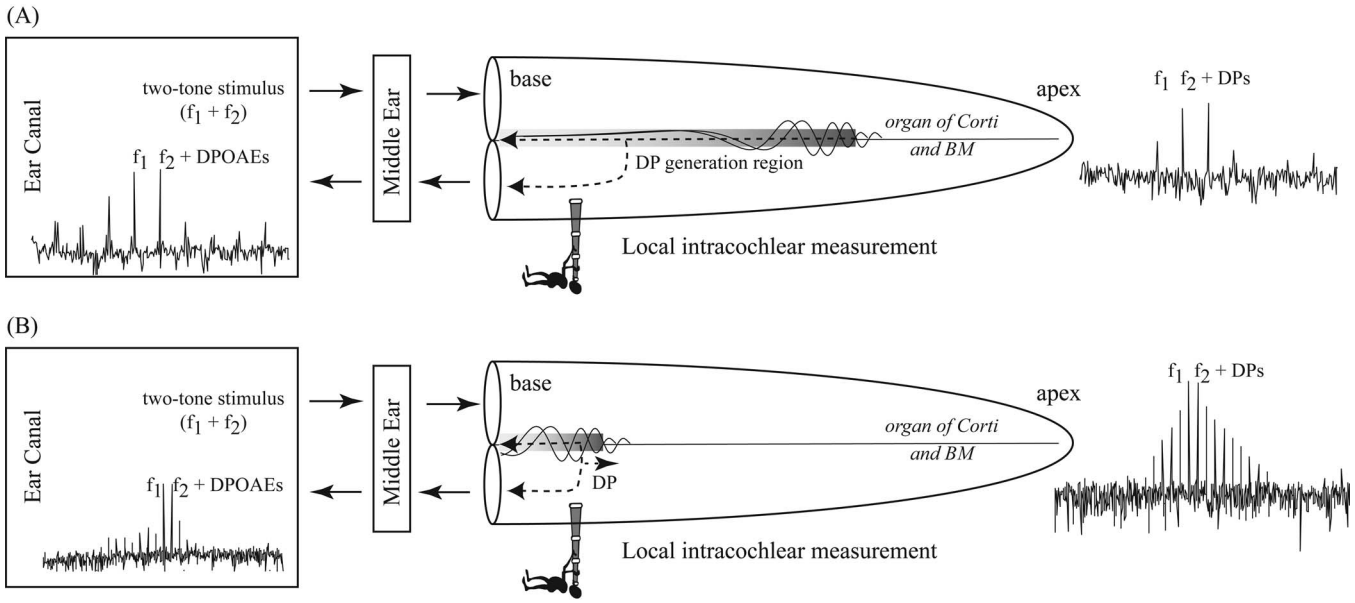


FIG. 14. Illustration of the generation and transmission of DPs. Upon two-tone stimulation, distortion products are generated by cochlear nonlinearity and can be detected within the cochlea as DPs and in the ear canal as DPOAEs. The pressure sensor was positioned at the basal turn of the cochlea where the BF was ~20 kHz. In (A) the primary f_1 and f_2 frequencies were well beneath the local BF. In this case the sub-BF DP is likely generated apical of our sensor location, and when detected, is on its way out of the cochlea. In (B) the primary f_1 and f_2 frequencies are in the BF region. In this case the DP was generated close to our sensor position, and is relatively large due to cochlear tuning of both the primaries and the DP. It has a forward-going character due to the forward-going character of the primaries.

DPs are transported out of the cochlea as a compression wave, where they excite the stapes and launch a forward-going wave at the DP frequency.

Many of our results are in accord with those of Ren. In particular, we show that the intracochlear DP has a forward-traveling behavior in a broad frequency region near BF. We also observed a DPOAE phase-frequency slope that was shallower than the DP phase-frequency slope. However, these results are not in disaccord with the reverse-traveling wave. We back up this statement below, starting by reviewing DPOAE and DP results (see also Shera *et al.*, 2005).

A. General characteristics of the DPOAE and DP

In the most general terms, the DPOAE is a complex, broad-frequency-band response. DPOAE amplitude data possessed fine structure: peaks and sharp minimums. The constructive and destructive summing of the two emission types (generator/wave fixed and reflector/place fixed) is thought to give rise to the amplitude fine structure (Stover *et al.*, 1996; Talmadge *et al.*, 2000; Kalluri and Shera, 2001). The DPOAE is diminished at frequencies above 25 kHz. These frequencies are represented in the fragile, basal region of the cochlea, and it is possible that the above-BF frequency rolloff occurred because cochlear nonlinearity had been compromised in this region.

Phase is key to this study. Emissions are known to exhibit two strikingly different phase behaviors (Kemp, 1986; Shera and Guinan, 1999). With fixed $f_2:f_1$ ratio, the DPOAE phase can be almost flat with frequency, or can be rapidly rotating through several cycles. These behaviors correspond, respectively, to the generator/wave-fixed and reflector/place-fixed emission types. It has been observed that the high-frequency-side DPOAEs (e.g., $2f_2-f_1$) are more apt to exhibit the steep phase-frequency slope, whereas the low-frequency-side DPOAEs (e.g., $2f_1-f_2$) tend to exhibit the flat slope (Knight and Kemp, 2000, 2001). It has also been observed that with increases in stimulus level, the DPOAE phase tends to flatten (Mauermann *et al.*, 1999; Mauermann and Kollmeier, 2004). We observed these characteristic DPOAE phase behaviors.

The DP was measured at a single position in the cochlea with BF ~20 kHz. The character of the DP was different for frequencies in the vicinity of the local BF and for frequencies below the local BF. Figure 14 is a cartoon illustration of the cochlear situation for sub-BF DPs [Fig. 14(A)] and approximately BF DPs [Fig. 14(B)], which is useful for understanding the dichotomy. At frequencies in a broad region around the BF the DPs were tuned, in some cases even more sharply than their parent primaries. The DP tuning in evidence supports the idea that even at high stimulus levels a DPOAE would be substantially contributed to by its own BF frequency region of the cochlea. In this frequency region, the phase was forward traveling. This forward-traveling phase is expected, since the DP is produced by the nonlinear response of the forward traveling primaries [Fig. 14(B)]. In addition, after being produced a DP would be expected to travel forward within the cochlea to its own best place (Robles *et al.*, 1997; Dong and Olson, 2005b).

At frequencies below the BF, the DP amplitude showed fine structure that was similar to the behavior of a DPOAE, suggesting that the DPs had been contributed to by more than one component. The wide ratio of 1.25 $2f_1-f_2$ results often showed a pronounced bimodal shape separated by a sharp minimum, as in Figs. 3(D), 5(B) and 10(H). The presence of this apparent interference notch also supports a dual

or multicomponent picture of the DP. In the sub-BF region, the phase departed from the forward-traveling-wave pattern, consistent with apical DP generation. It is reasonable to expect that sub-BF DPs are substantially contributed to by apically generated distortion, traveling through the more basal region of the cochlea to emerge from the cochlea as an emission [Fig. 14(A)]. Therefore the DPs are expected to be on their way out of the cochlea when we measure them. In support of this idea, robust DP nonlinearity does exist throughout the cochlea, including the apex, as illustrated by auditory nerve recordings (Kim and Molnar, 1979) and apical turn mechanics (Khanna and Hao, 1999; Cooper and Dong, 2002; Khanna, 2002; Cooper, 2006).

B. Relating DPs and DPOAEs

The signature of a traveling wave is in the phase response, and the signature of a reverse-cochlear traveling wave will lie in a DPOAE–DP phase that is similar to the well established single-tone response phase. This study was aimed at this result. We found that this pure result, the DPOAE–DP phase alone, was not definitive, but that by considering this phase in conjunction with the DP and DPOAE results, more compelling conclusions could be drawn.

1. At frequencies below BF

At substantially sub-BF stimulation frequencies, based on both amplitude and phase, DPs appeared to be generated apical of our basal measurement location. Therefore the DPs are expected to be on their way out of the cochlea when we measure them and it is reasonable to look for a reverse-wave characteristic in the DPOAE–DP phase. Note that it is the behavior of the DP that forms the basis for this reasonableness, and it applies regardless of the character of the DPOAE. If the measured DP travels out of the cochlea as a traveling wave then the phase, DPOAE–DP, will overlie the forward-traveling-wave phase. If on the other hand, the DP travels out of the cochlea instantaneously as a compressive pressure, the DPOAE–DP phase will overlie the reverse middle-ear phase. When their phases were related (DPOAE–DP), the phase-frequency relationship was similar to the forward-traveling-wave phase. However, in this sub-BF region, the forward-traveling-wave phase does not change rapidly, and therefore the DPOAE–DP phase does not demonstrate clear reverse-traveling-wave phase behavior. The strong similarity between single-tone transmission into the cochlea and DP transmission out of the cochlea as a traveling wave was nevertheless striking. This striking similarity, coupled with the qualitative similarity between the DP and the DPOAE, and the *lack of similarity* between the DP and the single-tone response in the low-frequency region, does lead to a parsimonious conclusion that these low-frequency DPs were passing through the basal region of the cochlea to give rise to the DPOAE.

2. At frequencies around BF

At frequencies around BF, the DP essentially always showed a forward-traveling-wave phase, and therefore our criterion above for "reasonable reverse-wave case" was never satisfied. However, the steeply phase-sloped "reflector/place-fixed" DPOAE is expected to reveal forward+reverse group delay from the local BF place. Therefore, the steep-phased DPOAE can be reasonably related to the DP in the region of the BF. Note that here it is the behavior of the DPOAE that forms the basis for this reasonableness. In Figs. 7(B) and 7(G) the DPOAE is steep phased in the BF region, and in the DPOAE–DP phase, a reverse wave is apparent, and similarly for Figs. 8(B), 9(B), and 9(G), and 10(B). The grouped data of Fig. 12 confirmed that the phase-frequency slope at 20 kHz of the place-fixed DPOAE, is roughly twice the phase-frequency slope of the forward-going wave, at its 20 kHz BF.

In Figs. 8(G) and 10(G), the high ratio 1.25 $2f_1-f_2$ case, the DPOAE never became steep phased, so the criterion for a "reasonable case" was never met. If we look for a reverse wave in the BF region by plotting the DPOAE–DP phase when the DPOAE was flat phased or had only a mild (middle-ear round-trip) phase slope, the results were like those in Fig. 8(J) at frequencies greater than 17 kHz, in Fig. 9(J) greater than 17 kHz (80 dB stimulus levels), and in Fig. 10(J). In these cases, the DPOAE–DP phase sloped upward, with almost a mirror image of the forward-traveling-wave phase, as though the DPOAE preceded the DP. A similar observation was one of the bases for the hypothesis that the DP traveled out of the cochlea as a compression wave (Ren, 2004). In light of the forward-traveling wave DP, and the flat-phased DPOAE, Ren interpreted this upward slope of the DPOAE–DP phase to mean that the DPOAE must have occurred before the DP. However, here we show that this upward slope is exactly what is to be expected for a forward-traveling-wave DP and a flat-phased DPOAE produced by scaling symmetry.

C. DP spatial variation

In Sec. IV B 2, we show and attempt to understand paradoxical results at frequencies around the BF in which, when a DP was related to a DPOAE via the phase of the two responses, the DPOAE appeared to precede the DP that is expected to be its precursor. We further showed that by looking at the DPOAE–DP phase in specific interpretable frequency regions, a reverse wave was in evidence. However, given the obvious pitfalls in relating phase responses, it is useful to be able to explore the question of reverse-traveling waves from a completely different vantage point. Spatial pressure variations provide this.

As introduced in Figs. 1(B) and 1(C), the intracochlear response to sound comprises two modes: the traveling-wave and compression wave. Theoretically, the pressure distributions of these two pressure modes within the cochlea are strikingly different. At frequencies in the region of the BF, the traveling-wave pressure, being governed by BM width and traveling-wave wavelength, is expected to vary over distances of tens of micrometers (Steele and Taber, 1979b; Andoh and Wada, 2004; Yoon *et al.*, 2006). At the same frequencies the compression pressure will vary over distances corresponding to quarter sound wavelengths—centimeters (Peterson and Bogert, 1950). The compression pressure can

be thought of as a time varying, approximately space filling pressure. Experimentally, these two pressure modes have been detected in the spatial variation of pressure with single-tone stimulation, where it was found that the slow traveling wave dominated close to the BM, ~100 μm from it the compression pressure dominated, producing a spatially invariant pressure (Olson, 1998, 1999, 2001; Dong and Olson, 2005b) in amplitude and phase. Therefore, a marked advantage of pressure measurements is that they can detect both the compression and slow-traveling-wave pressure modes. It has been proposed that DPs travel from their place of generation to the stapes through the cochlear fluid as a compression pressure, another possibility is that they are transmitted via a reverse cochlear traveling wave. The spatial pressure variation of the DPs is a criterion to distinguish the two possibilities: a rapid spatial variation argues for the reverse-traveling-wave mode, whereas if the DP pressure is spatially unvarying, the compression mode is supported. The results in Fig. 6 and results in Dong and Olson (2005b) showed that the DP pressure dropped sharply with distance from the BM. Therefore, the DP pressure distribution within the cochlea appears to be dominated by the traveling-wave pressure mode.

Given the observation that the DP pressure falls into the noise with distance from the BM, the DP compression pressure plateau, if one exists, must be less than 60 dB SPL in the cochlea (the sensor noise floor). Therefore, with the 40 dB loss from the middle-ear reverse transmission (Dong and Olson, 2006), the contribution to the DPOAE from the compression mode would be at most 20 dB SPL, which is much lower than the actual DPOAE we have measured in the EC. In summary, the spatial variation measurement does not rule out a contribution by a compression wave to the DPOAE, but does set the maximum limit of its contribution to DPOAE. Our results suggest that the reverse-traveling wave plays the more important role in DP reverse transmission.

V. SUMMARY

(1) Regarding intracochlear distortion: at frequencies in the broad vicinity of the BF, the DPs were tuned similarly to the primaries, with primarylike forward-traveling phase. This is consistent with distortion that was either generated locally or generated basally and carried to the measurement point by a DP traveling wave. At frequencies substantially lower than the local BF, the DPs did not behave like the primaries, and did not have the forward-traveling-wave phase of the primaries. They instead behaved more like a DPOAE.

(2) Regarding relating DPOAEs to DPs to attempt to observe the reverse wave directly, this can sometimes be done at sub-BF frequencies, for which the DP was not dominated by a forward-traveling DP. In these cases, the DPOAE–DP phase overlaid the forward phase of a primary response, consistent with a reverse cochlear traveling wave. The sub-BF region corresponds to a region of little phase variation, so does not amount to a compelling detection of reverse-traveling wave. Nevertheless, coupled to the similar characters of the DP and the DPOAE and the nonforward-traveling character of the DP, the combined results strongly favor the reverse-traveling-wave description of the sub-BF results. At frequencies in a broad-frequency region of the local BF, the DP was forward traveling. Here, a correlation between the DP and DPOAE can only be reasonably attempted very close to the BF, and only when the DPOAE had a rapidly varying phase. (More concretely, it cannot be attempted when the DPOAE phase was flat or nearly flat, as then the DPOAE phase frequency did not contain cochlear travel-time information.) When the DPOAE did have a rapidly varying phase, the reverse wave was detected, in that the DPOAE–DP phase did closely overlie the primary phase.

(3) The observed spatial dropoff in DP pressure with distance from the DP is inconsistent with a compression pressure DP, but consistent with a traveling-wave DP.

In summary, the data show that DPs are carried along the cochlea by traveling waves and provide compelling evidence against the hypothesis that high level DPs produce compression-pressure waves. The data offer strong supporting evidence for the reverse-traveling wave.

ACKNOWLEDGMENTS

We gratefully acknowledge the efforts of E. de Boer, R. A. Eatock, D. N. Sheppard, C. A. Shera, and two anonymous reviewers, who provided valuable comments on the manuscript. This work was funded by Grant No. R01 DC03130 from the NIH/NIDCD.

NOMENCLATURE

BF = best frequency
BM = basilar membrane
CAP = compound action potential
DP = distortion product
DPOAE = distortion product otoacoustic emission
EC = ear canal
ST = scala tympani
SV = scala vestibuli

Andoh, M., and Wada, H. (**2004**). "Prediction of the characteristics of two types of pressure waves in the cochlea: Theoretical considerations," J. Acoust. Soc. Am. **116**, 417–425.

Avan, P., Bonfils, P., Gilain, L., and Mom, T. (**2003**). "Physiopathological significance of distortion-product otoacoustic emissions at $2f1-f2$ produced by high- versus low-level stimuli," J. Acoust. Soc. Am. **113**, 430–441.

Avan, P., Magnan, P., Smurzynski, J., Probst, R., and Dancer, A. (**1998**). "Direct evidence of cubic difference tone propagation by intracochlear acoustic pressure measurements in the guinea-pig," Eur. J. Neurosci. **10**, 1764–1770.

Cooper, N. P. (**2003**). "Compression in the peripheral auditory system," in *Compression from Cochlear to Cochlear Implants*, edited by S. P. Bacon, R. R. Fay, and A. N. Popper (Springer, New York), pp. 18–61.

Cooper, N. P. (**2006**). "Mechanical preprocessing of amplitude-modulated sounds in the apex of the cochlea," Journal for Oto-Rhino-Laryngology and its Related Specialities **68**, 353–358.

Cooper, N. P., and Dong, W. (**2002**). "Baseline position shifts and mechanical compression in the apical turns of the cochlea," in *Biophysics of the Cochlea from Molecules to Models*, edited by A. W. Gummer, E. Dalhoff,

M. Nowotny, and M. P. Scherer (World Scientific, Titisee, Germany), pp. 261–267.
Cooper, N. P., and Shera, C. A. (**2004**). "Backward traveling waves in the cochlea? Comparing basilar membrane vibrations and otoacoustic emissions from individual guinea-pig ears," in *Proceedings of The Association for Research in Otolaryngology 2004 Midwinter Meeting*, Feb. 22–26, Daytona Beach, Florida.
Decraemer, W. F., de La Rochefoucauld, O., Dong, W., Khanna, S. M., Dirckx, J. J., and Olson, E. S. (**2007**). "Scala vestibuli pressure and three-dimensional stapes velocity measured in direct succession in gerbils," J. Acoust. Soc. Am. **121**, 2774–2791.
Dong, W., and Olson, E. S. (**2005a**). "Tuning and travel of two tone distortion in intracochlear pressure," in *Auditory Mechanisms: Processes and Models, The 9th International Symposium*, edited by A. L. Nuttall, T. Ren, P. Gillespie, K. Grosh, and E. de Boer (World Scientific, Portland, Oregon), pp. 56–62, July 23–28, Oregon.
Dong, W., and Olson, E. S. (**2005b**). "Two-tone distortion in intracochlear pressure," J. Acoust. Soc. Am. **117**, 2999–3015.
Dong, W., and Olson, E. S. (**2006**). "Middle ear forward and reverse transmission in gerbil," J. Neurophysiol. **95**, 2951–2961.
Goldstein, J. L. (**1967**). "Auditory nonlinearity," J. Acoust. Soc. Am. **41**, 676–689.
He, W., Nuttall, A. L., and Ren, T. (**2007**). "Two-tone distortion at different longitudinal locations on the basilar membrane," Hear. Res. **228**, 112–122.
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. (**1978**). "Stimulated acoustic emissions from within the human auditory system," J. Acoust. Soc. Am. **64**, 1386–1391.
Kemp, D. T. (**1986**). "Otoacoustic emissions, travelling waves and cochlear mechanisms," Hear. Res. **22**, 95–104.
Khanna, S. M. (**2002**). "Non-linear response to amplitude-modulated waves in the apical turn of the guinea pig cochlea," Hear. Res. **174**, 107–123.
Khanna, S. M., and Hao, L. F. (**1999**). "Nonlinearity in the apical turn of living guinea pig cochlea," Hear. Res. **135**, 89–104.
Kim, D. O., and Molnar, C. E. (**1979**). "A population study of cochlear nerve fibers: Comparison of spatial distributions of average-rate and phase-locking measures of responses to single tones," J. Neurophysiol. **42**, 16–30.
Kim, D. O., Molnar, C. E., and Matthews, J. W. (**1980**). "Cochlear mechanics: Nonlinear behavior in two-tone responses as reflected in cochlear-nerve-fiber responses and in ear-canal sound pressure," J. Acoust. Soc. Am. **67**, 1704–1721.
Knight, R. D., and Kemp, D. T. (**2000**). "Indications of different distortion product otoacoustic emission mechanisms from a detailed $f1, f2$ area study," J. Acoust. Soc. Am. **107**, 457–473.
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.
Lighthill, M. J. (**1981**). "Energy flow in the cochlea," J. Fluid Mech. **106**, 149–213.
Lukashkin, A. N., Lukashkina, V. A., and Russell, I. J. (**2002**). "One source for distortion product otoacoustic emissions generated by low- and high-level primaries," J. Acoust. Soc. Am. **111**, 2740–2748.
Mauermann, M., and Kollmeier, B. (**2004**). "Distortion product otoacoustic emission (DPOAE) input/output functions and the influence of the second DPOAE source," J. Acoust. Soc. Am. **116**, 2199–2212.
Mauermann, M., Uppenkamp, S., van Hengel, P. W., and Kollmeier, B. (**1999**). "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.
Mills, D. M. (**2002**). "Interpretation of standard distortion product otoacoustic emission measurements in light of the complete parametric response," J. Acoust. Soc. Am. **112**, 1545–1560.
Olson, E. S. (**1998**). "Observing middle and inner ear mechanics with novel intracochlear pressure sensors," J. Acoust. Soc. Am. **103**, 3445–3463.
Olson, E. S. (**1999**). "Direct measurement of intra-cochlear pressure waves," Nature (London) **402**, 526–529.
Olson, E. S. (**2001**). "Intracochlear pressure measurements related to cochlear tuning," J. Acoust. Soc. Am. **110**, 349–367.
Olson, E. S. (**2004**). "Harmonic distortion in intracochlear pressure and its analysis to explore the cochlear amplifier," J. Acoust. Soc. Am. **115**, 1230–1241.
Olson, E. S., and Dong, W. (**2006**). "Nonlinearity in intracochlear pressure," Journal for Oto-Rhino-Laryngology and Related Specialities **68**, 359–364.
Overstreet, E. H., 3rd, Richter, C. P., Temchin, A. N., Cheatham, M. A., and Ruggero, M. A. (**2003**). 'High-frequency sensitivity of the mature gerbil cochlea and its development," Audiol. Neuro-Otol. **8**, 19–27.
Peterson, L. C., and Bogert, B. P. (**1950**). "A dynamical theory of the cochlea," J. Acoust. Soc. Am. **22**, 369–381.
Ren, T. (**2004**). "Reverse propagation of sound in the gerbil cochlea," Nat. Neurosci. **7**, 333–334.
Robles, L., and Ruggero, M. A. (**2001**). "Mechanics of the mammalian cochlea," Physiol. Rev. **81**, 1305–1352.
Robles, L., Ruggero, M. A., and Rich, N. C. (**1997**). "Two-tone distortion on the basilar membrane of the chinchilla cochlea," J. Neurophysiol. **77**, 2385–2399.
Ruggero, M. A. (**2004**). "Comparison of group delays of $2f1-f2$ distortion product otoacoustic emissions and cochlear travel times," Acoustic Research Letters Online **5**, 143–147.
Ruggero, M. A., and Temchin, A. N. (**2007**). "Similarity of traveling-wave delays in the hearing organs of humans and other tetrapods," J. Assoc. Res. Otolaryngol. **8**, 153–166.
Ryan, A. (**1976**). "Hearing sensitivity of the mongolian gerbil, Meriones unguiculatis," J. Acoust. Soc. Am. **59**, 1222–1226.
Shera, C. A., and Guinan, J. J., Jr. (**1999**). "Evoked otoacoustic emissions arise by two fundamentally different mechanisms: A taxonomy for mammalian OAEs," J. Acoust. Soc. Am. **105**, 782–798.
Shera, C. A., and Guinan, J. J., Jr. (**2003**). "Stimulus-frequency-emission group delay: A test of coherent reflection filtering and a window on cochlear tuning," J. Acoust. Soc. Am. **113**, 2762–2772.
Shera, C. A., Guinan, J. J., Jr., and Oxenham, A. J. (**2002**). "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," Proc. Natl. Acad. Sci. U.S.A. **99**, 3318–3323.
Shera, C. A., Talmadge, C. L., and Tubis, A. (**2000**). "Interrelations among distortion-product phase-gradient delays: Their connection to scaling symmetry and its breaking," J. Acoust. Soc. Am. **108**, 2933–2948.
Shera, C. A., Tubis, A., and Talmadge, C. L. (**2005**). "Four counter-arguments for slow-wave OAEs," in *Auditory Mechanisms: Processes and Models, Proceedings of the 9th International Symposium*, edited by A. L. Nuttall, T. Ren, P. Gillespie, K. Grosh, and E. de Boer (World Scientific, Portland, Oregon), pp. 449–457, July 23–28.
Sokolich, W. G. (**1977**). "Improved acoustic system for auditory research," J. Acoust. Soc. Am. **62**, S12.
Steele, C. R., and Taber, L. A. (**1979a**). "Comparison of WKB and finite difference calculations for a two-dimensional cochlear model," J. Acoust. Soc. Am. **65**, 1001–1006.
Steele, C. R., and Taber, L. A. (**1979b**). "Comparison of WKB calculations and experimental results for three-dimensional cochlear models," J. Acoust. Soc. Am. **65**, 1007–1018.
Stover, L. J., Neely, S. T., and Gorga, M. P. (**1996**). "Latency and multiple sources of distortion product otoacoustic emissions," J. Acoust. Soc. Am. **99**, 1016–1024.
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.
Talmadge, C. L., Tubis, A., Long, G. R., and Piskorski, P. (**1998**). "Modeling otoacoustic emission and hearing threshold fine structures," J. Acoust. Soc. Am. **104**, 1517–1543.
Talmadge, C. L., Tubis, A., Long, G. R., and Tong, C. (**2000**). "Modeling the combined effects of basilar membrane nonlinearity and roughness on stimulus frequency otoacoustic emission fine structure," J. Acoust. Soc. Am. **108**, 2911–2932.
van Dijk, P., and Manley, G. A. (**2001**). "Distortion product otoacoustic emissions in the tree frog Hyla cinerea," Hear. Res. **153**, 14–22.
von Bekesy, G. (**1960**). *Experiments in Hearing* (McGraw Hill, New York).
Wilson, J. P. (**1980**). "Model for cochlear echoes and tinnitus based on an observed electrical correlate," Hear. Res. **2**, 527–532.
Yoon, Y. J., Puria, S., and Steele, C. R. (**2006**). "Intracochlear pressure and organ of corti impedance from a linear active three-dimensional model, Journal for Oto-Rhino-Laryngology and Related Specialities **68**, 365–372.
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.

Sample discrimination of frequency by hearing-impaired and normal-hearing listeners

Joshua M. Alexander[a)]
Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706

Robert A. Lutfi
Department of Communicative Disorders, University of Wisconsin, Madison, Wisconsin 53706

(Received 15 May 2006; revised 12 September 2007; accepted 25 October 2007)

In a multiple observation, sample discrimination experiment normal-hearing (NH) and hearing-impaired (HI) listeners heard two multitone complexes each consisting of six simultaneous tones with nominal frequencies spaced evenly on an ERB_N logarithmic scale between 257 and 6930 Hz. On every trial, the frequency of each tone was sampled from a normal distribution centered near its nominal frequency. In one interval of a 2IFC task, all tones were sampled from distributions lower in mean frequency and in the other interval from distributions higher in mean frequency. Listeners had to identify the latter interval. Decision weights were obtained from multiple regression analysis of the between- interval frequency differences for each tone and listeners' responses. Frequency difference limens (an index of sensorineural resolution) and decision weights for each tone were used to predict the sensitivity of different decision-theoretic models. Results indicate that low-frequency tones were given much greater perceptual weight than high-frequency tones by both groups of listeners. This tendency increased as hearing loss increased and as sensorineural resolution decreased, resulting in significantly less efficient weighting strategies for the HI listeners. Overall, results indicate that HI listeners integrated frequency information less optimally than NH listeners, even after accounting for differences in sensorineural resolution. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816415]

PACS number(s): 43.66.Ba, 43.66.Fe, 43.66.Sr [JHG] Pages: 241–253

I. INTRODUCTION

The objective of this study is to understand the factors that influence a hearing-impaired (HI) listener's ability to discriminate between sounds that require an integration of information about small frequency differences. Two factors are commonly discussed in the literature with regard to multiple observation, sample discrimination tasks. The first is a decision strategy that assigns relative importance or perceptual weight to the various channels of information. The second is internal noise, which is often used as a collective term for everything that cannot be explained by an observer's decision strategy. Sources of internal noise include sensorineural or peripheral processes associated with the precision of transduction and neural coding as well as central processes associated with observer variability or inconsistency in the execution of a decision strategy. As explained below, when the size of the differences to be discriminated is the same magnitude as the difference limens (DLs) and near the sensorineural noise floor, a delineation of the sources of internal noise is necessary for a proper interpretation of sensitivity relative to an ideal model. The primary aim of this paper is to offer a theoretical framework for investigating the influence of channel-specific sensorineural noise on decision weights and internal noise.

Most of the studies examining the effects of sensorineural hearing loss (SNHL) on decision weights and internal noise have used intensity discrimination tasks (Doherty and Lutfi, 1996, 1999; Lentz and Leek, 2002, 2003). The results from these experiments generally indicate that weighting strategies are variable across normal-hearing (NH) listeners and even more so across HI listeners. They also indicate that HI listeners adopt slightly different weighting strategies than NH listeners do. For example, Doherty and Lutfi (1996) found that HI listeners were apt to put the most weight on a 4000 Hz tone that was in the sloping region of hearing loss in a discrimination task where the signal was the sum of intensity level increments on six fixed-frequency tones. In contrast, the NH listeners as a group distributed their attention across the spectrum. A similar pattern of weights emerged for HI listeners in Doherty and Lutfi (1999) when the signal was a reliable level increment on only the 4000 Hz tone so that HI listeners were actually at an advantage relative to NH listeners. No differences were significant when the signal was at 250 and 1000 Hz, where hearing thresholds were more similar for the two groups.

In a spectral shape discrimination or "profile analysis" task in which listeners detected increments in the level of a 920 Hz tone relative to pairs of flanking tones above and below the signal frequency, Lentz and Leek (2002) found limited evidence that HI listeners used less optimal decision strategies than NH listeners. HI listeners were less likely to weight the flanking components when they should have been attended to, but more likely to weight the flanking compo-

[a)]Author to whom correspondence should be addressed. Present address: Boys Town National Research Hospital, Omaha, Nebraska, 63131. Electronic mail: alexanderj@boystown.org

nents when they should have been ignored. In a tone-detection task, Alexander and Lutfi (2004) also found that HI listeners as a group tended to put less weight on a 2000 Hz signal tone and more weight on flanking tones than NH listeners did. Differences between NH and HI listeners were also noted by Lentz and Leek (2003) when the spectral shape of the signal was a level increment in the three lowest or the three highest frequency components and a decrement in the other half. HI listeners as a group tended to put the most weight on the lowest and the highest frequencies of the complex while the NH listeners tended to put the most weight on the two centermost frequencies.

The above studies indicate that SNHL alters the way HI listeners weight information compared to NH listeners, although the differences are task dependent and do not always influence overall performance. They also suggest that internal noise is greater for HI listeners compared to NH listeners. In all of their conditions, Doherty and Lutfi (1996, 1999) found that an index of internal noise was significantly greater for HI listeners than for NH listeners. Lentz and Leek (2003) also indicated that HI listeners might have greater amounts of internal noise as demonstrated by a greater amount of variability in their weighting strategies across conditions. For these experiments, it is assumed that peripheral or sensorineural processes were not responsible for the observed differences between groups because intensity resolution is minimally affected by HI (Florentine *et al.*, 1993; Buus *et al.*, 1995). This suggests that central processes were responsible for the observed differences, but we cannot be sure because current analytic techniques do not allow for a separation of peripheral and central processes.

Challenges to traditional analyses originating from signal-detection theory arise when studying sensitivity for frequency discrimination in listeners with SNHL because differences within and between listeners can be more than a magnitude for individual tones (e.g., Freyman and Nelson, 1991). Because frequency distortion is a condition of many HI listeners' everyday lives and because current hearing aid technology does little to correct for it, it is of interest to examine how differences in the spectral profile of sensorineural noise influence the way listeners weight information as a function of frequency. In a decision-theoretic framework, best possible discrimination performance will occur when frequency information is weighted inversely proportional to the frequency-specific noise. Frequency regions where sensorineural noise is lower should be weighted more than regions where it is higher. Thus, for each sensorineural noise profile there is a corresponding optimal weighting strategy. HI listeners might listen differently than NH-listeners, but the differences might be systematic and follow each listener's individual ideal. In this case, comparing both NH and HI listeners to the stimulus ideal, where weights are proportional to the information sent instead of the information received, erroneously implies that HI listeners weight information inefficiently. In order to obtain a fair comparison of HI and NH listeners in conditions where sensorineural noise is not procedurally regulated, in what follows we will describe the concept of a *peripherally limited ideal observer* (PLIO), an analytic solution to separate the effects of sensorineural noise on decision weights and internal noise.

II. METHODS

A. Listeners

Thirteen NH listeners 20–45 years old (median age of 23), including the first author (NH 01), and 13 HI listeners 55–84 years old (median age of 77) participated in the study.[1] On standardized audiometric tests (ANSI S3.6-1996) NH listeners had pure-tone thresholds of 10 dB HL or better, except a few who had one threshold equal to 15 dB HL. The right ears of the NH listeners served as the test ear in the experiments. HI listeners had mild to moderate SNHL in the test ear and minimal conductive hearing loss (i.e., the differences between air and bone conduction thresholds were 10 dB or less and tympanometric tests were clinically normal). When both ears fit the inclusion criteria for the HI group, the right ear was used as the test ear.

B. Stimuli

Stimuli were generated with a 44.1 kHz sampling rate using the MATLAB® programming language and a 16 bit sound card. A programmable attenuator was used to control the overall level. Stimuli were presented monaurally through a Beyerdynamic DT 990 earphone. Unlike frequency discrimination, intensity discrimination follows the "near-miss to Weber's Law," which makes it convenient to use the dB scale to perturb the information for each tone in roughly equal units of the jnd (just noticeable difference). Difference limens for frequency (DLFs) do not vary systematically as a function of frequency for reasons beyond the current discussion (e.g., Moore, 1997). For this reason, we manipulated and measured frequency using a scale based on the equivalent rectangular bandwidth for *normal-hearing* listeners, ERB_N (Glasberg and Moore, 1990; Moore, 1997). The ERB_N is a log-based psychophysical scale for frequency where the ERB_N No. $=21.4 \log_{10}(4.37F+1)$ for frequency F in Hz. It should be noted that because no measure of frequency selectivity was obtained, it is unknown how the frequency manipulations mapped onto listeners' filter space, especially for the HI listeners. Furthermore, while many log-based scales could have been used, the important point is that the smallest and largest DLFs between 250 and 8000 Hz for NH listeners are usually within an order of magnitude in ERB_N (Sek and Moore, 1995). Tones were sampled from one of six different frequency regions whose center frequencies were logarithmically separated by five ERB_N: 257, 603, 1197, 2212, 3951, and 6930 Hz. Tones were 120 ms in duration with 12 ms cosine-squared ramps and random starting phase ($0-2\pi$ radians). In the discrimination experiments, each tone was presented at 82 dB sound pressure level (SPL) for all listeners.

C. Procedure

Listeners ran the experiments individually while seated in a double-walled, sound-attenuated chamber. Trials consisted of a two-interval, forced-choice task (2IFC) with a 1 s inter-stimulus interval, an initial warning, and visual feedback. The intervals were specified on a computer monitor

and listeners indicated which of the two intervals contained the signal by pressing a button. There were no time limits for making a response.

1. Tests of cochlear function: Adaptive staircase method

The purpose of these tests was to provide information about the degree of hearing loss and about the processing limits imposed by the auditory periphery. Detection thresholds in quiet (QTs) and DLFs were measured for each of the six center frequencies in isolation. Signal levels and frequency differences were adapted using a two-down, one-up decision rule which converges on the 70.7% point on the psychometric function (Levitt, 1971), or equivalently, the point where $d'_{(2IFC)}=0.77$ (Macmillan and Creelman, 2005). Each trial block consisted of 12 reversals in the adaptive track. The average values of the final eight reversals determined the QT or DLF for the block.

The initial step size of the adaptive tracks for QT was 4 dB and was reduced to 2 dB after the third reversal. Maximum presentation-level was limited to 85 dB SPL. QT for each tone was based on the mean across three trial blocks. QT blocks were completed before DLF blocks because DLFs were not tested for tones where QT $>$ 77 dB SPL (i.e., SL $<$ 5 dB).

For the DLFs, the first block of trials was used to give the listeners practice and to calibrate the starting frequency differences of the following adaptive tracks. Listeners had to identify the interval with the highest-frequency tone. For most listeners it was expected that pitch would be the most salient cue, although some HI listeners indicated that they sometimes used loudness for a cue especially when the frequencies were in a sloping region of hearing loss. The initial frequency difference, D, in the practice trial block was 0.2 $\mathrm{ERB_N}$ for the NH listeners and 0.6 $\mathrm{ERB_N}$ for the HI listeners. The lower-frequency tone in each interval was $D/2$ less than the nominal center frequency and the higher-frequency tone was $D/2$ greater so that regardless of the size of the frequency difference, the unbiased criterion was always the same. The starting frequency difference for subsequent blocks was 0.04 $\mathrm{ERB_N}$ above the running average of all previous blocks for that frequency. The initial step size of the adaptive tracks was 0.02 $\mathrm{ERB_N}$ and was reduced to 0.01 $\mathrm{ERB_N}$ after the third reversal. Before running the sample discrimination experiment, listeners completed two additional blocks for each frequency after the initial practice blocks. After running the sample discrimination experiment, two more blocks for DLF were run. The DLF for each frequency was based on the mean across four blocks.[2]

2. Sample discrimination of frequency

In a sample discrimination experiment the difference to be discriminated is nonadaptively varied from trial to trial by sampling the stimuli in each interval (indexed j) from one of two overlapping distributions with identical standard deviations, $\sigma_{S(j)}$, but different means (e.g., Lutfi, 1989, 1990, 1992; Lutfi *et al.*, 1996; Jesteadt *et al.*, 2003; Sorkin *et al.*, 1987; Neff and Odgaard, 2004). Stimuli in this sample discrimination task consisted of six simultaneous tones whose frequencies (X_i) were independently sampled from one of two distributions. As shown in Fig. 1, in the "noise" interval of a 2IFC task the frequencies of the six-tone complex were sampled from distributions lower in mean frequency and in the "signal" interval they were sampled from distributions higher in mean frequency. Listeners were told to consider all the tones in each interval and then to select the interval where the tones were "on average higher in pitch."

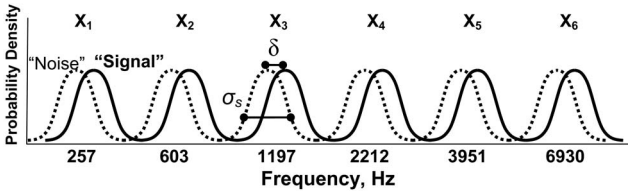


FIG. 1. Shown is a schematic of the two-interval, forced-choice sample discrimination experiment. The tones selected from the distributions higher in mean frequency are from the "signal" interval and those selected from the distributions lower in mean frequency are from the "noise" interval. The distributions for each interval were centered on six fixed frequencies that were separated by a logarithmic distance of five $\mathrm{ERB_N}$. The expected frequency difference between distribution means, δ, was constant for each of the six pairs of tones, but varied across ten different trial blocks. The standard deviation of the sampling distributions, σ_S, was a constant factor of δ.

The difference between the means of the signal and noise distributions, or the average difference to be discriminated is denoted δ_i, where i is the tone number. Across ten different 55 trial blocks, the expected frequency difference for each tone was fixed at $0.02, 0.04, \ldots$, or 0.2 $\mathrm{ERB_N}$ for the NH listeners and at $0.06, 0.12, \ldots$, or 0.6 $\mathrm{ERB_N}$ for the HI listeners. One block of trials was run for each condition (expected frequency difference). $\sigma_{S(ij)}$ was a constant factor of the expected frequency difference. For all listeners $\sigma_{S(ij)} = 1.155\delta_i$ across all tones and conditions. For ease of discussion, the ten conditions are nominally identified by the value of δ_i. Larger frequency manipulations were used for the HI listeners because it was anticipated that greater amounts of sensorineural noise (greater DLFs) would prevent them from performing the experiment above chance. As with the DLF procedure, the expected mean frequencies of the signal and noise distributions for each condition were $\delta_i/2$ above and below the nominal center frequencies, respectively, so that regardless of the condition, the unbiased criterion was constant.

To minimize the effects of practice, the first five trials of each block were discarded and the ten conditions were run in random order during one session. Furthermore, before data collection began at least one block of practice trials was run with $\delta_i=0.2$ for the NH listeners and $\delta_i=0.6$ for the HI listeners, with $\sigma_{s(ij)}=0.5\delta_i$. Listeners began the experiment if the attained d' was at least 1.5 on the first or second practice block or at least 1.0 after the third practice block.

D. Theoretical framework

We model the listener's task when discriminating between two multitone patterns as a univariate decision variable, DV, equal to the weighted sum of the information at each frequency. For our experiments, the information sent is

the difference in frequency between the pairs of tones across trial intervals. However, the actual information received is corrupted by frequency-specific sensorineural noise, which is modeled as random variables, ε_{ij}, with variance $\sigma^2_{P(ij)}$ and mean zero that add to each tone (X_{ij}). Variability in the use of received information is central noise and is modeled as a single random variable, e_C, with variance σ^2_C.

$$DV = e_c + \sum_{i=1}^{m} w_i[(X_{i2} + \varepsilon_{i2}) - (X_{i1} + \varepsilon_{i1})], \tag{1}$$

where m is the number of tones in each interval (indexed i), and w_i are the weights. On every trial, the listeners are assumed to respond "interval 2" when the DV exceeds some fixed, internal decision criterion (ideally, 0 in a 2IFC task).

Internal noise limits the ability to make trial-by-trial predictions; therefore, model predictions were made for sensitivity, d'_{DV}. An analytic prediction for the Z transform of the hit and false-alarm (FA) rates is given by the expected values of the DV when the signal is in interval 2 and interval 1, respectively.

$$Z(hit) = -Z(FA) = \sum_{i=1}^{m} w_i\delta_i \Bigg/ \left[\sigma_c^2 + \sum_{i=1}^{m} w_i^2(2\sigma^2_{S(i)} + 2\sigma^2_{P(i)})\right]^{1/2}. \tag{2}$$

Because the expected value of each sensorineural noise variable, ε_{i2} and ε_{i1}, is zero their expected difference is also zero. In addition, because the expected value of X_{i2} is $(\delta_i/2)$ and the expected value of X_{i1} is $-(\delta_i/2)$, the expected value of their difference is δ_i when the signal is interval 2 and $-\delta_i$ when the signal is interval 1. Therefore, $\Delta = 2\Sigma_{i=1}^{m} w_i\delta_i$, where Δ is the overall difference between the signal and noise distributions of the decision variable. For the unbiased observer, the hit and false-alarm rates are equal and $d'_{DV} = [Z(hit) - Z(FA)]_{2IFC}/\sqrt{2}$, (cf. Eqs. (7.2) and (7.11) in Macmillan and Creelman, 2005).[3] Therefore, the analytic prediction for sensitivity is

$$d'_{DV} = \sqrt{2} \times \sum_{i=1}^{m} w_i\delta_i \Bigg/ \left[\sigma_C^2 + \sum_{i=1}^{m} w_i^2(2\sigma^2_{S(i)} + 2\sigma^2_{P(i)})\right]^{1/2}. \tag{3}$$

Within our theoretical framework, increased DLFs (reduced resolution for frequency differences) associated with SNHL are simply modeled as a higher sensorineural noise floor, σ^2_P. Empirically, σ^2_P is estimated by the magnitude of hypothetical stimulus noise, σ^2_S, needed to describe the discriminability $(d'_{DL(i)})$ between two simple sounds free of external variability or experimental uncertainty: $d'_{DL(i)} = D_{DL(i)}/\sigma_P$ where $D_{DL(i)}$ is the difference limen or the smallest discriminable difference for the ith tone. For each frequency, the average DLF across four blocks, $D_{DL(i)}$, was converted to an estimate of sensorineural noise, $\sigma^2_{P(i)}$, using the expected sensitivity for the two-down/one-up, 2IFC task, $d'_{DL(i)} = 0.77$.

TABLE I. Differences between the various decision models arise from the way decision weights are derived and the source of internal noise.

Model	Weights	Internal noise
d'_{ideal}	$\delta_i/\sigma_{S(i)}$	None
d'_{PLIO}	$\delta_i/\sqrt{\sigma^2_{S(i)} + \sigma^2_{P(i)}}$	σ^2_P
d'_{wgt}	Regression	None
d'_{pred}	Regression	σ^2_P

The observed sensitivity of the listener, d'_{obs}, was computed from the difference between the Z-transformed hit and false-alarm rates for the 2IFC task (see above). For each listener and for each sample discrimination condition, four estimates of sensitivity, $d'_{\text{ideal}}, d'_{\text{PLIO}}, d'_{\text{wgt}}$, and d'_{pred} were analytically derived using Eq. (3). As shown in Table I, the four models differed in terms of the weights and sources of internal noise included in the model.

d'_{ideal} is the maximum expected sensitivity of a hypothetical listener free from internal noise, the stimulus ideal. Ideal sensitivity for a noiseless observer is achieved when weights are proportional to the information *sent* in each channel, $d'_i = \delta_i/\sigma_{S(i)}$. Because d'_i was kept constant (0.866) for all listeners and all sample discrimination conditions, ideal weights, w_{ideal}, for each tone were each equal to each other (1/6) and d'_{ideal} was equal to 2.12 (i.e., $d'_i\sqrt{m}$).

d'_{PLIO} is the maximum expected sensitivity given an individual listener's unique profile of sensorineural noise, what we term a "*peripherally limited ideal observer*" (PLIO). Sensitivity for a PLIO is achieved when weights are proportional to the information *received* in each channel. Information is less reliable in spectral regions where sensorineural noise is relatively large and should be weighted less in a listener's decision. Weights for the PLIO, w_{PLIO}, were proportional to the estimated $\delta_i/\sqrt{\sigma^2_{S(i)} + \sigma^2_{P(i)}}$ for each tone, where $\sigma^2_{P(i)}$ was estimated from the listener's DLFs. d'_{PLIO} each condition was derived by including $\sigma^2_{P(i)}$ in the denominator of Eq. (3) and by substituting w_{PLIO} for w_i.

d'_{pred} represents the predicted sensitivity after accounting for the listener's weighting strategy and peripheral limitations. Because d'_{pred} into account the stimulus statistics and all other listener variables, any remaining difference between d'_{pred} and d'_{obs} attributed to central noise *post hoc*. d'_{pred} was computed in the same way as d'_{PLIO} but using the listeners' regression weights, w_{obs}. This differs from traditional computations d'_{wgt} in that $\sigma^2_{P(i)}$ is included in the pooled variance of the expected decision variable. Observed weights (w_{obs}) across all ten conditions were computed by logistic regression. Listeners' interval 1 responses were coded as "0" and interval 2 responses as "1" and the probability of responding "interval 2" was regressed against the difference in ERB_N between each pair of tones in interval 2 and interval 1. Only statistically significant weights were used for the model estimates with nonsignificant weights set to zero in the equation for sensitivity [Eq. (3)]. So that the weights could be

TABLE II. Information for each of the six frequency regions tested in this study for the NH listeners (arranged from left to right in descending order of observed sensitivity). Section 1: mean quiet thresholds in dB SPL. Section 2: estimates of σ_P expressed as a proportion of ERB_N. Section 3: The weights for the PLIO, w_{PLIO}. Section 4: Observed weights, w_{obs}, that are statistically different from 0 are in bold and those that are not are un-bolded and in parentheses.

		NH 08	NH 10	NH 07	NH 06	NH 12	NH 02	NH 11	NH 14	NH 04	NH 09	NH 01	NH 13	NH 03	Mean	(SE)
QT (dB SPL)	**257**	11.7	12.8	7.5	9.2	9.5	13.3	12.5	13.5	12.3	15.8	7.5	22.4	9.4	**12.1**	(1.15)
	603	3.0	13.0	5.8	7.8	2.3	4.4	11.6	6.0	7.0	9.4	5.0	12.3	7.3	**7.3**	(0.99)
	1197	7.1	18.9	2.1	13.7	3.7	0.8	7.7	0.2	−0.1	13.5	2.8	2.8	7.3	**6.2**	(1.72)
	2212	7.3	14.0	−4.2	12.8	10.5	1.6	4.5	−0.2	1.5	8.6	−1.6	1.5	14.8	**5.5**	(1.82)
	3951	−3.1	7.6	6.0	−1.0	−3.8	−5.8	−2.3	−0.8	1.1	1.1	8.8	−3.2	−1.7	**0.2**	(1.32)
	6930	7.1	15.1	9.1	6.1	2.2	−2.9	8.7	10.2	8.7	10.1	18.3	11.3	17.0	**9.3**	(1.65)
σ_p (ERB_N)	**257**	0.07	0.12	0.14	0.03	0.09	0.06	0.04	0.04	0.05	0.12	0.03	0.05	0.08	**0.07**	(0.011)
	603	0.06	0.09	0.06	0.04	0.04	0.05	0.03	0.03	0.03	0.08	0.05	0.07	0.09	**0.05**	(0.006)
	1197	0.07	0.08	0.05	0.02	0.03	0.07	0.04	0.04	0.04	0.04	0.03	0.08	0.10	**0.05**	(0.006)
	2212	0.03	0.11	0.08	0.04	0.04	0.07	0.04	0.06	0.03	0.06	0.05	0.06	0.09	**0.06**	(0.007)
	3951	0.06	0.20	0.13	0.03	0.04	0.04	0.06	0.13	0.07	0.03	0.07	0.05	0.06	**0.08**	(0.014)
	6930	0.18	0.25	0.26	0.46	0.15	0.11	0.31	0.44	0.48	0.33	0.10	0.32	0.17	**0.27**	(0.038)
w_{PLIO}	**257**	0.17	0.17	0.16	0.19	0.13	0.16	0.18	0.21	0.19	0.16	0.21	0.18	0.17	**0.18**	(0.006)
	603	0.17	0.19	0.18	0.19	0.19	0.19	0.18	0.23	0.21	0.18	0.17	0.18	0.17	**0.19**	(0.005)
	1197	0.17	0.19	0.19	0.19	0.21	0.16	0.18	0.21	0.20	0.19	0.20	0.18	0.17	**0.19**	(0.005)
	2212	0.18	0.18	0.18	0.19	0.19	0.14	0.18	0.19	0.20	0.19	0.17	0.18	0.17	**0.18**	(0.004)
	3951	0.17	0.14	0.16	0.19	0.19	0.25	0.17	0.12	0.16	0.19	0.14	0.18	0.18	**0.17**	(0.010)
	6930	0.14	0.12	0.13	0.06	0.09	0.10	0.11	0.04	0.03	0.09	0.10	0.11	0.14	**0.10**	(0.010)
w_{obs}	**257**	**0.17**	**0.23**	**0.06**	**0.22**	**0.43**	**0.55**	**0.21**	**0.21**	**0.16**	**0.22**	**0.16**	**0.12**	**0.44**	**0.24**	(0.040)
	603	**0.33**	**0.19**	**0.30**	**0.31**	**0.16**	**0.19**	**0.31**	**0.45**	**0.36**	**0.18**	**0.17**	**0.11**	**0.20**	**0.25**	(0.029)
	1197	**0.21**	**0.32**	**0.31**	**0.33**	**0.13**	**0.14**	**0.34**	**0.33**	**0.26**	**0.49**	**0.19**	**0.67**	**0.37**	**0.31**	(0.042)
	2212	**0.15**	**0.08**	**0.18**	**0.08**	**0.18**	(0.05)	(0.04)	(0.04)	**0.09**	**0.11**	(0.01)	(0.01)	(0.09)	**0.09**	(0.017)
	3951	**0.14**	**0.12**	**0.15**	**0.07**	**0.10**	**0.12**	**0.14**	(−0.02)	**0.13**	(−0.01)	**0.15**	**0.10**	(0.08)	**0.10**	(0.016)
	6930	(0.05)	**0.07**	(−0.01)	(−0.01)	(0.00)	(0.00)	(0.02)	(0.01)	(−0.05)	(0.05)	**0.34**	(0.00)	(0.00)	**0.04**	(0.028)

interpreted as the proportion of attention devoted to each frequency region, they were normalized so that the sum of the absolute values of the significant weights was one (Berg, 1990; Lutfi, 1992; Doherty and Lutfi, 1996, 1999).

III. RESULTS

A. Sensorineural resolution and weights for the PLIO

A quantification of sensorineural resolution is necessary to determine the contribution of sensorineural noise to overall internal noise and its spectral profile is important for determining the relative weighting function for a PLIO (w_{PLIO}). The means of the DLF-transformed estimates of σ_P and standard errors of the mean (SE's) are provided in Tables II and III for the NH and HI listeners, respectively, and are plotted in Fig. 2 with 95% confidence intervals (CIs).

Sensorineural resolution estimates for HI 12 were excluded from the means and related statistical analyses because the estimates were unusually large and in some cases contributed to over 1/3 of the error variance. Lower values of σ_P indicate higher (better) sensorineural resolution. NH listeners clearly had better sensorineural resolution than the HI listeners did. On average, σ_P was about 3–10 times greater in HI listeners than in NH listeners. Across individual HI listeners, there was substantial variability in sensorineural resolution. The profile of σ_P ranged from normal to more than a magnitude greater than normal for a few listeners.

The absolute differences in σ_P had little influence on w_{PLIO} because the σ_S/σ_P ratios were about the same magnitude for the two groups of listeners. Plotted in the left panel of Fig. 3 are the mean PLIO weights for the NH and HI listeners (different symbols) with 95% CIs. Between group comparisons revealed that w_{PLIO} were significantly greater for the HI listeners at 257 and 603 Hz [$t(24)=2.4$, $p=0.03$; $t(24)=2.6$, $p=0.02$] but significantly greater for the NH listeners at 3951 Hz [$t(23)=3.9, p<0.001$]. These results demonstrate the advantage of incorporating sensorineural resolution into the model of the ideal listener. For best discrimination performance, HI listeners should listen differently than NH listeners. In particular, as a group, the observed weights should be slightly greater for the lowest frequency tones but less for the 3951 Hz tone. The important point is that individual listeners have unique ideal models based on information received and comparing them to the same ideal based on information sent instead of information received can penalize them for otherwise listening optimally.

B. Factors influencing observed decision weights

1. Hearing loss

The mean observed weights with 95% CIs for the NH and HI listeners are plotted in the right panel of Fig. 3. Observed weights for both groups of listeners clearly deviated from the weights for the PLIO. NH listeners relied primarily on the low-frequency tones (257–1197 Hz) for discrimina-

TABLE III. Hearing-impaired listeners (see Table II for description). Estimates σ_P for HI 12 were excluded from the means and SE's.

		HI 09	HI 01	HI 17	HI 04	HI 13	HI 11	HI 15	HI 10	HI 02	HI 18	HI 05	HI 12	HI 03	Mean	(SE)
QT	**257**	23.6	33.3	28.7	26.4	26.4	12.1	28.6	19.1	28.3	57.1	44.6	32.0	36.3	**30.5**	(3.24)
(dB SPL)	**603**	24.5	12.5	27.9	27.8	26.8	7.5	18.9	21.5	49.1	44.5	54.8	31.1	33.2	**29.2**	(3.96)
	1197	20.9	12.8	40.8	41.1	26.8	7.0	14.4	48.4	45.7	26.6	51.5	31.3	41.5	**31.4**	(4.22)
	2212	55.8	14.7	53.7	54.5	36.7	60.6	40.7	57.0	58.5	18.0	68.9	42.0	45.1	**46.6**	(4.66)
	3951	62.5	36.8	58.1	61.1	[a]	59.1	51.8	62.8	59.6	17.9	63.7	44.6	50.0	**52.3**	(4.11)
	6930	62.2	53.1	72.3	74.9	[a]	50.8	72.6	71.4	[a]	36.4	[a]	56.1	74.2	**62.4**	(4.32)
σ_p	**257**	0.07	0.04	0.09	0.15	0.17	0.08	0.21	0.30	0.14	0.46	0.26	0.89	0.35	**0.19**	(0.039)
(ERB_N)	**603**	0.06	0.03	0.06	0.06	0.12	0.05	0.16	0.23	0.07	0.53	0.22	0.73	0.36	**0.16**	(0.046)
	1197	0.08	0.05	0.06	0.09	0.09	0.09	0.15	0.32	0.08	0.18	0.67	0.98	0.70	**0.21**	(0.070)
	2212	0.29	0.06	0.08	0.23	0.09	0.40	0.10	0.23	0.37	0.29	0.88	0.71	0.63	**0.30**	(0.074)
	3951	1.35	0.12	0.36	0.73	[a]	0.23	0.24	1.09	0.68	1.01	1.39	2.32	1.32	**0.77**	(0.153)
	6930	1.10	1.02	0.49	1.19	[a]	0.42	0.90	0.35	[a]	1.83	[a]	1.24	1.41	**0.97**	(0.173)
w_{PLIO}	**257**	0.20	0.20	0.18	0.20	0.18	0.19	0.18	0.18	0.21	0.19	0.28	0.18	0.26	**0.20**	(0.010)
	603	0.20	0.20	0.18	0.20	0.25	0.19	0.19	0.21	0.33	0.18	0.30	0.22	0.25	**0.22**	(0.014)
	1197	0.20	0.20	0.18	0.20	0.29	0.19	0.19	0.17	0.32	0.23	0.18	0.17	0.15	**0.21**	(0.014)
	2212	0.19	0.19	0.18	0.19	0.29	0.14	0.20	0.21	0.09	0.22	014	0.22	0.17	**0.19**	(0.014)
	3951	0.09	0.17	0.15	0.12	[a]	0.17	0.17	0.06	0.05	0.12	0.10	0.07	0.09	**0.11**	(0.013)
	6930	0.11	0.04	0.13	0.09	[a]	0.13	0.07	0.16	[a]	0.07	[a]	0.13	0.08	**0.10**	(0.013)
w_{obs}	**257**	**0.49**	**0.47**	**0.38**	**0.68**	**0.44**	**0.61**	**0.49**	**0.74**	**0.23**	**0.86**	**1.00**	**0.45**	**0.64**	**0.58**	(0.060)
	603	**0.26**	**0.30**	**0.37**	**0.22**	**0.27**	**0.10**	**0.24**	**0.26**	**0.69**	(−0.09)	(0.07)	**0.36**	**0.36**	**0.2**	(0.053)
	1197	**0.19**	**0.23**	**0.15**	**0.10**	**0.29**	**0.29**	**0.26**	(0.04)	**0.09**	**0.14**	(0.06)	**0.19**	**0.03**	**0.16**	(0.027)
	2212	(0.03)	(0.05)	(0.07)	(0.01)	(0.08)	(0.08)	(0.06)	(0.04)	(0.03)	(−0.04)	(−0.03)	(0.06)	(−0.14)	**0.02**	(0.018)
	3951	**−0.07**	(−0.01)	**0.10**	−(0.014)	[a]	(−0.07)	(0.02)	(0.04)	(−0.03)	(0.05)	(−0.07)	(−0.13)	(0.06)	**−0.01**	(0.020)
	6930	(0.30)	(−0.05)	(−0.04)	(0.06)	[a]	(0.00)	(−0.06)	(0.01)	[a]	(0.11)		(−0.12)	(−0.10)	**−0.02**	(0.024)

[a]Denotes tones where QT >85 dB SPL and where the DLF (σ_P) was subsequently not tested.

tion to the exclusion of the high-frequency tones (2212–6930 Hz). As indicated by the bolded entries in Table II, which indicate statistical significance, only two NH listeners (NH 01 and NH 10) put significant weight on the highest-frequency tone. One of these listeners was the first author, a highly trained listener on this task, who put the greatest weight on the 6930 Hz tone. The HI listeners also put very little weight on the high-frequency tones. As shown in Table III, only two HI listeners put significant weight on a high-frequency tone (3951 Hz for HI 09 and HI 17). Unlike the NH listeners who put about equal emphasis on the low-frequency tones, the HI listeners put substantially greater weight on the lowest-frequency tone at 257 Hz with progressively less weight on the 603 and 1197 Hz tones.

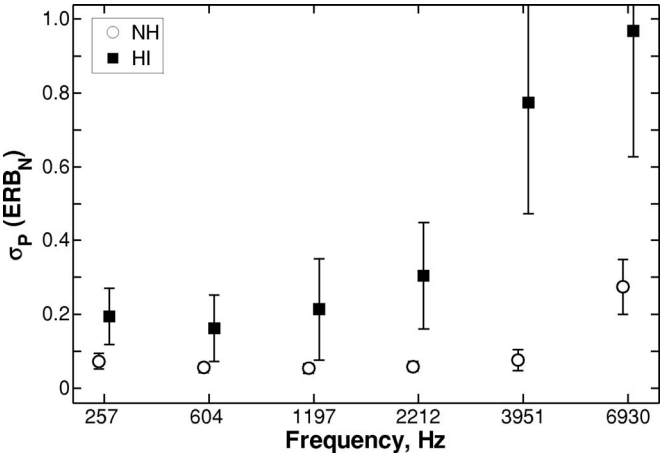


FIG. 2. Displayed are mean estimates of sensorineural resolution for each frequency region for the NH and HI listeners (open circles and filled squares, respectively) as obtained on an ERB_N scale. Error bars represent the 95% confidence intervals. (Estimates σ_P for HI 12 are excluded from the means and CI's.)

The above observations were qualified by within-subjects analyses of variance (ANOVAs) conducted separately for each group of listeners. Weights for the NH listeners were significantly greater than zero for every tone except the 6930 Hz tone. None of the lower-frequency weights was significantly different from each other and none of the higher-frequency weights was significantly different from each other. Each paired comparison across the lower- and higher-frequency groups was significant. For the HI listeners only the lower-frequency weights were significantly different from zero, especially the 257 Hz weight, which was significantly greater than all the other weights ($p<0.001$). The 603 Hz weight was significantly greater than the 2212 Hz weights and above and the 1197 Hz weight was significantly greater than the 3951 and 6930 Hz weights.

To see if there was a systematic relationship between the observed weights and the amount of hearing loss, the weights for each frequency were correlated with the QTs across both groups of listeners. The 257 Hz weight had a significant positive relationship with QT $[R(24)=0.68, p<0.001]$ and the 1197 and 3951 Hz weights had significant negative relationships with QT $[R(24)=0.62,\ p<0.001$ and

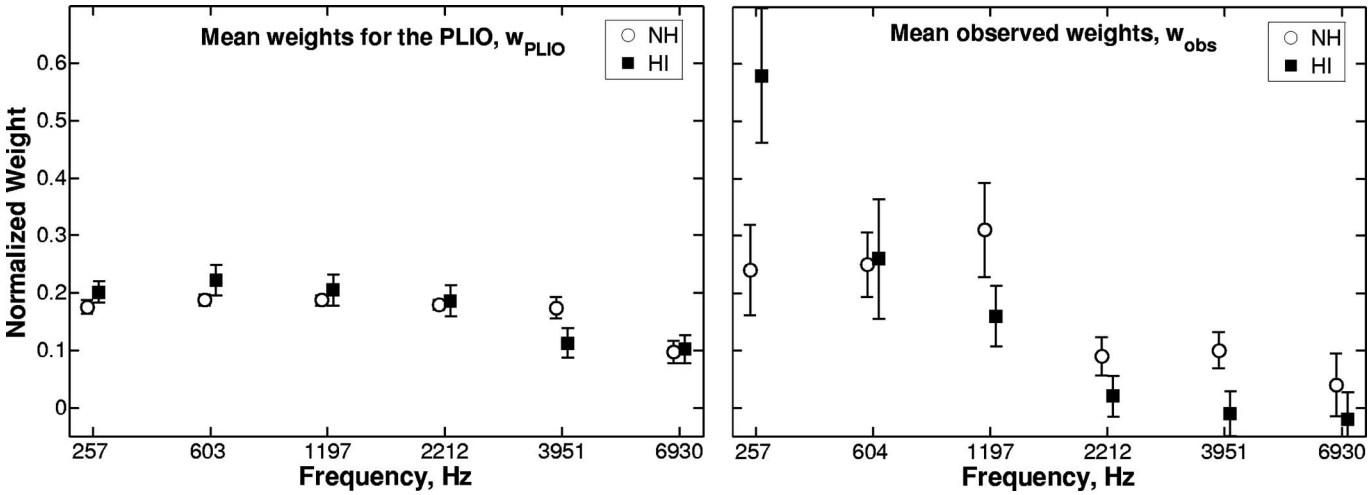


FIG. 3. Mean normalized weights with 95% confidence intervals are shown for each frequency region for the NH and HI listeners (open circles and filled squares, respectively). Weights for the peripherally limited ideal observer are plotted in the left panel and the observed weights in the right panel.

$R(23)=0.67$, $p<0.001$, respectively]. The 2212 Hz weight had a marginally significant negative relationship with QT [$R(24)=0.39, p=0.05$]. These results indicate that the low-frequency emphasis in the profile of observed weights increased as hearing loss increased.

2. Sensorineural resolution

The correlations between the observed weights and the weights for the PLIO for the NH listeners [$R(76)=0.44, p<0.001$] and for the HI listeners [$R(72)=0.50, p<0.001$] were moderate but also indicate that neither group adopted optimal decision rules derived from estimates of sensorineural resolution. w_{PLIO} are roughly equivalent to normalized values of σ_P for each listener. Perhaps weights were influenced by un-normalized values of σ_P, the degree of sensorineural distortion. To test this, the weights for each frequency were correlated with the logarithm of σ_P across both groups of listeners. There was a significant positive relationship between sensorineural distortion and the 257 Hz weights [$R(23)=0.61, p=0.001$], but significant negative relationships between sensorineural distortion and the 1197, 2212, and 3951 Hz weights [$R(23)=0.40$, $p<0.05$; $R(23)=0.61$, $p=0.001$; and $R(23)=0.51$, $p=0.01$, respectively]. The 603 Hz weight had a marginally significant negative relationship with sensorineural distortion [$R(23)=0.37, p=0.069$]. This is the same general pattern observed with QT. As overall sensory information becomes distorted, the 257 Hz weight takes on greater importance at the expense of the weights higher in frequency.

In summary, weighting functions for the NH and HI groups were quite variable across listeners. There was an overall tendency for the lower frequencies to be given the greatest weight by both groups. A clear pattern emerged with how HI listeners weighted the lowest-frequency tones centered around 257 Hz. As overall hearing loss increased and sensorineural resolution decreased, the 257 Hz weight increased at the expense of a decrease in the higher-frequency weights. A more formal assessment of whether NH and HI listeners differed in terms of how efficiently they weighted information will require that the weights and sensorineural noise estimates be integrated into the theoretical framework of the PLIO.

3. Sensitivity

The theoretical framework above introduced four decision models that progressively account for factors that influence listeners' sensitivity when discriminating stimuli in a multiple-observation, sample discrimination task. The stimulus ideal observer model, d'_{ideal}, accounts for the stimulus microstructure—the mean frequency differences across trial intervals for each pair of tones and the trial-by-trial perturbations in frequency. The PLIO model, d'_{PLIO}, accounts for sensorineural noise in both the derivation of the weights (w_{PLIO}) and in the variance of the expected decision variable. The models for d'_{wgt} and d'_{pred} account for listeners' unique weighting strategies, with the latter also accounting for the effects of limited sensorineural resolution in the standard deviation of the expected decision variable. Any remaining difference between d'_{pred} and observed sensitivity is attributed to central noise. The top panels in Fig. 4 display the means of these four estimates of sensitivity for the NH and HI listeners as a function of the expected frequency difference, δ_i. Predictions for each model are plotted as different line types and the observed sensitivities, d'_{obs}, are plotted as data points with 95% CIs.

From Fig. 4, it is clear that the observed sensitivities for both groups of listeners are substantially less than the stimulus ideal, d'_{ideal} (dotted line). The differences between these two reflect the combined influences of listeners' inefficient weighting strategies and internal noise. Compared to d'_{ideal}, the predictions for d'_{wgt} (dashed line) are much closer to d'_{obs}, which indicates that a majority of listeners' less-than-ideal sensitivity can be attributed to the inefficient weighting strategies discussed in the previous section. Still, estimates of d'_{wgt} are consistently greater than d'_{obs}, especially when the expected frequency differences are relatively small. This indicates that internal noise also contributes to the reduction in observed sensitivity. The differences in sensitivity between

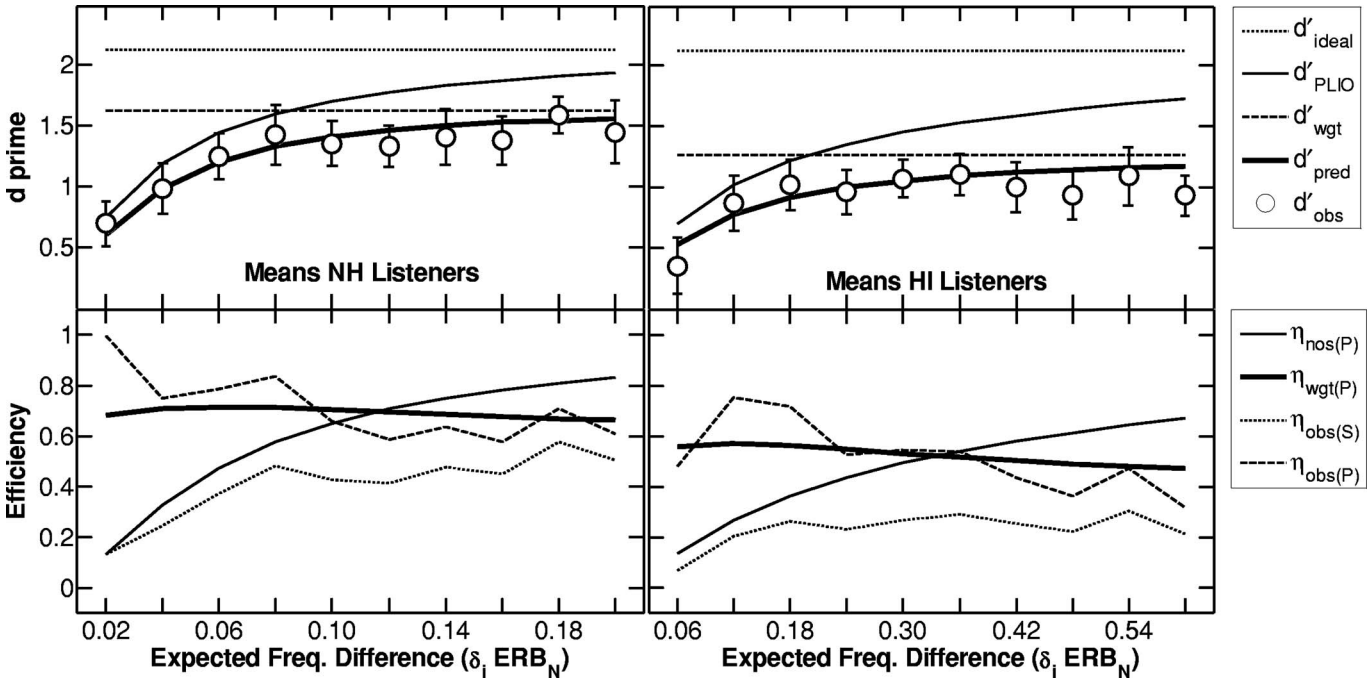


FIG. 4. For each decision model, the average estimates of sensitivity as a function of the expected frequency difference between intervals are displayed as different lines for the NH and HI listeners (left and right panels, respectively) in the top panels. The observed sensitivities are plotted as data points with 95% confidence intervals. Efficiencies, which compare different pairs of *d*-prime estimates from the top panels, are displayed in the bottom panels.

the stimulus ideal observer and the PLIO, d'_{PLIO} (thin solid line), reflects the hypothesized contribution of sensorineural noise to the observed sensitivity. As expected, the effects of limited frequency resolution are greatest when the frequency manipulations are smallest. As the size of the frequency manipulations increase relative to the sensorineural noise, the mean sensitivity for the PLIO increases and approaches sensitivity for the stimulus ideal. The predicted sensitivity for d'_{pred} (thick solid line) combines the contributions of listeners' inefficient weighting strategies with sensorineural noise. Estimates for d'_{pred} are lower than d'_{PLIO} which reflects the fact that even when compared to a PLIO most listeners employ inefficient weighting strategies. Estimates for d'_{pred} very closely approximate d'_{obs}. This indicates that for this experiment, sensorineural noise alone seems to account for most of the reduction in observed sensitivity that would otherwise be attributed to internal noise generally or to central noise specifically.

The left half of Table IV presents the means of the sensitivity estimates for the different decision models across the 10 trial blocks for each listener and for each group as a whole. Sensitivity for the stimulus ideal is not displayed because d'_{ideal} was a constant 2.12 for each listener. As displayed in the last row of the table, sensitivity estimates for each model (excluding d'_{ideal}) were significantly less for the HI listeners compared to the NH listeners. The significant difference in d'_{obs} indicates that HI listeners were less sensitive than NH listeners even though frequency manipulations were three times greater for the former. The significant differences between the groups in d'_{wgt}, d'_{pred}, and d'_{PLIO} indicate that both less efficient weighting strategies and greater sensorineural noise in the HI listeners were responsible for the differences in observed sensitivity.

C. Efficiency

1. Theoretical overview

To isolate the different factors that influence listeners' observed sensitivities, the above model estimates for sensitivity need to be compared. For this purpose, we adopt a normalizing statistic akin to R^2 with a long history in the literature. With origins in energy-detector analogs for observers, efficiency is defined as a ratio of variances and is computed by taking the ratio of squared *d* primes (Tanner and Birdsall, 1958). Within our theoretic treatment of the data, this property makes the efficiency measure more attractive than simply taking the difference between two *d* primes. For example, an efficiency of 0.5 could indicate that the pooled variance of the expected decision variable differs by a factor of 2 between the two models or that the expected values of the decision variables differ by a factor of $\sqrt{2}$. Most often, the expected decision variables will differ in both the numerator (mean) and denominator (variance). Efficiency, denoted η, is 1 when the two models under consideration have the same sensitivity.

Efficiency analysis as outlined by Berg (1990) provides a useful framework for partitioning the effects of decision weights (η_{wgt}) and internal noise (η_{nos}) on a listener's sensitivity relative to the stimulus ideal observer (η_{obs}). We expand on Berg's convention to (1) account for the fact that the ideal weighting strategy will differ for listeners and conditions as a function of information received, (2) attempt to partial out the sensorineural and central components of internal noise, and (3) isolate as much as possible the effects sensorineural noise on the observed efficiency.

Table V lists the different efficiencies used in our analysis along with the ratios (model comparisons) that define

TABLE IV. Displayed are the means of the sensitivity estimates (d') for the different decision models and efficiency measures (η) for the NH and HI listeners collapsed across the ten trial blocks, rank ordered for each group according to d'_{obs} (see text). Note that $d'_{ideal}=2.12$ for all listeners. The means and standard errors of the mean for each group are also provided.

	d'_{obs}	d'_{wgt}	d'_{PLIO}	d'_{pred}	$\eta_{obs(S)}$	$\eta_{wgt(P)}$	$\eta_{nos(P)}$	$\eta_{nos(C)}$	$\eta_{obs(P)}$
NH 08	1.59	1.84	1.66	1.50	0.56	0.82	0.61	1.12	0.92
NH 10	1.55	1.87	1.31	1.26	0.53	0.91	0.38	1.53	1.39
NH 07	1.42	1.76	1.44	1.38	0.45	0.91	0.46	1.07	0.98
NH 06	1.41	1.70	1.70	1.56	0.44	0.83	0.64	0.82	0.69
NH 12	1.37	1.67	1.72	1.26	0.42	0.54	0.66	1.18	0.63
NH 02	1.36	1.42	1.70	1.15	0.41	0.46	0.64	1.39	0.64
NH 11	1.35	1.65	1.68	1.46	0.41	0.76	0.63	0.86	0.65
NH 14	1.29	1.44	1.56	1.31	0.37	0.71	0.54	0.97	0.69
NH 04	1.26	1.74	1.64	1.57	0.35	0.91	0.60	0.65	0.59
NH 09	1.16	1.50	1.55	1.21	0.30	0.61	0.54	0.92	0.56
NH 01	1.00	1.82	1.76	1.37	0.22	0.61	0.69	0.53	0.32
NH 13	0.95	1.24	1.57	0.96	0.20	0.38	0.55	0.98	0.37
NH 03	0.95	1.44	1.51	1.06	0.20	0.49	0.51	0.81	0.40
Mean	**1.28**	**1.62**	**1.60**	**1.31**	**0.38**	**0.69**	**0.57**	**0.99**	**0.68**
(SE)	(0.06)	(0.06)	(0.04)	(0.05)	(0.03)	(0.05)	(0.03)	(0.08)	(0.08)
HI 09	1.19	1.27	1.46	1.06	0.31	0.53	0.47	1.25	0.66
HI 01	1.14	1.44	1.81	1.40	0.29	0.61	0.72	0.66	0.40
HI 17	1.14	1.54	1.76	1.41	0.29	0.64	0.69	0.65	0.42
HI 04	1.12	1.19	1.50	1.02	0.28	0.47	0.50	1.19	0.56
HI 13	1.06	1.46	1.54	1.25	0.25	0.66	0.53	0.72	0.48
HI 11	1.03	1.27	1.70	1.18	0.24	0.48	0.64	0.77	0.37
HI 15	1.01	1.42	1.60	1.14	0.23	0.51	0.57	0.79	0.40
HI 10	0.95	1.10	1.37	0.78	0.20	0.32	0.42	1.51	0.48
HI 02	0.95	1.19	1.49	1.11	0.20	0.56	0.49	0.74	0.41
HI 18	0.79	1.03	1.14	0.59	0.14	0.26	0.29	1.80	0.48
HI 05	0.76	0.87	1.01	0.63	0.13	0.40	0.23	1.44	0.57
HI 12	0.53	1.43	0.72	0.55	0.06	0.59	0.12	0.92	0.54
HI 03	0.44	1.20	0.98	0.78	0.04	0.64	0.21	0.31	0.20
Mean	**0.93**	**1.26**	**1.39**	**0.99**	**0.20**	**0.51**	**0.45**	**0.98**	**0.46**
(SE)	(0.07)	(0.06)	(0.10)	(0.09)	(0.03)	(0.04)	(0.06)	(0.12)	(0.03)
t	4.0[a]	4.7[a]	2.1[c]	3.3[b]	4.2[a]	2.8[b]	2.1[c]	0.1	2.6[c]

The last row displays the results of t tests with 24 degrees of freedom for each estimate between the NH and HI listeners
$p \leqslant 0.001$;
$p \leqslant 0.01$;
$p \leqslant 0.05$.

them and the theoretical constructs they isolate. Observer efficiency, $\eta_{obs(S)}$, is the same as Berg's original η_{obs} but is subscripted (S) to denote that the observed sensitivity is compared to the stimulus ideal. Differences between the two sensitivities are attributed to observed weights that deviate from the stimulus ideal (w_{ideal}), to sensorineural noise, and to central noise. Unlike Berg's original η_{wgt}, which compares the observed weighting strategy to one based on an index of information sent, we are interested in how efficiently listeners weight information compared to an ideal model that weights information proportional to the reliability of information received, w_{PLIO}. This comparison is particularly important for the present study because the reliability of information received varies considerably between listeners, conditions, and frequencies. We introduce a modified weight efficiency, $\eta_{wgt(P)}$, that is obtained by comparing two estimates of sensitivity that include the listener's unique profile of sensorineural noise. One of these estimates uses the listener's unique weighting strategy, d'_{pred}, and the other estimate uses the ideal weighing strategy based on information received, d'_{PLIO}.

In an attempt to separate the effects of sensorineural and central noise, we define $\eta_{nos(P)}$ and $\eta_{nos(C)}$, respectively (cf. Appendix). By $\eta_{nos(P)}$ we mean the loss of efficiency due to

TABLE V. Analytic definitions and constructs of interest for each of the efficiency measures.

Efficiency	Notation	Analytic definition	Constructs
Observer	$\eta_{obs}(S)$	$(d'_{obs}/d'_{ideal})^2$	w_{obs} *vs.* w_{ideal} σ_P^2 and σ_C^2
Weight	$\eta_{wgt}(P)$	$(d'_{pred}/d'_{PLIO})^2$	w_{obs} *vs.* w_{PLIO}
Sensorineural noise	$\eta_{nos}(P)$	$(d'_{PLIO}/d'_{ideal})^2$	w_{PLIO} *vs.* w_{ideal} σ_P^2
Central noise	$\eta_{nos}(C)$	$(d'_{obs}/d'_{pred})^2$	σ_C^2
Adjusted observer	$\eta_{obs}(P)$	$(d'_{obs}/d'_{PLIO})^2$	w_{obs} *vs.* w_{PLIO} σ_C^2

sensorineural noise. Again, the effects of sensorineural noise are twofold. First, sensorineural noise alters how an ideal observer should weight information. Second, sensorineural noise adds to the variance of the expected decision variable. To index the cumulative effects of sensorineural noise, we compare the sensitivity of the PLIO (d'_{PLIO}) to that of the stimulus ideal (d'_{ideal}). By our definition, central noise is the residual variance in the expected decision variable after accounting for sensorineural noise and the listener's weighting strategy. Therefore, $\eta_{\mathrm{nos}(C)}$ is obtained by comparing the observed sensitivity of the listener (d'_{obs}) to the sensitivity of a hypothetical listener who shares the same weighting strategy and sensorineural limitations (d'_{pred}).

With the above analytic definitions in place, the efficiency of the observed sensitivity relative to the stimulus ideal, $\eta_{\mathrm{obs}(S)}$, can be partitioned as follows:

$$\eta_{\mathrm{obs}(S)} = \eta_{\mathrm{wgt}(P)} \times \eta_{\mathrm{nos}(P)} \times \eta_{\mathrm{nos}(C)}. \tag{4}$$

To isolate as much as possible the effects of sensorineural noise on observer efficiency for a fair comparison between NH and HI listeners, the listener's observed sensitivity (d'_{obs}) should be compared not to the stimulus ideal observer (d'_{ideal}), but to a peripherally limited observer who weights the information received at each frequency optimally (d'_{PLIO}). We call this comparison the adjusted observer efficiency, $\eta_{\mathrm{obs}(P)}$, which is obtained by dividing both sides of Eq. (4) by $\eta_{\mathrm{nos}(P)}$. The resultant equation, Eq. (5), is reminiscent of Berg's original efficiency analysis

$$\eta_{\mathrm{obs}(P)} = \eta_{\mathrm{wgt}(P)} \times \eta_{\mathrm{nos}(C)}. \tag{5}$$

2. Results

The relationships between the sensitivity estimates of the different decision models are shown as efficiencies in the bottom panels of Fig. 4. Observed sensitivity for both NH and HI listeners were limited by sensorineural noise, which diminished the reliability of information received when the frequency manipulations were relatively small. This is indicated by $\eta_{\mathrm{nos}(P)}$, which was about 0.2 at the smallest frequency manipulations for both groups. At the largest frequency manipulations, $\eta_{\mathrm{nos}(P)}$ was about 0.8 for the NH listeners and about 0.65 for the HI listeners. Because $\eta_{\mathrm{nos}(P)}$ was much less than 1 for the HI listeners, this indicates that there is still potential for improvement in observed sensitivity at larger frequency manipulations than what was tested. After accounting for the effects of sensorineural noise in the listeners' observed sensitivity, $\eta_{\mathrm{obs}(P)}$ was much higher compared to the traditional assessment of observed sensitivity, $\eta_{\mathrm{obs}(S)}$, especially when the expected frequency differences were relatively small compared to sensorineural resolution.

The right half of Table IV presents the efficiency measures derived from the means of the different sensitivity estimates for each listener and for each group as a whole. As displayed in the last row of the table, all efficiency measures were significantly less for the HI listeners compared to the NH listeners, except $\eta_{\mathrm{nos}(C)}$. Observer efficiency (re: the stimulus ideal), $\eta_{\mathrm{obs}(S)}$, was significantly lower for HI listeners compared to the NH listeners. A significant contributing factor to this difference was sensorineural noise efficiency. However, even when differences in sensorineural noise were accounted for, observer efficiency (re: the PLIO), $\eta_{\mathrm{obs}(P)}$, was still significantly lower for the HI listeners. This is because the exaggerated low-frequency emphasis in the weighting strategies of the HI listeners resulted in significantly lower weighting efficiencies compared to the NH listeners, even though differences in sensorineural resolution were accounted for in the analysis of $\eta_{\mathrm{wgt}(P)}$. $\eta_{\mathrm{nos}(C)}$ was not statistically different between the NH and HI listeners indicating that SNHL did not influence the variability or consistency with which the decision strategies were used across trials and conditions.[4]

IV. DISCUSSION

One of the primary objectives of this study was to adapt the theory of ideal observers to include the effects of frequency-specific sensorineural noise, what we term the peripherally limited ideal observer, or PLIO, decision model. This allows us to shift the research focus from information sent to information received when trying to understand differences in observed sensitivity between listeners. This is important because frequency resolution at the auditory periphery can differ substantially as a function of hearing loss and frequency, which has implications for how internal noise and ideal weights are conceptualized within the decision theoretic framework of the PLIO.

A. Sensorineural and central noise

Internal noise is often used as a general umbrella term that includes both the effects of imprecise coding of the physical properties of the stimulus at the auditory periphery and the effects of noisy decision processes arising from inconsistency and inattention on the part of the listener. We identify these two sources of internal noise as sensorineural and central noise, respectively. Using listener and frequency specific estimates of DLFs, the effects of sensorineural noise on observed sensitivity were modeled as additional sources of variance in the expected decision variable. When these effects were combined with a listener's unique weighting strategy, a specific prediction for observed sensitivity, d'_{pred}, was obtained. The intention was that any residual variance needed to bring this prediction closer to the observed sensitivity would be modeled as a nonintegrated constant associated with central noise. As it turned out, the inclusion of sensorineural noise was adequate at bringing the estimates of d'_{pred} very near the observed sensitivities. This indicates that SNHL did not influence central processes related to the implementation of a decision strategy across trials and conditions and that sensorineural noise was the primary component of the overall internal noise in this discrimination task for both groups of listeners.

B. Decision weights

Within our theoretical framework for the PLIO, ideal weights should be proportional to the reliability of information received, which takes into account each listener's unique profile of sensorineural noise. In this study, most listeners did

not employ optimal weighting strategies. Instead, higher-frequency tones (≥2212 Hz) were given very little weight by individuals from either group of listeners compared to lower-frequency tones (≤1197 Hz). Furthermore, as hearing loss and sensorineural noise increased there was a tendency for the 257 Hz weight to increase at the expense of a decrease in the other weights. This resulted in significantly lower weighting efficiencies for the HI listeners.

There are at least a few explanations for why the lower-frequency tones were on average given significantly greater weight than the higher-frequency tones.[5] First, it is worth noting that there were a number of instances where the weights in the high-frequency region were significantly greater than zero for the NH listeners. This was especially true for the first author, a highly practiced listener on this task, who put the greatest weight on the highest-frequency tone. This indicates that attentional mechanisms influence the magnitude of the high-frequency weights to some extent, rather than a simple failure to hear the individual tones due to upward spread of masking, for example. A preference for low-frequency information, especially on a relative scale, has been demonstrated elsewhere.

In a series of experiments, Neff and Odgaard (2004) had listeners engage in sample discrimination of frequency experiments similar in nature to the present study except listeners were asked to focus selectively on a single frequency difference near 2000 Hz instead of integrating information across six simultaneous differences. Additional frequency information above and below the target frequency served as distracters and consisted of either random-frequency tones, fixed-frequency tones with and without varying level, or noise bands. For random-frequency tones, weighting analyses revealed a dominance of the lower-frequency distracters over the higher-frequency distracters. This relative patten persisted even when the distracters and targets were shifted two octaves higher in frequency. Furthermore, the critical feature of the low-frequency distracters was informational (variation) rather than energetic (merely being present) since fixed-frequency tones and noise bands produced little to no negative effects on target discrimination.

Alexander (2004, pp. 110–154; Alexander and Lutfi, 2003) reported that in tone detection tasks involving random-frequency distracters and fixed-frequency targets at 800, 2000, or 5000 Hz that (1) detection thresholds increased with increases in the informational content of distracters below the target frequency compared to those above and (2) distracters received significantly greater weight when they were immediately below the target frequency compared to when they were immediately above. Alexander also found that for a 2000 Hz target, randomly turning on and off the below-target distracters from trial to trial (informational masking) while keeping the above-target distracters always on significantly increased thresholds for target detection compared to the opposite when the below-target distracters were static (energetic masking) and the above-target distracters were random.

We carried out two related experiments using the same listeners and almost the same conditions. In one experiment, the high-frequency tones were randomly selected as before except that they were selected from the exact same distribution for the signal and noise intervals so that the expected frequency difference for these tone pairs was zero. Since the high-frequency tones had no informational value across the experiment (i.e., w_{PLIO} equaled zero), they were distracters and listeners were reminded to ignore them. Likewise, in a second experiment, the low-frequency tones had expected frequency differences of zero and served as distracters. Performance in the high-frequency distracter experiment yielded weighting functions and sensitivity estimates very close to the current experiment. However, in the low-frequency distracter experiment d'_{obs} and d'_{pred} were very close to zero for most of the listeners from both groups since the weighting patterns were still low-frequency dominant. Like the Neff and Odgaard (2004) and Alexander (2004; Alexander and Lutfi, 2003) studies, even when it was highly disadvantageous to attend to the low-frequency information, listeners could not ignore it.

It seems clear that listeners enter the laboratory setting with a preset weighting pattern for pitch-like discriminations. Relatively low frequencies tend to dominate the perception of tonal complexes. Perhaps only with a lot of practice and motivation can this bias be overcome. This preset weighing strategy is determined by a lifetime of experience attending to real-world sounds, especially the most functionally relevant sounds like speech. It is known that the energy of real environments and especially of speech and music is distributed much like pink noise and decreases as a function of frequency (e.g., Voss and Clarke, 1975, 1978). A person with a longstanding high-frequency sensorineural hearing loss experiences an exaggerated form of this low-frequency dominance. While the above explanation is speculative, the fact remains that people will bring to the laboratory setting an ingrained listening bias that may not conform to our task demands. An ecologically informed ideal observer model would appreciate the statistics of the acoustics of the everyday world. Within this framework, predictions for sensitivity will better accord with listeners' performance. Future research can benefit from exploring how experience with hearing impairment shapes the way people listen to complex sounds that vary along several acoustic dimensions, like speech. Casting performance in terms of an ecologically informed ideal observer that includes peripheral limitations as well as environmental experience will improve on models of speech intelligibility, including the effects of noise, the spatial distribution of multiple talkers, and various hearing aid and cochlear implant processing algorithms.

ACKNOWLEDGMENTS

This work was part of a dissertation submitted by the first author in partial fulfillment of the requirements for the Ph.D. in Communicative Disorders at the University of Wisconsin-Madison. The authors would like to thank Karen Malott and Laurie Canadeo for their assistance during data collection. The authors would especially like to thank Dr. Bruce Berg for his generous comments that helped to clarify the technical details of the decision theory equations and notation and an anonymous reviewer whose comments on

earlier drafts proved extremely helpful. This research was supported by a grant to the first author from The National Organization for Hearing Research Foundation and a grant to the second author from the NIDCD (R01 DC1262-10). This manuscript was written while supported by NIDCD grants R01 DC04072 (Keith R. Kluender) and T32 DC000013 (BT-NRH)

APPENDIX: THE PARTITION OF NOISE EFFICIENCY

For the sake of simplicity, assume that a hypothetical listener has equal σ_P for each tone so that the weights for d'_{PLIO} are equal to each other and are identical to the weights for d'_{ideal}. Furthermore, assume that the listener employs an optimal listening strategy so that the weights can be treated as constants and disregarded altogether. Therefore, from Eq. (3) we have

$$\eta_{\mathrm{nos}(P)} = \left(\frac{d'_{\mathrm{PLIO}}}{d'_{\mathrm{ideal}}}\right)^2 = \frac{2(m\delta)^2}{m2\sigma_S^2 + m2\sigma_P^2} \times \frac{m2\sigma_S^2}{2(m\delta)^2} = \frac{m2\sigma_S^2}{m2\sigma_S^2 + m2\sigma_P^2}, \tag{A1}$$

$$\eta_{\mathrm{nos}(C)} = \left(\frac{d'_{\mathrm{obs}}}{d'_{\mathrm{pred}}}\right)^2 = \frac{2(2m\delta)^2}{m2\sigma_S^2 + m2\sigma_P^2 + \sigma_C^2} \times \frac{m2\sigma_S^2 + m2\sigma_P^2}{2(m\delta)^2} = \frac{m2\sigma_S^2 + m2\sigma_P^2}{m2\sigma_S^2 + m2\sigma_P^2 + \sigma_C^2}, \tag{A2}$$

$$\eta_{\mathrm{nos}} = \eta_{\mathrm{Pnos}} \times \eta_{\mathrm{Cnos}} = \frac{m2\sigma_S^2}{m2\sigma_S^2 + m2\sigma_P^2 + \sigma_C^2}. \tag{A3}$$

[1]Because there was no overlap in age between the NH and HI listeners, we cannot be sure to what extent the obtained results, such as estimates of frequency resolution, reflect processes related to aging *per se*. Fortunately, this aspect is accounted for because sensorineural resolution is input for the different decision models. It is also possible that age-related factors influenced how listeners weighted information in the multiple-observation, sample discrimination experiment, although, at this time, it is not entirely clear how or why.

[2]Sometimes over the course of the four blocks there was improvement for one or more frequencies. The blocks and frequencies for which this occurred did not seem systematic within or between listeners. Additional blocks were run until there was no substantial improvement for four successive blocks. Paired *t* tests comparing DLFs obtained from the two trial blocks immediately before the sample discrimination experiment and from the last two trial blocks obtained after the sample discrimination experiment were not statistically significant for any of the six tones in either group ($p \geq 0.10$ for all comparisons). This indicates that there was a lack of a practice effect over the course of the experimental protocol.

[3]An index of sensitivity can at least be calculated by (l) the difference in the *Z*-transformed hit and false-alarm rates, or (2) the difference in the means of the signal and noise distributions of the decision variable (Δ) divided by their common standard deviation (σ). For a yes-no (y/n) task these two computations are the same. However, for a 2IFC task the mean difference (the numerator) is integrated by a factor of 2, while the standard deviation (the denominator) is integrated by $\sqrt{2}$. This leads to an overall increase of $\sqrt{2}$ between the difference in the *Z*-transformed hit and false-alarm rates for the 2IFC task compared to the y/n task. Because it is desirable to have an index of sensitivity that is task independent, the Δ/σ ratio is the preferred index for d', meaning that the difference in the *Z*-transformed hit and false-alarm rates for the 2IFC task must divided by $\sqrt{2}$. The difference in the *Z*-score computation of sensitivity between the two tasks reflects the fact that greater sensitivity is needed to achieve the same performance or percent correct on a y/n task compared to a 2IFC task.

[4]While at first it seems that central noise played little role in this multiple-observation, sample discrimination task, it is likely that each estimate of sensorineural noise collected during the DLF procedure already included some central noise. The actual central noise during the sample discrimination experiment might be only some small factor of the central noise collected during the estimate of one DLF. Therefore, analytically integrating sensorineural noise (and central noise) across the six frequency regions might lead to an over estimate of the total amount of sensorineural noise. In this regard, d'_{pred} and d'_{wgt} might be considered as lower and upper bounds, respectively, for modeling the influence of sensorineural noise on observed sensitivity. This combined with the fact that predictions for sensorineural noise and sensitivity were analytically derived from expected values might explain the few instances where d'_{pred} under-predicts d'_{obs} and where $\eta_{\mathrm{nos}(C)}$ is greater than one.

[5]One review noted that the level of the broadband stimuli in this experiment (about 50 and 85 dB SL for the HI and NH listeners, respectively) possibly elicited the acoustic reflex in some or most listeners. Because maximal attenuation from the graded acoustic reflex occurs between 500 and 1000 Hz (about 20–25 dB with 10 dB/oct. slopes), it could be responsible in part for the observed low-frequency dominance in the weighting patterns. Presumably, their reduced loudness would cause them to "pop-out" perceptually (unlikely) or their sensorineural resolution would be enhanced because the broadening of auditory filters associated with high-presentation levels would be somewhat mitigated by the attenuation. The following results suggest that the acoustic reflex was not responsible for the low-frequency dominance of the observed weights. Weights were also obtained from a subset of listeners ($n=10$ for each group) who replicated this study using a shorter stimulus duration (40 ms versus 120 ms) in Alexander (2004). Because the acoustic reflex is dependent on a temporal summation of energy, its effects should have been diminished at the shorter duration. This being said, paired-sample *t* tests revealed that no comparisons were statistically significant except for the HI listeners for the 257 and 603 Hz weights. The shorter duration was associated with significantly lower weights at 257 Hz and with significantly higher weights at 603 Hz compared to the longer duration ($p<0.05$). However, this might be because sensorineural resolution was significantly worse at 257 Hz for the shorter duration compared to the longer duration.

Alexander, J. M., and Lutfi, R. A. (**2003**). "Upward spread of informational masking in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **113**, 2287.

Alexander, J. M., and Lutfi, R. A. (**2004**). "Informational masking in hearing-impaired and normal-hearing listeners: Sensation level and decision weights," J. Acoust. Soc. Am. **116**, 2234–2247.

Alexander, J. M. (**2004**). Decision Factors in Multitone Detection and Discrimination by Listeners with Normal and Impaired Hearing (Unpublished dissertation, University of Wisconsin-Madison).

ANSI. (**1996**). ANSI S3.6-1996, "American National Standards specification for audiometers," (American National Standards Institute, New York).

Berg, B. G. (**1990**). "Observer-efficiency and weights in a multiple observation task," J. Acoust. Soc. Am. **88**, 149–l58.

Buus, S., Florentine, M., and Zwicker, T. (**1995**). "Psychometric functions for level discrimination in cochlearly impaired and normal listeners with equivalent-threshold masking," J. Acoust. Soc. Am. **98**, 853–861.

Doherty, K. A., and Lutfi, R. A. (**1996**). "Spectral weights for overall level discrimination in listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **99**, 1053–1057.

Doherty, K. A., and Lutfi, R. A. (**1999**). "Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **105**, 1831–1840.

Florentine, M., Reed, C. M., Rabinowitz, W. M., Braida, L. D., Durlach, N. I., and Buus, S. (**1993**). "Intensity perception. XIV. Intensity discrimination in listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **94**, 2575–2586.

Freyman, R. L., and Nelson, D. A. (**1991**). "Frequency discrimination as a function of signal frequency and level in normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **34**, 1371–1386.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Jesteadt, W., Nizami, L., and Schairer, K. S. (**2003**). "A measure of internal

noise based on sample discrimination," J. Acoust. Soc. Am. **114**, 2147–2157.
Lentz, J. J., and Leek, M. R. (**2002**). "Decision strategies of hearing-impaired listeners in spectral shape discrimination," J. Acoust. Soc. Am. **111**, 1389–1398.
Lentz, J. J., and Leek, M. R. (**2003**). "Spectral-shape discrimination by hearing-impaired and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 1604–1616.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Lutfi, R. A. (**1989**). "Informational processing of complex sound. I: Intensity discrimination," J. Acoust. Soc. Am. **86**, 934–944.
Lutfi, R. A. (**1990**). "Informational processing of complex sound. II: Cross-dimensional analysis," J. Acoust. Soc. Am. **87**, 2141–2148.
Lutfi, R. A. (**1992**). "Informational processing of complex sound. III: Interference," J. Acoust. Soc. Am. **91**, 3391–3401.
Lutfi, R. A., Doherty, K. A., and Oh, E. (**1996**). "Psychometric functions for the discrimination of spectral variance," J. Acoust. Soc. Am. **100**, 2258–2265.
Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide*, 2nd ed. (Erlbaum, Mahwak, NJ).
Moore, B. C. J. (**1997**). *An Introduction to the Psychology of Hearing*, 4th ed. (Academic, San Diego).
Neff, D. L., and Odgaard, E. C. (**2004**). "Sample discrimination of frequency differences with distracters," J. Acoust. Soc. Am. **116**, 3051–3061.
Sek, A., and Moore, B. C. J. (**1995**). "Frequency discrimination as a function of frequency, measured in several ways," J. Acoust. Soc. Am. **97**, 2479–2486.
Sorkin, R. D., Robinson, D. E., and Berg, B. G. (**1987**). "A detection theory method for the analysis of visual and auditory displays," in *Proc. Human Factors Soc., 31st Annual Meeting*, pp. 1184–1188.
Tanner, W. P., and Birdsall, T. G. (**1958**). "Definitions of d' and η as psychological measures," J. Acoust. Soc. Am. **30**, 922–928.
Voss, R. F., and Clarke, J. (**1975**). "'$1/f$ noise' in music and speech," Nature (London) **258**, 317–318.
Voss, R. F., and Clarke, J. (**1978**). "'$1/f$ noise' in music: Music from $1/f$ noise," J. Acoust. Soc. Am. **63**, 258–263.

The effect of masker level uncertainty on intensity discrimination

Emily Buss[a)]
Department of Otolaryngology/Head and Neck Surgery, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599

(Received 2 April 2007; revised 15 October 2007; accepted 15 October 2007)

Thresholds were measured for detection of an increment in level of a 60-dB SPL target tone at 1 kHz, either in quiet or in the presence of maskers at 0.5 and 2 kHz. Interval-by-interval level rove applied independently to remote masker tones substantially elevated thresholds compared to intensity discrimination in quiet, an effect on the order of 10+dB $[10\log(\Delta I/I)]$. Asynchronous onset and stimulus envelope mismatches across frequency reduced but did not eliminate masking. A preinterval cue to signal frequency had no effect, but cuing masker frequency reduced thresholds, whether or not masker level was also cued. About 1 to 2 dB of threshold elevation in these conditions can be attributed to energetic masking. Decreasing the overall presentation level and increasing masker separation essentially eliminates energetic masking; under these conditions masker level rove elevates thresholds by approximately 7 dB when the target and masker tones are gated synchronously. This masking persists even when the flanking masker tones are presented contralateral to the target. Results suggest that observers tend to listen synthetically, even in conditions when this strategy reduces sensitivity to the intensity increment.
 [DOI: 10.1121/1.2812578]

PACS number(s): 43.66.Dc, 43.66.Fe [AJO] Pages: 254–264

I. INTRODUCTION

The term *masking* can be used to describe any elevation in signal threshold due to the presence of a masker. Masking is said to be *energetic* when the peripheral response to the masker interferes with response to the signal; this occurs when the auditory channel or channels best representing the signal are also excited or suppressed by concurrent masker energy. However, psychophysical thresholds are not determined solely by energy in the channel associated with the signal frequency. Across-channel effects have been shown to introduce masking under some conditions, such as comodulation-detection difference (Cohen and Schubert, 1987; McFadden, 1987) and across-channel masking (Moore *et al.*, 1990). Under other stimulus conditions, across-channel effects are thought to improve sensitivity, as demonstrated in the paradigms of profile analysis (Green, 1988) and comodulation masking release (Hall *et al.*, 1984). These across-channel effects are often discussed in terms of central auditory processing cues, including those thought to underlie grouping (Hall and Grose, 1990), but could also involve peripheral mechanisms, such as suppression (e.g., Oxenham and Plack, 1998; Moore and Borrill, 2002). Threshold elevation that cannot be attributed to *energetic* masking is sometimes described as *informational* masking. Informational masking is typically assumed to occur central to the cochlea and is thought to be due to stimulus uncertainty and/or perceptual similarities between masker and signal (for a review, see Durlach *et al.*, 2003).

The experiments described here examine an effect reported by Fantini and Moore (1994). That study compared the conditions under which different classes of across-channel cues improve thresholds. In one control condition, observers were asked to detect a level increment in one tone in the presence of remote masker tones for which level was roved. An optimal listening strategy in this task would be to monitor a narrow frequency region around the signal frequency and to ignore auditory channels associated with the masker tones. However, the presence of roved masker tones elevated thresholds by approximately 4 dB $[10\log(\Delta I/I)]$ despite the fact that energetic masking was argued to be negligible for the conditions tested. Fantini and Moore called this effect *across channel interference* (ACI), and while they did not describe it in this way, ACI can be thought of as a form of informational masking associated with masker level uncertainty.

The literature on informational masking has traditionally focused on the detrimental effects of frequency uncertainty. In one common paradigm, pure tone detection for a signal at a fixed frequency is estimated in the presence of a masker composed of pure tones with randomly selected frequencies, excluding a protected region around the signal frequency. Thresholds under these conditions can be elevated by 40 dB or more relative to fixed masker frequency conditions (Kidd *et al.*, 1994; Neff and Dethlefs, 1995). Using a frequency uncertainty paradigm, Neff and Callaghan (1988) assessed the effects of roving masker frequency and/or amplitude. While frequency rove elevated detection thresholds substantially, amplitude rove had little or no effect, suggesting that amplitude uncertainty was not associated with informational masking for these stimuli. Later modeling work by Oh and Lutfi (1998) bolstered the conclusion that masker amplitude rove does not introduce informational masking for tone detection. On the face of it this conclusion may seem at odds

a)Electronic mail: ebuss@med.unc.edu

with the ACI result of Fantini and Moore (1994), which can be described as significant masking in the face of masker amplitude (but not frequency) uncertainty. The key difference between the paradigms used in these studies may be the task used to quantify masking; amplitude rove may have very different effects for tone detection and intensity discrimination, with substantial informational masking in the latter case.

Results of several recent studies are consistent with the conclusion that there is substantial informational masking for intensity discrimination in the presence of masker level rove. Both Doherty and Lutfi (1999) and Stellmack *et al.* (1997) estimated spectral weights for intensity discrimination of one tone in the presence of remote masker tones; applying level rove introduced substantial across-channel masking. While these results are broadly consistent with the previous ACI data, it is difficult to compare them in detail. Notably, in these studies level rove was applied to both the masker and the target tones, and task difficulty was manipulated by adjusting the variance of the associated rove distributions. The fact that the level of the target itself is uncertain in this paradigm could increase task complexity, perhaps by way of preventing the observer from forming an accurate template of the standard (no-signal) interval.

The effect of masker uncertainty on the processing of intensity information at the target frequency has also been studied in the context of the profile analysis paradigm (Green, 1988). Whereas masker level is unrelated to the presence of a signal in the ACI and informational masking paradigms discussed earlier, masker level is incorporated into the optimal strategy for detecting an increment in target level for a typical profile analysis task. Because both target and masker tones are roved together in this paradigm, synthetic listening incorporating stimulus components across frequency can improve thresholds. Introduction of stimulus uncertainty into a profile analysis paradigm interferes with the regular relationship between target and masker stimuli, and this in turn elevates threshold. This effect has been demonstrated for both frequency uncertainty (Richards *et al.*, 1989; Gockel and Colonius, 1997) and amplitude uncertainty (Kidd *et al.*, 1986; Lentz and Richards, 1998). Independently, roving profile components in either level or frequency elevates thresholds more than predicted based on optimal use of the information provided to the observer (Kidd *et al.*, 1986; 1988; Lentz and Richards, 1998; Richards and Zeng, 2001). This finding is consistent with the ACI result and with the informational masking studies discussed earlier: It demonstrates that observers tend to incorporate level information across frequency regardless of whether or not this is an optimal strategy.

The present series of experiments was designed to provide information about the conditions under which ACI occurs. One major motivation was to evaluate the hypothesis that ACI is largely driven by the operation of a synthetic mode of listening, wherein across-channel cues are combined. This was assessed via manipulation of stimulus parameters thought to modulate the degree of synthetic listening. Experiment 1 uses two stimulus segregation cues, onset asynchrony and envelope mismatches across frequency, to test the hypothesis that segregation cues which reduce synthetic listening can also reduce the magnitude of ACI. Experiment 2 introduces preinterval cues; this manipulation may be used to promote analytic listening by cuing observers to particular aspects of a stimulus. In addition to exploring aspects of ACI related to synthetic listening, another goal of this research was to explore the extent to which ACI may be influenced by energetic masking. Therefore, the third experiment estimates the contribution of energetic masking to previous ACI results.

II. EXPERIMENT 1

Fantini and Moore (1994) asked observers to detect an increment in level of a 60-dB SPL, 2-kHz pure tone. Thresholds rose by approximately 4 dB with inclusion of a set of roved-level maskers, defined as tones at 1.02, 1.43, 2.80, and 3.29 kHz, with masker levels randomly assigned without replacement from the set of 0, −7, −14, and −21 dB re: 60 dB SPL. Those authors noted that when the maskers were present, the dominant percept in the face of random changes in the masker amplitude was a change in overall pitch or timbre. This observation suggests that observers were not able to focus attention at the signal frequency to the exclusion of the masker. Therefore, it was hypothesized here that manipulations promoting analytic listening, such as onset asynchrony and incoherence of amplitude modulation across frequency, could improve intensity discrimination thresholds in the presence of remote maskers. Asynchronous onset has been shown to reduce informational masking (Neff, 1995), and to reduce across-channel effects, both those which elevate thresholds and those which reduce thresholds (e.g., Hall and Grose, 1991; Green and Dai, 1992; Grose and Hall, 1993). While there are fewer data on the grouping effects of amplitude modulated (AM) coherence as such in informational masking, it has been argued to play a significant role in stream segregation (Bregman, 1978) and in comodulation masking release (Grose and Hall, 1993).

A. Methods

1. Observers

Observers were six adults, ranging in age from 23 to 42 years (mean of 29.7 years). All had thresholds of 20 dB HL or less at octave frequencies 250–8000 Hz (ANSI, 1996), and none reported a history of chronic ear disease. All observers were practiced in psychoacoustical tasks at the outset of the experiment, having participated in at least one prior experiment unrelated to the current research.

2. Stimuli

Stimuli were made up of a target and two maskers. The target was centered at 1000 Hz and was either a pure tone (*steady*) or tone that had been AM. In a no-signal interval, the target was 60 dB SPL, and in a signal-present interval that level was elevated from baseline. The target was always 500 ms in duration, with 20-ms $\cos^2$ onset and offset ramps. Maskers were pairs of tones or AM tones at 500 and 2000 Hz. These tones were nominally 60 dB SPL, with a rove of ±10 dB (drawn from a uniform distribution) applied

on an interval-by-interval basis and determined independently for the two maskers. In synchronous gating conditions, maskers were gated on and off with the target, for a total duration of 500 ms, including 20-ms $\cos^2$ ramps. In the asynchronous gating conditions, maskers were gated on 500 ms prior to the target and gated off synchronously with the target, for a total duration of 1 s. Amplitude modulation of target and/or masker tones was achieved via multiplication with a raised 10-Hz sinusoid, with phase set to $-\pi/2$ at the beginning of stimulus onset. As such, all components receiving AM were coherently modulated, beginning in a modulation minimum, with 100% modulation depth. Conditions in which both target and masker components shared the same temporal envelope (either *steady* or *AM*) will be referred to as *matched* and those with different temporal envelopes as *unmatched*.

In a pair of supplemental conditions, maskers were always assigned a level of 70 dB (+10 dB re: 60-dB SPL standard level), the highest level possible in the roved condition. This *no-rove* manipulation was designed to eliminate effects due to amplitude uncertainty, with level set at the top of the rove range to measure fixed-level performance in the "worst case" of energetic masking. Envelopes were matched and gating was synchronous across target and masker components in the *no-rove* conditions.

All stimuli were generated in software (RPvds; TDT), played out of one channel of a DAC (RP2; TDT), routed through a headphone buffer (HB7; TDT), and presented to the left ear with circumaural headphones (Sennheiser, HD 265).

3. Procedures

Stimuli were presented in a two-alternative forced-choice procedure. In one interval the target component was 60 dB SPL, and in the other (randomly chosen) interval its level was greater than 60 dB SPL. Observers responded via a hand-held response box and received visual feedback indicating the correct response. The "signal-present" interval was generated by in-phase addition of a 1000-Hz pure tone or AM tone to the 60-dB target. The level of this added tone was adjusted according to a three-down, one-up tracking rule, estimating the signal level associated with 79% correct (Levitt, 1971). Initial signal level adjustments were made in steps of 8 dB, reduced to 4 dB after the second reversal, and reduced to 2 dB after the fourth reversal. Each track continued until ten reversals were obtained. Threshold estimates were computed as the average level at each track reversal after the first four. Between three and five replications were run in each condition. Thresholds improved by more than 10 dB in 2 out of 36 cases (6 observers × 6 conditions); in those cases the poor thresholds were omitted, leaving only the last three estimates. Thresholds were averaged to produce a final estimate. Data were obtained in blocks, completed in quasirandom order.

B. Results

Thresholds are reported in units of $10\log(\Delta I/I)$. While there were individual differences in sensitivity, both in masked and unmasked conditions, the trends discussed in the following were evident in all observers' data. Figure 1 shows mean thresholds, with signal condition indicated on the abscissa and masker condition indicated with symbols. Error bars indicate standard error of the mean associated with each estimate.

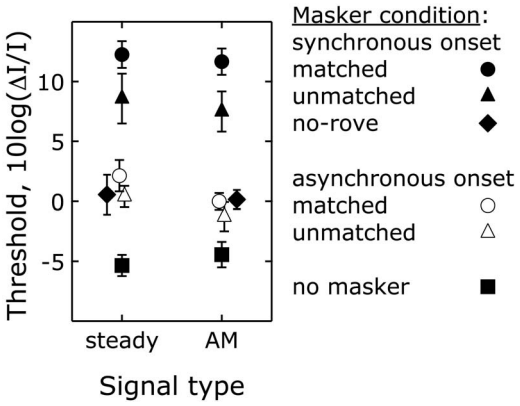


FIG. 1. Mean thresholds are plotted in units of $10\log(\Delta I/I)$ as a function of signal type, as indicated on the abscissa. Symbols specify the masker condition, as indicated in the legend. Error bars show one standard error of the mean across the six observers' data.

Thresholds in the absence of masker tones (closed squares) were similar for both the steady and the 10-Hz AM targets, with mean values of −5.4 and −4.4 dB, respectively. This is in good agreement with pure tone intensity discrimination thresholds for a 60-dB SPL standard reported in the literature, which span −5 to 0 dB in units of $10\log(\Delta I/I)$ (Viemeister, 1972; Penner *et al.*, 1974; Neff and Jesteadt, 1996). Thresholds rose with the introduction of synchronously gated roved-level maskers, as indicated by the closed symbols. The largest elevations in threshold were obtained for conditions where the target and maskers had matched envelopes (closed circles), with masking of 17.6 and 16.1 dB, respectively. Introduction of maskers with unmatched stimulus envelopes elevated thresholds more modestly, with masking of 13.9 dB for a steady signal and 11.9 dB for an AM signal (closed triangles). A similar pattern of thresholds was obtained for conditions in which the target and maskers were asynchronously gated, but with approximately 10-dB improved sensitivity overall, as indicated by open symbols. In the matched conditions, masking was 7.5 dB for the steady signal and 4.5 dB for the AM signal (open circles). Masking was reduced by 1–1.5 dB for the unmatched target/masker envelope condition (open triangles). These results suggest that manipulations designed to facilitate analytic listening improved performance, with larger effects of gating asynchrony (>10 dB) than envelope mismatch (~5 dB).

These observations were evaluated by a repeated measures analysis of variance (ANOVA), with two levels of SIGNAL (steady, AM), two levels of MATCH (matched, unmatched across target/masker envelope) and two levels of GATING (synchronous, asynchronous). This analysis resulted in a main effect of MATCH ($F_{1,5}=17.5, p<0.01$) and a main effect of GATING ($F_{1,5}=136.4, p<0.0001$). There was no main effect of SIGNAL ($F_{1,5}=4.1, p=0.10$), and

none of the interactions approached significance ($p>=0.25$). To assess the elevation in threshold under conditions of combined segregation cues, two paired t-tests were performed comparing each no-masker threshold with the associated masked threshold under conditions of envelope and gating segregation cues (i.e., the unmatched/asynchronous onset condition). Thresholds for the steady signal were significantly higher with asynchronously gated, AM maskers than in the absence of maskers ($t_5=6.89$, $p<0.001$ one-tailed). Likewise, thresholds for the AM signal were significantly higher with asynchronously gated, steady maskers than in the absence of maskers ($t_5=2.27$, $p<0.05$ one-tailed).

Results of the supplemental *no-rove* conditions employing the maximum value of rove on every trial are shown with closed diamonds in Fig. 1. Thresholds in these conditions were 0.56 and 0.15 dB for the matched steady and the matched AM stimuli, respectively. These thresholds are reduced relative to the associated roved-masker conditions by 11.7 dB ($t_5=10.3$, $p<0.0001$, two-tailed) and 11.5 dB ($t_5=24.6$, $p<0.0001$, two-tailed). However, they are also significantly different from the associated no-masker conditions by 5.9 dB ($t_5=3.83$, $p<0.05$, two-tailed) and 4.6 dB ($t_5=4.19$, $p<0.01$, two-tailed). This result suggests that masker amplitude uncertainty likely plays a dominant role in threshold elevation, but does not entirely account for the effects observed.

C. Discussion

The basic finding of ACI reported by Fantini and Moore (1994) was broadly replicated in Experiment 1. That is, inclusion of masker tones with independently roved level interfered with intensity discrimination at the target frequency even though these maskers were well removed from the signal in frequency. This effect was on the order of 15 dB, larger than the 4-dB effect noted by Fantini and Moore, but comparable to previous data on the effects of random amplitude perturbation in profile analysis (e.g., Lentz and Richards, 1998). The largest threshold elevation due to the presence of maskers was observed when the target and maskers were all steady pure tones or AM tones. This effect was reduced by approximately 4–5.5 dB when the target and maskers had unmatched envelopes. The introduction of target/masker onset asynchrony had a larger effect, reducing the ACI effect by approximately 10 dB. The effects of these manipulations combined reduced ACI by 13–14.5 dB but did not eliminate it, leaving 3–6.5 dB of masking. Both asynchronous onset and introduction of envelope mismatches across frequency are often discussed in the literature as facilitating sound source determination and analytic listening. The results obtained here are consistent with the interpretation that analytic processing reduces the ACI masking effect.

Thresholds in the supplemental conditions, with masker level consistently assigned as the maximum possible in roved conditions (70 dB SPL), were similar to those in the asynchronous gating conditions. This is consistent with the hypothesis that stimulus uncertainty played a large role in threshold elevation in the roved-level, synchronous onset conditions, and that asynchronous gating largely counteracted those effects. Spiegel *et al.* (1981) showed poorer intensity discrimination with the inclusion of fixed-level tonal maskers, a finding that is consistent with the current result. In contrast, Fantini and Moore (1994) report that thresholds improved slightly with the inclusion of fixed-level maskers. Significant differences across studies exist, but it is unclear which factors are responsible for the different results. The fact that the thresholds were elevated in the *no-rove* condition relative to the no-masker baseline in the present study suggests that amplitude uncertainty may not be the sole source of masking in these conditions.

It is frequently assumed that maskers an octave removed from the signal introduce essentially no energetic masking to the processing of that signal. Glasberg *et al.* (1984), for example, measured pure tone detection thresholds (as opposed to intensity discrimination) in the presence of pairs of masker tones up to 400 Hz above and below a 1-kHz signal. Auditory filters fitted to these data suggest that excitation associated with a masker tone at 500 Hz is attenuated by 40 dB in the auditory filter centered on 1 kHz. Based on these results, energetic masking would be negligible for the 60-dB, 1-kHz target, even at the highest masker level used in the current experiment: A 500-Hz masker tone at 70 dB SPL would change excitation at 1 kHz by approximately −30 dB in units of $10\log(\Delta I/I)$, and the change associated with a 2-kHz masker tone would be even less. These effects are well below thresholds measured experimentally. This line of reasoning does not rule out energetic masking for intensity discrimination, however. It is widely believed that intensity discrimination for a tone is based on cues distributed across the spectral range encompassing spread of excitation of that tone (e.g., Florentine and Buus, 1981). If some of those cues originate near the region of significant masker excitation, then even remote maskers could elevate thresholds via energetic masking. In contrast to this energetic masking explanation, it is also possible that fixed-level maskers may be associated with informational masking based on stimulus similarity attributes, as opposed to stimulus uncertainty attributes. There is some precedent for this hypothesis in the literature. Leibold and her colleagues (Leibold *et al.*, 2005; Leibold and Neff, 2007) report evidence of informational masking even in conditions of little or no external stimulus uncertainty. Experiment 3 will evaluate these two hypotheses regarding the small threshold elevation in the presence of fixed-level maskers. The next experiment focuses on the relatively large threshold elevation in the presence of roved maskers, which is susceptible to masking release based on segregation cues, and so is likely to be informational rather than energetic.

III. EXPERIMENT 2

Data collected in Experiment 1 demonstrated that remote masker tones significantly elevate thresholds for detecting an increment in target tone intensity, particularly in the presence of level rove and in the absence of target/masker segregation cues. The exact mechanism for the masker rove effect is unclear, however. One possibility is that masker

variability draws attention away from the (less variable) target stimulus. Another possibility is that masker variability produces a changing overall stimulus timbre or pitch, a factor that would interfere with performance if the increment detection cue were based on an obligatory synthesis of the target and masker tones.

Experiment 2 attempted to identify the mechanism of masking in ACI by presenting preinterval cues designed to reduce the possible sources of masking. For example, loss of focus on the target frequency with masker onset might be ameliorated by a cue to signal frequency, while masker level uncertainty should be greatly reduced by a preinterval cue indicating the masker tone levels associated with the subsequent interval. A similar cue-based approach was used by Richards and Neff (2004). In that study the task was to detect a tone in the presence of a multitone masker, with frequencies randomly assigned to each tone on each interval or on each trial. Informational masking was reduced by preinterval masker frequency cues, a result that can be interpreted in terms of reduced stimulus uncertainty in the listening interval. In some cases thresholds were also reduced by signal-frequency cues, even in the absence of signal frequency rove. This effect cannot be attributed to reduction in stimulus uncertainty, but may relate to allocation of attention to frequency-specific cues. The utility of different preinterval cues in ACI may shed light on the mechanisms of masking at work in the uncued case.

A. Methods

1. Observers

Nine observers completed this experiment, ranging in age from 24 to 50 years (mean of 33.7 years). All had thresholds of 20 dB HL or less at octave frequencies 250–8000 Hz (ANSI, 1996) in the test ear, and none reported a history of chronic ear disease. All had previously participated in a study of ACI, including Observers 1–5 from Experiment 1. Observers 7–10 had previously participated in ACI protocols not reported here.

2. Stimuli

Stimuli were identical to those used in Experiment 1 with the exception of the inclusion of preinterval cues. A cue interval was presented prior to each of the two listening intervals. Stimuli in both the cue and the listening interval were 500 ms in duration, including 20-ms $\cos^2$ onset and offset ramps. The cue/interval pairs were separated by a 300-ms delay, and there was a 500-ms delay between stimulus pairs.

There were five conditions. The first condition presented a cue interval consisting of a 500-ms silence, and so is referred to as the *no-cues* condition. In the *signal-standard* condition the cue was a 60-dB standard target tone, identical to that presented in no-signal intervals. In the *masker-freq* condition the cue consisted of the two masker tones played at their median level of 60 dB SPL. In the *masker-freq&lev* condition the cue consisted of the two masker tones played at the levels associated with rove values chosen for the subsequent listening interval. Finally, in the *full-standard* condition the cue was the sum of the 60-dB target tone and the pair of masker tones, with masker levels corresponding to the subsequent listening interval. In this condition, stimuli in the cue and listening intervals differed only when there was an intensity increment (i.e., on a signal-present interval) and were identical for no-signal intervals.

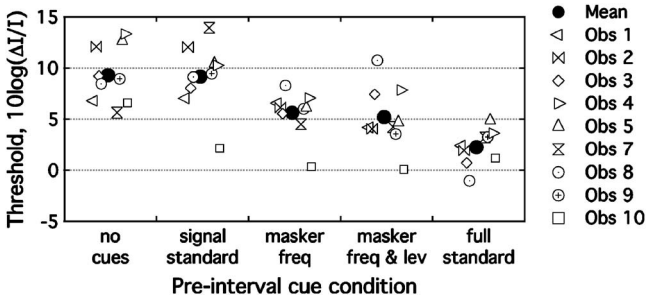


FIG. 2. Mean thresholds in the presence of a preinterval cue are plotted in units of $10\log(\Delta I/I)$ as a function of cue condition. Open symbols show mean thresholds for individual observers, and closed circles indicate the mean across observers.

3. Procedures

Observers completed the *no-cue* condition first; the four remaining conditions were then completed in random order, with all thresholds completed in blocks by condition. As in Experiment 1, thresholds were obtained in a two-alternative forced-choice, three-down one-up track. Other procedures were likewise identical, with the following exception. Because of variability in these data, additional steps were taken to obtain stable and reliable threshold estimates. Observers completed up to six replications in each condition, depending on the volatility of estimates and observer availability. In cases where more than four data points were available, the lowest and the highest estimates were omitted from mean data. This procedure reduced the across-observer variance but did not change the overall pattern of mean data across conditions.

B. Results

Figure 2 shows thresholds plotted as a function of cue condition for each of the nine observers. Open symbols indicate each individual observer's data, and closed circles show the mean across observers. These data were subjected to a repeated measures ANOVA, with five levels of CUE, as indicated on the abscissa of Fig. 2. This analysis resulted in a significant effect of CUE ($F_{4,32}=16.15, p<0.0001$). Mean threshold in the *signal-standard* cue condition was nearly identical to that in the *no-cues* condition, differing by just 0.1 dB ($t_8=0.11, p=0.91$). Planned comparisons for the series of masker cues indicate that each of the masker cue conditions aided performance compared to the no-cue baseline. Mean threshold improved by 3.5 dB with the introduction of masker tones in the *masker-freq* condition, dropped an additional 0.4 dB with introduction of masker level information in the *masker-freq&lev* condition, and improved by a further 3.0 dB in the *full-standard* condition. Each of these conditions was incrementally better than the last with a one-tailed t-test ($\alpha=0.05$) with the exception of introducing masker level information: Thresholds in the *masker-freq&lev*

condition were not significantly different from those in the *masker-freq* condition. Data of Observer 10 were somewhat different from the mean. This observer appeared to benefit from all four of the cues to a similar extent, including the *signal-standard* cue. Repeating the ANOVA without data from Observer 10 did not change the pattern of significance reported above.

The maximal informational masking (*no-cues*) condition was associated with a mean threshold of 9.3 dB, somewhat lower than the 12.2-dB threshold obtained under analogous conditions in Experiment 1. It is unclear whether to attribute this difference to individual differences, practice effects, or the longer interstimulus interval (from 500 ms in Experiment 1 to 800 ms in the current paradigm), but in any case this value provides substantial masking to compare against preinterval cue masking release conditions. The best-cued thresholds (*full standard*) were on average 2.3 dB (with a 1.8 dB s.d.). As such, these thresholds are still significantly greater than those expected in the absence of maskers (approximately −5.3 dB, based on results of Experiment 1) and comparable to those obtained with pure tone signal and maskers with asynchronous onset (2.1 dB).

C. Discussion

It was hypothesized prior to this experiment that if masker tones deflected attention from the signal frequency, then presenting the standard target prior to the listening interval could help focus attention on the signal frequency to the exclusion of remote masker tones. This was not borne out in the data, where no masking release was associated with the *signal-standard* condition, with the possible exception of Observer 10. While unexpected, this result is not without precedent in the literature; Richards and Neff (2004) reported a wide range of signal cue results across observers and across paradigms, including additional masking with inclusion of a signal cue in some cases.

Masker cues were predicted to improve performance to the extent that they reduced stimulus uncertainty in the listening interval. That is, cues to masker level were predicted to reduce masking, but cues comprised of masker tones presented at the median (60-dB) level were not predicted to impact performance. Again, this expectation was not borne out in the data, where both the *masker-freq* and *masker-freq&lev* conditions were associated with similar reductions in masking. This finding merits further investigation, but one possible explanation for this result is that presenting the maskers prior to the listening interval highlights the target as the "new" aspect of the stimulus. This idea is similar to the basis of the CoRE model (Lutfi, 1993), where the perception of a stimulus is dominated by stimulus features with the largest trial-to-trial variance. This result could also be related to auditory streaming (Bregman and Pinker, 1978).

While all masker-based cues significantly improved thresholds, the *full-standard* cue was the most effective cue. From an information theoretic perspective this is an odd result; the *full-standard* and *masker-freq&lev* conditions differ only in the inclusion of the 60-dB tone at 1 kHz, a stimulus feature that is constant across all trials and so does not add information. Better sensitivity in the *full-standard* condition suggests that observers may be unable to listen to the target analytically, basing their decision instead on the combination of target and masker tones. This might be the case if the percept associated with the three-tone complex were akin to a chord. In this case information about the masker tones alone would not convey information about the interaction of target and masker tones, and so might not be predictive of the overall percept. This hypothesis is consistent with the observation that the target tone alone was not an effective cue (*signal-standard*), but that the combination of the target and masker tones significantly improved performance over the case of masker tones alone (*masker-freq&lev* versus *full-standard*).

IV. EXPERIMENT 3

Experiments 1 and 2 showed that intensity discrimination for a 60-dB tone at 1 kHz can be significantly impaired by the presence of roved-level masker tones an octave above and an octave below that target, elevating thresholds as much as 15 dB relative to threshold for the same task in the absence of maskers. This masking can be significantly reduced by inclusion of segregation cues (Experiment 1) or by preinterval cues that foreshadow features of the masker or target/masker complex (Experiment 2). Thresholds in these conditions are not reduced to the baseline (no-masker) condition, however. Even in the presence of the most effective segregation cues or preinterval cues, thresholds are elevated by approximately 3 dB or more with respect to the baseline condition, similar to the masking obtained in the no-rove conditions of Experiment 1. One possible source of this masking is energetic. Intensity discrimination is widely believed to rely on off-frequency changes in excitation pattern (Florentine and Buus, 1981), including excitation an octave removed from the signal frequency (Viemeister, 1971); cues from these off-frequency regions could be energetically masked in the present paradigm.

Data from Moore and Raab (1974) lend credence to the idea that intensity discrimination at 1 kHz could be affected by energetic masking for tones an octave removed. Masking effects in that study were on the order of 5 dB, similar to the *no-rove* effect seen for steady tones in Experiment 1; however, a higher target standard level was used in the Moore and Raab study, and greater spread of excitation at high than low stimulus levels would be associated with a wider distribution of cues across frequency. The purpose of Experiment 3 was to assess the role of energetic masking in the ACI effect reported above. Three strategies were used to estimate the role of energetic masking independent of informational masking. First, intensity discrimination thresholds were measured at three fixed masker levels—the bottom, middle, and top of the rove range. It was hypothesized that if thresholds with the fixed-level masker in Experiment 1 were due to energetic masking, then those thresholds should be sensitive to increases in masker level. Second, thresholds were also measured in the presence of a pair of 400-Hz wide noise bands configured to produce energetic masking comparable to that of the tonal maskers. If intensity discrimination in the

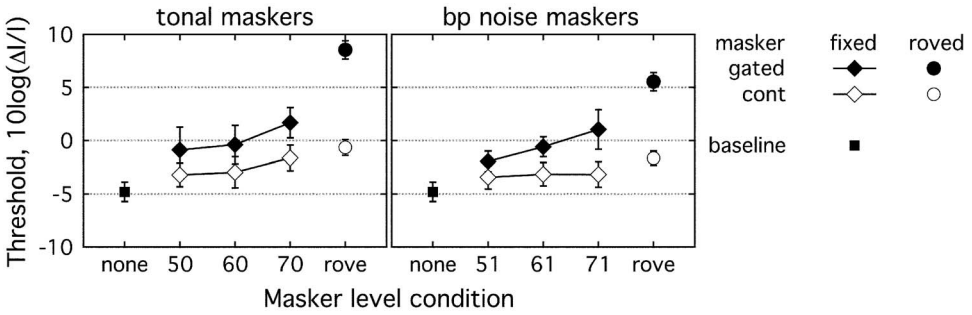


FIG. 3. Mean thresholds are plotted in units of $10\log(\Delta I/I)$ as a function of masker level condition. Symbols specify masker gating, and error bars indicate one standard error of the mean across the eight observers' data. In all cases the standard target stimulus was a 60-dB tone at 1000 Hz. In the left panel the masker was a pair of tones at 500 and 2000 Hz, and in the right panel maskers were a pair of 400-Hz wide bands of noise centered on 400 and 2100 Hz.

presence of fixed-level tonal maskers is due to similarity-based informational masking, then introducing a qualitative mismatch across target and maskers (tone versus noise band) should reduce masking. Third, stimulus onset synchrony was manipulated to either discourage or facilitate analytic listening. Supplemental conditions also explored the effects of contralateral masker presentation, where no energetic masking would occur.

A. Methods

1. Observers

Observers were eight adults, ages 18–54 years (mean 32 years). All had thresholds of 20 dB HL or less at octave frequencies 250–8000 Hz (ANSI, 1996), and none reported a history of chronic ear disease. All observers were practiced in psychoacoustical tasks at the outset of the experiment, having participated in at least one prior experiment unrelated to the current research. In addition, Observer 1 had previously completed both Experiments 1 and 2, Observer 7 had completed just Experiment 2, and a third observer had previously completed another ACI protocol.

2. Stimuli

Stimuli were based on those used in Experiment 1. The observer's task was to detect an increment in the level of a 1000-Hz tone above the 60-dB SPL standard level. This target was 500 ms in duration, including 20-ms $\cos^2$ ramps. In the tonal masker conditions, the masker was a pair of pure tones at 500 and 2000 Hz; masker levels were either fixed at 50, 60, or 70 dB SPL, or the level was roved independently on an interval-by-interval basis using a pair of uniform draws from a distribution 50–70 dB. In the narrow-band noise masker conditions the masker was a pair of 400-Hz wide bands of noise centered on 400 and 2100 Hz; masker levels were either fixed at 51, 61, or 71 dB SPL, or the level of each band was roved uniformly over this span. Excitation pattern simulations[1] suggested that these narrow-band masker frequencies and levels produce comparable excitation in the region of the signal frequency as that associated with the tonal maskers. In gated conditions the two maskers were gated on and off with the target, and in the continuous conditions they played throughout a threshold estimation track, effectively introducing a target/masker onset asynchrony. When the masker level was roved in the continuous presentation conditions it was adjusted in the interstimulus interval; the transition was smoothed via convolution with a 20-ms boxcar function.

3. Procedures

As in Experiment 1, thresholds were obtained in a two-alternative forced-choice procedure, with a 500-ms duration interstimulus interval. Signal level was adjusted in a three-down one-up track, with track parameters identical to those described earlier. Observers completed three threshold estimates, with a fourth estimate taken in cases where prior estimates varied by 3 dB or more. Data reported in the following are the mean of all estimates obtained. Observers completed all narrow-band noise conditions prior to beginning the tonal masker conditions. Additional data were collected at the end of the experiment to spot check for practice effects. In no case was there sufficient evidence of improvement to prompt replacement of data.

B. Results

The pattern of results was consistent across the eight observers, so only the mean data are shown. Figure 3 shows mean thresholds plotted as a function of condition for tonal masker conditions (left panel) and narrow-band noise conditions (right panel). As in Fig. 1, squares indicate baseline performance of intensity discrimination in the absence of maskers. Diamonds indicate performance with fixed-level maskers and circles show roved-masker data; in both cases synchronous gating data are shown with closed symbols, and continuous conditions are shown with open symbols. Error bars show standard error of the mean, and in some cases error bars are occluded by the symbols.

In the course of this experiment baseline performance for intensity discrimination in the absence of maskers was estimated twice, once at the beginning of the narrow-band masker conditions and again prior to tonal masker conditions, with mean thresholds of −5.5 and −4.1 dB, respectively. These estimates were not significantly different ($t_7=2.0, p=0.08$), so the mean of −4.8 dB is plotted in both panels for comparison with the masked thresholds.

Data collected in the fixed masker conditions were submitted to a repeated measures ANOVA, with two levels of MASKER (tone, noise), three levels of LEVEL (low, mid, high), and two levels of SYNCHRONY (gated, continuous). There was a main effect of LEVEL ($F_{2,14}=30.54, p<0.0001$) and a main effect of SYNCHRONY ($F_{1,7}=16.53, p<0.005$). There was no main effect of MASKER ($F_{1,7}=1.20, p=0.31$). None of the interactions with MASKER approached significance ($p>=0.5$), but the interaction between LEVEL and SYNCHRONY approached significance ($F_{2,14}=2.70, p=0.10$). These results support the

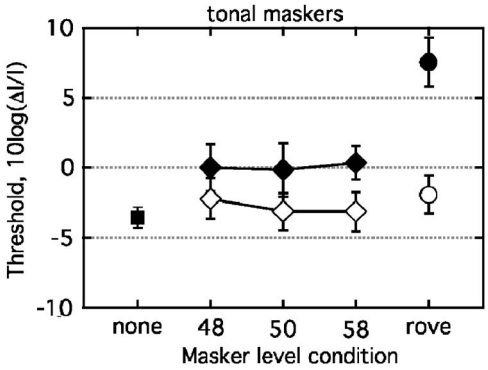


FIG. 4. Mean thresholds are plotted in units of $10\log(\Delta I/I)$ as a function of masker level condition for the supplemental conditions using reduced masker level and wider masker spacing. Plotting conventions follow those of Fig. 3. The standard target was a 50-dB SPL tone at 948.7 Hz, and the masker was a pair of tones at 300 and 3000 Hz.

conclusion that masker level was positively related to threshold, as would be expected if masker tones introduced energetic masking. Playing the maskers continuously reduced threshold, and there was a nonsignificant trend for a larger gating effect at the high masker levels.

Comparisons between fixed-level and roved data are somewhat complicated by the effect of level within the fixed-level conditions. Using the maximum fixed-level threshold as a reference, however, provides a liberal estimate of the energetic masking present in the roved conditions. The difference in thresholds between fixed and roved level conditions is greater for the gated than the continuous presentation modes, with differences on the order of 4.5–7 and 1.5 dB, respectively. The roved level data were submitted to a repeated measures ANOVA, with two levels of MASKER (tone, noise) and two levels of SYNCHRONY (gated, continuous). There was a main effect of MASKER ($F_{1,7}=18.76, p<0.005$) and a main effect of SYNCHRONY ($F_{1,7}=186.34, p<0.0001$). The interaction fell short of significance ($F_{1,7}=3.45, p=0.11$).

These results suggest that energetic masking was likely small but significant for these stimuli; supplemental conditions were run to see if ACI could be demonstrated under conditions associated with no evidence of energetic masking. These conditions employed lower levels (50 dB, with rove range of ±8 dB) and wider masker spacing (300 and 3000 Hz). The standard was a 948.7-Hz tone, geometrically centered between the two masker tones. Results of the supplemental conditions are shown in Fig. 4. Baseline intensity discrimination in the absence of maskers for the 50-dB, 948.7-Hz standard was compared with the mean from the previous two conditions using a 60-dB, 1-kHz standard frequency (with means of −3.6 and −4.8 dB, respectively). This difference was significant ($t_7=3.28, p<0.05$), indicating slightly reduced sensitivity for increments to the 50-dB standard as compared to the 60-dB standard employed previously. A repeated measures ANOVA was performed for the fixed level maskers with three levels of LEVEL (low, mid, high) and two levels of SYNCHRONY (gated, continuous). There was a main effect of SYNCHRONY ($F_{1,7}=29.0, p<0.001$), but no effect of LEVEL ($F_{2,14}=0.86, p=0.44$) and no interaction ($F_{2,14}=0.43, p=0.65$). Because there was no significant effect of level, thresholds for the continuous presentation were averaged across the three fixed-level conditions. That mean (−2.8 dB) was not significantly different from the no-masker baseline ($t_7=1.09; p=0.31$), but it was significantly lower than the associated roved condition ($t_7=2.47, p<0.05$), an effect of only 0.9 dB. The comparable comparison in gated conditions resulted in a 3.6-dB effect of introducing fixed-level maskers ($t_7=3.70, p<0.01$) and a further 7.5-dB effect of introducing masker level rove ($t_7=8.27, p<0.001$).

In addition to the monaural stimulus presentation used up to this point, thresholds in supplemental conditions were also obtained in the presence of roved-level masker tones presented contralateral to the target tone. This condition was completed several weeks after the previously reported conditions, at which point one of the eight observers was no longer available for testing. Contralateral presentation tended to improve thresholds relative to the roved-level ipsilateral data presented earlier, with mean improvement of 1.3 dB in the gated condition and 0.5 dB in the continuous condition. Paired one-tailed t-tests comparing these thresholds with thresholds in the no-masker condition indicated significant contralateral masking for the gated ($t_6=2.73, p<0.05$) but not the continuous ($t_6=0.70, p=0.25$) masker presentation.

C. Discussion

At the outset of this experiment it was hypothesized that if fixed-level masking was energetic in nature, then thresholds should increase with increasing masker level. Thresholds in the two conditions employing a 60-dB target were found to increase with increasing masker level, and there was some indication that this effect may be greater for gated than continuous masker presentation. It is sometimes argued that gating effects obtained with long duration signals reflect informational rather than just energetic masking (as in Neff, 1995). It has also been suggested that attention bands for the detection of a ~300-ms tone are more sharply tuned in frequency when the masker is continuous than when it is gated (Dai and Buus, 1991; Wright and Dai, 1994). This effect has been described in terms of the masker onset capturing the observer's attention and introducing a bias to monitor a family of auditory filters rather the just the optimal filter(s). If the masker onset in the current paradigm broadens or otherwise modifies spectral weighting of intensity cues, this could introduce threshold elevation independent of energetic masking. As such, higher thresholds in gated as compared to continuous conditions could be interpreted as a form of informational masking.

Recently Jesteadt *et al.* (2007) reported that thresholds across a range of paradigms could be fitted using the excitation-based loudness model of Moore *et al.* (1997).[2] In one portion of that study, intensity discrimination thresholds of Viemeister (1972) were fitted using a criterion change in partial loudness of 4 phons. Predictions were quite accurate over a range of stimulus levels for a 950-Hz pure tone, gated with 160-ms duration, but thresholds were underpredicted in conditions incorporating high-pass noise. Stimuli in the pri-

mary no-masker and fixed-level conditions were (with a 1000–Hz signal) submitted to this model. Using a 4-phon criterion, the predicted threshold in the no-masker condition is −7.4 dB, somewhat lower than the −4.8 dB obtained in the present study. Increasing the criterion change in partial loudness to 8 phons, as used in the modeling of Leibold and Jesteadt (2007), increases predicted threshold to −5.0 dB. Including fixed-level, 70-dB SPL masker tones increased thresholds by 1.7 dB, to −3.3 dB. The same prediction is made using the 71-dB SPL noise bands. This predicted masking is similar to the threshold elevation of 1.6 dB observed with narrow-band noise maskers and is within the confidence interval (±2 sem) of the 3.2-dB effect observed with tonal maskers. These results lend support to the hypothesis that threshold elevation in the fixed-masker, continuous conditions reflects energetic masking. Because this model does not make use of temporal cues it is not feasible to model the different gating conditions in the context of the model. One parsimonious explanation for the data, however, is that thresholds in the continuous condition represent the effects of energetic masking and those in the gated conditions reflect additional informational masking due to attentional capture associated with masker onset.

The partial loudness model also provides a framework for thinking about the detection process in roved-level conditions. In broad terms, this model is based on loudness as a function of frequency, similar to an excitation pattern. This function is computed for a masker alone and then again for a signal-plus-masker stimulus. The difference in loudness is computed and that difference is integrated across frequency. This process assumes that the observer has some internal representation of the masker alone stimulus that serves as a template. In the fixed-level conditions the masker alone is presented frequently—once in every 2AFC trial. In the roved-level conditions the masker alone reference is changing on every interval. When the masker is playing continuously and the target is gated on during the listening interval the observer can use the masker fringe, occurring after the change in level and before onset of the signal, as the basis for a masker alone template. Thresholds in these conditions are only slightly elevated relative to those in the fixed-level conditions, suggesting that information provided by the fringe is only slightly less helpful than fixing the masker level. In the gated conditions the observer is never presented with an example of the masker-alone stimulus associated with a particular listening interval. In this case, the only strategy remaining would be to use features of the full stimulus—including maskers, target and possibly the signal—to form a template. Theoretically all the information necessary to do this task at the limits of energetic masking are present; for example, knowing that the masker is always a pair of tones at 500 and 2000 Hz, a "perfect" template could be computed based on the excitation at 500 and 2000 Hz. The fact that thresholds are more severely elevated in the gated roved-level conditions suggests that this is an error-prone and inaccurate process, which might be limited by memory or stimulus-driven attentional capture (Egeth and Yantis, 1997).

A simulation of the roved level, tonal masker conditions was undertaken to estimate thresholds from energetic masking alone for the primary roved-level conditions of Experiment 3, assuming that the observer is able to construct an accurate no-signal template as described earlier. The MATLAB script used to collect psychophysical thresholds was adapted to calculate "responses" based on partial loudness. On each interval a single stimulus was generated, with independent values of rove selected for each masker tone. The partial loudness associated with addition of a signal tone was calculated. If that value exceeded the 8-phon criterion then the procedure correctly identified the signal-present interval, but if not then the procedure randomly selected either the signal-present or the no-signal interval. This process was repeated for 50 track reversals, and three such tracks were completed. This procedure predicts a mean masked threshold of approximately −4.0 dB and no-masker threshold −5.3 dB. This informal simulation suggests that energetic masking with the introduction of roved-level maskers may elevate thresholds by about 1 to 2 dB, substantially less than the 11-dB masking effect obtained psychophysically in the gated condition. By exclusion, the remaining ~10 dB can be categorized as informational masking.

The supplemental data collected with a 50-dB standard and more widely spaced tonal maskers (at 300 and 3000 Hz) resembled those in the primary conditions, but thresholds in the fixed-level conditions did not increase with increases in masker level. The partial loudness model predicts no energetic masking in these conditions; threshold is predicted to be constant across the no-masker and all three fixed-masker conditions. While the mean fixed-masker thresholds are 0.8 dB higher than those in the no-masker condition, this difference is small and nonsignificant, suggesting that energetic masking does not have an appreciable effect in these conditions. As in the previous data, gating the fixed-level stimuli elevated thresholds by 2.2–3.5 dB, and roving level elevated thresholds substantially, particularly in the gated condition. These results demonstrate that gating and roved-level effects occur even in the absence of energetic masking. The finding of ACI for gated maskers presented contralateral to the target tone further confirms that energetic masking is not a precondition for demonstrating this effect. Similar findings were reported by Shub *et al.* (2005), who argued that their results could be modeled in terms of a binaural intensity summation model.

V. GENERAL CONCLUSIONS

The results of Experiment 1 showed that the masking effect of across-channel interference (ACI) reported by Fantini and Moore (1994) can be reliably obtained with either steady or amplitude modulated tones. Intensity discrimination thresholds for a 1000-Hz, 60-dB SPL standard were increased substantially with inclusion of masker tones at 500 and 2000 Hz, played at 60 dB SPL ±10 dB. Conditions incorporating unmatched envelope patterns across frequency were associated with a reduction in the ACI effect. Asynchronous target/masker onset reduced thresholds to a greater extent, and the combination of envelope and onset manipulations was associated with a combined release from masking. These results are consistent with the hypothesis that ACI

is reduced under stimulus conditions facilitating analytic listening. Thresholds in conditions of onset asynchrony were comparable to those with synchronous onset and masker level set consistently at the top of the rove range. Eliminating masker level uncertainty did not eliminate masking, leaving approximately 6 dB of masking in the tonal masker conditions that cannot be accounted for by masker level uncertainty.

Experiment 2 showed that preinterval cues to signal frequency and standard level were ineffective at lowering threshold, suggesting that memory for the target frequency does not limit performance in the presence of masker tones. Preinterval cues incorporating the masker tones were effective in reducing threshold, even when the cue consisted of tones at 60 dB SPL rather than the random levels of the subsequent listening interval. The most effective cue incorporated both masker and target tones, foreshadowing the subsequent stimulus in all respects other than the presence of an intensity increment at the target frequency associated with a signal interval. Even in these conditions there was some evidence of residual masking: Average thresholds were on average 7 dB above those measured in Experiment 1 in the absence of masker tones.

Experiment 3 was designed to assess the possible contribution of energetic masking in the ACI effect, particularly that portion of the effect that cannot be eliminated with stimulus features promoting analytic listening or with cuing. Results of this study are consistent with the conclusion that energetic masking is responsible for approximately 1 to 2 dB of masking for the stimuli used in Experiments 1 and 2. Tonal roved-level maskers were slightly more effective than matched bandpass noise maskers, suggesting that similarity-based informational masking may play some role in ACI. Additional conditions using a lower stimulus level and wider masker spacing resulted in essentially no evidence of energetic masking, but significant threshold elevation of 7.5 dB under roved-level, gated masker conditions. This effect persisted even when the masker tones were presented contralateral to the target tone. These results suggest that the ACI effect does not depend on energetic masking.

Taken together, the results presented here suggest that observers have difficulty ignoring roved level maskers under conditions favoring synthetic listening, such as when stimulus components share a common onset and coherent temporal envelope. Spiegel *et al.* (1981) interpreted analogous findings under conditions of masker frequency uncertainty as suggesting that observers attend to the profile of intensity as a function of frequency. This tendency to judge the intensity of a target relative to adjacent masker tones is quite beneficial under some listening conditions (Green, 1988). This beneficial effect can be reduced or eliminated with stimulus manipulations promoting analytic listening, such as introduction of envelope mismatches across frequency (Green and Nguyen, 1988) or asynchronous onset (Green and Dai, 1992). The ACI effects described in the current study may reflect the same across-channel processes, but under conditions where such processes are not advantageous.

ACKNOWLEDGMENTS

This work was supported by a grant from the NIH NIDCD (No. RO1 DC007391). Thanks are due to Joe Hall, John Grose, Lori Leibold, Andy Oxenham, Walt Jesteadt, and an anonymous reviewer for helpful comments on this work.

[1]Excitation patterns were based on Moore *et al.* (1997). Software for making these calculations (excite2005.exe) is available for download from: http://hearing.psychol.cam.ac.uk/Demos/demos.html.

[2]In that study loudness was calculated using the software partloud.exe, available for download from: http://hearing.psychol.cam.ac.uk/Demos/demos.html.

ANSI. (**1996**). *ANSI S3-1996, American National Standards Specification for Audiometers* (American National Standards Institute, New York).

Bregman, A. S. (**1978**). "The formation of auditory streams," in *Attention and Performance* (Erlbaum, Hilldale, NJ).

Bregman, A. S., and Pinker, S. (**1978**). "Auditory streaming and the building of timbre," Can. J. Psychol. **32**, 19–31.

Cohen, M. F., and Schubert, E. D. (**1987**). "The effect of cross-spectrum correlation on the detectability of a noise band," J. Acoust. Soc. Am. **81**, 721–723.

Dai, H. P., and Buus, S. (**1991**). "Effect of gating the masker on frequency-selective listening," J. Acoust. Soc. Am. **89**, 1816–1818.

Doherty, K. A., and Lutfi, R. A. (**1999**). "Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **105**, 1831–1840.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368–379.

Egeth, H. E., and Yantis, S. (**1997**). "Visual attention: Control, representation, and time course," Annu. Rev. Psychol. **48**, 269–297.

Fantini, D. A., and Moore, B. C. (**1994**). "A comparison of the effectiveness of across-channel cues available in comodulation masking release and profile analysis tasks," J. Acoust. Soc. Am. **96**, 3451–3462.

Florentine, M., and Buus, S. (**1981**). "An excitation-pattern model for intensity discrimination," J. Acoust. Soc. Am. **70**, 1646–1654.

Glasberg, B. R., Moore, B. C., and Nimmo-Smith, I. (**1984**). "Comparison of auditory filter shapes derived with three different maskers," J. Acoust. Soc. Am. **75**, 536–544.

Gockel, H., and Colonius, H. (**1997**). "Auditory profile analysis: Is there perceptual constancy for spectral shape for stimuli roved in frequency?," J. Acoust. Soc. Am. **102**, 2311–2315.

Green, D. M. (**1988**). *Profile Analysis: Auditory Intensity Discrimination* (Oxford University Press, New York).

Green, D. M., and Dai, H. (**1992**). "Temporal relations in profile comparisons," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner (Pregamon, Oxford), pp. 471–477.

Green, D. M., and Nguyen, Q. T. (**1988**). "Profile analysis: Detecting dynamic spectral changes," Hear. Res. **32**, 147–163.

Grose, J. H., and Hall, J. W. (**1993**). "Comodulation masking release: Is comodulation sufficient?," J. Acoust. Soc. Am. **93**, 2896–2902.

Hall, J. W., and Grose, J. H. (**1990**). "Comodulation masking release and auditory grouping," J. Acoust. Soc. Am. **88**, 119–125.

Hall, J. W., and Grose, J. H. (**1991**). "Some effects of auditory grouping factors on modulation detection interference (MDI)," J. Acoust. Soc. Am. **90**, 3028–3035.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.

Jesteadt, W., Tan, H., Khaddam, S., and Leibold, L. J. (**2007**). "Prediction of behavioral thresholds using a model of partial loudness," paper presented at the 30th Midwinter Research Meeting of the Association for Research in Otolaryngology.

Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Kidd, G., Jr., Mason, C. R., and Green, D. M. (**1986**). "Auditory profile analysis of irregular sound spectra," J. Acoust. Soc. Am. **79**, 1045–1053.

Kidd, G., Jr., Mason, C. R., and Hanna, T. E. (**1988**). "Evidence for sensory-trace comparisons in spectral shape discrimination," J. Acoust. Soc. Am. **84**, 144–149.
Leibold, L. J., and Jesteadt, W. (**2007**). "Use of perceptual weights to test a model of loudness summation," J. Acoust. Soc. Am. **122**, EL69-EL73.
Leibold, L. J., and Neff, D. L. (**2007**). "Effects of masker-spectral variability and masker fringes in children and adults," J. Acoust. Soc. Am. **121**, 3666–3676.
Leibold, L. J., Neff, D. L., and Jesteadt, W. (**2005**). "Effects of reduced spectral uncertainty and masker fringes with multi-tonal maskers," J. Acoust. Soc. Am. **118**, 1893–1894.
Lentz, J. J., and Richards, V. M. (**1998**). "The effects of amplitude perturbation and increasing numbers of components in profile analysis," J. Acoust. Soc. Am. **103**, 535–541.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Lutfi, R. A. (**1993**). "A model of auditory pattern analysis based on component-relative-entropy," J. Acoust. Soc. Am. **94**, 748–758.
McFadden, D. (**1987**). "Comodulation detection differences using noise-band signals," J. Acoust. Soc. Am. **81**, 1519–1527.
Moore, B. C., and Borrill, S. J. (**2002**). "Tests of a within-channel account of comodulation detection differences," J. Acoust. Soc. Am. **112**, 2099–2109.
Moore, B. C., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness and partial loudness," J. Audio Eng. Soc. **45**, 224–240.
Moore, B. C., Glasberg, B. R., and Schooneveldt, G. P. (**1990**). "Across-channel masking and comodulation masking release," J. Acoust. Soc. Am. **87**, 1683–1694.
Moore, B. C., and Raab, D. H. (**1974**). "Pure-tone intensity discrimination: Some experiments relating to the near-miss' to Weber's law," J. Acoust. Soc. Am. **55**, 1049–1054.
Neff, D. L. (**1995**). "Signal properties that reduce masking by simultaneous, random-frequency maskers," J. Acoust. Soc. Am. **98**, 1909–1920.
Neff, D. L., and Callaghan, B. P. (**1988**). "Effective properties of multicomponent simultaneous maskers under conditions of uncertainty," J. Acoust. Soc. Am. **83**, 1833–1838.
Neff, D. L., and Dethlefs, T. M. (**1995**). "Individual differences in simultaneous masking with random-frequency, multicomponent maskers," J. Acoust. Soc. Am. **98**, 125–134.
Neff, D. L., and Jesteadt, W. (**1996**). "Intensity discrimination in the presence of random-frequency, multicomponent maskers and broadband noise," J. Acoust. Soc. Am. **100**, 2289–2298.
Oh, E. L., and Lutfi, R. A. (**1998**). "Nonmonotonicity of informational masking," J. Acoust. Soc. Am. **104**, 3489–3499.
Oxenham, A. J., and Plack, C. J. (**1998**). "Suppression and the upward spread of masking," J. Acoust. Soc. Am. **104**, 3500–3510.
Penner, M. J., Leshowitz, E., Cudahy, E., and Ricard, G. (**1974**). "Intensity discrimination for pulsed sinusoids of various frequencies," Percept. Psychophys. **15**, 568–570.
Richards, V. M., and Neff, D. L. (**2004**). "Cuing effects for informational masking," J. Acoust. Soc. Am. **115**, 289–300.
Richards, V. M., Onsan, Z. A., and Green, D. M. (**1989**). "Auditory profile analysis: Potential pitch cues," Hear. Res. **39**, 27–36.
Richards, V. M., and Zeng, T. (**2001**). "Informational masking in profile analysis: Comparing ideal and human observers," J. Assoc. Res. Otolaryngol. **2**, 189–198.
Shub, D. E., Pogal-Sussman, T., and Colburn, H. S. (**2005**). "The effects of distractor frequency on monaural intensity discrimination under monotic and dichotic conditions," paper presented at the Association for Research in Otolaryngology, New Orleans, LA.
Spiegel, M. F., Picardi, M. C., and Green, D. M. (**1981**). "Signal and masker uncertainty in intensity discrimination," J. Acoust. Soc. Am. **70**, 1015–1019.
Stellmack, M. A., Willihnganz, M. S., Wightman, F. L., and Lutfi, R. A. (**1997**). "Spectral weights in level discrimination by preschool children: Analytic listening conditions," J. Acoust. Soc. Am. **101**, 2811–2821.
Viemeister, N. F. (**1971**). "Intensity discrimination of pulsed sinusoids: The effects of filtered noise," J. Acoust. Soc. Am. **51**, 1265–1269.
Viemeister, N. F. (**1972**). "Intensity discrimination of pulsed sinusoids: The effects of filtered noise," J. Acoust. Soc. Am. **51**, 1265–1269.
Wright, B. A., and Dai, H. (**1994**). "Detection of unexpected tones in gated and continuous maskers," J. Acoust. Soc. Am. **95**, 939–948.

Across-channel interference in intensity discrimination: The role of practice and listening strategy

Emily Buss[a)]
Department of Otolaryngology/Head and Neck Surgery, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599

(Received 20 April 2007; revised 15 October 2007; accepted 29 October 2007)

Pure tone intensity discrimination thresholds can be elevated by the introduction of remote maskers with roved level. This effect is on the order of 10 dB [$10\log(\Delta I/I)$] in some conditions and can be demonstrated under conditions of little or no energetic masking. The current study examined the effect of practice and observer strategy on this phenomenon. Experiment 1 included observers who had no formal experience with intensity discrimination and provided training over 6 h on a single masked intensity discrimination task to assess learning effects. Thresholds fell with practice for most observers, with significant improvements in six out of eight cases. Despite these improvements significant masking remained in all cases. The second experiment assessed trial-by-trial effects of roved masker level. Conditional probability of a "signal-present" response as a function of the rove value assigned to each of the two masker tones indicates fundamental differences among observers' processing strategies, even after 6 h of practice. The variability in error patterns across practiced listeners suggests that observers approach the task differently, though this variability does not appear to be related to sensitivity. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816569]

PACS number(s): 43.66.Dc, 43.66.Fe, 43.66.Ba [RAL] Pages: 265–272

I. INTRODUCTION

In energetic masking, the presence of a masker is thought to corrupt the neural encoding of the signal, thereby limiting further processing of that signal. In contrast, in informational masking the signal is thought to be well represented at the periphery, but it is assumed that the central auditory system is not able to make optimal use of that peripheral information. This effect is frequently studied in the context of pure tone detection in the presence of masker tones with randomly selected frequencies (e.g., Kidd *et al.*, 1994; Neff and Dethlefs, 1995). Introducing a masker level rove in this paradigm has little or no additional effect on pure tone detection threshold provided those maskers are sufficiently remote from the signal frequency to minimize energetic masking (Neff and Callaghan, 1988; Oh and Lutfi, 1998). Masker level uncertainty is associated with substantial informational masking for masked intensity discrimination, however (Buss, 2008; Doherty and Lutfi, 1999; Fantini and Moore, 1994; Stellmack *et al.*, 1997). For example, a recent study by Buss (2008) showed that intensity discrimination threshold of a 50-dB SPL standard tone at 948.7 Hz was elevated by approximately 10 dB ($10\log(\Delta I/I)$) with the inclusion of masker tones at 300 and 3000 Hz, roved in level on each interval (50 dB SPL±8 dB). This effect was shown to be independent of energetic masking, and thus (by exclusion) was attributed solely to informational masking. Following the convention of Fantini and Moore (1994), this effect will be referred to as across-channel interference (ACI).

Previous work on ACI has shown that segregation cues improve thresholds (Buss, 2007). In one manipulation the target and the maskers were either pure tones or tones that had been amplitude modulated via multiplication with a raised 10-Hz sinusoid; thresholds were highest when envelopes were coherent across all three tones and fell when the target and maskers had mismatched envelopes. Gating the masker on prior to the onset of the target also reduced thresholds. These effects were interpreted as showing that segregation cues can facilitate analytic listening, but that in the absence of these cues observers were adopting a synthetic listening strategy, incorporating information about the masker tone level into the discrimination decision despite the fact that this strategy reduces sensitivity. There was a small but significant threshold elevation in conditions with fixed-level remote maskers which could likewise be reduced by segregation cues, suggesting that some synthesis of information across frequency took place even under conditions of minimal uncertainty.

The current studies sought to more carefully characterize the practice effects and underlying perceptual processes associated with ACI. It is commonly assumed that informational masking reflects more than just inattention or confusion regarding the psychoacoustic task (e.g., Durlach *et al.*, 2003). In the context of ACI, this distinction would differentiate between a percept lacking robust cues to a target level independent of masker level (informational masking), as opposed to confusion regarding which of a set of robust cues are most predictive of a signal-present interval (task confusion). Experiment 1 tested observers who were naïve with respect to ACI stimuli on a sequence of six 1-h sessions to assess whether task confusion played a role in the initial results. One goal of the present study was to determine whether extended practice would allow development of an analytic listening strategy in the absence of stimulus cues associated with segregation. If observers are able to volun-

[a)]Electronic mail: ebuss@med.unc.edu

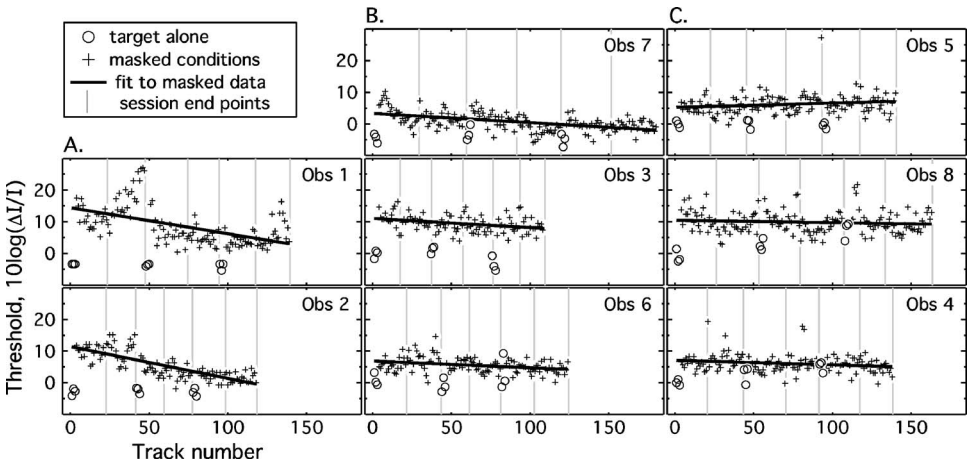


FIG. 1. Thresholds are plotted in units of $10\log(\Delta I/I)$ as a function of trial number for each observer in Experiment 1. Estimates of thresholds for intensity discrimination in quiet are indicated with open circles, while those in the presence of maskers are indicated with plus signs. The dashed vertical lines segmenting each panel indicate the beginning and end of each of six test sessions. Panels are grouped by magnitude of the practice effect: data in column A show robust practice effects, those in column B show more modest improvements, and data in column C show the smallest effects.

tarily adopt an analytic processing strategy that minimizes the contribution of the masker tones, then it seems likely that this would be evident within a 6-h series of practice trials. The second experiment characterized the listening strategy adopted after extensive practice in terms of the probability of a "signal-present" response conditional on the level of each masker tone. This approach will discriminate among a family of processing strategies that could elevate thresholds in these conditions.

II. EXPERIMENT 1: CHANGES IN ACI WITH PRACTICE

Informational masking for tone detection under conditions of masker frequency uncertainty has been shown to be reduced with practice for some listeners (Neff and Callaghan, 1988; Neff and Dethlefs, 1995). Studies of this training effect indicate that most benefits are obtained in the first 600 trials (Neff and Callaghan, 1988). The question considered here is whether practice has a comparable effect for intensity discrimination in the presence of roved-level maskers, or whether extended practice with feedback eliminates the effect. Asymptotic performance on ACI is of practical importance in determining how much training to provide prior to data collection in future studies. More importantly, it is also of theoretical significance; the implications of the ACI effect would be quite different if it were due merely to task confusion regarding what aspects of the percept are the most reliable indicators of a signal-present interval, as opposed to reflecting immutable limitations of the percept.

A. Methods

1. Observers

Observers were eight adults, from 17 to 50 years old (mean of 27.9 years). All had thresholds of 20 dB HL or less at octave frequencies 250–8000 Hz. (ANSI, 1996), and none reported a history of chronic ear disease. About half of these observers had previously participated in psychoacoustical experiments, and none had prior experience listening to ACI or other stimuli designed to assess informational masking.

2. Stimuli

The target was a 948.7-Hz pure tone. This component had a standard level of 50 dB SPL, and the task was to detect an increment in this level. The maskers, when present, were synchronously gated tones at 300 and 3000 Hz. The level of each masker was roved independently based on draws from a uniform distribution spanning 42–58 dB SPL. Both target and masker tones were 220 ms in duration, including 10 ms $\cos^2$ onset and offset ramps, and the interstimulus interval was 500 ms. All stimuli were generated in software (RPVDs; TDT), played out of one channel of a digital-analog converter (RP2; TDT), routed through a headphone buffer (HB7; TDT), and presented to the left ear with circumaural headphones (Sennheiser, HD 265).

3. Procedures

Thresholds were estimated by way of a two-alternative forced-choice, three-down one-up track estimating the 79% correct point (Levitt, 1971). Target level increments were made by in-phase addition of a pure tone at the target frequency of 948.7 Hz. At the outset of the track, the level of this tone was adjusted in steps of 4 dB, and this step was reduced to 2 dB after the second track reversal. The track continued until a total of eight reversals was obtained, and the threshold was computed as the average level at the last six reversals. Lights on a handheld response box indicated listening intervals and provided feedback. At the outset of each track the signal level was 10 dB above the most recent threshold (or anticipated threshold, in the case of the first track). Observers provided data over 3 weeks, listening in a total of six 1-h sessions.

On days 1, 3, and 5 the test session began with three consecutive threshold estimates in the no-masker condition (i.e., intensity discrimination in quiet). During the remainder of these sessions observers ran sequential blocks of the masker-present condition. On days 2, 4, and 6 observers just ran sequential blocks of the masker-present condition. Observers were encouraged to complete as many tracks as possible during a 1-h session, with a 5–10 min break offered at the midpoint of each session.

B. Results

Thresholds as a function of trial number are plotted in Fig. 1, with each observer's data in a separate panel. Thresholds in the no-masker condition are shown with open circles, and those in the masker-present condition with pluses. Grey vertical lines indicate break points between data collection sessions, and thick black lines show the line fits to the masker-present conditions, as described below. Observers completed a total of 109–180 tracks, including the nine no-masker tracks. On average, each track included 50 trials, so the total number of trials completed by each observer ranged from approximately 5500 to 9000.

Three estimates of intensity discrimination in quiet were obtained on three occasions for each observer: averaging across test session and across observers, the mean threshold was −0.6 dB. A repeated measures analysis of variance (ANOVA) with three levels of TIME (days 1, 3, and 5) indicated no difference in thresholds as a function of measurement time point ($F_{2,14}=0.82, p=0.46$). In contrast, for the masker-present conditions mean thresholds improved from 8.1 dB on day 1 to 5.1 dB on day 6.

Data in the masker-present condition were fit with a line to characterize the change in threshold as a function of trial number. While some of the data would be better fitted with a more complex function, these fits do seem to capture the general trends of interest. Inspection of Fig. 1 suggests that there are marked individual differences in the extent to which practice reduces thresholds in the masker-present condition. The data in column A show evidence of robust practice effects in the masker-present condition. For Obs 1 and Obs 2, thresholds changed by −0.08 and −0.10 dB/trial, respectively ($p<0.00001$). Fits to their data estimate about 11.5 dB improvement over the course of training in both cases, though closer inspection of the line fits suggests that this may be an overestimate, as thresholds appear to have asymptoted prior to the end of day 6. The data in column B show more modest evidence of practice effects. For these observers, improvement was −0.02 to −0.03 dB/trial ($p<0.0005$), and line fits estimate improvement in thresholds from 2.6 to 5.3 dB. Data in column C show the weakest evidence of improvement as a function of trial number. The data of Obs 4 are consistent with a modest threshold improvement, with the line fit estimating a significant slope of −0.01 dB/trial ($p<0.05$) and threshold reduction on the order of 2 dB. The line fitted to the data of Obs 8 was not significantly different from a slope of zero ($p=0.21$). The data of Obs 5 were fitted with a line indicating worsening in performances over time, on the order of 2 dB over all trials, with a slope of 0.01 dB/trial ($p<0.05$); this result was qualitatively unchanged when data were refitted omitting the single high threshold at the end of day 4.

Among the observers who showed improvement with practice, the smallest ACI effect at asymptote was estimated as 2.2 dB for Obs 7. The significance of this effect was assessed by way of a single-sample *t*-test assessing whether the mean of the no-masker thresholds is lower than the best threshold predicted by the line fitted to the masker-present data (−1.95 dB). This test indicated a significant difference ($t_8=3.29, p<0.05$, two-tailed). The significance of ACI was tested in a similar manner for the remaining seven observers and found to be significant in all cases ($p<0.05$). This result supports the conclusion that the ACI was significant after 6 h of practice for all observers.

C. Discussion

Results of these extended practice trials suggest that practice with ACI stimuli improves performance for most observers. Of the eight observers tested, two showed improvement on the order of 10 dB, while others made more modest gains or failed to benefit from training. Visual inspection of Fig. 1 suggests that those observers who did show marked improvement made the fastest gains in the first 2–3 days of training. By the end of day 3 most observers had completed between 3000 and 4000 trials. This period of improvement is longer than the 600 trials reported by Neff and her colleagues (Neff and Callaghan, 1988; Neff and Dethlefs, 1995) as the point at which most observers had reached asymptotic performance for detection of a tone in the presence of remote maskers of uncertain frequency and amplitude.

Mean thresholds collected on day 1 indicate 8 dB of ACI, similar to the approximately 10 dB of ACI under comparable stimulus conditions of Buss (2008). That study employed stimuli with the same frequencies and levels as used here, but with a longer duration signal (500 ms instead of 220 ms). That study concluded that the roved masker ACI was functionally free from energetic masking and so could be attributed solely to informational masking. It was argued that observers processed these stimuli synthetically, weighting masker level information despite the fact that these components do not convey task-related information. In this sense ACI resembles profile analysis under conditions where the across-frequency cue has been corrupted by independent perturbations of masker tones (Kidd *et al.*, 1986). The results presented here suggest that while the mean effect size may decrease with training, it is not eliminated with 6 h of practice. The reduction in ACI with training could be interpreted as an increase in the degree to which the observer can voluntarily engage in analytic listening.

Thresholds in quiet did not show any signs of improvement over the course of this experiment. In fact, mean thresholds rose over the course of the study for two listeners (Obs 4 and Obs 8). While this result may have been due to failure to sustain attention or motivation over the many hours of listening, it could also represent a shift in the strategy. Such an effect may be related to the group effects noted by Green and Mason (1985) when comparing intensity discrimination in naïve observers and observers with extensive training in profile analysis. In that study observers who had previously practiced in profile analysis tended to perform more poorly on intensity discrimination in quiet than those observers who had not received such training; while this difference could have been due to training effects, the authors also noted that selection criteria for previous profile analysis studies could have identified groups with different sensitivity prior to stimulus exposure. In contrast to profile analysis, a

synthetic listening strategy based on across-frequency level comparisons is markedly nonoptimal for the masked intensity discrimination task considered here. However, hours of exposure to ACI stimuli may have biased Obs 4 and 8 to adopt a listening strategy similar to that suited to profile analysis, which could in turn adversely affect thresholds in quiet.

III. EXPERIMENT 2: ERROR PATTERNS

The second experiment sought to investigate the underlying perceptual factors associated with ACI by estimating contributions of the low- and high-frequency masker tones. The paradigm used here shares some features with conditional on single sample (COSS) analysis (Berg, 1989), where weights describing the combination of information are derived based on the relationship between random variability in some aspect of the stimulus and the probability of a signal-present response. This general approach has been used in previous studies of informational masking (Doherty and Lutfi, 1999; Neff and Odgaard, 2004; Stellmack *et al.*, 1997). The model underlying the COSS analysis assumes that independent information is combined linearly across weighted channels. In contrast, the current paradigm restricted stimulus variability (in this case, rove) to a family of five levels and computed the conditional probability of a "signal-present" response for each of 25 possible combinations of rove (5 low- $\times$ 5 high-frequency masker tone levels). Reporting the probability for all possible combinations of rove has the advantage that interactions of low and high masker rove values can be assessed. This approach resembles that used by Dye *et al.* (1994) to assess synthetic versus analytic listening in a binaural task.

Whereas the previous studies of informational masking cited above have reported significant individual differences in weighting strategies, the present study was undertaken to test specific predictions regarding the processing underlying ACI. Pure tone intensity discrimination has been shown to make use of information distributed across the auditory filters stimulated by that tone (e.g., Viemeister, 1972), so even in the absence of maskers this task is based weighted information across frequency. Buss (2008) modeled intensity discrimination in the presence of fixed-level maskers based on the change in partial loudness (Moore *et al.*, 1997) with the addition of a signal tone; while this approach did a reasonable job of predicting thresholds for fixed-level maskers, it severely underestimated thresholds in the presence of roved-level maskers. Failure of the model to account for thresholds in the presence of a pair of roved-level masker tones could be attributed to a poor internal representation of the excitation pattern associated with the no-signal stimulus in the observer's decision process. An inaccurate representation of the no-signal stimulus could result in stimulus energy associated with a masker tone being mistaken for energy associated with a signal. Such a mechanism would result in a positive correlation between masker level and probability of a "signal-present" response.

Alternatively, Buss (2008) hypothesized that intensity discrimination in the presence of roved-level maskers could involve obligatory synthetic processing similar to that underlying profile analysis. In the profile analysis paradigm the relative levels of tones distributed across frequency provide a very potent cue to the presence of a signal; in the typical paradigm, the signal can be defined as an increment in level of one tone (the target) relative to a family of flanking tones. If ACI is based on a similar type of processing, this might be reflected in a higher probability of a "signal-present" response when both maskers are at a low level relative to the target tone, with reduced probability of a "signal-present" response as the maskers increase in level relative to the target. Such a strategy might be based on relative levels of target and masker tones for a suprathreshold signal where, on average, energy at the target frequency exceeds that at either masker frequency during the signal-present listening interval. If an observer adopted this strategy, the probability of a "signal-present" response would be negatively correlated with masker tone level.

A. Methods

1. Observers

Observers were seven normal-hearing adults, between 19 and 50 years old (mean of 37 years). All met the inclusion criteria of the previous experiment, and all had participated in a study of ACI prior to this experiment. Observers 1–5 had previously completed Experiment 1. Observers 9 and 10 had prior experience in another ACI study not included in the current report; that study also spanned approximately six 1-h sessions and included a range of ACI conditions.

2. Stimuli

As in Experiment 1, the target was a tone at 948.7 Hz and maskers were tones at 300 and 3000 Hz. The task was to detect an increment in the 50-dB standard level of the target. In contrast with the previous study, maskers were independently roved with uniform draws from a restricted set of possible levels: including −8, −4, 0, +4, and +8 dB, re: 50 dB SPL. Thus, there were 25 possible combinations of low- and high-frequency masker levels.

3. Procedures

Stimuli were presented in a two-alternative, forced-choice paradigm, with a three-down one-up stepping rule estimating 79% correct (Levitt, 1971). As in Experiment 2, initial signal level adjustments for the tone added to the target were made in steps of 4 dB, reduced to 2 dB after the second track reversal. In contrast to Experiment 1, the track continued for a total of 12 track reversals, and the final threshold estimate was computed as the mean level at the last ten track reversals; this relatively large number of track reversals was employed to increase the number of trials using a threshold-level signal. The rove value assigned to each masker in each interval and the observer's response were recorded after each trial. This information was saved to disk for later analysis. All testing was performed in one condition,

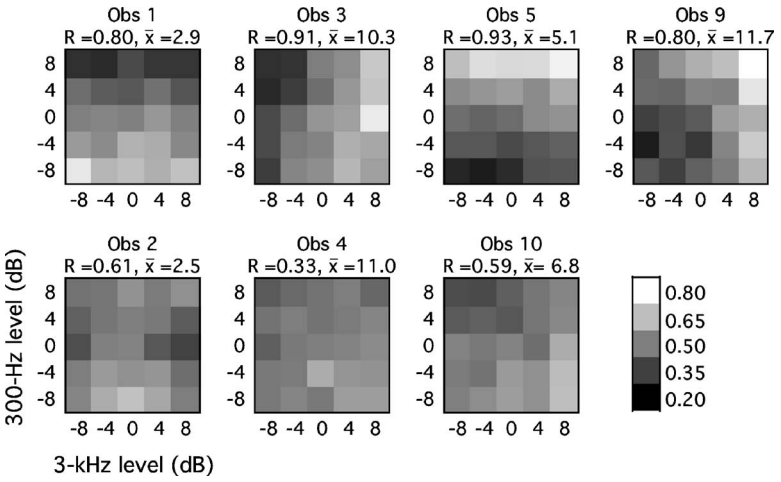


FIG. 2. Results for Experiment 2 are plotted separately for each observer. Panels show probability of a "signal-present" response for each of 25 combinations of masker level, with rove assignment of the low-frequency masker indicated on the ordinate and that of the high-frequency masker on the abscissa. Shading indicates conditional probability of a "signal-present" response, as indicated in the key.

with masker tones present. Three 1-h sessions were completed within a two-week span.

B. Results

Over the course of this experiment observers completed between 39 and 58 threshold estimation tracks. Mean thresholds for all observers spanned 2.5–11.0 dB, with an across-observer mean of 7.2 dB. These thresholds are comparable to the those obtained in Experiment 1 using rove values drawn from a continuous uniform distribution spanning ±8 dB, re: 50 dB SPL, supporting the assumption that limiting masker rove values did not substantially change the task.

The relationship between a "signal-present" response and each value of masker rove was assessed based on the record of trial-by-trial stimulus characteristics for each observer. For each threshold track the point associated with the second track reversal was identified. Data prior to this point were excluded from further analysis, as the signal itself was likely to have dominated responses at the beginning of the track. Information from the remaining trials was used to compute two matrices: (1) the total number of times each of the 25 possible masker rove configurations was presented, considering stimuli in both intervals of each trial, and (2) the total number of times each configuration was identified by the observer as containing the signal, regardless of whether the observer response was correct. Dividing the second matrix by the first produces the probability of a "signal-present" response for each of the 25 masker rove combinations. Using this procedure for the current data set, each cell in the matrix was based on between 135 and 281 data points for each observer.

Stability of the probability matrices was assessed in the following manner. Each observer's data were reanalyzed in two blocks, one based on the first half of trials and the other based on the second half of trials, and the correlation between the resulting pair of matrices was computed. Confidence intervals for these correlations were calculated with a Bonferroni correction for multiple comparisons ($n=7$): the criterion for significance (one-tailed) is $r=0.48$ for $\alpha=0.05$. Correlations for individual observer's data ranged from $r=0.93$ to 0.33. Only Obs 4's data failed to reach the criterion of significance, indicating that the probability matrices reported here capture trends that are consistent over the course of the experiment for all but one observer.

Figure 2 shows the conditional probabilities for each of the 25 masker rove configurations for each observer. Shading in each cell represents the probability of a "signal-present" response, with lighter shading indicating higher probability, as indicated in the key. Because this is a two-alternative task, values above 0.5 reflect positive weights, values below 0.5 reflect negative weights, and values near 0.5 can be interpreted as weights at or near zero. The text above each panel indicates the associated split-half correlation, as well as the mean threshold across trials associated with each data set. Recall that masker level was not predictive of the presence of a signal, so the pattern of dependencies shown here does not reflect any aspect of optimal processing.

The top row of panels shows data for which a nonuniform pattern of probabilities was evident based on both visual inspection and on high values of split-half correlation. Observers 5 and 9 tended to select intervals with high values of masker rove as the signal interval, as evidenced by the white shading in the upper right of each panel. This pattern of results is consistent with the hypothesis that ACI is due to overly broad spectral integration of level information around the target frequency, such that the masker energy is mistaken for spread of excitation due to the presence of a signal. In contrast, data of Obs 1 show the opposite trend, with low values of masker rove associated with the highest probability of a "signal-present" response. This pattern is consistent with a strategy based on a spectral profile, with the target tone higher in level than the masker tones. Observers do not seem to attend equally to the low- and high-frequency masker tones. The results of Obs 3 and 9 appear to be dominated by the level of the 3-kHz masker tone, as evidenced by the vertical trends in the data, while those of Obs 1 and 5 appear to be more influenced by the 300-Hz masker tone, as evidenced by the horizontal trends in the data.

The bottom row of panels in Fig. 2 show results of Obs 2, 4, and 10; these data tend towards more uniform probabilities as a function of masker level. Such a pattern might be obtained if the decision process was unaffected by masker tone levels. If that were the case, then one might expect

thresholds to be lower for this group than in the group with less uniform weighting. Comparison of mean thresholds for the two groups does not bear this out, however. As noted above each panel in Fig. 2, thresholds for both groups include examples of relatively good performance and relatively poor performance (2.9–11.7 dB in the group with variable weights versus 2.5–11.0 dB in the group with more uniform weights).

C. Discussion

The probability of a "signal-present" response conditional on the level of the low- and high-frequency masker tone rove values suggests that some observers were consistently incorporating masker level into their processing strategy despite the fact that this information is not predictive of the correct response. The consistency of this nonoptimal weighting does not appear to be related to sensitivity; those observers with relatively uniform weights across stimulus components were no more sensitive to intensity increments at the target frequency than those observers who incorporated the masker level in a reliable way. A likely explanation of uniform probability in this group is that masker tone level was weighted inconsistently, due to volatility of strategy over the course of the experiment or due to a strategy that cannot be captured in terms of fixed weights. This finding is consistent with the report of Lutfi *et al.* (2003) which showed individual differences in the extent to which a fixed-weight model was able to predict the form of a psychometric function for tone detection in the presence of frequency- and level-roved masker tones. The wide range of individual differences for thresholds obtained here is also consistent with previous data on the contribution of informational masking components to intensity discrimination in the presence of roved-level maskers (Doherty and Lutfi, 1999).

Only one observer's data (Obs 1) resembled the pattern predicted based on the analogy between ACI and profile analysis. For this observer, the probability of a "signal-present" response was negatively correlated with masker level. Such a pattern might result from the observer monitoring the level of the tone at the target frequency relative to the level of both maskers and responding "signal-present" when the target exceeds the masker level. Other observers clearly did not employ this strategy, including two (Obs 5 and 9) who appeared to respond based on the opposite rule—"signal-present" if the level of the target tone is below that of the masker tones. In the typical profile analysis paradigm a "signal-present" interval is cued by a target level which exceeds that of the masker tones. However, spectral profile discrimination has also been demonstrated for other profile features, including a decrement in the target component (Ellermeier, 1996). Analogously, it is possible observers in the present ACI task could be "listening for" a profile characterized by a relatively low target tone level or some other relationship across tones. It is unclear how such a strategy would originate given the statistics of the stimuli used here. However, given that *any* weighting of the masker tones is nonoptimal in this task, it would not be too surprising for the pattern of weights to be unrelated to the statistics of the stimuli.

TABLE I. Two-fit weight estimates for the low (300-Hz) and high (3000-Hz) masker tones. These values can be interpreted as the change in probability of a "signal present" response for a 1-dB change in masker level. The final column shows the correlation between the probability matrix computed based on data and that reconstructed based on the pair of weights.

	Low frequency	High frequency	Correlation
Obs 1	−0.0198	−0.0009	0.96
Obs 2	−0.0049	−0.0008	0.42
Obs 3	−0.0039	0.0194	0.92
Obs 4	−0.0046	0.0041	0.79
Obs 5	0.0229	0.0050	0.97
Obs 9	0.0099	0.0187	0.90
Obs 10	−0.0086	0.0074	0.89

The results obtained here can also be compared to those of Neff and Odgaard (2004). In that study, frequency discrimination was measured in the presence of maskers composed of tones with roved frequency, and observers were shown to put more weight on low-frequency masker tones than high-frequency masker tones. In the current study some of the observers' responses were correlated more strongly with the low- than the high-frequency masker level, with either positive or negative correlation (Obs 5 and Obs 1, respectively). This was not the case for all observers, however.

One potential advantage of representing conditional probabilities of a signal-present response for each roved masker level is that this method can capture nonlinear interactions between low- and high-frequency maskers. For example, if large level discrepancies between the two maskers were associated with increased probability of a "signal present" response, this would be reflected in dark shading along the negative diagonal, and would not be modeled well in terms of the linear combination of weights from the two maskers. Data for each observer were reanalyzed to determine a weight for each masker tone based on the correlation between rove level and the probability of a "signal-present" response collapsed across all values of the opposing masker tone. Weights associated with each masker are reported separately for each observer in Table I. A matrix of conditional probabilities for all 25 combinations of rove values was then estimated based on these two weights. This estimate closely resembled the original matrix in most cases. This was quantified by computing the correlation between the original probability matrix and the two-weight matrix. These correlations are comparable to or higher than the split-half correlations reported in Fig. 2 in all cases but one. For Obs 2 the split-half correlation was higher than the correlation between the original and the two-fit matrices ($r=0.61$ and $r=0.42$, respectively), raising the possibility that nonlinear interactions could play a role in this observer's results. In order to provide additional insight into the results of this observer, the associated data from Fig. 2 are also shown in table form (Table II). The marginal means shown here give some indication of the response contingencies characterizing the strategy used by this observer: the probability of a "signal-

TABLE II. Conditional probabilities associated with the data of Obs 2 are shown as a function of the level of the 300- and 3-kHz masker tone, reported in dB re: 50 dB SPL. The final column and the final row show the associated mean probabilities.

	300-Hz level					
300-Hz level	−8	−4	0	4	8	Mean
8	0.47	0.48	0.55	0.49	0.54	*(0.51)*
4	0.43	0.48	0.50	0.49	0.42	*(0.46)*
0	0.39	0.51	0.51	0.40	0.36	*(0.43)*
−4	0.50	0.57	0.54	0.55	0.46	*(0.52)*
−8	0.52	0.61	0.65	0.59	0.50	*(0.57)*
Mean	*(0.46)*	*(0.53)*	*(0.55)*	*(0.50)*	*(0.46)*	

present" response was elevated at the extremes of the 300-Hz component rove range and in the middle of the 3-kHz component rove range. This pattern suggests that Obs 2 was using a cue based on the magnitude of deviations from the 50-dB standard level rather than absolute component levels. This evidence of a nonlinear effect of masker level for the data of Obs 2 stands in contrast to the good fits achieved with the two-weight linear fit for the remaining observers' data.

IV. CONCLUSIONS

Experiment 1 showed that the ACI effect can be reduced substantially with practice. Significant improvements were obtained in six of the eight observers tested, but ACI was still significantly greater than zero for all observers after 6 h of practice. The time course of training appeared to be prolonged relative to previous reports of training effects in studies of tone detection in the presence of an informational masker (Neff and Dethlefs, 1995). The pattern of improvement observed here is perhaps more consistent with the report of Kidd *et al.* (1986) which documented a prolonged period of practice for profile analysis, with threshold improvement out to 3000 trials. If the ACI effect is due to a synthetic listening strategy, then the current results suggest that this strategy is adaptable for some listeners, but that even extended practice does not equip an observer to adopt a wholly analytic listening strategy.

In Experiment 2 the probability of a "signal-present" response was computed for each of a family of possible low- and high-frequency masker rove values. The results indicated a range of weighting strategies, including both positive and negative correlation between masker level and "signal-present" response. As such, some observers' data were consistent with overly broad spectral integration of the level cue and others with a cue based on greater energy at the target than masker frequencies (positive and negative correlations, respectively). There were also individual differences in degree to which the low- or the high-frequency maskers contributed to performances. There appears to be a great deal of latitude in the exact mechanism by which information is combined across frequency, including whether a "signal-present" response is positively or negatively correlated with masker level. Most of the data were modeled accurately by assuming a linear combination of independent weights applied to the two masker tones, as previously assumed, suggesting that the masker effects combine linearly for most but not all observers.

One interpretation of the present data is that ACI is due to different mechanisms: broadly integrating spectral level cues in some listeners and an across-channel comparison akin to spectral profile analysis in others. Alternatively, observers could be comparing levels across frequency in the manner of a spectral profile task in all cases, but with different expectations regarding the spectral profile associated with a signal-present interval. On average, addition of a signal tone elevates energy at the target as compared to the masker frequencies, so it was hypothesized *a priori* that the probability of a "signal-present" response would be negatively correlated with masker level. That is, it was expected that observers would be listening for a target tone exceeding the level of the maskers. However, individual observers could form spurious expectations regarding the relationship between tones in the presence of a signal, perhaps based on early experience or inherent bias. It is unclear how such expectations would come about, but given that the best strategy in the ACI task is to ignore masker level, any profile strategy is in some sense spurious. The range of response patterns obtained is consistent with the hypothesis that synthetic listening is somewhat obligatory for the current stimuli, but that the across-channel profile associated with the signal interval appears to be arbitrary across observers.

ACKNOWLEDGMENTS

This work was supported by a grant from the NIH NIDCD (Grant No. RO1 DC007391). Thanks are due to Robert Lutfi, Walt Jesteadt, Lori Leibold, and Joseph Hall for helpful comments and discussion of this material.

[1]Weights were estimated based on the probability of a "signal present" response in the following manner. The matrix associated with each observer's data can be defined as $P_{i,j}$, where subscripts indicate the index associated with the low (i) and high (j) frequency masker tone levels. The levels associated with both i and j dimensions are defined as: $X=[-8,-4,0,4,8]$. The contribution of each masker tone alone was quantified by averaging probabilities across either i or j dimensions and subtracting 0.5 (chance performance for the 2AFC task). Weights for the low (W_L) and high (W_H) frequency maskers were then defined as the slope of the line fitted to these adjusted averages as a function of X. Probability matrices for the combined effects of W_L and W_H were estimated based on these weights using the following procedure, where a is the contribution of

high-frequency maskers, b is the contribution of low-frequency maskers, and Q is the combination, incremented by 0.5 for comparison with the data.

$$a=[W_H(X(1)\ldots X(n))]*\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix},$$

$$b=\begin{bmatrix}W_L(X(1))\\ \vdots \\ W_L(X(n))\end{bmatrix}*[1\ldots 1]$$

$$Q=0.5+a+b$$

ANSI. (**1996**). *ANSI S3-1996, American National Standards Specification for Audiometers* (American National Standards Institute, New York).

Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks," J. Acoust. Soc. Am. **86**, 1743–1746.

Buss, E. (**2008**). "The effect of masker level uncertainty on intensity discrimination," J. Acoust. Soc. Am. **123**, in press.

Doherty, K. A., and Lutfi, R. A. (**1999**). "Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **105**, 1831–1840.

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.

Dye, R. H., Jr., Yost, W. A., Stellmack, M. A., and Sheft, S. (**1994**). "Stimulus classification procedure for assessing the extent to which binaural processing is spectrally analytic or synthetic," J. Acoust. Soc. Am. **96**, 2720–2730.

Ellermeier, W. (**1996**). "Detectability of increments and decrements in spectral profiles," J. Acoust. Soc. Am. **99**, 3119–3125.

Fantini, D. A., and Moore, B. C. (**1994**). "A comparison of the effectiveness of across-channel cues available in comodulation masking release and profile analysis tasks," J. Acoust. Soc. Am. **96**, 3451–3462.

Green, D. M., and Mason, C. R. (**1985**). "Auditory profile analysis: frequency, phase, and Weber's law," J. Acoust. Soc. Am. **77**, 1155–1161.

Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Kidd, G., Jr., Mason, C. R., and Green, D. M. (**1986**). "Auditory profile analysis of irregular sound spectra," J. Acoust. Soc. Am. **79**, 1045–1053.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Lutfi, R. A., Kistler, D. J., Callahan, M. R., and Wightman, F. L. (**2003**). "Psychometric functions for informational masking," J. Acoust. Soc. Am. **114**, 3273–3282.

Moore, B. C., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness and partial loudness," J. Audio Eng. Soc. **45**, 224–240.

Neff, D. L., and Callaghan, B. P. (**1988**). "Effective properties of multicomponent simultaneous maskers under conditions of uncertainty," J. Acoust. Soc. Am. **83**, 1833–1838.

Neff, D. L., and Dethlefs, T. M. (**1995**). "Individual differences in simultaneous masking with random-frequency, multicomponent maskers," J. Acoust. Soc. Am. **98**, 125–134.

Neff, D. L., and Odgaard, E. C. (**2004**). "Sample discrimination of frequency differences with distracters," J. Acoust. Soc. Am. **116**, 3051–3061.

Oh, E. L., and Lutfi, R. A. (**1998**). "Nonmonotonicity of informational masking," J. Acoust. Soc. Am. **104**, 3489–3499.

Stellmack, M. A., Willihnganz, M. S., Wightman, F. L., and Lutfi, R. A. (**1997**). "Spectral weights in level discrimination by preschool children: Analytic listening conditions," J. Acoust. Soc. Am. **101**, 2811–2821.

Viemeister, N. F. (**1972**). "Intensity discrimination of pulsed sinusoids: the effects of filtered noise," J. Acoust. Soc. Am. **51**, 1265–1269.

Nonconscious control of fundamental voice frequency

Honorata Zofia Hafke[a)]
Institute of Acoustics, Adam Mickiewicz University, Umultowska 85, 61-114 Poznan, Poland

(Received 4 December 2006; revised 19 September 2007; accepted 2 November 2007)

The aim of this paper is to answer the question whether "perception-action" dissociation, which is well documented in vision, may also be found in auditory information processing. Trained singers were asked to produce vowel sounds into a microphone. The sound that each singer produced was fed back to their ears via headphones. Two seconds after the sound production had begun, the auditory feedback was shifted in pitch by a certain degree (9, 19, 50, or 99 cents in either direction). In every set of sounds, instances without any pitch shifts also appeared. After each trial, participants reported whether they were aware of a pitch change or not. It was found that even though the participants were unaware of subtle pitch changes, the fundamental frequency of their vowel production was found to shift slightly in the opposite direction to the pitch shift. These results show that auditory information is processed by two separate systems: one for perception and one for action. They also show that the function of the auditory control system differs from the visual control system. The latter is used to control bodily movements while the function of the former is a nonconscious, instant control of vocalization. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2817357]

PACS number(s): 43.66.Hg, 43.70.Bk [DD] Pages: 273–278

I. INTRODUCTION

Milner and Goodale (1995) postulate two separate visual systems in the human brain. Evidence from studies of both humans and other primates has shown that there is a distinction between vision for perception and vision for action, which is reflected in the organization of the visual pathways in the cerebral cortex of primates (Aglioti *et al.*, 1995; Goodale *et al.*, 1991; Goodale and Milner, 1992; Goodale and Milner, 2004; Haffenden *et al.*, 2001; Kroliczak *et al.*, 2006). Each stream uses visual information in a different way. The ventral stream transforms visual information into perceptual representations of objects and their relations to each other. Such representations enable us to identify objects and attach various properties to them. In contrast, the transformations carried out by the dorsal stream provide moment-to-moment information concerning the location and disposition of objects, with respect to the effector being used, and thereby mediate the visual control of skilled actions directed toward those objects. In recent years, researchers have attempted to find a similar dissociation between action and perception in human audition (Repp, 2000, 2005). They tried to find an exact auditory analogue for the visual system which controls action. The idea that audition can mimic vision in controlling motor actions such as grasping or reaching objects is difficult to defend. Therefore, if we want to retain the general idea of two separate auditory systems: one for conscious perception and the other for nonconscious control of action, first we have to find out what kind of control could be effectively performed by the system that is responsible for the nonconscious processing of auditory information. The present author assumed that audition is mainly used to control different forms of vocal action. The experiments were designed to test if this control is performed consciously or nonconsciously. If the experiments show that vocal utterances are controlled nonconsciously, then the hypothesis that there are two separate auditory systems will gain solid support.

Activities whose motor reactions are controlled by audition are for the most part based on the presence of acoustic feedback information. This information is picked up from sounds we generate ourselves, including our own voice, sounds from musical instruments, and sounds produced by other forms of our activity. The information obtained by the auditory system helps to correct the process of producing these sounds so as to meet our expectations. We are constantly informed of the results of our motor activities by acoustic signals generated during these activities. Acoustic waves intercepted by our auditory system allow the appropriate muscles to be activated in order to achieve the desired acoustic effect. The key role of feedback in the process of learning speech, mastering a musical instrument, and keeping a stable rhythm is well known and documented (Smith, 1975; Finney and Palmer, 2003; Repp, 2000; Pfordresher and Palmler, 2006). But, until now, authors have only focused on the functioning of our auditory control and the appropriate reactions to changes in feedback, thereby overlooking the nature of these reactions: they seem to assume that our actions are controlled consciously and engage the system of auditory perception.

The hypothesis tested in this paper is that fundamental voice frequency is tracked and controlled nonconsciously by an auditory system which controls vocal production. The effects of this control are used to correct vocal productions nonconsciously.

It should be stressed that the proposed approach goes a step further than research on the role of the "what" and "where" pathways in the auditory domain (Rauschecker,

[a)]Electronic mail: honorata@ia.spl.amu.edu.pl.

1998; Rauschecker, 2000; Alain *et al.*, 2001; Arnott *et al.* 2004). The advocates of the latter approach claim that the "what" pathway, or the ventral stream, is responsible for the recognition of sound sources; while the "where" pathway, or dorsal stream, is responsible for the perception of location and of the arrangement of sound sources in space. The present author accepts Milner and Goodale's (1995) hypothesis that the ventral stream processes information for perception while the dorsal processes information for action performance. However, this hypothesis had to be adapted to the sphere of audition by assuming that the dorsal stream in the auditory cortex is primarily responsible for the control of the production of sounds, and in particular for the control of phonation.

Some recent findings support the hypothesis that motor reactions are separated in the auditory modality. Repp (2000) showed that in a synchronized tapping task, subliminal pulse changes in a tone sequence were compensated for. This is an automatic, nonconscious process that is sensitive to the temporal order of information below the perceptual threshold. Action control is also insensitive to perceptual illusions (Repp, 2005). An auditory illusion, such as the influence of an intensity change on perceived timing, does not affect online action control. The results obtained in the work of Hickok and Poeppel (2003) suggest that in the case of speech perception a ventral stream is involved in mapping sound onto meaning, while a dorsal stream is involved in mapping sound onto articulatory-based representations.

To observe the dissociation between perception and motor control, a subliminal experimental situation was created. In a psychoacoustic experiment, values below the perceptual threshold were used, i.e., values which were not processed through the normal channels or apparatus of perception. The hypothesis was that a subliminal modification to auditory voice feedback would cause an appropriate correction as a response, even if this change was not actually perceived.

On the assumption that the auditory system functions in the same way as the visual one and processes the information vital for motor reactions in real time, a reaction that would compensate for such a modification should be expected. Two experimental conditions were used: $F0$ shifted up and $F0$ shifted down. Thanks to recent research (Burnett *et al.*, 1998; Larson *et al.*, 1996; Larson, 1998; Burnett and Larson, 2002; Natke *et al.* 2003; Yi Xu *et al.*, 2004; Sivasankar *et al.*, 2005), it has been established that during vocalization subjects react to perturbations in the pitch of voice feedback by changing their fundamental voice frequency ($F0$), usually to compensate for the pitch shift. Responses indicating a negative feedback system which stabilizes $F0$ occur with a latency of 100–160 ms (Burnett *et al.*, 1998; Hain *et al.*, 2000; Natke and Kalveram, 2001). The experiments did not focus on the way the listeners perceived the modifications. Previous research (Klatt and Zue, 1971; Pape and Mooshammer, 2006) determined the just-noticeable difference (JND) of fundamental voice frequency contours for digitally synthetic stimuli. Depending on the experimental conditions and the stimuli used (i.e., different vowels), it was found that pitch shifts ranging from 4 to 98 cents constituted the averaged JND. Under the experimental conditions of this study, and using the vowel "u," the averaged threshold value below which pitch shifts could not be observed was 26 cents. Throughout all the trials, changes in the subjects' fundamental voice frequency were examined. Reactions to subliminal pitch perturbations of voice feedback were also checked by an analysis of changes in fundamental voice frequency ($F0$).

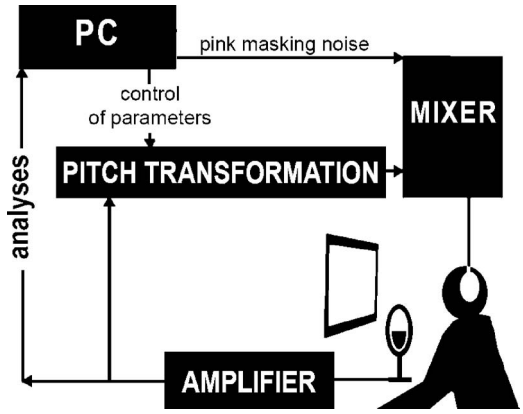


FIG. 1. Schema of the experimental acoustic feedback setup.

II. METHOD

A. Subjects

Nine adults between 21 and 28 years of age participated in this study. All the listeners qualified as having normal hearing, which was defined as the audiometric threshold of 20 dB hearing level, or better, for a range from 250 to 8000 Hz (ANSI, 1996). They reported no neurological defects and had no speech or voice disorders. All of them were trained singers. The subjects were seated in a sound-treated room.

B. Stimuli

The subjects' voices were recorded with a Shure SM 58 microphone (with a 6 cm mouth-to-mike distance), amplified with a Behringer mixer model 802A, and processed for auditory feedback pitch shifting through a DP2 Ensoniq ultraharmonizer. The pitch-shift processing introduced a small delay of 10 ms. The output of the harmonizer was mixed with pink masking noise [75 dB sound pressure level (SPL)] and presented to the subject over Sennheiser HD 600 headphones. The experiment schema is presented in Fig. 1.

Because previous studies (Burnett *et al.*, 1998) showed no relationship between voice $F0$ response and voice intensity, the intensity was not under strict control in this study.

C. Procedure

In the experiment subjects were instructed to vocalize the vowel /u/ for 5 s. They were asked to maintain constant pitch and loudness. They heard their own voice via headphones. The voice signal was mixed with pink noise (70 dB SPL) to partially mask bone-conducted auditory feedback. The experiment schema is presented in Fig. 2.

At the beginning of the third second of each vocalization, voice feedback was pitch shifted by 9, 19, 50, or 100 cents, respectively, in both directions. In every set of sounds,

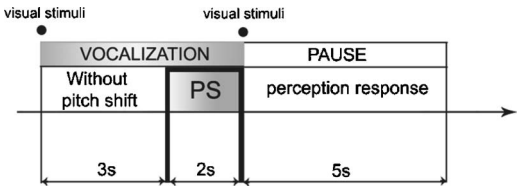


FIG. 2. Schematic illustration of stimulus presentation.

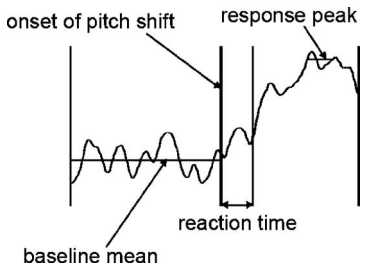


FIG. 3. Example of $F0$ response of a subject to downward pitch stimuli.

instances without any pitch shifts also appeared. Pitch shifts in the auditory feedback signal were controlled by a MIDI based control system. During a 5 s break between each vocalization, subjects were required to report whether the trial had contained a pitch shift and, if yes, what direction it was. Both the beginning and the ending of phonation were marked with visual stimuli. Pitch shift values were randomly ordered. Every pitch shift value was repeated 30 times.

At the same time, the listener's voice was digitally recorded and changes in their fundamental voice frequency were examined in order to check subject's actual reactions to pitch shifts.

In order to get information about threshold perception in pitch shift, the results from all the subjects had to be divided into two categories: results where subjects heard pitch changes during the vocalization, and results where subjects did not perceive any pitch changes during the vocalization. The results were presented in the form of psychometric curves, with a threshold value of 75% for correct detections. Additionally, the individual results were analyzed to see whether the direction of the fundamental voice frequency shift was appropriately detected. For each of the listeners, percentage values representing the number of correct answers to the pitch shift stimuli were calculated. The threshold was set at 50% correct answers, in order to intensity the criteria of the experiment and verify that the subject was actually able or not able to determine the pitch shift direction.

D. Analysis

The signal was low-pass filtered at 500 Hz in order to remove high frequency harmonics. The fundamental frequency of vocalization during each trial was calculated using an algorithm incorporated in the Praat software (Boersma, 1993). The data was then analyzed using a special application developed in the MATLAB environment. For each trial, an average value of $F0$ and a standard deviation were calculated for the period before the onset of the pitch shift. Subsequently, a reaction check was performed. A positive reaction was defined as a deviation in $F0$, which had a latency of at least 60 ms, a magnitude of more than 2 SDs of the 1000 ms pre-stimulus mean, and a duration of at least 120 ms in a maximum of 800 ms after the pitch shift (Fig. 3). The application automatically determined the time when the averaged signal departed and reentered the 2 SD response criterion, and calculated the corresponding values in cents for valid responses only.

Changes in the fundamental frequency were converted to cents using the following formula (Hain *et al.*, 2000):

$$Cents = 100 * \left(\frac{10 * \mathrm{Log}(P/M)}{\mathrm{Log}(2)} \right),$$

where P is response peak $F0$ in hertz, and M is the base line mean $F0$ in hertz.

III. RESULTS

More than 80% of the signals fully corresponded to the predefined criteria. Among the analyzed signals, in 89% of the answers there was compensation for the applied change (opposing responses), and in 11% there was no such compensation (following responses). In the display of results, the results of control trials (sets of sounds without pitch shifts) were omitted, because in the trials more than 70% of motor reactions were classified as invalid.

The results are discussed at two levels: as group data and individual data. The averaged results of the detection task and the averaged results of the motor response are presented in Figs. 4(a) and 4(b), respectively.

Figure 4(a) represents the averaged results for the perception task of the experiment. Percentage values represent the number of detections of the pitch shift. The average motor response values (in cents) are presented in Fig. 4(b). From the equation of the accumulative standard distribution curve, threshold values for the probability of 75% were calculated. The averaged threshold values were −21 cents for the downward pitch shift of fundamental voice frequency, and 30 cents for the upward pitch shift. Therefore the values −19, −9, 9, and 19 cents were under the 75% threshold, and were thus in the same range as those gathered in previous experiments of a similar nature (Klatt and Zue, 1971; Klatt 1973; Pape and Mooshammer, 2006). However, in the previous studies thresholds were determined in different experimental conditions.

The results were also presented on an individual level in order to show the correct detections of the direction of pitch shift. Individual results of the detection task, along with the corresponding motor reactions, are shown in Fig. 5. The column on the left represents the results of the perception part of the experiment. Percentage values represent the number of correct answers to the pitch shift. Only one subject perceived all the changes in voice pitch correctly. For seven subjects pitch shift values −19, −9, and 9 were under the 50% threshold. This means that, for these values, the listeners were unable to perceive the direction of the pitch shifts. Therefore, these values can be identified as subliminal for perception.

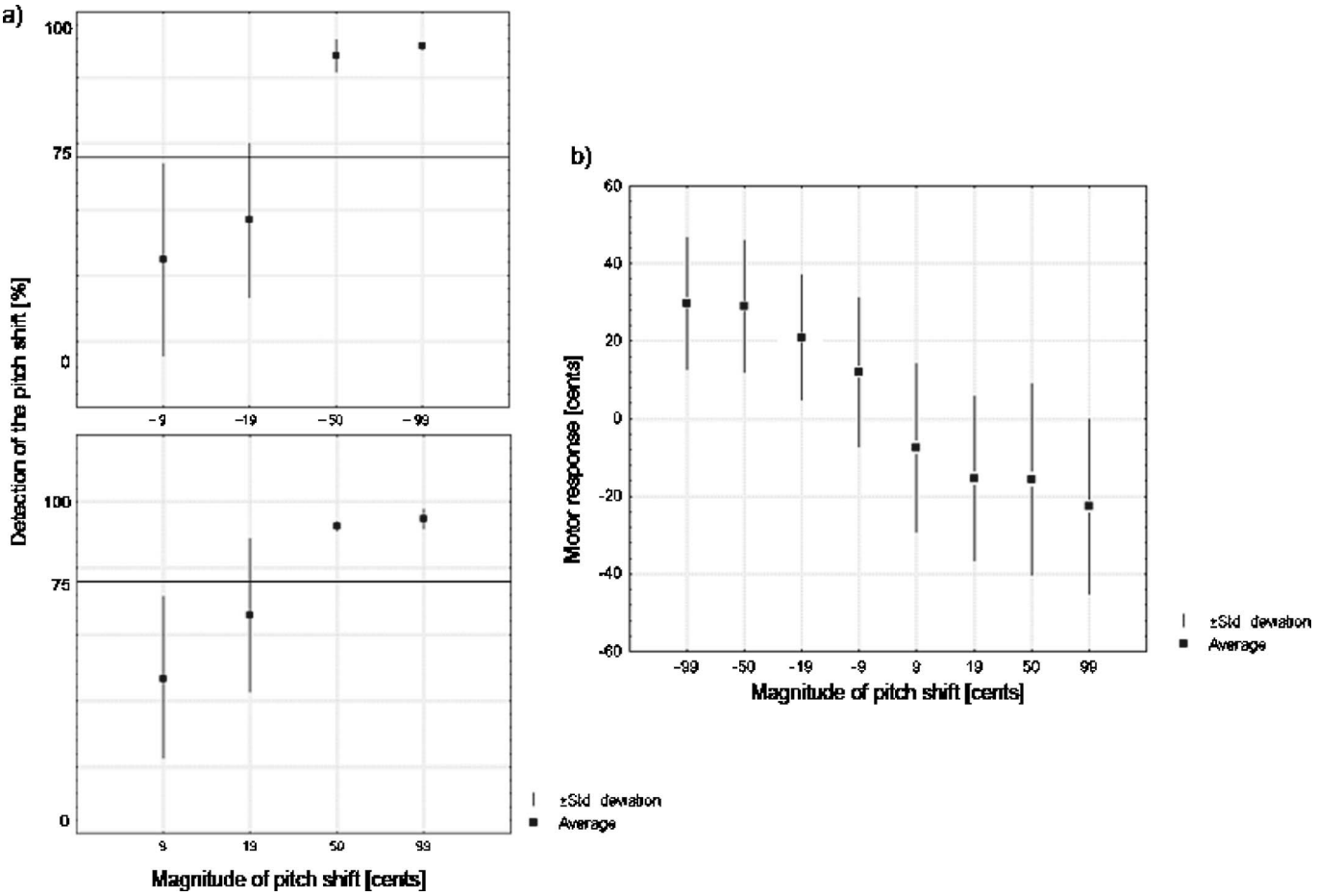


FIG. 4. The averaged results of the detection part of the experiment. Percentage values show the number of detections of the pitch shift (a). The averaged results of the motor part of the experiment (b).

An interesting asymmetry of the results can be observed: it was easier to perceive upward pitch shifts than downward pitch shifts.

The individual motor response values (in cents) for each subject are presented in the right-hand column. It can be clearly seen that when the pitch feedback is shifted down, the subjects raise their pitch, which contrasts with the pitch feedback being shifted up in order to compensate for the pitch change in the auditory feedback. The same response can also be observed for shifts which are subliminal to perception. The post-hoc T Tukey (HSD) test was carried out for stimuli that were not noticed perceptually. The test showed the statistical significance of differences between reaction values to upward and downward pitch shifts for the majority of conditions in all cases.

IV. DISCUSSION

Recent studies (Burnett *et al.*, 1998; Larson *et al.*, 1996; Larson, 1998; Burnett and Larson, 2002, Hain *et al.*, 2000; Jones and Munhall, 2000; Natke *et al.*, 2003; Yi Xu *et al.*, 2004; Sivasankar *et al.*, 2005) in the field of fundamental voice frequency control have clearly shown that the audio-vocal system controls the stability of fundamental voice frequency by compensating for the changes applied to feedback. It has been proven that the so-called pitch shift reflex process is of automatic nature (Burnett *et al.*, 1998), although its neural mechanism is still poorly understood.

The results of the experiment suggest that the stream utilized to control fundamental voice frequency could be both independent of the perception stream and cooperate with it at the same time. In the results of the experiment presented in this study there are vast differences between reported perceptions and motor reactions. In many cases, listeners reacted appropriately to voice pitch shifts, although they did not perceive the change. This fact suggests that a certain part of the information contained in the acoustic signal is also processed subconsciously for calculations necessary for appropriate motor reactions. The present experiment was carried out on a group that consisted only of trained singers. The decision to choose this type of group was based on the fact that, because of the subjects' training in singing, random changes in voice pitch before the onset of pitch shift were not significant to the results of the experiment. It can be presumed, however, that in the case of people not trained in singing, the same mechanism of fundamental voice frequency control operates. The results of previous studies (Hain, 2000) revealed that the process of controlling fundamental voice frequency could be independent of our will. It also suggests a possible division into perception and motor streams.

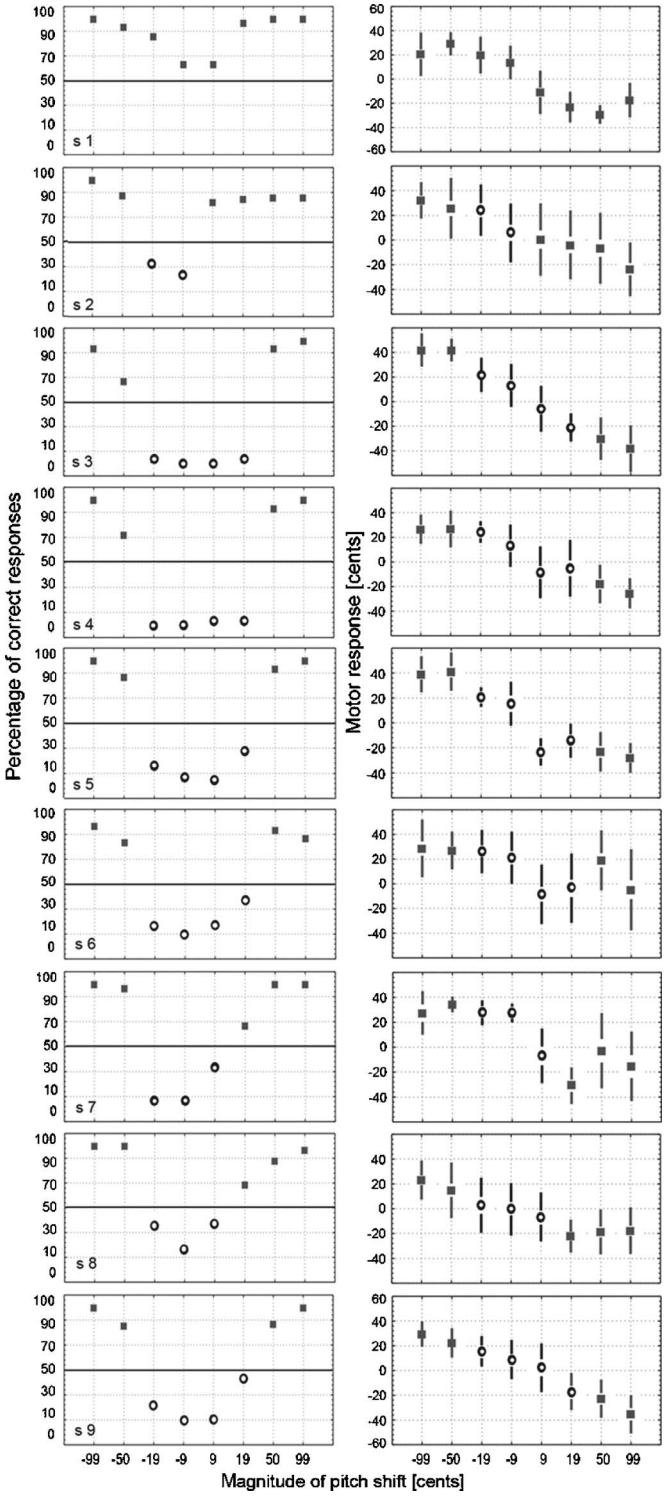


FIG. 5. Results with appropriate standard deviations for all subjects. The column on the left represents the results of the perception part of the experiment. Percentage values show the number of correct answers to the pitch shift. Circles aligned under the line corresponding to the perception threshold stand for values that were not registered perceptually. The column on the right represents the results of the motor part of the experiment. Reactions that were not registered in the perception part of the experiment were singled out.

In their neurological studies, Hickok and Poeppel (2003) discovered the division of two separate streams in cortical processing. One of those streams attached a certain meaning to the sound event, while the other articulated representations of the sound itself. The latest studies (Purcell and Munhall, 2006) show that auditory feedback during speech production plays a significant role in the maintenance of accurate articulation. Any real time changes that modified the frequency of the first formant caused a reaction which compensated for those changes by adjusting the parameters of the vocal track. Such a reaction is strictly a motor reaction. All the experiments show that acoustic feedback is necessary for the regulation of the parameters of produced sounds, as it triggers appropriate motor reactions. In this study, by arranging a specific experimental situation, we were able to separate the functioning of the two streams. The use of stimuli situated beneath the threshold of perception allowed us to isolate the motor stream in order to observe its functioning. A similar situation was arranged by Repp in his study of the perception of rhythm with simultaneous finger tapping. In his study he showed that listeners are capable of reacting to subliminal changes in a rhythmic sequence, even though such changes are not registered by their perception. The intent of our study was to verify the possibility of there being motor reactions beyond conscious control which are used to control fundamental voice frequency. The experiment dedicated to this purpose verified that this is the case for subliminal changes in acoustic feedback. Such a reaction also took place with distinctive changes, but according to the previous studies (Burnett *et al.*, 1998; Larson *et al.*, 1996; Larson, 1998, Burnett and Larson, 2002; Hain *et al.*, 2000) it did not fully compensate for the changes introduced into feedback.

V. CONCLUSION

1. The results of the experiment carried out suggest that the motor control of fundamental voice frequency functions properly, even when the pitch shifts are not being consciously perceived.

2. The obtained results suggest that in the case of the motor control of phonation, a dissociation between perception and action for auditory information takes place.

ACKNOWLEDGMENTS

Grateful thanks to Professor Anna Preis and Professor Andrzej Klawiter for extensive assistance, many helpful comments and discussions.

Alain, C., Arnott, S. R., Hevenor, S. J., Graham, S., and Grady, C. L. (**2001**). "'What' and 'where' in the human auditory system," Proc. Natl. Acad. Sci. U.S.A. **98**, 12301–12306.

ANSI. ANSI S3.6-1996, (**1996**). Specifications for Audiometers, American National Standards Institute, New York.

Aglioti, S., DeSouza, J., and Goodale, M. A. (**1995**). "Size-contrast illusions deceive the eyes but not the hand," Curr. Biol. **5**, 679–685.

Arnott, S. R., Binns, M. A., Grady, C. L., and Alain, C. (**2004**). "Assessing the auditory dual-pathway model in humans," Neuroimage **22**, 401–408.

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice F0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**(6), 3153–3161.

Burnett, T. A., and Larson, C. R. (**2002**). "Early pitch shift response is active in both steady and dynamic voice pitch control," J. Acoust. Soc. Am. **112**, 1058–1063.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proceedings of the Inst. of Phonetic Sciences **17**, 97–110.

Finney, S. A., and Palmer, C. (**2003**). "Auditory feedback and memory for

music performance: Sound evidence for an encoding effect," Mem. Cognit. **31**, 51–64.
Goodale, M. A., Milner, A. D., Jakobson, L. S., and Carey, D. P. (**1991**). "A neurological dissociation between perceiving objects and grasping them," Nature (London) **349**, 154–156.
Goodale, M. A., and Milner, A. D. (**1992**). "Separate visual pathways for perception and action," Trends Neurosci. **15**, 20–25.
Goodale, M. A., and Milner, M. A. (**2004**). *Sight Unseen: An Exploration of Conscious and Nonconscious Vision*, Oxford: (Oxford University Press, Oxford), p. 140.
Haffenden, A. M., Schiff, K. C., and Goodale, M. A. (**2001**). "The dissociation between perception and action in the Ebbinghaus illusion: Nonillusory effects of pictorial cues on grasp," Curr. Biol. **11**, 177–181.
Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., and Kenney, M. K. (**2000**). "Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex," Exp. Brain Res. **130**, 133–141.
Hickok, G., and Poeppel, D. (**2003**). "Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language," Cognition **92**, 67–99.
Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**(3), 1246–1251.
Klatt, D. H., and Zue, V. W. (**1971**). "Just—noticeable differences for selected aspects of synthetic fundamental—frequency contours," J. Acoust. Soc. Am. **49**, 86–87.
Klatt, D. H. (**1973**). "Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception," J. Acoust. Soc. Am. **53**, 8–16.
Kroliczak, G., Heard, P., Goodale, M. A., and Gregory, R. L. (**2006**). "Dissociation of perception and action unmasked by the hollow-face illusion," Brain Res. **1080**, 9–16.
Larson, C. R., White, J. P., Freedland, M. B., and Burnett, T. A. (**1996**). "Interactions between voluntary modulations and pitch-shifted feedback signals: Implications for neural control of voice pitch," in *Vocal Fold Physiology: Controlling Complexity and Chaos*, edited by P. J. Davis and N. H. Fletcher (Singular, San Diego), pp. 279–289.
Larson, C. R. (**1998**). "Cross-modality influences in speech motor control: The use of pitch shifting for the study of F0 control," J. Commun. Disord. **31**, 489–503.
Milner, A. D., and Goodale, M. A. (**1995**). *The Visual Brain in Action* Oxford (Oxford University Press, Oxford).
Natke, U., Donath, T. M., and Kalveram, K. T. (**2003**). "Control of voice fundamental frequency in speaking versus singing," J. Acoust. Soc. Am. **113**(3), 1587–1593.
Natke, U., and Kalveram, K. T. (**2001**). "Fundamental frequency under frequency shifted auditory feedback of long stressed and unstressed syllables," J. Speech Lang. Hear. Res. **44**, 577–584.
Pape, D., and Mooshammer, C. (**2006**): "Is intrinsic pitch language-dependent?—evidence from a cross-linguistic vowel pitch experiment (with additional screening of the listeners' DL for music and speech)," in *MULTILING-2006*, paper No. 018.
Pfordresher, P. Q., and Palmer, C. (**2006**). "Effects of hearing the past, present, or future during music performance," Percept. Psychophys. **68**, 362–376.
Purcell, W. D., and Munhall, G. K. (**2006**). "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," J. Acoust. Soc. Am. **120**, 966–977.
Rauschecker, J. P. (**1998**). "Cortical processing of complex sounds," Curr. Opin. Neurobiol. **8**(4), 516–521.
Rauschecker, J. P., and Biao, T. (**2000**). "Mechanisms and streams for processing of "what" and "where" in auditory cortex," Proc. Natl. Acad. Sci. U.S.A. **97**(22), 11800–11806.
Repp, B. H. (**2000**). "Compensation for subliminal timing perturbations in perceptual-motor synchronization," Psychol. Res. **63**, 106–128.
Repp, B. H. (**2005**). "Does an auditory perceptual illusion affect on-line auditory action control? The case of (de)accentuation and synchronization," Exp. Brain Res. **168**, 493–504.
Sivasankar, M., Bauer, J. J., Babu, T., and Larson, Ch. R. (**2005**). "Voice responses to changes in pitch of voice or tone auditory feedback," J. Acoust. Soc. Am. **117**(2), 850–857.
Smith, C. R. (**1975**). "Residual hearing and speech production in deaf children," J. Speech Hear. Res. **18**, 795–811.
Yi, Xu, Larson, C. R., Bauer, J. J., and Hain, T. C. (**2004**). "Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences," J. Acoust. Soc. Am. **116**(2), 1168–1178.

Listeners' sensitivity to "onset/offset" and "ongoing" interaural delays in high-frequency, sinusoidally amplitude-modulated tones

Thomas N. Buell
Department of Neuroscience and Department of Surgery (Otolaryngology), University of Connecticut Health Center, Farmington, Connecticut 06030, USA

Sarah J. Griffin
Cellular and Molecular Neuroscience, MRC Toxicology Unit, University of Leicester, Lancaster Road, Leicester LE1 9HN, United Kingdom

Leslie R. Bernstein[a)]
Department of Neuroscience and Department of Surgery (Otolaryngology), University of Connecticut Health Center, Farmington, Connecticut 06030, USA

(Received 13 April 2007; revised 24 October 2007; accepted 24 October 2007)

The relative potency of onset/offset and envelope-based ongoing interaural time delays (ITDs) was assessed using high-frequency stimuli. A two-cue, two-alternative, forced-choice adaptive task was employed to measure threshold ITDs with 100% sinusoidally amplitude-modulated tones centered at 4 kHz. Modulation rates of 125, 250, and 350 Hz were tested with durations of 32, 90, or 240 ms. In the first experiment, ITDs to be detected were imposed only at the onset/offset, only within the ongoing portion, or within both the onset/offset and ongoing portions of the stimuli. Results indicated that ongoing ITDs dominated onset/offset ITDs. The relative potency of ongoing ITDs was directly proportional to duration and inversely proportional to modulation rate. Quantitative analysis suggested that listeners effectively combine onset/offset and ongoing ITDs. Furthermore, the data could be largely accounted for by assuming that listeners attend to the interaural decorrelation of the stimulus resulting from onset/offset and/or ongoing ITDs. A second experiment showed that, (1) overall, an ongoing ITD of one-half period of the envelope had little impact on listeners' sensitivity to delays of the onset/offset and (2) sensitivity to delays within the onset/offset portion of the waveform was reduced by roving the delay within the ongoing portion of the waveform. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816399]

PACS number(s): 43.66.Pn, 43.66.Ba [RLF] Pages: 279–294

I. INTRODUCTION

McFadden and Pasanen (1976) carefully distinguished among the types of interaural time delay (ITD) that could be present and could be useful to listeners when localizing sounds in space or when lateralizing sounds presented via headphones. The first of these is the interaural delay between the onsets (and offsets) of the waveform at the two ears. This type of delay has, historically, been referred to as the "onset," "gating," or "transient" interaural delay (e.g., Tobias and Schubert, 1959; McFadden and Pasanen, 1976; Abel and Kunov, 1983; Buell *et al.*, 1991; Zurek, 1993). In addition to the onset delay, an interaural delay can also be embedded within the steady-state, ongoing portion of the waveform at the two ears. Such ongoing interaural delays could be conveyed either by the cycle-by-cycle time differences in the fine structure of the waveform and/or by the time-varying envelope of complex waveforms such as sinusoidally amplitude modulated (SAM) tones or Gaussian bands of noise. Thus, McFadden and Pasanen identified three different types of interaural delay: onset/offset delays, ongoing fine-structure-based delays, and ongoing envelope-based delays. Note that at low frequencies (below about 1500 Hz), ongoing interaural delays would be conveyed almost exclusively by the fine structure of the waveform. At higher frequencies, however, where the binaural auditory system is insensitive to interaural delays within the fine structure, the ongoing interaural delay could only be conveyed by the envelope.[1]

Several studies have been directed toward assessing the relative salience or potency of a given magnitude of the "onset/offset" or "ongoing" delay. These experiments involved either (1) measurement of the discriminability of changes in one type of delay while holding constant or obscuring the other (Hafter, Dye, and Gilkey, 1979; Perrott and Baars, 1974) or (2) measures of the relative magnitude of one type of delay required to compensate for the other when the two were placed in opposition (e.g., Tobias and Schubert, 1959) or (3) measures of the degree of laterality produced when the two types of delay were presented singly or in combination (Buell *et al.*, 1991). The general outcome of those studies is that ongoing delays are substantially more potent than onset/offset delays with the *relative* strength of the ongoing delay being directly proportional to the duration

a)Author to whom correspondence should be addressed. Electronic mail: les@neuron.uchc.edu

of the stimulus.[2] In fact, even with signals as short as 10 ms, the ongoing delay was found to be dominant.

Extending the results from previous "discrimination" experiments, Buell *et al.*, (1991) measured the extents of laterality produced by combinations of onset/offset and ongoing delays. Their results confirmed the dominance of ongoing delays found in the earlier studies. They found, however, that when the ongoing delay was "ambiguous," in that it did not clearly favor either ear (it contained an interaural delay corresponding to a phase shift at or near 180°), the onset/offset delay dominated perception and determined the laterality of the intracranial image.

In all of the studies discussed above, the investigators employed either stimuli restricted to low frequencies or stimuli that were broadband. In our view, the results of those studies addressed only the relative potencies of the first two of the types of interaural delays identified by McFadden and Pasanen (1976), namely, onset/offset and ongoing *fine-structure-based* delays. This interpretation is based on two observations. First, for stimuli restricted to low frequencies, sensitivity to ongoing interaural delay has been shown to stem almost exclusively from the processing of timing information within the fine structure as opposed to from timing information within the envelope (e.g., Henning, 1980; Henning and Ashton, 1981; Bernstein and Trahiotis, 1985). Second, for broadband stimuli, interaural delays within the low-frequency portions of the stimulus (which are conveyed by the fine structure) have been shown to dominate binaural processing (e.g., Blauert, 1982, 1983; Wightman and Kistler, 1992). In summary, what the prior studies cited above have assessed is the relative potency of onset/offset versus ongoing fine-structure-based delays conveyed by low-frequency information.

To our knowledge, there has been no direct and systematic investigation of the relative potency of onset/offset versus ongoing delays at *high* frequencies, where the ongoing information is conveyed not by the fine structure but, rather, by the envelopes of the waveforms.[3] It occurred to us that any dominance of ongoing delays measured with high-frequency stimuli might be relatively weak, as compared to that measured with low-frequency stimuli. Our reasoning was as follows. Listeners are known to be substantially more sensitive to ongoing interaural delays conveyed by the fine structure of low-frequency waveforms than they are to ongoing interaural delays conveyed by the envelopes of high-frequency waveforms such as SAM tones, two-tone complexes, and bands of noise (for reviews, see Blauert, 1982, 1983; Grantham, 1995; Bernstein, 2001). A primary question was, given the poorer sensitivity to envelope-based interaural delays, to what degree would ongoing delays still dominate over onset/offset delays at high frequencies? We chose to investigate this question using high-frequency, SAM tones because, for such stimuli, the ongoing interaural delays would be conveyed only by their envelopes. In addition, measuring the relative potency of onset/offset and ongoing envelope delays using high-frequency SAM tones would aid in the assessment and interpretation of classic studies in which sensitivity to interaural delays within high-frequency waveforms was measured. In some of those studies, the stimuli were presented with "whole waveform delays" for which *both* onset/offset and ongoing delays were present in equal amounts (e.g., David *et al.*, 1958, 1959, Yost *et al.*, 1971), in others, the type(s) of delay employed cannot be determined unequivocally from the published reports (Henning, 1974; Klumpp and Eady, 1956). In the absence of information regarding the relative potency of the different types of delay at high frequencies, it is not clear the extent to which the thresholds measured in those studies reflect listeners' sensitivity to the onset/offset delays, the ongoing envelope-based delays, or both.

A. General approach

The experiments reported here were designed to assess the relative salience of onset/offset and ongoing envelope-based interaural delays in high-frequency sinusoidally amplitude modulated tones centered at 4 kHz. In the first experiment, threshold interaural delays were measured for conditions in which the delay was imposed on only the onset and offset of the waveform ($\Delta|0$), only the ongoing portion of the waveform, ($0|\Delta$) or both the onset/offset and ongoing portions of the waveform ($\Delta|\Delta$). These three conditions of interaural delay are depicted in panels A, B, and C, respectively, of Fig. 1. As exemplified by Fig. 1, in order to refer to the different combinations of interaural delay employed, we adopt the nomenclature where the symbol to the left of the bar indicates the status of the onset/offset interaural delay and the symbol to the right of the bar indicates the status of the ongoing interaural delay. A "Δ" indicates that a change in the corresponding interaural delay was manipulated as the independent variable; a "0" indicates that the corresponding interaural delay was zero. It is important to note that the $\Delta|0$ and $0|\Delta$ delay types effectively contain inconsistent values across the types of interaural delay of the waveform. Specifically, when the delay is imposed on only the onset/offset portion of the waveform, the ongoing portion contains an interaural delay fixed at a value of zero. Similarly, when the delay is imposed on only the ongoing portion of the waveform, the onset/offset portion contains an interaural delay fixed at a value of zero. No such inconsistency exists when the interaural delay is imposed on both the onset/offset and ongoing portions of the waveform (the $\Delta|\Delta$ delay type).[4]

Another manipulation involved varying the duration of the SAM tones. This parameter has been shown to affect sensitivity to ongoing interaural envelope-based delays at high frequencies such that thresholds decrease as the duration of the stimulus is increased up to values as large as 400 ms or so (e.g., Nuetzel and Hafter, 1976; McFadden and Moffitt, 1977). In addition, recall that in earlier studies that focused on assessing the relative salience of onset/offset and ongoing interaural delays at low frequencies, the degree to which the ongoing interaural delay dominated listeners' judgments was directly proportional to the duration of the stimulus.

Yet another manipulation involved varying the rate of modulation of the SAM tone. This parameter was chosen because it has also been shown to affect greatly listeners' sensitivity to ongoing envelope-based interaural delays. Spe-

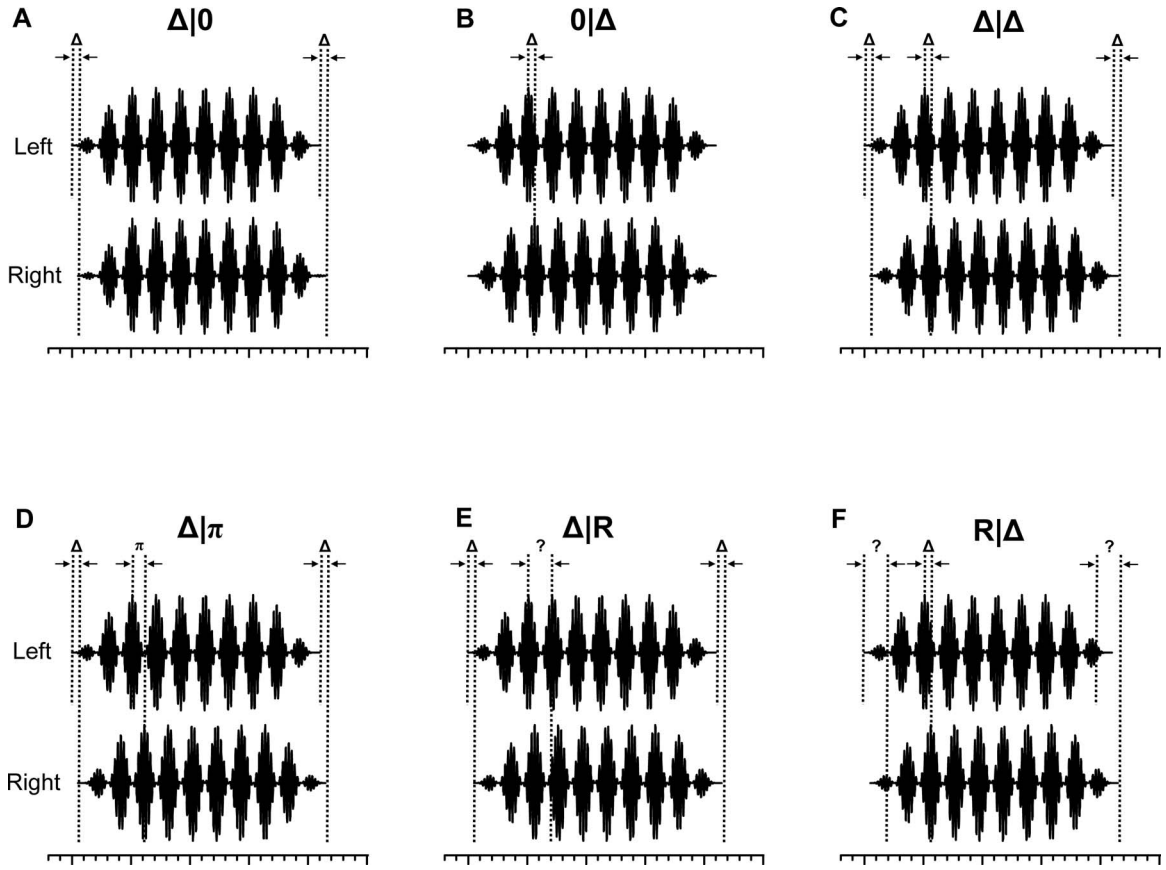


FIG. 1. Depictions of the delay-types employed in Experiments 1 (panels A–C) and 2 (panels D–F). The labels centered above each panel (e.g., Δ|0) indicate the status of the interaural delay in the onset/offset and ongoing portions of the waveform, respectively. A "Δ" indicates the presence of the interaural delay to be detected; a "0" indicates that the particular aspect of the waveform contained no interaural delay; a "π" indicates that a delay was imposed that was equal to one-half the period of the sinusoidal modulator; an "R" indicates that the interaural delay was "roved" from interval to interval within the behavioral task. Vertical dotted lines and arrows act as guides to aid in visualizing the effects produced by each type of interaural delay. A "?" above a pair of vertical lines indicates that the interaural delay was roved within the specified aspect of the waveform.

cifically, for rates of modulation between about 30 and 500 Hz, sensitivity to ongoing interaural envelope delay at high frequencies is characterized by a low-pass function such that thresholds typically rise as rate of modulation is increased above about 150 Hz and are generally extremely large or unmeasureable for rates as high as 500 Hz (e.g., Nuetzel and Hafter, 1981; Henning and Ashton, 1981; Bernstein and Trahiotis, 2002). In summary, our goal was to assess whether and to what degree any dominance of the ongoing interaural delay over the onset/offset interaural delay at high frequencies would be affected by manipulating the sensitivity to the ongoing interaural delay. This was accomplished by varying the duration or the rate of modulation of the 4-kHz-centered SAM tones.

In the second experiment reported here, conditions were employed in which the listener's task was to detect an onset/offset interaural delay that was paired with an ongoing delay corresponding to 180° (or π radians) of the sinusoidal envelope (the Δ|π delay type, depicted in panel D of Fig. 1). Recall that Buell *et al.* (1991) showed that listeners' performance was dominated by the onset/offset interaural delay, rather than by the ongoing interaural delay, when such an "ambiguous," 180° phase shift was employed. We wished to determine whether a similar result would occur in our discrimination paradigm employing high-frequency SAM tones. Finally, a set of conditions was employed in which thresholds were measured for detection of onset/offset interaural delay while the ongoing interaural delay was varied randomly, or "roved" between observation intervals rather than being fixed at zero (the Δ|R delay type, depicted in Panel E of Fig. 1). Our hypothesis was that, to the degree the ongoing interaural delay contributes to the listeners' judgments, roving its value would "disrupt" the ability of listeners to discriminate the presence of an onset/offset interaural delay. In this fashion, the roving condition served as another means to evaluate the relative contribution of the types of interaural delay. In a complementary set of roving conditions, the roles of the onset/offset and ongoing delays just described were reversed (the R|Δ delay-type depicted in panel F of Fig. 1).

II. EXPERIMENT 1

A. Procedure

Detection of interaural time delay was measured using sinusoidally amplitude modulated (SAM) tones with carrier frequencies of 4 kHz. The carriers were 100% modulated at rates of 125, 250, or 350 Hz. Three durations were tested: 32, 90, and 240 ms (including 5 ms $\cos^2$ rise-decay ramps). All stimuli were generated digitally with a sampling rate of 100 kHz (TDT AP2), were low-pass filtered at 8.5 kHz

(TDT FLT2), and were presented via Etymotic ER-2 insert earphones at a level of 75 dB sound pressure level (SPL). A continuous diotic noise low-pass filtered at 1300 Hz (spectrum level equivalent to 30 dB SPL) was presented to preclude the listeners' use of any information at low spectral frequencies (e.g., Nuetzel and Hafter, 1976, 1981; Bernstein and Trahiotis, 1994).

Threshold interaural delays were determined using a two-cue, two-alternative, forced choice, adaptive task. Each trial consisted of a visual warning interval (200 ms), followed after 500 ms by four observation intervals each separated by 350 ms. Each interval was marked visually on a computer monitor. Feedback was provided for 400 ms after the listener responded. The stimuli in the first and fourth intervals were diotic. The listener's task was to detect the presence of an interaural delay (left-ear leading) that was presented with equal *a priori* probability in either the second or the third interval. The remaining interval, like the first and fourth intervals, contained diotic stimuli. The exact waveform destined for each observation interval was determined by selecting, at random, the starting point within a single pre-computed "master" waveform that was twice the duration of the signal ultimately presented to the listener. This sampling strategy ensured that the starting phases of the sinusoidal envelope assigned to each interval of each trial were drawn uniformly across the entire period of the sinusoidal modulator.

As outlined above, three separate conditions of interaural delay were tested. In the first condition, the interaural delay to be detected was restricted to the onset and offset of the waveform. That is, the ongoing, steady-state portion of the waveform contained no interaural delay [the $\Delta|0$ delay-type, Fig. 1(a)]. This delay-type was implemented by gating asynchronously between the two earphones an otherwise diotic waveform.

In the second condition, the interaural delay to be detected was restricted to the ongoing portion of the waveform [the $0|\Delta$ delay-type, Fig. 1(b)]. This delay-type, for which both the fine-structure and the envelope were delayed, was created by applying a time-delay to the waveform (implemented via a linear phase-shift of its spectral components) and then synchronously gating the resulting waveforms to the two earphones.

Finally, in the third condition, the interaural delay was applied to both the onset/offset and ongoing portions of the waveform [the $\Delta|\Delta$ delay-type, Fig. 1(c)]. This delay-type was created by gating asynchronously between the two earphones an interaurally time-delayed waveform.

The interaural delay chosen for a particular trial was determined adaptively in a manner designed to estimate 70.7% correct detection (Levitt, 1971). The initial step-size for the adaptive track corresponded to a factor of 1.584 (equivalent to a 2 dB change of ITD) and was reduced to a factor of 1.122 (equivalent to a 0.5 dB change of ITD) after two reversals. A run was terminated after 14 reversals and threshold was defined as the geometric mean of the ITD across the last 12 reversals. Because the techniques used to create the interaural delays involved "shifting" the waveforms by the appropriate number of samples, the resolution of the psychophysical procedure was ± 5 μs (one half the sampling period).

Four normal-hearing adults served as listeners. Three of them had participated in previous binaural experiments in our laboratory. Prior to the formal collection of data, all listeners received substantial practice on a number of the conditions ultimately employed in the experiment. The ordering of the first set of conditions was determined by choosing randomly among the three types of interaural delay ($\Delta|0, 0|\Delta, \Delta|\Delta$), rates of modulation of 125 and 250 Hz, and durations of 32 and 240 ms. The same ordering of conditions was employed for all four listeners. Three consecutive estimates of threshold were obtained for each condition until all 12 conditions had been exhausted. Then, three more consecutive thresholds were obtained by revisiting the same conditions in reverse order. This entire process was repeated once more so that 12 thresholds were obtained for each condition.

Upon inspection of the data collected in the set of conditions specified above, it was decided to collect data for a second set of conditions in order to explore more fully the effects of rate of modulation. The three types of delay and two durations used in the first set of conditions were employed in this second set but with the rate of modulation set at 350 Hz. Likewise, in order to explore more fully the effects of duration, a third set of conditions was tested in which the three types of delay were combined with the three rates of modulation (125, 250, 350 Hz) with the duration of the signal set at 90 ms. The ordering of the conditions composing the second and third sets was chosen randomly and repeated in the manner described above for the first set of conditions.

Upon completion of the collection of data for each of the three sets of conditions, the obtained estimates of threshold were examined. In 10 out of the 108 cases across the three sets of conditions (27 total conditions $\times$ four listeners) the original set of 12 estimates contained extremely high estimates of threshold (e.g., 2000–4000 μs) among substantially lower values. For those cases, two to three additional sets of three estimates of threshold were collected. In rare cases (three of the 108 cases), the listener was simply unable to perform the task and collection of data for that particular condition was terminated.

For each combination of listener and condition, only the final 12 estimates of threshold were considered. Of those 12, any estimate exceeding 1500 μs was excluded. That value was chosen because (1) it exceeds half the period of the highest rate of modulation employed (350 Hz) and, therefore could have provided an inconsistent cue to the listener regarding which earphone contained the "leading" or the "lagging" interaural delay (e.g., Domnitz and Colburn, 1977); (2) when applied to the onset and offset of a stimulus a delay of this magnitude could foster "secondary," potentially confusing intracranial images (Perrott and Baars, 1974); (3) it greatly exceeds the maximum interaural delay that can occur "naturally" in a sound field. Estimates exceeding 1500 μs occurred in only 18 of the 108 cases.

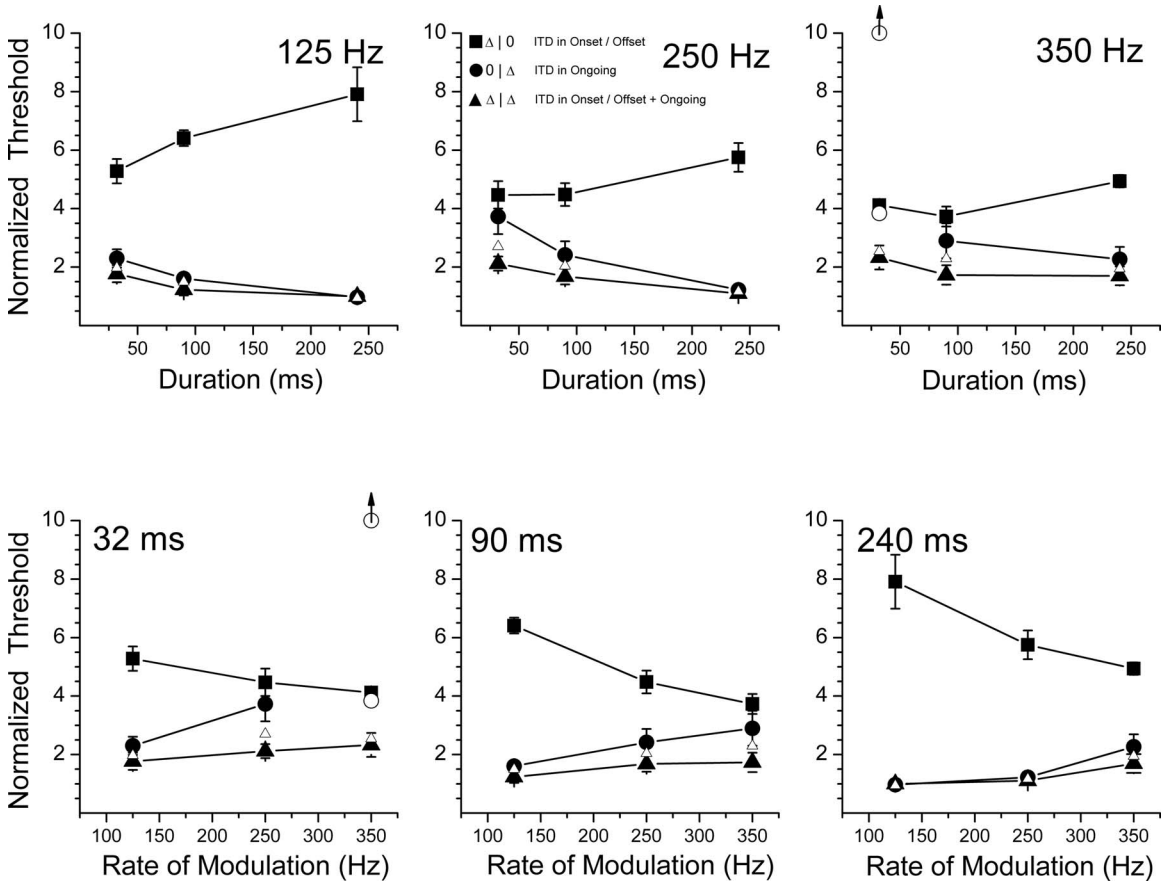


FIG. 2. Normalized thresholds (see text) as a function of the duration (upper panels) or rate of modulation (lower panels) of the SAM waveforms centered at 4 kHz. The values shown are the means across the data obtained from the four listeners. Error bars represent ± one standard error of the mean. The parameter within each plot is the delay-type employed. The open circles within two of the plots represent the mean value of normalized threshold for the two listeners who could perform the task with the 32 ms/350 Hz pairing of duration and rate of modulation. The open circles with upward pointing arrows plotted for that condition serve as reminders that two of the listeners could not perform the task. Open triangles represent predicted values for the Δ|Δ delay type derived from the quantitative information-based description of detection (see text).

After excluding estimates of threshold that were greater than 1500 μs, the two highest and the two lowest of the remaining estimates were excluded to prevent individual "outlying" estimates of threshold from influencing the final estimates. The mean of the resulting set of estimates (typically numbering eight but no fewer than four) served as the final estimate of threshold for the particular listener and stimulus condition.

B. Results

The thresholds for each listener in each condition were divided by his or her threshold obtained with the Δ|Δ delay type when the duration of the stimulus was 240 ms and the rate of modulation was 125 Hz. That condition was chosen as the "reference" because it yielded thresholds that were among the smallest for all four listeners. This "normalization" allows an examination of the effects of the independent variables across listeners while removing inter-individual differences in absolute sensitivity. The reference thresholds obtained from the individual listeners were 148 μs for listener BT, 177 μs for listener RC, 64 μs for listener KM, and 134 μs for listener SZ.

Figure 2 contains the means of the normalized thresholds across the four listeners. The error bars represent ± one standard error of the mean. In the top row, the normalized threshold ITDs are plotted as a function of the duration of the stimulus. The parameter of each plot is the type of ITD imposed on the waveform (see Fig. 1). The left-hand, middle, and right-hand panels display data obtained for rates of modulation of 125, 250, and 350 Hz, respectively. Note that in the right-most panel, the mean for the 0|Δ delay type (circles), obtained when the rate of modulation was 350 Hz and the duration was 32 ms, is plotted as two open circles. For this condition, in which the rate of modulation was the highest and the duration was the shortest, thresholds could only be obtained from two of the four listeners and the lower of the two open circles is plotted at their mean of 3.84. The open circle and arrow plotted at the top of the ordinate symbolize that thresholds for the other two listeners could not be measured.

In the bottom row, the normalized ITD thresholds are re-plotted as a function of rate of modulation of the SAM tone. The parameter of each plot is, once again, the type of interaural delay imposed on the waveform. The left-hand, middle, and right-hand panels display data obtained for stimulus durations of 32, 90, and 240 ms, respectively. The open circles plotted in the left-most panel represent the single condition (described above) in which thresholds for

two of the four listeners could not be measured. The small open triangles are predicted values of threshold for the $\Delta|\Delta$ delay type and will be discussed after the data are presented.

The data in Fig. 2 were subjected to a three-factor (three delay-types × three durations × three rates of modulation), within-subjects analysis of variance (ANOVA). The error terms for the main effects and for the interactions were the interaction of the particular main effect (or the particular interaction) with the subject "factor" (Keppel, 1973).[5] The data in Fig. 2 indicate that, in general, regardless of the rate of modulation or the duration, thresholds obtained with the $\Delta|0$ delay-type (squares) were highest having a mean (collapsed across both duration and modulation frequency) of 5.2. Thresholds obtained with the $\Delta|\Delta$ delay-type (triangles) were the lowest, with a mean of 1.6. Excluding the single condition, discussed above, in which two subjects could not perform the task, thresholds obtained with the $0|\Delta$ delay-type (circles) were also relatively low, having a mean value of 2.2. The main effect of delay-type was found to be significant [$F(2,6)=69.93$, $p<0.001$].

These outcomes indicate that, for high-frequency SAM tones, the listeners were substantially more sensitive to interaural delays within the ongoing envelope of the waveform than they were to interaural delays restricted to the onsets/offsets of the waveforms. Recall that the $\Delta|0$ and the $0|\Delta$ delay-types are conditions in which the onset/offset and ongoing interaural delays are inconsistent. In the former condition, the listener's task is to detect the presence of a nonzero interaural delay in only the onset/offset portions of the waveform in the face of an interaural delay of zero in the ongoing portion. In the latter case, the listener's task is to detect the presence of a nonzero interaural delay in only the ongoing portion of the waveform in the face of an interaural delay of zero in the onset/offset of the waveform. The fact that thresholds were substantially lower for the $0|\Delta$ delay type as compared to the $\Delta|0$ delay type indicates that ongoing interaural delays dominated the onset/offset delays. That dominance, however, while substantial, is not absolute. This is evidenced by the fact that thresholds obtained with the $\Delta|0$ delay-type (squares), although relatively large, were still measurable. The dominance of ongoing interaural envelope delays over onset/offset delays is consistent with a similar dominance of ongoing *fine-structure* delays measured by previous investigators who used stimuli containing energy at low frequencies (e.g., Tobias and Schubert, 1959; Perrott and Baars, 1974; Hafter, Dye, and Gilkey, 1979; Kunov and Abel, 1981; Abel and Kunov, 1983).

Focusing on the top row of Fig. 2 and the $\Delta|0$ delay-type (squares), the data indicate that normalized thresholds *increased* with duration. Specifically, normalized thresholds obtained with a duration of 240 ms were 1.2–1.5 times greater (depending on the rate of modulation) than those obtained with a duration of 32 ms. On the other hand, for both the $0|\Delta$ (circles) and $\Delta|\Delta$ delay–types (triangles), thresholds *decreased* with increases in duration. Specifically, thresholds measured with the duration of 240 ms were about 0.4–0.7 of those measured when the duration was 32 ms. Thus, the curves within each panel of the top row of Fig. 2 tend to diverge with increasing duration indicating that the degree of dominance of the ongoing interaural envelope delays increases with the duration of the waveform. This patterning of the data is also consistent with the earlier studies which showed that the degree of dominance of ongoing *fine-structure-based* delays increased with the duration of the stimulus. The outcomes based on this visual inspection of the data were borne out by the ANOVA which indicated a significant interaction between delay–type and duration [$F(4,12)=17.86$, $p<0.001$]. There was also a significant main effect of duration [$F(2,6)=6.22$, $p=0.034$].

The finding that the dominance of interaural delays within the ongoing portions of the waveform occurs, not only when the ongoing interaural delays are conveyed by the low-frequency, fine-structure, but also when the ongoing interaural delays are conveyed by the envelopes of the waveforms is particularly noteworthy. This is so because thresholds measured when ongoing interaural delays are restricted to the envelopes of SAM waveforms centered at 4 kHz are substantially poorer than those measured when the ongoing interaural delays are restricted to the fine structure of low-frequency waveforms (e.g., Blauert, 1982, 1983; Grantham, 1995; Bernstein, 2001). Despite this difference in acuity, ongoing interaural delays conveyed by the envelopes of our high-frequency SAM waveforms still greatly dominated interaural delays within the onset/offsets.

In the bottom row of Fig. 2, the thresholds are re-plotted in terms of normalized interaural delay as a function of the rate of modulation of the stimulus. Considering the $\Delta|0$ delay type (squares), the data indicate that normalized thresholds *decreased* with increases in rate of modulation. Specifically, thresholds obtained with a rate of modulation of 350 Hz were about 0.6–0.8 (depending on the duration) of those measured when the rate of modulation was 125 Hz. On the other hand, for both the $0|\Delta$ (circles) and $\Delta|\Delta$ delay-types (triangles), thresholds *increased* with increases in rate of modulation. Specifically, thresholds measured with the rate of modulation of 350 Hz were about 1.3–2.3 greater than those measured with a rate of modulation of 125 Hz. The ANOVA indicated that the main effect of rate of modulation was not significant [$F(2,6)=0.62$, $p=0.571$]. That main effect was, apparently, obscured by the highly significant interaction between delay-type and rate of modulation [$F(4,12)=14.15$, $p<0.001$].

The convergence of the curves with increasing rate of modulation within each panel of the bottom row of Fig. 2 indicates that the degree of dominance of the ongoing interaural envelope delays decreased with rate of modulation. This finding, while novel, was not unexpected because previous studies have shown that thresholds for ongoing envelope-based ITDs within high-frequency waveforms increase substantially when the rate of modulation is increased beyond about 150 Hz.

C. Discussion

The data in Fig. 2 appear to be amenable to an intuitive, qualitative account. They indicate that thresholds measured when changes in interaural delay were restricted to the onset/

offset portion of the waveform ($\Delta|0$, squares) were affected by manipulations (duration and rate of modulation) of only the steady-state, diotic *ongoing* portion of the waveform, the portion that contained no binaural information relevant to solving the behavioral task. Thus, while listeners were able to utilize interaural delays within the onset/offset portions of the waveform, their judgments were still influenced by the steady-state, diotic ongoing portions. Furthermore, when the interaural delay to be detected was present within *both* the onset/offset and the ongoing portions of the waveform ($\Delta|\Delta$, triangles), thresholds were somewhat smaller than those obtained when the interaural delay was restricted to the ongoing portion of the waveform ($0|\Delta$, circles). These two outcomes support the notion that information stemming from the two types of delay was combined.

It occurred to us that that the notion that information from the two types of delay is combined could be evaluated via a formal, quantitative information-based description of detection. Specifically, we sought to determine whether thresholds measured with the $\Delta|\Delta$ delay type could be predicted on the basis of a combination of the thresholds obtained with the $\Delta|0$ and the $0|\Delta$ delay types at each pairing of duration and rate of modulation tested, assuming that each type of information is derived from an "independent channel." Following Green and Swets (1974), we assume that

$$d'_{i+j} = \sqrt{d_i'^2 + d_j'^2}, \tag{1}$$

where i refers to the onset/offset interaural delay, j refers to the ongoing interaural delay and $i+j$ refers to the condition in which both cues are present (the $\Delta|\Delta$ delay-type).

Taking into account the fact that a single value of ITD was applied to the onset/offset and ongoing interaural delays in the $\Delta|\Delta$ condition, Eq. (1) can be re-written as

$$\frac{ITD}{\sigma_{i+j}} = \sqrt{ITD^2 * \left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right)}. \tag{2}$$

$$\frac{ITD^2}{\sigma_{i+j}^2} = ITD^2 * \left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right). \tag{3}$$

$$\frac{1}{\sigma_{i+j}^2} = \frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}. \tag{4}$$

$$\sigma_{i+j} = \frac{1}{\sqrt{\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}}}. \tag{5}$$

Keeping in mind that thresholds measured in each condition are estimates of the corresponding values of σ, Eq. (5) becomes

$$ThreshITD_{i+j} = \frac{1}{\sqrt{\frac{1}{ThreshITD_i^2} + \frac{1}{ThreshITD_j^2}}} \tag{6}$$

Recall that, for all the delay-types tested, both onset/offset and ongoing interaural delays were present (whether zero or nonzero). Thus, it was not possible to obtain estimates of thresholds for *only* onset/offset or *only* ongoing interaural delays as called for by the right-hand side of Eq. (6) (the terms $ThreshITD_i$ and $ThreshITD_j$, respectively). In the case of the onset/offset ($\Delta|0$) delay-type, we took as the best estimate of $\text{ThreshITD}_\text{i}$, the measurement obtained for the $\Delta|0$ delay type at a duration of 32 ms and a rate of modulation of 350 ms. This condition was chosen because it was expected that the influence of the ongoing delay would be minimized at this shortest duration and highest rate of modulation. This expectation was borne out by the data in Fig. 2 indicating that obtained thresholds at this pairing of duration and rate of modulation were among the smallest for the $\Delta|0$ delay type (squares) and among the largest for the $0|\Delta$ delay type (circles).

For the ongoing delay type ($0|\Delta$), we did not use a single estimate for ThreshITD_j in Eq. (6). Rather, for each pairing of duration and rate of modulation, we substituted the threshold obtained in the ($0|\Delta$) condition. This was done based on the reasonable assumption that the changes in threshold depicted in Fig. 2 that occur with changes in either duration or rate of modulation, and which were expected, stem from changes in the processing variance associated with the interaural delays within the ongoing portion of the waveform [σ_j in Eq. (5)]. In addition, because thresholds with the $0|\Delta$ delay type were, overall, about three times smaller than those obtained with the $\Delta|0$ delay type, thresholds obtained with the $0|\Delta$ delay type are, themselves, relatively pure estimates of $\text{ThreshITD}_\text{j}$.

Equation (6) was used to derive predictions of threshold for the $\Delta|\Delta$ delay type separately for each listener. Each of those predictions was normalized in the same manner as described for the data in Fig. 2. Finally, the normalized predictions were averaged across the four listeners. For the single condition in which two of the listeners could not perform the task ($0|\Delta$ delay-type with 32 ms/350 Hz pairing), the prediction was based on the thresholds obtained from the remaining two listeners.

The predictions for the $\Delta|\Delta$ delay type are plotted as small open triangles within each plot in Fig. 2. Overall, across the durations and rates of modulations tested, the predictions capture the trends in the data quite well but slightly overestimate the thresholds obtained. One important characteristic of the predicted thresholds for the $\Delta|\Delta$ delay type is that they, like the thresholds obtained, are smaller for every combination of duration and rate of modulation tested, than either of the corresponding thresholds obtained with the $\Delta|0$ and $0|\Delta$ delay types. This is most easily observed for the 32 ms/250 Hz pairing of duration and rate of modulation where the thresholds obtained with the $\Delta|0$ and $0|\Delta$ delay-types were most comparable.

The overestimation of thresholds, which was consistently observed when similar analyses were performed on the data from individual listeners, may result from our inability to derive "pure" estimates of the processing variance associated with the onset/offset interaural delay *in isolation* and the ongoing interaural delay *in isolation*. In particular, recall that the threshold measured for the $\Delta|0$ delay-type and the 32 ms/350 Hz pairing was used as the estimate of $\text{ThreshITD}_\text{i}$. That condition was chosen because it was expected that the influence of the ongoing delay would be

minimized at that shortest duration and highest rate of modulation. As discussed earlier, the data in Fig. 2 supported that expectation. Specifically, at that pairing of duration and rate of modulation, thresholds for the $\Delta|0$ delay type were among the lowest, while thresholds for the $0|\Delta$ delay type were among the largest. It is the case, however, that for the two listeners for whom thresholds could be obtained for the $0|\Delta$ delay-type, those thresholds were comparable to those measured for the $\Delta|0$ delay type (open circle and solid square in top-right panel of Fig. 2). Thus, at least for those two listeners, thresholds measured for the $\Delta|0$ delay type, may have been somewhat inflated by the presence of the steady-state, diotic ongoing portion of the waveform. To the degree that was the case, our estimate of ThreshITD_i would also be inflated. That could account for the overestimations of thresholds by the model for the $\Delta|\Delta$ delay type.

Finally, it is important to note that had the underlying assumption of independence between the two types of interaural-delay-based information been essentially incorrect, then our predicted thresholds would have *underestimated* the thresholds obtained. Across all conditions tested, the predictions account for 37% of the variance in the data.[6]

In a second analysis, we considered to what extent all of the data could be accounted for by simply assuming that the listener requires a constant criterion reduction in the interaural correlation of the otherwise diotic stimulus in order for detection to occur. Such a change could be brought about by sufficiently large interaural delays of the onset/offset, the ongoing, or both portions of the waveform. Because each type of delay contributes to the overall interaural correlation of the waveform, it is also possible that the listeners' apparent combination of interaural onset/offset delays and interaural ongoing delays was a manifestation of their use of that single metric in solving the behavioral task. To explain, assume that the interaural correlation is computed across the entire duration of the stimulus. Because onset/offset interaural delays are relatively short lived compared to ongoing interaural delays, the latter would be expected to be more potent, or dominant, in terms of determining the interaural correlation of the waveform. This is consistent with our empirical finding that thresholds were greatest for the $\Delta|0$ delay type (Fig. 2).

Furthermore, by that same logic, the relative dominance of the ongoing interaural delay would be expected to be directly related to the duration of the stimulus. This is consistent with the data indicating that thresholds for the $0|\Delta$ delay type decrease with duration while thresholds for the $\Delta|0$ delay type generally increase with duration. In the latter case, the greater contribution, with duration, of the steady-state, diotic ongoing portion of the stimulus would serve to dilute the contribution of the interaural delay within the short-lived onset/offset portion.

The qualitative explanations offered above were evaluated quantitatively via the same correlation-based model used successfully to account for a variety of other binaural data (Bernstein *et al.*, 1999; Bernstein and Trahiotis, 2002; 2003; Trahiotis *et al.*, 2001). Briefly, the model incorporates bandpass filtering, implemented via a gammatone filter (CF =4 kHz), half-wave, square-law rectification, envelope compression (exponent=0.23), and two stages of low-pass filtering. The first stage, with a cutoff of 425 Hz, was designed to capture effects stemming from the loss of neural synchrony to the fine-structure of the stimulus that occurs as the center frequency is increased (Bernstein and Trahiotis, 1996). The second stage of low-pass filtering, with a cutoff of 150 Hz, was designed to capture effects stemming from an apparent envelope "rate limitation" that serves to "smooth" the fluctuations of the envelopes and to limit monaural and binaural performance at high rates of modulation (see Bernstein and Trahiotis, 2002).

In order to derive predictions for the data, it was necessary to determine, for each stimulus utilized in the experiments, the function relating ITD to the normalized interaural correlation. Fifty tokens of each stimulus were employed for which the starting phases of the carrier and of the modulator were each chosen randomly. Numerical measures were obtained by implementing the peripheral stages of the model within MATLAB© and then computing the normalized interaural correlation between the model's "left" and "right" outputs for ITDs between 0 and 1000 μs in 50 μs steps. Then, using a least-squares criterion, polynomials were fit to paired values consisting of the mean value of the normalized correlation (across the 50 tokens) and ITD.

For each listener, we sought the single value of normalized interaural correlation that maximized the amount of variance accounted for between predicted and obtained values of threshold-ITD across all conditions. Then, again, for each listener, predicted *normalized* thresholds were computed by dividing each predicted value of threshold-ITD by the threshold-ITD obtained for the condition that served as the reference in Fig. 2 ($\Delta|\Delta$ delay-type, 240 ms/125 Hz pairing of duration and rate of modulation). Finally, the normalized predicted thresholds were averaged, condition by condition, across the four listeners. As was the case for the predicted thresholds plotted in Fig. 2, for the single condition for which two of the four listeners could not perform the task ($0|\Delta$ delay type, 32 ms/350 Hz pairing of duration and rate of modulation), their data were excluded from the analysis.

Figure 3 displays the obtained normalized thresholds (top row, re-plotted from Fig. 2) and the corresponding thresholds predicted by the interaural correlation-based model (bottom row). The left-hand, middle, and right-hand columns contain the plots for the $\Delta|0$, $0|\Delta$, and $\Delta|\Delta$ delay types, respectively. In each plot, thresholds are plotted as a function of rate of modulation. The parameter in each plot is the duration of the stimulus. Asterisks plotted at the left-most portion of each abscissa indicate the mean of the values within each corresponding plot. The particular organization of Fig. 3 was chosen because it clearly reveals the strengths and weaknesses of the interaural correlation-based model which accounted for 42% of the variance across the entire set of 27 conditions.

Comparing the positioning of the asterisks in the top and bottom rows of Fig. 3 reveals that the model predicts quite well both the absolute value and ordering of the mean values of the obtained data for the $\Delta|0$, $0|\Delta$, and $\Delta|\Delta$ delay types. The model's success in this regard can be understood intuitively by noting that as the portion of stimulus containing

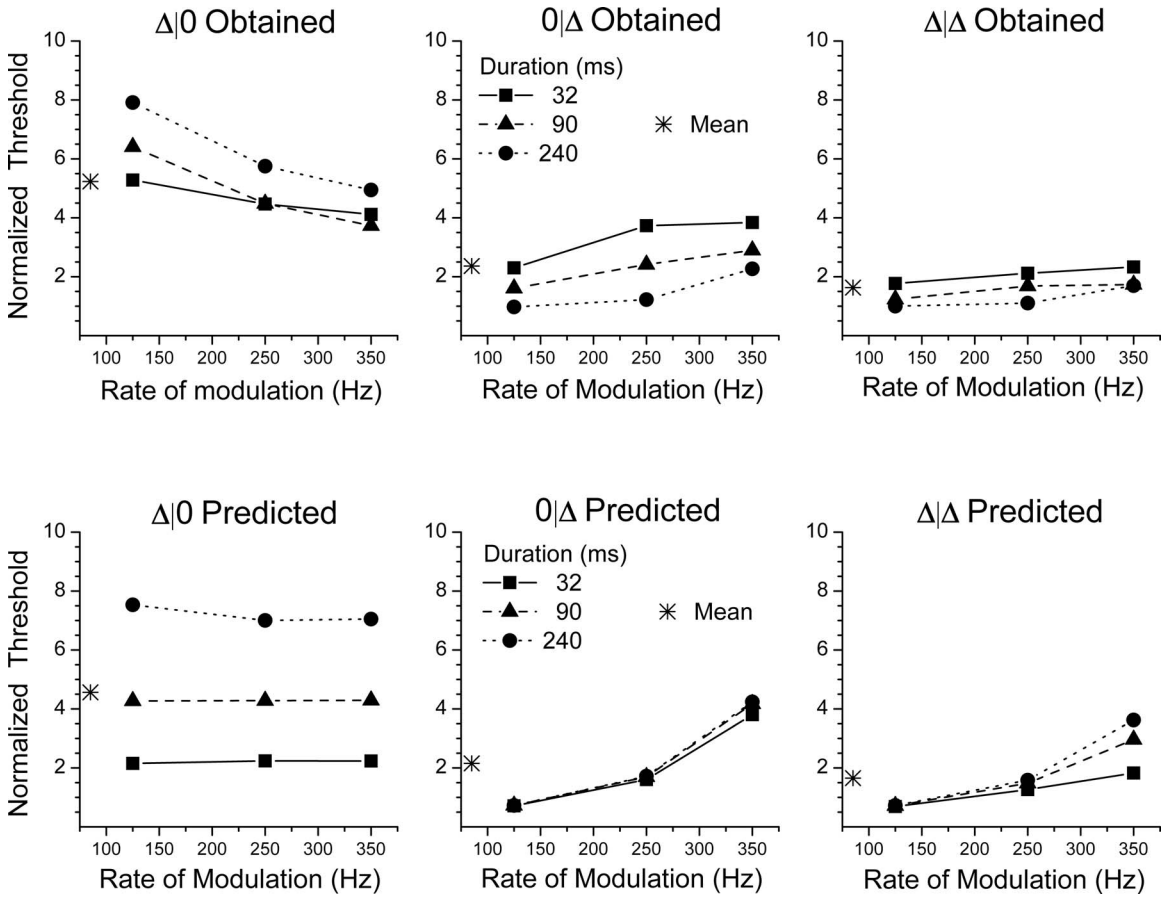


FIG. 3. Obtained normalized thresholds (upper row, re-plotted from Fig. 2) and normalized thresholds predicted by the interaural correlation-based model (lower row). The left-hand, middle, and right-hand columns contain the plots for the $\Delta|0$, $0|\Delta$, and $\Delta|\Delta$ delay-types, respectively. The parameter in each plot is the duration of the stimulus. Asterisks plotted at the left-most portion of each abscissa indicate the mean of the values within each corresponding plot.

diotic information is reduced (resulting in less "dilution" of the ITD imposed), thresholds are correctly predicted to decrease.

The left-hand upper and lower panels reveal that, while the model does a less-than-satisfactory job of capturing the absolute value of the thresholds for the $\Delta|0$ delay-type, it does correctly predict the trend that thresholds decrease with decreases in the duration of the stimulus. When the interaural delay is restricted to the onset/offset portion of the waveform, decreasing the duration of the stimulus decreases the proportion of the diotic information contained in the ongoing portion and so, within the model, leads to a given value of ITD producing a greater amount of interaural decorrelation.

The middle and right columns reveal that, for both the $0|\Delta$ and $\Delta|\Delta$ delay types, the model fails to capture the decreases in threshold observed with increases in the duration of the stimulus. Effectively, what the model fails to capture is the increase in sensitivity with duration to ITDs within the ongoing portion of the waveform. In our view, it would be necessary to incorporate into the model specific assumptions concerning the time over which the listener integrates binaural information and some form of internal noise in order to account for that trend.

Finally, we turn to an assessment of the ability of the model to account for the changes in threshold that were observed with changes in rate of modulation. For the $0|\Delta$ and $\Delta|\Delta$ delay types (middle and right-hand columns), the model correctly predicts the overall increase in thresholds as rate of modulation was increased. The model, however, predicts a much steeper rise in thresholds than was observed when the rate of modulation was increased from 250 to 350 Hz. Recall that the correlation-based model employs a 150 Hz low-pass filter designed to capture listeners' decreased sensitivity to ongoing interaural envelope delays at high rates of modulation. Overall, our listeners were less affected by increasing the rate of modulation with the $0|\Delta$ and $\Delta|\Delta$ delay types (both of which contain ongoing interaural delays) than the model predicted.

Turning to the $\Delta|0$ delay-type (left-hand column), it can be seen that the model fails to account for the decrease in thresholds that were observed when the rate of modulation was increased. This failure occurs because, within the model, the effective interaural correlation within the ongoing portion of the waveform remains at 1.0, regardless of the rate of modulation. To explain, when the ongoing portion of the waveform contains a *nonzero ITD*, the effect of the 150 Hz low-pass filter within the model is to drive the "internal" interaural correlation toward 1.0 in direct proportion to the extent that the rate of modulation exceeds 150 Hz. If the physical ITD imposed within the ongoing portion of the waveform is zero, the interaural correlation within that portion will be 1.0 and will remain so regardless of the rate of

TABLE I. The entries within the body of the table are the obtained normalized thresholds, their corresponding values predicted via an interaural-correlation-based model, and the differences between them. The right-most column and bottom-most row indicate the percentage of variance accounted for by the model when the entries are collapsed across the three rates of modulation or durations, respectively. The entry "N/A" indicates that the percentage of the variance accounted for by the model was less than the percentage that would be accounted for if the grand mean were used as a single predictor.

			Rate of Modulation									
			125 Hz			250 Hz			350 Hz			
			Obtained	Predicted	Obtained-Predicted	Obtained	Predicted	Obtained-Predicted	Obtained	Predicted	Obtained-Predicted	% Var. Acc:
Duration	32 ms	Δ|0	5.28	2.16	3.13	4.47	2.24	2.23	4.11	2.24	1.88	31
		0|Δ	2.30	0.71	1.59	3.73	1.60	2.13	3.84	3.80	0.04	
		Δ|Δ	1.77	0.69	1.08	2.12	1.26	0.85	2.33	1.83	0.50	
	90 ms	Δ|0	6.41	4.27	2.14	4.48	4.28	0.20	3.73	4.29	−0.56	59
		0|Δ	1.61	0.73	0.88	2.42	1.70	0.72	2.90	4.17	−1.27	
		Δ|Δ	1.23	0.72	0.52	1.68	1.48	0.20	1.73	2.96	−1.23	
	240 ms	Δ|0	7.91	7.53	0.38	5.75	7.00	−1.25	4.94	7.05	−2.11	55
		0|Δ	0.97	0.73	0.25	1.22	1.72	−0.50	2.27	4.24	−1.97	
		Δ|Δ	1.00	0.72	0.28	1.10	1.58	−0.48	1.70	3.62	−1.93	
% Var. Acc:				65			42			N/A		

modulation. It would appear that specific assumptions concerning processing noise would be required in order for the model to correctly predict the observed changes in threshold with rate of modulation for the Δ|0 delay type.

Table I contains additional detail regarding the predictions of the interaural-correlation-based model. For each of the 27 conditions, the table displays the mean obtained normalized threshold, the mean predicted normalized threshold, and the difference between the two (obtained threshold minus predicted threshold). Positive values of the difference represent underestimates of obtained thresholds; negative values represent overestimates of obtained thresholds.

In summary, our notion was that the entire set of data in Fig. 2 could be accounted for by simply assuming that listeners' detection is based on a constant criterion reduction in the interaural correlation of the otherwise diotic stimulus. We reasoned that this might be true regardless of the type(s) of interaural delay (onset/offset, ongoing, or both) imposed. The fact that, across all conditions, the model accounts for as much as 42% of the variance supports that notion. In our view, the modest success of the model is especially noteworthy considering that its parameters, as applied here, were developed on the basis of previously published sets of binaural data, it contains no specific assumptions regarding the time over which listeners integrate the binaural cues, and it does not incorporate any form of "internal noise." Such enhancements to the model would appear necessary in order to redress its shortcomings.

III. EXPERIMENT 2

The results of Experiment 1 showed that the degree of dominance of the ongoing interaural delay can be varied by changes in the duration and rate of modulation of the ongoing portion of the waveform. It appears that changes in the degree of dominance can be viewed as arising from changes in the relative sensitivity to *ongoing* interaural delays that, in turn, affect the degree to which they dominate onset/offset delays. This notion is supported by the data in Fig. 2 showing that the changes in duration and rate of modulation that result in *decreased* thresholds for the ongoing delay type are the same changes that lead to *increased* thresholds for the onset/offset delay type.

Buell *et al.* (1991) measured the extents of laterality produced by combinations of onset/offset and ongoing delays using low-frequency tones. They also observed dominance of ongoing interaural delays over onset/offset interaural delays. Interestingly, Buell *et al.* found that when the ongoing delay was "ambiguous," in that it did not clearly favor either ear (it contained an interaural delay corresponding to a phase shift at or near 180° or π radians), it was the *onset/offset interaural delay* that dominated the laterality of the intracranial image. In order to assess the generality of that finding, we sought to determine whether an ambiguous interaural phase-shift imposed on the *ongoing envelope* of our high-frequency SAM tones would also enhance listeners' sensitivity to changes in onset/offset interaural delays.

An additional technique employed in this second experiment was to measure listeners' sensitivity to changes in onset/offset interaural delay while roving the value of the ongoing delay from interval to interval. The goal was to determine whether and to what degree adding variability to the value of the ongoing delay would alter sensitivity to interaural delays in the onset/offset portions of the waveform. Finally, a condition was employed in which listeners' sensitivities to changes in ongoing interaural delay was measured while roving the value of the onset/ongoing delay from interval to interval. Only two pairings of duration and rate of modulation were employed: 32 ms/350 Hz and 240 ms/125 Hz. These were chosen because, as demonstrated by the results of Experiment 1, the former pairing combines the duration and rate of modulation that yielded the least dominance of the ongoing interaural delay over the onset/offset interaural delay, while the latter pairing combines the duration and rate of modulation that yielded the greatest dominance.

A. Procedure

Using the same techniques employed in Experiment 1, detection of interaural time delay was measured using sinusoidally amplitude modulated (SAM) tones with carrier frequencies of 4 kHz. Rates of modulation were 125 and 350 Hz and were coupled with durations of 240 and 32 ms, respectively. For the $\Delta|\pi$ delay-type (panel D of Fig. 1), the interaural delay to be detected was restricted to the onset/offset portion of the waveform while the ongoing sinusoidal envelope contained a time-delay equivalent to an interaural phase-shift of π radians. This delay-type was created by first applying a time-delay to the waveform equivalent to one-half the period of the envelope. Then, the onset/offset interaural delay was created by gating asynchronously the waveforms to the two earphones. Within the two-cue, two-alternative, forced choice, adaptive task, the stimuli within all four intervals contained the time-delayed ongoing envelope. Only the stimulus within the "signal" interval contained the additional onset/offset interaural delay that was to be detected. Thus, in contrast to what was the case in Experiment 1, the "nonsignal" and cueing intervals did not contain diotic stimuli.

A second delay-type, $\Delta|R$ was also employed. Once again, the interaural delay to be detected was restricted to the onset/offset portion of the waveform. This delay-type was created in the same manner as the $\Delta|\pi$ delay-type with the following exception. The value of the ongoing interaural envelope delay was chosen randomly (roved) from interval to interval rather than being fixed at half the period of the envelope (Fig. 1, panel E). The value of the ongoing interaural envelope delay was chosen with equal *a priori* probability from a range of ± 500 μs in 10 μs steps. The result was that, within the two-cue, two-alternative, forced choice, adaptive task, the value of the ongoing interaural envelope delay was randomly chosen anew for each of the four intervals. Only the stimulus within the "signal" interval contained the additional onset/offset interaural delay that was to be detected.

A third delay-type was employed in which the interaural delay to be detected was restricted to the *ongoing* portion of the waveform. For this $R|\Delta$ delay-type, (Fig. 1, panel F), the roles of the onset/offset and ongoing interaural delays were reversed relative to those described immediately above for the $\Delta|R$ delay-type. The $R|\Delta$ delay-type was created by applying the desired ongoing interaural delay to the waveform and then gating asynchronously the waveforms at the two ears to create the onset/offset interaural delay. The magnitude of the onset/offset interaural delay was drawn randomly for each interval with equal *a priori* probability from a range of ± 500 μs in 10 μs steps.

Data were obtained from the same four listeners who participated in Experiment 1. Thresholds for the $R|\Delta$ delay-type were obtained along with and had been interleaved with the conditions explored in Experiment 1. For the $\Delta|\pi$ and $\Delta|R$ delay types, the ordering of conditions was determined by choosing randomly among the two types of interaural delay, and the two pairings of rate of modulation and duration. The same ordering of conditions was employed for all four listeners. Twelve thresholds for each condition were measured using the same strategy described for Experiment 1. In three out of 24 cases (six conditions $\times$ four listeners) the original set of 12 estimates contained extremely high estimates of threshold (e.g., 2000–4000 μs) among substantially lower values. For those cases, one to two additional sets of three estimates of threshold were collected. In seven of the 24 cases, the listener was unable to perform the task and collection of data for that particular condition was terminated. As was the case for Experiment 1, for each combination of listener and condition, only the final 12 estimates of threshold were considered. Of those 12, any estimate exceeding 1500 μs was excluded. After excluding estimates of threshold that were greater than 1500 μs, the two highest and the two lowest of the remaining estimates were also excluded. The mean of the resulting set of estimates (typically numbering eight but no fewer than six) served as the final estimate of threshold for the particular listener and stimulus condition. Finally, the data were normalized by using the same reference thresholds as those used in Experiment 1.

B. Results and discussion

Each of the four panels of Fig. 4 contains a pair of plots depicting mean normalized thresholds obtained from a single listener. The error bars represent $\pm$one standard deviation about the mean. The data are plotted in this manner because, (1) in three out of the ten conditions plotted, one or more listeners could not perform the task and (2) there were substantial inter-individual differences across conditions. This heterogeneity across listeners meant that the average of their thresholds would not have been representative of the outcomes. The upper plot within each panel contains the data (hatched bars) obtained when the interaural delay to be detected was restricted to the onset/offset portion of the waveform. The lower plot within each panel contains the data (open bars) obtained when the interaural delay to be detected was restricted to the ongoing portion of the waveform.

Normalized thresholds obtained with the duration/rate combinations of 32 ms/350 Hz and 240 ms/125 Hz are plotted as groups of bars in the left and right portions of each plot, respectively. An arrow atop a bar indicates that the threshold for that condition was unmeasurable. The symbol below each bar represents the type of interaural delay employed. The left-most bar within each group ($\Delta|0$ or $0|\Delta$ delay-types) represents data obtained from the individual listeners in Experiment 1.

We begin by focusing on the data obtained when the interaural delay to be detected was restricted to the onset/offset portion of the waveform (hatched bars, upper plots of each panel). Thresholds for detecting a change in the onset/offset interaural delay were greater when the stimulus was relatively long (240 ms) and the rate of modulation was relatively low (125 Hz) as compared to when the stimulus was relatively short (32 ms) and the rate of modulation was high (350 Hz). This outcome is consistent with the results of Experiment 1 and the notion that the ongoing interaural delay was more dominant for the 240 ms/125 Hz pairing than for the 32 ms/350 Hz pairing of duration and rate of modulation. It should be recognized, however, that all of the thresh-

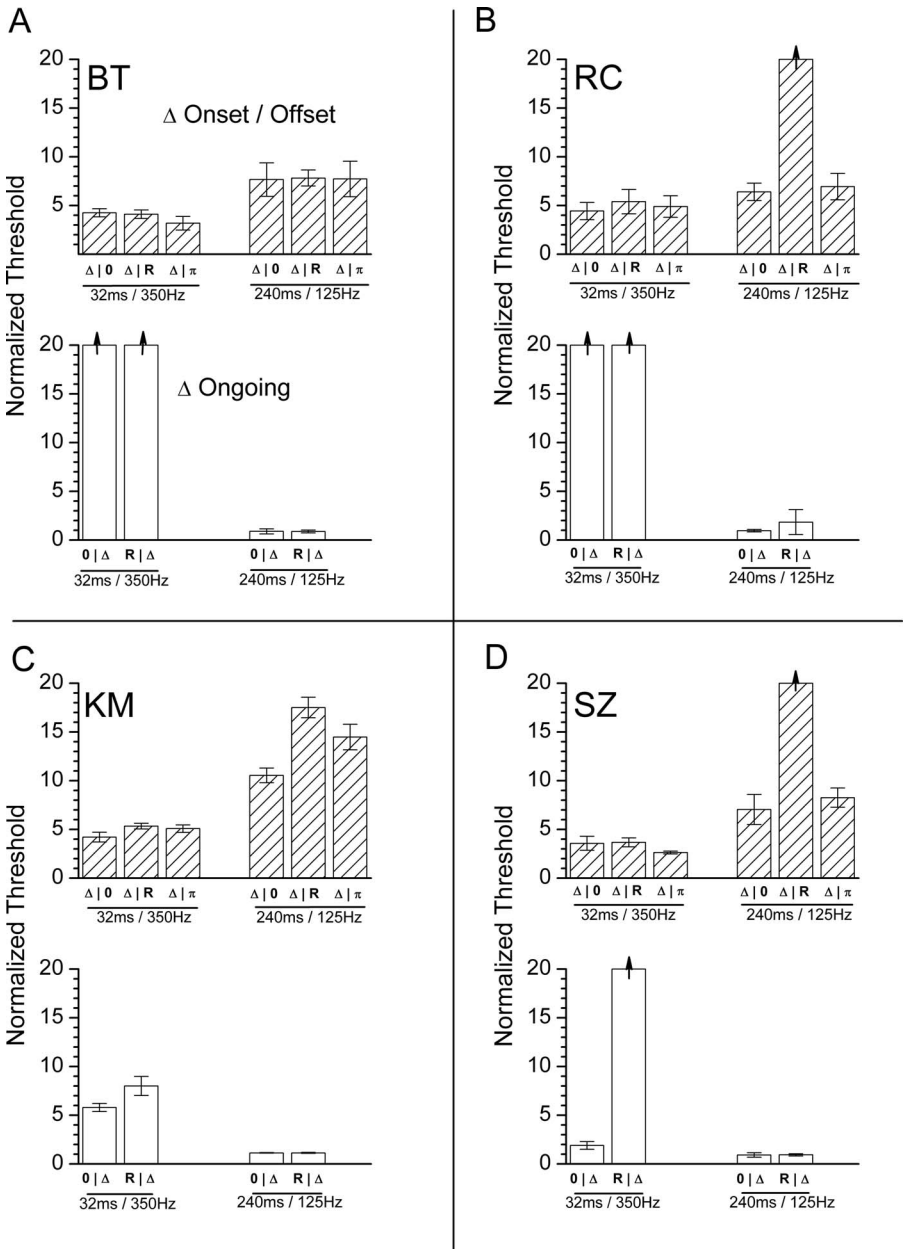


FIG. 4. Normalized thresholds plotted as a function of stimulus condition. Each of the four panels displays the averaged data obtained from an individual listener. Error bars represent ± one standard deviation about the mean. The upper plot and lower plots contain the data obtained when the interaural delay to be detected was restricted to the onset/offset portion of the waveform (hatched bars) or the ongoing portion of the waveform (open bars), respectively. The left-most bar within each group ($\Delta|0$ or $0|\Delta$ delay types) represents data obtained from the individual listeners in Experiment 1. Thresholds in units of microseconds can be derived by multiplying the normalized thresholds depicted for each listener by that listener's reference threshold obtained in experiment 1 for the $\Delta|\Delta$ delay-type and with the 240 ms/125 Hz pairing of duration and rate of modulation. The reference thresholds obtained from the individual listeners were 148 μs for listener BT, 177 μs for listener RC, 64 μs for listener KM, and 134 μs for listener SZ.

olds obtained when the interaural delay to be detected was restricted to the onset/offset portion of the waveform were relatively poor being, at least, about four times the thresholds measured in the reference condition.

For the 32 ms/350 Hz pairing of duration and rate of modulation and for each listener, thresholds for detecting onset/offset interaural delays were essentially equivalent regardless of whether the ongoing portion of the waveform contained no interaural delay, was roved, or was interaurally phase-shifted by π radians. This outcome is not surprising in light of the apparently diminished dominance of the ongoing interaural delay that was demonstrated in Experiment 1 for this short duration and high rate of modulation (see Fig. 2). Perhaps the very short duration of the stimulus coupled with its high rate of modulation reduced sufficiently listeners' sensitivities to binaural cues within the ongoing portion of the waveform such that manipulations of its interaural characteristics were essentially inconsequential.

Such an outcome would not be expected when a relatively long duration and low rate of modulation were employed (e.g., the 240 ms/125 Hz pairing). This is so because, under such conditions, the dominance of interaural delays within the ongoing portion of the waveform is relatively enhanced. Thus, manipulations of its interaural characteristics might be more effectively conveyed. Indeed, for two of the four listeners, (RC and SZ), thresholds were immeasurable for the ($\Delta|$R) delay-type when the ongoing interaural delay was roved ($\Delta|$R—middle bar within a trio). For listener KM, the threshold measured with the $\Delta|$R delay-type was substantially elevated relative to that obtained with the $\Delta|0$ delay-type. For listener BT, however, the thresholds obtained with the ($\Delta|$R) delay-type were essentially identical to that obtained with the $\Delta|0$ delay-type.

Thresholds measured with the $\Delta|\pi$ delay-type (right-most bar within a trio) for the 240 ms/125 Hz pairing were,

for three of the four listeners, similar to those obtained with the $\Delta|0$ delay-type (left-most bar within a trio). On the other hand, for listener KM, the normalized threshold measured with the $\Delta|\pi$ delay-type was somewhat larger than that measured in the $\Delta|0$ delay-type (14.5 vs.10.6). Thus, counter to what one might have expected on the basis of Buell *et al.*'s (1991) measures of the laterality of low-frequency tones, imposing an "ambiguous" interaural delay on the ongoing envelope of our high-frequency SAM tones did not, overall, enhance listeners' sensitivity to onset/offset interaural delays, much less render them dominant.

We now turn our attention to the data obtained when the interaural delay to be detected was restricted to the ongoing portion of the waveform (open bars, lower plots of each panel). Thresholds for detecting a change in the ongoing interaural delay were, overall, markedly smaller when the stimulus was relatively long (240 ms) and the rate of modulation was relatively low (125 Hz—right-most pair of bars in each plot) as compared to when the stimulus was relatively short (32 ms) and the rate of modulation was high (350 Hz—left-most pair of bars in each plot). Once again, this is consistent with the results of Experiment 1 and with the notion that the ongoing interaural delay was more dominant for the 240 ms/125 Hz pairing than for the 32 ms/350 Hz pairing of duration and rate of modulation.

For the 240 ms/125 Hz pairing of duration and rate of modulation and for each listener, thresholds for detecting ongoing interaural delays were essentially equivalent regardless of whether the interaural delay within the onset/offset portion contained no interaural delay $(0|\Delta)$ or was roved $(R|\Delta)$. This finding is not surprising in light of the expected and demonstrated enhanced dominance of the ongoing interaural delay under these conditions (see Fig. 2). Perhaps the long duration of the stimulus coupled with its low rate of modulation resulted in so great a dominance of the interaural delays within the ongoing portion of the waveform that manipulations of the interaural characteristics of the onset/offset portion were essentially inconsequential.

Such an outcome would not necessarily be expected when a relatively short duration and high rate of modulation was employed (e.g., the 32 ms/350 Hz pairing). This is so because, as demonstrated by Experiment 1, under such conditions, the dominance of interaural delays within the ongoing portion of the waveform is reduced. Listener SZ's data obtained with the 32 ms/350 Hz pairing reveal that her threshold was relatively small when the onset/offset interaural delay was zero $(0|\Delta)$ but was unmeasurable when the onset/offset interaural delay was roved $(R|\Delta)$. For listener KM, the elevation of threshold measured with the roved onset/offset $(R|\Delta)$ delay-type was more modest (8.0 versus 5.8). For listeners BT and RC, a similar comparison cannot be made because thresholds for both conditions (zero-valued and roved onset/offset delays) were unmeasurable. Note that it is the heterogeneity of thresholds in the $(0|\Delta)$ condition measured in Experiment 1 that led to the plotting of separate means (open circles) in two panels of Fig 2.

Overall, the results of Experiment 2 indicate that for high-frequency SAM waveforms, imposing an "ambiguous" ongoing interaural delay equivalent to one-half period of the envelope had little impact on listeners' sensitivity to interaural delays within the onset/offset portion of the waveform. This finding was contrary to expectations based on the findings of Buell *et al.* (1991). A number of factors may underlie this apparent discrepancy in the outcomes across the two studies. First, it is important to recall that Buell *et al.* measured *extents of laterality* produced by combinations of onset/offset and ongoing ITDs and not thresholds for discrimination of a change of ITD as was done here. The lack of a consistent outcome between our data and those of Buell *et al.*, may represent yet another example of the finding that thresholds for *detection* of an ITD cannot be accurately predicted on the basis of its potency in terms of the extent of laterality it produces (e.g., Stern and Colburn, 1978; Trahiotis *et al.*, 2001). In addition, it may be the case that the manner in which onset/offset and ongoing ITDs interact at magnitudes that are supra-threshold (as was the case in the study of Buell *et al.*) may differ substantially from the manner in which they interact when the magnitude of one of the two types of delay is near its threshold for detection (as was the case in the present study).

Second, it is important to recognize that the stimuli employed by Buell *et al.* were low-frequency tones of either 500 or 1000 Hz. One possibility is that the differences observed between our data and those of Buell *et al.* are related to the fact that the "ambiguous" delays in our study were conveyed by the envelopes of high-frequency waveforms rather than by the fine-structure of low-frequency tones. Another possibility is that the apparent inconsistency across the two studies is related to differences in the magnitudes of the ambiguous ongoing interaural delays employed. To explain, when Buell *et al.* imposed phase-shifts of π radians within the ongoing portions of their 500 or 1000 Hz tones, those phase-shifts were equivalent to ITDs of 1.0 and 0.5 ms, respectively. In contrast, phase shifts of π radians within the ongoing portion of our SAM tones corresponded to ITDs that were all substantially larger than those imposed by Buell *et al.* For the rate of modulation of 125 Hz, the corresponding ITD was 4 ms. As discussed, that rate of modulation was expected to convey the most potent envelope-based ITD information. The failure of such large ambiguous ongoing interaural delays within the envelopes of high-frequency SAM tones to have any discernable effect on onset/offset ITD thresholds might result from a relative inability of the binaural system to code "accurately" such large ITDs (e.g., Mossop and Culling, 1998; van der Heijden and Trahiotis, 1999). In any case, additional research would be required to determine the source of the different outcomes obtained by us and by Buell *et al.*

In contrast to the outcomes observed with the $\Delta|\pi$ delay-type, sensitivity to interaural delays with the onset/offset (or ongoing) portions of the waveform were impacted by roving the interaural delay within the ongoing (or onset/offset) portion of the waveform. Such effects were only observed when the pairing of duration and rate of modulation was such that the salience of the aspect of the waveform containing the noninformation-bearing, roved interaural delay was expected to be relatively high.

IV. SUMMARY

Overall, the results of the experiments presented here, in which the ongoing interaural delays were derived from the envelopes of high-frequency SAM waveforms, are consistent with the findings and interpretations of classic studies in which ongoing interaural delays were derived from the fine-structure (e.g., Tobias and Schubert, 1959; Perrott and Baars, 1974; Hafter, Dye, and Gilkey, 1979; Kunov and Abel, 1981; Abel and Kunov, 1983). Specifically, ongoing delays in the present study were found to be substantially more potent than onset/offset interaural delays with the relative salience of the ongoing delay being directly proportional to the duration of the stimulus. The dominance of the ongoing delay conveyed by the envelopes of the high-frequency SAM tones employed in this study occurred despite the well-known poorer sensitivity of listeners to those interaural delays as compared to their sensitivity to interaural delays within the fine-structure of low-frequency stimuli. In addition, the data presented here demonstrated that the salience of the ongoing delay was inversely proportional to the rate of modulation of the SAM waveform over the range of rates of modulation studied.

The results of two different quantitative analyses, an information-based description of detection and an interaural-correlation-based model of binaural processing, were applied to the data from Experiment 1. Both analyses were moderately successful and support the notion that binaural information stemming from onset/offset and ongoing interaural delays is effectively combined in an independent fashion. The appeal of the interaural-correlation approach is that it capitalizes on an existing model that uses a single metric computed across the "effective" stimulus subsequent to the application of transformations designed to mimic auditory processing.

Finally, our findings aid in the interpretation of classic studies in which sensitivity to interaural delays within high-frequency waveforms was measured (e.g., David *et al.*, 1958, 1959; Yost, 1971; Henning, 1974; Klumpp and Eady, 1956). The data suggest that for high-frequency stimuli of sufficiently long duration (>100 ms or so) and with (effective) rates of modulation between 100 and 200 Hz, the contribution to thresholds of interaural delay of onset/offset interaural delays when whole waveform interaural delays are imposed is so negligible that the sensitivity can, for practical purposes, be considered to stem solely from interaural delays conveyed by the ongoing envelope of the waveform. For very brief stimuli, however, or for stimuli characterized by envelopes having high rates of fluctuation, onset/offset interaural delays can contribute substantially to listeners' sensitivities to such whole waveform delays. Thus, the data presented here lend further support to McFadden and Pasanen's (1976) caution that experimenters take care to distinguish among the types of interaural delay present and potentially available to listeners both at low and at high frequencies.

ACKNOWLEDGMENTS

The work reported here followed directly from a pilot study conducted by Dr. Sarah J. Griffin while visiting the University of Connecticut Health Center under the auspices of a University College London Bogue Fellowship and a Wellcome Trust studentship. The authors thank Ms. Chantal Turner who carried out the experiments on a day-to-day basis and who performed the initial summary statistics. The authors also thank Dr. Richard Freyman, Dr. Wes Grantham, and an anonymous reviewer for their insightful and helpful comments. This research was supported by research Grants NIH DC-04147 and DC-04073 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

[1]It is important to note that our focus on the relative potency of "ongoing" interaural delay versus the "onset/offset," "gating," or "transient" interaural delay mirrors that of previous investigators (e.g., Tobias and Schubert, 1959; Perrott and Baars, 1974; Kunov and Abel, 1981; Abel and Kunov, 1983; Buell *et al.*, 1991). Our focus should be distinguished from that of studies aimed at determining the relative potency of interaural delays within relatively "early" portions of the waveform (which could include *both* onset and ongoing interaural delays) versus the potency of interaural delays within relatively "later" portions of the waveform (which could include *both* ongoing and offset interaural delays). Studies with this latter focus are rather wide ranging in terms of the stimuli and tasks employed but could be considered as having a common goal of assessing the weighting of binaural information as a function of its temporal position within the stimulus (e.g., Wallach *et al.*, 1949; Franssen, 1962; Clifton, 1987; Zurek, 1980; Hafter and Buell, 1990; Saberi and Perrott, 1990; Houtgast and Aoki, 1994; Freyman *et al.*, 1997; Akeroyd and Bernstein, 2001). It is not clear to us that a common mechanism can account for this diverse set of data and associated phenomena, much less be applied to results obtained by directly manipulating the onset/offset and ongoing interaural delays of the waveform as was done in the current study. We believe onset/offset and ongoing interaural delays as we have defined them do not necessarily correspond to "early" and "later" arriving interaural delays as defined by others. Our definition is in keeping with our interpretation of what was meant by McFadden and Pasanen (1976) in their discussion of "onset" interaural delays. Specifically, our interpretation, which was validated via personal communication with Dr. McFadden, is that when referring to onset delays, those authors were, in fact, referring to "gating" or "transient" interaural delays and not to interaural delays within the "early" (onset and ongoing) portions of the waveform. Considering all of the above, we find it appropriate to limit the scope of our presentation by not placing it within the context of fundamentally different research concerning the relative potency of "early" and "later" arriving interaural delays. Interestingly, Tobias and Schubert (1959) adopted a similar stance after attempting to reconcile data concerning the relative salience of ongoing versus onset/offset interaural delays with those obtained by Wallach *et al.* in their study of the "precedence effect."

[2]Kunov and Abel (1981) and Abel and Kunov (1983) also conducted studies aimed at assessing the relative potency of onset/offset and ongoing interaural delays by placing the two cues in opposition. Those authors concluded that ongoing interaural delays were not substantially more potent than onset/offset interaural delays. As discussed by Buell *et al.* (1991), however, in both studies, Kunov and Abel gated their low-frequency, pure-tone stimuli prior to imposing an interaural delay. This resulted in onset/offset and ongoing interaural delays that were of the same physical magnitude. Because of the periodic nature of their pure-tone stimuli, however, delays corresponding to greater than one-half the period of the tone resulted in *effective* onset/offset and ongoing delays that not only favored opposite ears but that had different magnitudes such that the onset/offset interaural delay could be substantially larger than the ongoing interaural delay.

[3]Within their study of the perceived lateral position of "virtual" auditory stimuli, Macpherson and Middlebrooks (2002) attempted to alter the relative salience of onset/offset and ongoing interaural time delays within noise high-passed between 4 and 16 kHz. While their results suggested that, for these broadband stimuli, ongoing interaural delays were more potent than onset/offset interaural delays, the primary focus of their study was on the relative salience of ITDs versus interaural intensitive disparities (IIDs).

[4]Examination of Fig. 1 also reveals that, for the $\Delta|0$ and $0|\Delta$ delay types,

the magnitudes of the initial and final portions of the envelopes of each waveform differ across a left/right pair. For the $\Delta|0$ delay-type (panel A), the magnitude of the envelope will be *greater* toward the beginning of the stimulus and *smaller* near the end of the stimulus for the waveform that is turned on earlier. Thus, there will exist short-lived interaural intensitive differences (IIDs) favoring the leading and lagging ears during the initial and final portions of the stimuli, respectively. These opposing IIDs are a necessary consequence of turning on and off asynchronously two otherwise identical waveforms. For the $0|\Delta$ delay type (panel B), the two waveforms are turned on and off synchronously. Still, the magnitudes of the initial and final portions of the envelopes of each waveform can differ across a left/right pair. In this case, the relative magnitudes of the envelopes of the two waveforms near their respective beginning and end, and the magnitude and direction of any concomitant IIDs will be a function of both the starting phase of the sinusoidal envelope and the magnitude of the ongoing interaural delay. In the experiments conducted here (see Sec. II A.), a sampling strategy was implemented such that the starting phase of the sinusoidal envelope was chosen randomly for each and every observation interval. This ensured that the IIDs discussed above for the $0|\Delta$ delay type (panel B) would not consistently favor one ear or the other at the beginning or end of the stimuli. These IIDs are a necessary consequence of turning on and off synchronously two waveforms containing an ongoing interaural delay. It should be recognized that the existence of the types of IIDs discussed above is part and parcel of studying the effects of onset/offset and ongoing interaural delays.

[5]As discussed, two of the listeners could not perform the task when the $0|\Delta$ delay type was employed with the 50ms/350 Hz pairing of duration and rate of modulation. For the purposes of the ANOVA, the normalized thresholds entered into the appropriate "cells" for those two listeners were equal to 1500 μs divided by the individual listener's reference threshold. This approach was chosen because (1) a value of 1500 μs represents the largest value of interaural delay that was accepted as valid in the computation of normalized thresholds (see text) and (2) because it is conservative in terms of the evaluation of the significance of the main effects and interactions.

[6]The formula used to compute the percentage of the variance for which our predicted values of threshold accounted was $100\times(1-[\Sigma(O_i-P_i)^2]/[\Sigma(O_i-\bar{O})^2])$ where O_i and P_i represent individual observed and predicted values of threshold, respectively, and $\bar{O}$ represents the mean of the observed values of threshold (e.g., Bernstein and Trahiotis, 1994).

Abel, S. M., and Kunov, H. (**1983**). "Lateralization based on interaural phase differences: Effects of frequency, amplitude, duration, and shape of rise/decay," J. Acoust. Soc. Am. **73**, 955–960.

Akeroyd, M. A., and Bernstein, L. R. (**2001**). "The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses," J. Acoust. Soc. Am. **110**, 2516–2526.

Bernstein, L. R. (**2001**). "Auditory processing of interaural timing information: New insights," J. Neurosci. Res. **66**, 1036–1046.

Bernstein, L. R., and Trahiotis, C. (**2002**). "Enhancing sensitivity to interaural delays at high frequencies by using 'transposed stimuli'," J. Acoust. Soc. Am. **112**, 1026–1036.

Bernstein, L. R., and Trahiotis, C. (**1985**). "Lateralization of low-frequency, complex waveforms: The use of envelope-based temporal disparities," J. Acoust. Soc. Am. **77**, 1868–1880.

Bernstein, L. R., and Trahiotis, C. (**1994**). "Detection of interaural delay in high-frequency SAM tones, two-tone complexes, and bands of noise," J. Acoust. Soc. Am. **95**, 3561–3567.

Bernstein, L. R., and Trahiotis, C. (**1996**). "The normalized correlation: Accounting for binaural detection across center frequency," J. Acoust. Soc. Am. **100**, 3774–3784.

Bernstein, L. R., and Trahiotis, C. (**2003**). "Enhancing interaural-delay-based extents of laterality at high frequencies by using 'transposed stimuli'," J. Acoust. Soc. Am. **113**, 3335–3347.

Bernstein, L. R., Par, Steven van de, and Trahiotis, C. (**1999**). "The normalized correlation: Accounting for NoSπ thresholds obtained with Gaussian and 'low-noise' masking noise," J. Acoust. Soc. Am. **106**, 870–876.

Blauert, J. (**1982**). "Binaural localization: Multiple images and applications in room- and electroacoustics," in *Localization of Sound: Theory and Application*, edited by R. W. Gatehouse (Amphora Press, Groton, CT).

Blauert, J. (**1983**). *Spatial Hearing* (MIT Press, Cambridge, MA).

Buell, T. N., Trahiotis, C., and Bernstein, L. R. (**1991**). "Lateralization of low-frequency tones: Relative potency of gating and ongoing interaural delay," J. Acoust. Soc. Am. **90**, 3077–3085.

Clifton, R. K. (**1987**). "Breakdown of echo suppression in the precedence effect," J. Acoust. Soc. Am. **82**, 1834–1835.

David, E. E., Guttman, N., and van Bergeijk, W. A. (**1958**). "On the mechanism of binaural fusion," J. Acoust. Soc. Am. **30**, 801–802.

David, E. E., Guttman, N., and van Bergeijk, W. A. (**1959**). "Binaural interaction of high-frequency complex stimuli," J. Acoust. Soc. Am. **31**, 774–782.

Domnitz, R. H., and Colburn, H. S. (**1977**). "Lateral position and interaural discrimination," J. Acoust. Soc. Am. **61**, 1586–1598.

Franssen, N. V., (**1962**). *Sterophony*, "Phillips Technical Lecture, Einhoven, The Netherlands, English translation," **1964**.

Freyman, R. L., Zurek, P. M., Balakrishnan, U., and Chiang, Y. C. (**1997**). "Onset dominance in lateralization," J. Acoust. Soc. Am. **101**, 1649–1659.

Grantham, D. W., (**1995**). "Spatial hearing and related phenomena," in *Handbook of Perception and Cognition: Hearing*, edited by B. C. J. Moore (Academic, San Diego).

Green, D. M., and Swets, J. A., (**1974**). *Signal Detection Theory and Psychophysics*, (Wiley, New York).

Hafter, E. R., and Buell, T. N. (**1990**). "Restarting the adapted binaural system," J. Acoust. Soc. Am. **88**, 806–812.

Hafter, E. R., Dye, R. H., Jr., and Gilkey, R. H., (**1979**). "Lateralization of tonal signals which have neither onsets nor offsets," J. Acoust. Soc. Am. **65**, 471–477.

van der Heijden, M., and Trahiotis, C., (**1999**). "Masking with interaurally delayed stimuli: The use of 'internal' delays in binaural detection," J. Acoust. Soc. Am. **105**, 388–399.

Henning, G. B., (**1974**). "Detectability of interaural delay in high-frequency complex waveforms," J. Acoust. Soc. Am. **55**, 84–90.

Henning, G. B. (**1980**). "Some observations on the lateralization of complex waveforms," J. Acoust. Soc. Am. **68**, 446–453.

Henning, G. B., and Ashton, J. (**1981**). "The effect of carrier and modulation frequency on lateralization based on interaural phase and interaural group delay," Hear. Res. **4**, 186–194.

Houtgast, T., and Aoki, S. (**1994**). "Stimulus-onset dominance in the perception of binaural information," Hear. Res. **72**, 29–36.

Keppel, G., (**1973**). *Design and Analysis: A Researchers Handbook* (Prentice–Hall, Englewood Cliffs, NJ).

Klumpp, R. G., and Eady, H. R. (**1956**). "Some measurements of interaural time difference thresholds," J. Acoust. Soc. Am. **28**, 859–860.

Kunov, H., and Abel, S. M. (**1981**). "Effects of rise/decay time on the lateralization of interaurally delayed 1 kHz tones," J. Acoust. Soc. Am. **69**, 769–773.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Macpherson, E. A., and Middlebrooks, J. C. (**2002**). "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," J. Acoust. Soc. Am. **111**, 2219–2236.

McFadden, D., and Moffitt, C. M. (**1977**). "Acoustic integration for lateralization at high frequencies," J. Acoust. Soc. Am. **61**, 1604–1608.

McFadden, D., and Pasanen, E. G. (**1976**). "Lateralization at high frequencies based on interaural time differences," J. Acoust. Soc. Am. **59**, 634–639.

Mossop, J. E., and Culling, J. F. (**1998**). "Lateralization of large interaural delays," J. Acoust. Soc. Am. **104**, 1574–1579.

Nuetzel, J. M., and Hafter, E. R. (**1976**). "Lateralization of complex waveforms: Effects of fine-structure, amplitude, and duration," J. Acoust. Soc. Am. **60**, 1339–1346.

Nuetzel, J. M., and Hafter, E. R. (**1981**). "Discrimination of interaural delays in complex waveforms: Spectral effects," J. Acoust. Soc. Am. **69**, 1112–1118.

Perrott, D. R., and Baars, B. J. (**1974**). "Detection of interaural onset and offset disparities," J. Acoust. Soc. Am. **55**, 1290–1292.

Saberi, K., and Perrott, D. R. (**1990**). "Lateralization thresholds obtained under conditions in which the precedence effect is assumed to operate," J. Acoust. Soc. Am. **87**, 1732–1737.

Stern, R. M., and Colburn, H. S. (**1978**). "Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position," J. Acoust. Soc. Am. **64**, 127–140.

Tobias, J. V., and Schubert, E. D. (**1959**). "Effective onset duration of auditory stimuli," J. Acoust. Soc. Am. **31**, 1595–1605.

Trahiotis, C., Bernstein, L. R., and Akeroyd, M. A. (**2001**). "Manipulating the "straightness" and "curvature" of patterns of interaural cross-correlation affects listeners' sensitivity to changes in interaural delay,"

J. Acoust. Soc. Am. **109**, 321–330.
Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**) "The precedence effect in sound localization," Am. J. Psychol. **52**, 315–336.
Wightman, F. L., and Kistler, D. J. (**1992**). "The dominant role of low-frequency interaural time differences in sound localization," J. Acoust. Soc. Am. **91**, 1648–1661.
Yost, W. A., Wightman, F. L., and Green, D. M. (**1971**). "Lateralization of filtered clicks," J. Acoust. Soc. Am. **50**, 1526–1530.
Zurek, P. M. (**1980**). "The precedence effect and its possilbe role in the avoidance of interaural ambiguitites," J. Acoust. Soc. Am. **67**, 952–964.
Zurek, P. M. (**1993**). "A note on onset effects in binaural hearing," J. Acoust. Soc. Am. **93**, 1200–1201.

Influences of auditory object formation on phonemic restoration[a)]

Barbara G. Shinn-Cunningham[b)]
Hearing Research Center, Department of Cognitive and Neural Systems and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02421 and Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139

Dali Wang
Hearing Research Center and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02421

(Received 17 August 2007; revised 6 October 2007; accepted 8 October 2007)

In phonemic restoration, intelligibility of interrupted speech is enhanced when noise fills the speech gaps. When the broadband envelope of missing speech amplitude modulates the intervening noise, intelligibility is even better. However, this phenomenon represents a perceptual failure: The amplitude modulation, a noise feature, is misattributed to the speech. Experiments explored whether object formation influences how information in the speech gaps is perceptually allocated. Experiment 1 replicates the finding that intelligibility is enhanced when speech-modulated noise rather than unmodulated noise is presented in the gaps. In Experiment 2, interrupted speech was presented diotically, but intervening noises were presented either diotically or with an interaural time difference leading in the right ear, causing the noises to be perceived to the side of the listener. When speech-modulated noise and speech are perceived from different directions, intelligibility is no longer enhanced by the modulation. However, perceived location has no effect for unmodulated noise, which contains no speech-derived information. Results suggest that enhancing object formation reduces misallocation of acoustic features across objects, and demonstrate that our ability to understand noisy speech depends on a cascade of interacting processes, including glimpsing sensory inputs, grouping sensory inputs into objects, and resolving ambiguity through top-down knowledge. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804701]

PACS number(s): 43.66.Pn, 43.71.Rt, 43.66.Ba, 43.71.An [JCM] Pages: 295–301

I. INTRODUCTION

In everyday settings, the speech we hear is often partially masked by other sound sources, such as other talkers and events (Cherry, 1953). Our ability to communicate using noisy, ambiguous speech can be attributed in part to the redundancy in meaningful speech, which allows us to fill in masked or missing portions of the attended signal (Cooke, 2006). For instance, over a range of interruption rates, listeners are able to understand speech relatively well when half of the speech signal is replaced by silence (Miller and Licklider, 1950; Powers and Wilcox, 1977).

Speech intelligibility is even better when the speech gaps are filled in by unmodulated, steady-state noise, presumably because perceptual "filling in" of an interrupted speech signal is more automatic and complete than when there are sudden, audible silences (Warren, 1970; Powers and Wilcox, 1977; Bashford *et al.*, 1992). This filling in is informed by our expectations of the likely content of meaningful speech at every level of analysis, from continuity of spectrotemporal energy in the sound to lexical, linguistic, and semantic constraints (Warren, 1970; Bashford *et al.*, 1992; Warren *et al.*, 1994, 1997; Petkov *et al.*, 2007). While some of these expectations are learned (e.g., filling in a missing phoneme to generate a meaningful word in a given sentence), others may be hard-wired (e.g., perceiving a frequency glide interrupted by noise as if the glide is continuous; see Bashford *et al.*, 1992; Bailey and Herrmann, 1993; Darwin, 2005; Petkov *et al.*, 2007).

When the broadband temporal amplitude of missing speech is used to amplitude modulate the noise presented in speech gaps (henceforth referred to as speech-modulated noise), intelligibility is enhanced compared to when the noise is unmodulated (Bashford *et al.*, 1996). At first glance, this result is unsurprising—providing more speech-derived information in the input stimulus enhances intelligibility.

However, the speech and speech-modulated noise are perceived as distinct auditory objects (Bregman, 1990). Evidence suggests that listeners actively attend to one auditory object at a time in most situations (e.g., see Best *et al.*, 2006), consistent with the biased-competition model of visual attention (Desimone and Duncan, 1995). Thus, the improvement in speech intelligibility must come about because a feature of the noise (its modulation) is incorrectly bound with a competing object (the speech), an example of an "illusory conjunction" (Treisman and Gelade, 1980; Dyson and

[a)]Portions of this work were presented at the 2007 Mid-Winter Meeting of the Association for Research in Ogo-larynology, Denver, Colorado.
[b)]Author to whom correspondence should be addressed. Electronic mail: shinn@bu.edu

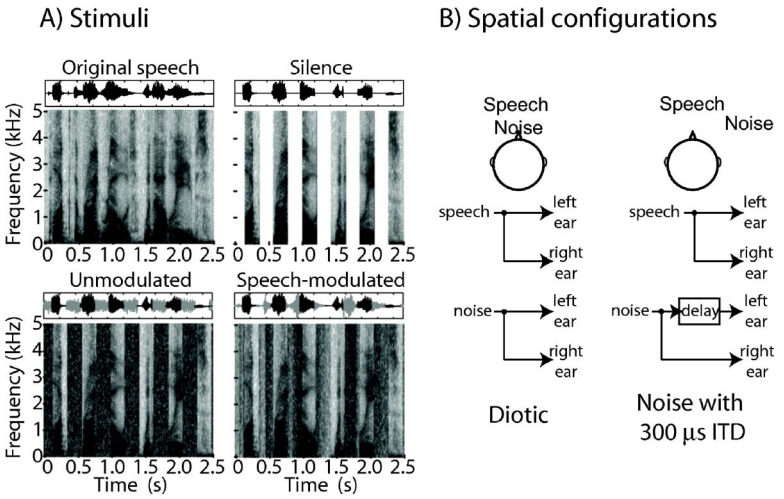


FIG. 1. (Color online) Stimuli and stimulus configurations. (A) Example of original speech and periodically interrupted speech when interruptions are filled with silence, unmodulated noise, and speech-modulated noise. Gray-scaled panels are spectrograms, showing the relative energy as a function of time (abscissa) and frequency (ordinate), with black corresponding to high energy and white to low energy. Small insets above each spectrogram plot the time domain wave forms, with black showing the unprocessed signal and gray showing the inserted noise wave forms. (B) Spatial configurations in Experiment 2 varied the ITD of the noise so that it was either heard at midline (left) or to the right of midline (right).

Quinlan, 2003). Illusory conjunctions of features in nonspeech stimuli are most likely to occur when the cues driving auditory object formation are ambiguous (Hall *et al.*, 2000). This suggests that manipulating acoustic grouping cues to increase the perceptual segregation of the speech and the speech-modulated noise might affect speech intelligibility.

We reasoned that strengthening the perceptual segregation of the noise and speech should reduce the likelihood that a noise feature would be misattributed to the speech. Increased segregation of speech and noise should not have any impact on intelligibility of interrupted speech presented with unmodulated noise in the gaps, which provides no information about how to complete missing speech. Instead of possessing a feature derived from the speech, unmodulated noise simply serves as a plausible masker of the missing speech, encouraging perceptual filling in (Warren, 1970; Hall *et al.*, 2000; Darwin, 2005). In contrast, if low-level auditory object formation affects speech perception, improved segregation of the speech-modulated noise and interrupted speech should decrease the likelihood of integrating the modulation (a noise feature) with the speech, and should therefore degrade speech intelligibility.

The dominant cues driving auditory object formation are spectrotemporal (Bregman, 1990; Darwin and Carlyon, 1995); however, spatial cues play a larger role in object formation when other cues are ambiguous (Darwin and Hukin, 1997, 1998; Freyman *et al.*, 2001; Shinn-Cunningham *et al.*, 2007), as when interrupted speech is presented with speech-modulated noise. We therefore manipulated the perceived spatial separation of interrupted speech and noise to affect the perceptual segregation of the speech and noise, and to see if segregation affected speech intelligibility.

We found that these low-level cues affected speech understanding when the interfering noise was modulated by the speech envelope, but not when the noise was unmodulated. These results demonstrate that low-level auditory cues affect speech just as they affect other, less specialized acoustic signals.

II. METHODS

A. Subjects

Nine normal-hearing subjects performed the tasks (five in Experiment 1 and four in Experiment 2). Subjects were recruited through on-campus advertisement, and all were students at Boston University (between ages 23 and 35). None had prior experience with psychophysical tasks, or with the corpus of test materials employed. All participants had pure-tone thresholds of 20 dB HL or better at all frequencies in the range from 250 to 8000 Hz, in both ears, and their threshold at 500 Hz was 15 dB HL or better. All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board.

B. Equipment

Stimuli were processed in MATLAB (Mathworks, Natick, MA) using a sampling rate of 25 kHz. The stimuli were processed in MATLAB and sent to Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over Sennheiser HD580 headphones. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. MATLAB was used to control the stimulus presentation, to record responses, and to analyze results.

C. Stimuli

Speech sentences were from the Harvard IEEE corpus (IEEE, 1969). Sentences were periodically interrupted so that 50% of each signal was replaced by silence. This was accomplished by multiplying each sentence with a square wave (ranging between zero and one) with a 50% duty cycle. The periodicity of the periodic square wave was chosen to match rates that in past studies elicited large improvements in speech intelligibility when unmodulated noise filled in the speech gaps (Powers and Wilcox, 1977). Three rates, equal to 1.5, 2.2, and 3.0 Hz, were used.

In some conditions, the silent speech gaps were filled in, either with unmodulated white noise or speech-modulated noise. The speech-modulated noise was generated by multiplying unmodulated white noise by the Hilbert envelope of the speech that was missing in the gap (see Fig. 1). The average long-term, broadband root-mean-square intensity of the speech and noise were matched across the stimulus set; however, the spectra were not matched.[1]

The sentences used in each condition were chosen randomly from the 720 sentences making up the corpus. No sentence was presented more than once to any subject.

D. Spatial cues and stimulus conditions

In all conditions of both experiments, the interrupted speech was presented diotically.

Experiment 1 compared intelligibility of interrupted speech with silent gaps, unmodulated noise, and speech-modulated noise [see Fig. 1(a)]. Like the interrupted speech, both unmodulated noise and speech-modulated noise were presented diotically in Experiment 1, so that both speech and intervening noise were heard in the center of the head.

Experiment 2 compared intelligibility of interrupted speech with unmodulated and speech-modulated noise, with the speech and noise either collocated or perceived from different directions. In the collocated conditions, all stimuli were diotic, while in the spatially separated conditions, the noise (either unmodulated or speech modulated) was presented with an interaural time difference of 300 μs leading to the right ear. Thus, in the collocated configurations, both speech and noise were heard at the same, midline location [see the left-hand side of Fig. 1(b)]. In the spatially separated configurations, the speech was perceived at midline and the noise to the right of the listener [see the right-hand side of Fig. 1(b); subjectively, the off-midline perceptual locations of the unmodulated and speech-modulated noises with the 300 μs ITD were indistinguishable, and to the right of midline].

E. Procedure

Each listener performed four experiment sessions (at most one per day), each of which lasted approximately 1 h. Within each session, multiple experimental blocks were presented. In each block, the stimulus type and spatial configuration were fixed, but the interruption rate was randomly chosen on a trial-by-trial basis (with all three rates presented an equal number of times in a block). In each session, listeners performed one block of trials for each combination of stimulus type and spatial configuration, in random order (different in each session and for each listener). Thus, in each session, a listener performed an equal number of all possible combinations of stimulus type, spatial configuration, and interruption rate. There was no evidence that performance improved from session to session, and the randomization of the order of conditions across subjects and sessions ensured that any such learning effects would be averaged out, if they did exist.

In Experiment 1, which had three stimulus types and one spatial configuration, each listener performed three blocks of 60 trials each in each of the four sessions. In each block, listeners performed 20 trials at each interruption rate. Across the four experimental sessions, each subject performed a total of 80 repetitions (20 trials/session $\times$ 4 sessions) for each combination of stimulus type (three—silent gaps, unmodulated noise, and speech modulated noise), spatial configuration (only collocated in Experiment 1), and repetition rate (three: 1.5, 2.2, and 3.0 Hz).

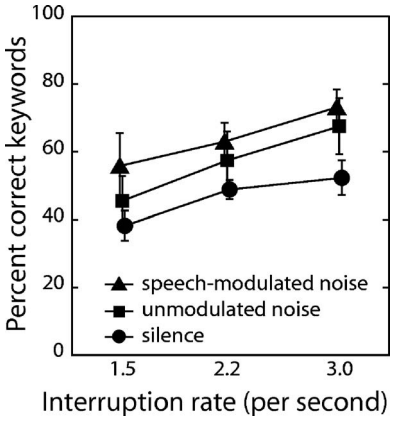


FIG. 2. Speech intelligibility is best when speech gaps are filled with speech-modulated noise and worst when gaps are silent. Average percent-correct performance on key words is plotted as a function of speech interruption rate for silent gaps, gaps filled with unmodulated noise, and gaps filled with speech-modulated noise. Error bars show the 95% confidence interval around the across-subject mean in performance.

In Experiment 2, there were two stimulus types (unmodulated and speech-modulated noise) and two spatial configurations (collocated and separated), for a total of four different combinations of stimulus and configuration. In this experiment, listeners performed four blocks of 45 trials each in each of the four sessions. In each block, listeners performed 15 trials of at each interruption rate. Across the four experimental sessions, each subject performed a total of 60 repetitions (15 trials/session $\times$ 4 sessions) for each combination of stimulus type (two—unmodulated and speech modulated noise), spatial configuration (two—collocated and separated), and repetition rate (three—1.5, 2.2, and 3.0 Hz).

F. Scoring

Each sentence contained three to five key words (adjectives, adverbs, nouns, and verbs). After each sentence was presented, listeners typed in the words they heard. The percentage of key words correctly reported was scored as a measure of speech intelligibility for each combination of stimulus type, spatial configuration, and repetition rate.

III. RESULTS

A. Experiment 1: Speech-modulated noise, unmodulated noise, or silence

Experiment 1 was designed to replicate previous findings showing that intelligibility is enhanced when speech-modulated noise fills in the speech gaps compared to silent gaps and to unmodulated noise (Bashford *et al.*, 1996). The across-subject average of the raw percent-correct key words is plotted in Fig. 2 (error bars show the 95% confidence interval around the across-subject average, in percent correct). Because individual subjects showed the same pattern as the across-subject average, only the average is shown.

Raw results verify that the interrupted speech is least intelligible when the speech gaps are silent (circles fall below other symbols in Fig. 2), most intelligible when the gaps are filled with speech-modulated noise (triangles fall above other symbols), and intermediate when the gaps are filled

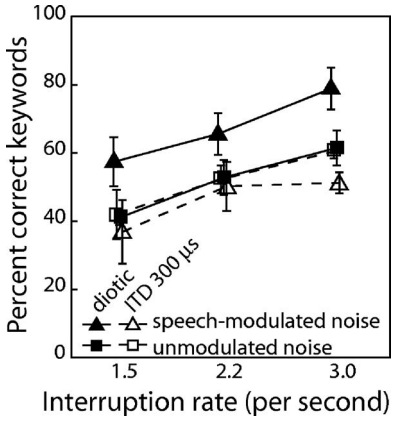


FIG. 3. Spatial configuration does not affect speech intelligibility for interrupted speech with unmodulated noise, but affects performance when the speech envelope modulates the intervening noise. Percent-correct performance on key words is plotted as a function of speech interruption rate. Speech gaps are filled with unmodulated or speech-modulated noise that are either at the same midline location as the interrupted speech or to the right of midline, with a 300 μs ITD. Error bars show the 95% confidence interval around the across-subject mean in performance.

with unmodulated noise (squares fall in between circles and triangles). As in previous studies, performance improves as the interruption rate increases (Powers and Wilcox, 1977).

A two-way ANOVA with factors of interruption rate and stimulus type supported the above-mentioned observations. Both the main effects of stimulus condition and interruption rate were significant [$F(2,36)=26.4$, $p<0.0001$ for stimulus condition; $F(2,36)=27.2$, $p<0.0001$ for interruption rate]. However, the interaction term was not statistically significant [$F(4,36)=0.673$, $p=0.62$]. Post-hoc tests with Bonferroni correction showed that compared to the silent-gap condition, performance was significantly better for both the unmodulated noise ($p<0.0001$) and the speech-modulated noise conditions ($p<0.0001$). In addition, performance with speech-modulated noise was significantly better than with unmodulated noise ($p=0.014$). Performance improved significantly with increasing interruption rate, and all three pairwise comparisons of interruption rates were significant ($p<0.0001$ for 1.5 Hz vs 2.2 Hz; $p<0.0001$ for 1.5 Hz vs 3.0 Hz; $p=0.007$ for 2.2 Hz vs 3.0 Hz).

B. Experiment 2: Effects of spatial separation for different noise types

Experiment 2 investigated the hypothesis that the across-object misallocation of amplitude modulation that produced better performance for speech-modulated noise than for unmodulated noise occurs only when the perceptual separation of the speech and noise objects is weak. We used perceived location to affect the perceptual segregation of the speech and noise by manipulating the interaural time difference (ITD) of the noise signals so that the noise was either perceived at the same midline location as the interrupted speech or to the right of midline.

Figure 3 plots mean percent correct scores (error bars represent 95% confidence interval of the across-subject mean).

Consistent with our hypothesis, perceived spatial separation between the interrupted speech and the noise did not have a noticeable impact on results for unmodulated noise: Percent-correct performance was essentially equal in both spatial configurations using unmodulated noise (compare closed and open square symbols in Fig. 3). As in Experiment 1 and previous reports (Bashford *et al.*, 1996), performance was better for speech-modulated noise than unmodulated noise when speech and noise collocated (the closed triangles are above the squares in Fig. 3). Finally, intelligibility was degraded when the interrupted speech was diotic and the speech-modulated noise was perceived to the right (the open triangles fall below the squares in Fig. 3).

Again, these conclusions were supported by statistical tests. A two-way ANOVA with factors of spatial condition and interruption rate found that both main effects were significant [$F(3,36)=42.8$, $p<0.0001$ for spatial condition; $F(2,36)=64.9$, $p<0.0001$ for interruption rate]. However, the interaction between interruption rate and condition was not statistically significant [$F(6,36)=0.777$, $p=0.593$ for the interaction]. Post-hoc tests with Bonferroni correction revealed that there was no significant difference between performance for the unmodulated noise, collocated condition and the unmodulated noise, spatially separated condition ($p>0.999$). All other conditions were significantly different at the $p=0.05$ level. Performance was significantly better in the speech-modulated noise, collocated condition than in the speech-modulated noise, spatially separated condition ($p<0.0001$); in the speech-modulated noise, collocated condition than in the unmodulated noise, collocated condition ($p<0.0001$); in the speech-modulated noise, collocated condition than in the unmodulated noise, spatially separated condition ($p<0.0001$); in the unmodulated noise, spatially separated condition than in the speech-modulated noise, spatially separated condition ($p=0.0457$); and in the unmodulated noise, collocated condition than in the speech-modulated noise, spatially separated condition ($p=0.0186$).

IV. DISCUSSION

Experiment 1 confirmed that intelligibility of interrupted speech is enhanced when the broadband envelope of the speech is used to amplitude modulate noise in the speech gaps. Although it may not initially appear surprising that adding information about the speech improves intelligibility, a more careful consideration of what is taking place is warranted. The information provided by the broadband speech envelope in the speech-modulated noise is rudimentary, providing no information about the spectral content of the missing speech. The only speech information present in the speech-modulated stimuli is crude prosodic and voicing information, which co-vary with overall speech amplitude. That such reduced information provides any improvement in speech intelligibility is a testament to how efficiently listeners use any snippet of evidence they hear in order to resolve ambiguity about the content of noisy, interrupted speech.

Moreover, it is especially surprising that these simple amplitude modulation cues aid speech intelligibility given that all listeners report that they perceive the modulation as

part of the noise. In other words, the modulation is heard as an attribute of the noise, yet still contributes to the intelligibility of the collocated interrupted speech. It is likely that the noise modulation contributes to speech intelligibility because it partially matches the modulations that are expected to be present in the speech during the gaps, based on the speech glimpses the listener hears. That is, the current results suggest that knowledge of the likely spectrotemporal structure of speech causes a form of perceptual competition between the speech and noise, each of which has some evidence that it is the proper "owner" of the modulation.

Perceived location has little effect on speech intelligibility when the intervening noise simply serves as a plausible masker of the missing speech. Only when a feature of the noise provides partial information about the missing speech does the spatial relationship between the speech and noise affect intelligibility. Importantly, in our spatial manipulations, we only manipulated ITDs (e.g., we did not simulate changes in the level or spectral content of the stimulus at the ears that arise with changes in source location), which should have no direct effect on intelligibility other than strengthening the low-level auditory organization of the scene and the perceptual segregation of the competing speech and noise. Indeed, because only ITD was manipulated, there were no differences in the relative energy of the speech and noise signals in the collocated and spatially separated conditions. Nonetheless, perceived location had a large effect on intelligibility when the noise contained a feature derived from the missing speech.

In the current experiment, it appears that the perceptual grouping of the scene is ambiguous when speech-modulated noise and interrupted speech are spatially collocated because the modulation of the noise fits expectations of what should be present in the missing speech. As a result of competition for the modulation, the modulation is perceived as a feature of the noise, yet still contributes to the intelligibility of the interrupted speech. Perceived location can tip the balance for how to resolve the ambiguity about how to perceptually allocate the modulation in the noise, simply by providing additional evidence that the modulation belongs to the noise rather than the interrupted speech.

When ITDs promote hearing the speech-modulated noise and interrupted speech as separate objects, intelligibility is actually worse than in the two (collocated and spatially separated) conditions using unmodulated noise. This result likely reflects the fact that the modulations reduce the overall energy in the speech-modulated noise compared to the unmodulated noise. When the modulations do not contribute to perception of the interrupted speech because ITDs better segregate noise and speech, enhancements in intelligibility come about because the noise serves as a plausible masker of the missing speech. However, the speech-modulated noise is less effective as a possible masker of the missing speech, and thus produces less automatic filling in of the missing speech. If the modulations are perceived as coming from the same direction as the interrupted speech, they are partially attributed to the speech and enhance intelligibility, so the "plausibility" of the intervening noise as a masker of the missing speech is irrelevant. However, when interrupted speech and speech-modulated noise are spatially distinct, the modulation is perceptually allocated only to the speech-modulated noise, and the only role of the noise in speech intelligibility is to act as a plausible masker of the missing speech. Because the noise is modulated and contains temporal gaps and less overall energy, it is less effective in this role than the unmodulated noise. Thus, the fact that the intelligibility for speech-modulated noise is worse than for unmodulated noise when the interrupted speech and noises are perceived as coming from different directions further emphasizes the fact that the way a scene is organized into objects has a direct impact on the ability to understand the interrupted speech.

Many past studies suggest that the automatic filling in of missing speech caused by intervening noise only occurs when the noise is sufficiently intense to ensure that it would have masked the speech if it were continuous (e.g., see Verschuure and Brocaar, 1983). In the current study, informal reports of the subjects suggest that the speech was not perceived as continuous (although we did not test this formally). Consistent with these subjective reports, we set the noise level to have the same broadband rms as the missing speech, and used a white (not speech-shaped) spectrum. As a result, the noise is unlikely to have masked the missing speech (if it had been present) at all frequencies;[1] the speech-to-noise ratio is 0 dB, averaged across frequency, which is greater than is typically required to achieve perceived continuity. Thus, we observe perceptual filling in of the missing speech, even though the missing speech would have been audible, if it were present. It is likely the improvement in intelligibility afforded by the unmodulated noise would have been even greater with a more intense, speech-shaped noise that did produce an illusion of continuity in the speech (Verschuure and Brocaar, 1983). Nonetheless, the unmodulated noise was sufficient to encourage automatic filling in, leading to improvements in speech intelligibility over interrupted speech presented alone.

A handful of past studies have explored the degree to which perceived continuity of an interrupted signal is affected by spatial attributes of the signal and the plausible masking signal (Hartmann, 1984; Kashino and Warren, 1996; Darwin *et al.*, 2002). These studies show that perceived continuity is stronger when the signal and plausible masker have the same spatial cues rather than different spatial cues, consistent with binaural processing reducing the level of the signal that would have been masked when the interaural cues in the interrupted signal and candidate masker differ from one another. The fact that the unmodulated noise was equally effective in improving intelligibiilty when it was diotic and when it was to the side is somewhat surprising in light of these studies. Specifically, one might expect poorer speech intelligibility for the unmodulated noise with the nonzero ITD than for the diotic, unmodulated noise. However, as discussed earlier, our listeners did not perceive the speech as continuous even in the collocated, unmodulated noise condition. These results suggest that there is a less direct link between perceived continuity of an interrupted signal (which appears to be sensitive to the binaural parameters of the interrupted signal and candidate masker) and the amount of automatic filling in of the missing speech content (which

appears to be less sensitive to the exact level of the candidate masker employed, at least in the current study) than some past studies suggest.

The influence of low-level auditory processes on speech perception has been a source of some debate. Some argue that the general rules governing auditory object formation do not apply to the perceptual organization of speech because there is a special, speech-specific phonetic system that is independent of the auditory system (Bentin and Mann, 1990; Whalen and Liberman, 1996; Remez, 2005). The phenomenon of "duplex perception" (Repp *et al.*, 1983), in which listeners integrate information from a frequency glide with information from other elements defining a vowel while still perceiving the glide as a distinct auditory object, is often cited as evidence supporting this kind of specialized phonetic processor (Liberman *et al.*, 1981; Repp *et al.*, 1983; Repp, 1984; Bentin and Mann, 1990; Whalen and Liberman, 1996).

In this sense, the current results resemble those of duplex perception experiments. In both paradigms, a spectrotemporal element (here, the amplitude modulation in the noise; in duplex perception, the frequency glide) could logically belong to either a speech object or a competing object, and ends up contributing perceptually to both. However, the same phenomenon is observed for an ambiguous element that logically could belong to two different nonspeech objects (e.g., Darwin and Ciocca, 1992; Bailey and Herrmann, 1993; Darwin, 1995; Hukin and Darwin, 1995; Hill and Darwin, 1996; Darwin and Hukin, 1997; 1998).

In all of these examples, ambiguous or conflicting grouping cues appear to lead to an element contributing to two different objects (e.g., when the spectrotemporal structure of a speech object supports hearing an element as part of the speech element, while other grouping cues support hearing the element as part of a separate object). These results suggest that the perceptual organization of both speech and nonspeech sounds depends on the majority of all evidence available to the listener, from low-level features (common onsets, harmonicity, comodulation, etc.) to higher-order cues such as expectations about speech structure. Any manipulation of grouping or streaming cues that alters the balance of competition for an ambiguous element can change the degree to which that element contributes to the objects in the mixture, while "sharing" of an element typically occurs only if the evidence is conflicting or ambiguous.

How an ambiguous feature or element is allocated across the objects in a sound mixture does not obey intuitively appealing rules of energy trading, wherein the total perceived content of an element is divided between competing objects (McAdams, 1989; Darwin, 1995; Shinn-Cunningham *et al.*, 2007). This seemingly paradoxical result is consistent with the idea that attention alters how an auditory scene is perceptually organized (Carlyon *et al.*, 2001; Sussman *et al.*, 2007), as if how an ambiguous scene is organized into objects depends on what object is the focus of attention (Shinn-Cunningham *et al.*, 2007). Current results are consistent with the view that perceptual organization of an ambiguous auditory scene depends on high-level factors, including listener expectations and goals.

In everyday settings, the problem of how to determine what sound energy belongs to what sound source is a significant challenge. It is often claimed that human listeners are very good at separating sound sources. Yet, often, as in the current experiments, there is a great deal of perceptual uncertainty about how to separate the sound energy in a mixture into constituent sources.

Ultimately, the goal for listeners is not segregating the sources, but understanding them. Rather than being good at estimating exactly what source produced what sound energy, listeners may simply excel at analyzing a noisy, ambiguous source and extracting its meaning, using all available evidence including a range of cues that affect source separation.

We conclude that our robust ability to understand noisy signals does not derive from an exceptional ability to perceptually separate sound sources in a mixture. Instead, our ability to understand noisy signals relies on integrating bottom-up sensory information with top-down knowledge of the likely source content (including knowledge of speech structure), taking into account all kinds of evidence that a particular sound feature belongs to a particular object. In this view, separating a source from a mixture and understanding it are intrinsically linked, rather than stages in a single, hierarchical, feed-forward process.

V. CONCLUSIONS

When interrupted speech and noise objects are imperfectly segregated, plausible modulations in the noise, derived from the missing speech, can be "borrowed" by the speech to enhance intelligibility. However, when perceptual segregation of speech and noise is strengthened through a manipulation of ITDs, this enhancement disappears. In contrast, ITD had no effect on perception when the intervening noise does not contain any speech-derived attributes.

These results show a direct interaction between auditory object formation and speech perception, at odds with claims that low-level auditory processes do not affect perception of speech. However, the current results also hint that expectations about speech content influence grouping. Together, these results suggest that perception of any complex auditory signal presented in a sound mixture depends on reciprocal, competitive interactions between low-level auditory processes and high-level knowledge about the likely spectrotemporal content of the sources making up the mixture.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation. Virginia Best, Chris Darwin, John Middlebrooks, and an anonymous reviewer provided helpful comments on the manuscript.

[1]The intervening noise was white, rather than speech shaped in its spectra. As a result, and because the speech is sparse in frequency, the intervening unmodulated noise presented during the speech gaps was not optimal for eliciting perceived continuity of the speech. Specifically, within a given narrow-band frequency range that contained significant speech energy going into a speech gap, there was typically a drop of energy at the gap onset even when noise was presented in the gap. This choice was made in part because we were more interested in how intelligibility was affected by the presence of different intervening noises than in any illusory continuity of

the speech induced by the noises. During piloting, intelligibility enhancements were obtained using white noise, so we used white noise during our formal tests.

Bailey, P. J., and Herrmann, P. (**1993**). "A reexamination of duplex perception evoked by intensity differences," Percept. Psychophys. **54**, 20–32.

Bashford, J. A., Jr., Riener, K. R., and Warren, R. M. (**1992**). "Increasing the intelligibility of speech through multiple phonemic restorations," Percept. Psychophys. **51**, 211–217.

Bashford, J. A., Jr., Warren, R. M., and Brown, C. A. (**1996**). "Use of speech-modulated noise adds strong 'bottom-up' cues for phonemic restoration," Percept. Psychophys. **58**, 342–350.

Bentin, S., and Mann, V. (**1990**). "Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes," J. Acoust. Soc. Am. **88**, 64–74.

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.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).

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.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

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 J. C. Sorin, H. Meloni, and J. Schoenigen, Elsevier, Amsterdam, the Netherlands, pp. 1–12.

Darwin, C. J. (**2005**). "Simultaneous grouping and auditory continuity," Percept. Psychophys. **67**, 1384–1390.

Darwin, C. J., Akeroyd, M. A., and Hukin, R. W. (**2002**). "Binaural factors in auditory continuity," International Conference on Auditory Displays, Kyoto, Japan.

Darwin, C. J., and Carlyon, R. P. (**1995**). "Auditory grouping," in *Hearing*, edited by B. C. J. Moore, Academic Press, San Diego, CA, pp. 387–424.

Darwin, C. J., and Ciocca, V. (**1992**). "Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component," J. Acoust. Soc. Am. **91**, 3381–3390.

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.

Desimone, R., and Duncan, J. (**1995**). "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222.

Dyson, B. J., and Quinlan, P. T. (**2003**). "Feature and conjunction processing in the auditory modality," Percept. Psychophys. **65**, 254–272.

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.

Hall, M. D., Pastore, R. E., Acker, B. E., and Huang, W. (**2000**). "Evidence for auditory feature integration with spatially distributed items," Percept. Psychophys. **62**, 1243–1257.

Hartmann, W. M. (**1984**). "A search for central lateral inhibition," J. Acoust. Soc. Am. **75**, 528–535.

Hill, N. I., and Darwin, C. J. (**1996**). "Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning," J. Acoust. Soc. Am. **100**, 2352–2364.

Hukin, R. W., and Darwin, C. J. (**1995**). "Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification," Percept. Psychophys. **57**, 191–196.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Kashino, M., and Warren, R. M. (**1996**). "Binaural release from temporal induction," Percept. Psychophys. **58**, 899–905.

Liberman, A. M., Isenberg, D., and Rakerd, B. (**1981**). "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," Percept. Psychophys. **30**, 133–143.

McAdams, S. (**1989**). "Segregation of concurrent sounds. I. Effects of frequency modulation coherence," J. Acoust. Soc. Am. **86**, 2148–2159.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.

Petkov, C. I., O'Connor, K. N., and Sutter, M. L. (**2007**). "Encoding of illusory continuity in primary auditory cortex," Neuron **54**, 153–165.

Powers, G. L., and Wilcox, J. C. (**1977**). "Intelligibility of temporally interrupted speech with and without intervening noise," J. Acoust. Soc. Am. **61**, 195–199.

Remez, R. E. (**2005**). "Perceptual organization of speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez, Blackwell, Oxford, UK, pp. 28–50.

Repp, B. H. (**1984**). "Against a role of 'chirp' identification in duplex perception," Percept. Psychophys. **35**, 89–93.

Repp, B. H., Milburn, C., and Ashkenas, J. (**1983**). "Duplex perception: Confirmation of fusion," Percept. Psychophys. **33**, 333–337.

Shinn-Cunningham, B. G., Lee, A. K., and Oxenham, A. J. (**2007**). "A sound element gets lost in perceptual competition," Proc. Natl. Acad. Sci. U.S.A. **104**, 12223–12227.

Sussman, E. S., Horvath, J., Winkler, I., and Orr, M. (**2007**). "The role of attention in the formation of auditory streams," Percept. Psychophys. **69**, 136–152.

Treisman, A. M., and Gelade, G. (**1980**). "A feature-integration theory of attention," Cogn. Psychol. **12**, 97–136.

Verschuure, J., and Brocaar, M. P. (**1983**). "Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise," Percept. Psychophys. **33**, 232–240.

Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.

Warren, R. M., Bashford, J. A., Jr., Healy, E. W., and Brubaker, B. S. (**1994**). "Auditory induction: Reciprocal changes in alternating sounds," Percept. Psychophys. **55**, 313–322.

Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. A., Jr., and Healy, E. W. (**1997**). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," Percept. Psychophys. **59**, 275–283.

Whalen, D. H., and Liberman, A. M. (**1996**). "Limits on phonetic integration in duplex perception," Percept. Psychophys. **58**, 857–870.

The role of spectral modulation cues in virtual sound localization[a)]

Jinyu Qian
Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

David A. Eddins[b)]
Department of Otolaryngology, University of Rochester, Rochester, New York 14642, and International Center for Hearing and Speech Research, Rochester Institute of Technology, Rochester, New York 14623

(Received 15 August 2006; revised 8 August 2007; accepted 5 October 2007)

Sound localization cues generally include interaural time difference, interaural intensity difference, and spectral cues. The purpose of this study is to investigate the important spectral cues involved in so-called head related transfer functions (HRTFs) using a combination of HRTF analyses and a virtual sound localization (VSL) experiment. Previous psychoacoustical and physiological studies have both suggested the existence of spectral modulation frequency (SMF) channels for analyzing spectral information (e.g., the spectral cues coded in HRTFs). SMFs are in a domain related to the Fourier transform of HRTFs. The relationship between various SMF regions and sound localization was tested here by filtering or enhancing HRTFs in the SMF domain under a series of conditions using a VSL experiment. Present results revealed that azimuth localization was not significantly affected by HRTF manipulation. Applying notch filters between 0.1 and 0.4 cycles/octave or between 0.35 and 0.65 cycles/octave resulted in significantly less accurate elevation responses at low elevations, while spectral enhancement in these two SMF regions did not produce a significant change in sound localization. Likewise, low-pass filtering at 2 cycles/octave did not significantly influence localization accuracy, suggesting that the major cues for sound localization are in the SMF region below 2 cycles/octave. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804698]

PACS number(s): 43.66.Qp, 43.66.Ba, 43.66.Pn [RAL] Pages: 302–314

I. INTRODUCTION

Sound localization is a fundamental aspect of auditory system. The important acoustic cues for sound localization are generally believed to be included in head related transfer functions (HRTFs). The interaural differences among HRTFs include interaural time differences (ITDs) and interaural intensity differences information, which are critical cues for azimuth localization, while the monaural spectral features and interaural spectral differences among HRTFs are particularly important for elevation perception (e.g., Strutt, 1907; Kuhn, 1977; Yost, 1981; Blauert, 1983; Musicant and Butler, 1985; Middlebrooks and Green, 1990). One way to obtain HRTFs is empirical measurement (e.g., Wightman and Kistler, 1989a; Wenzel *et al.*, 1990; Kistler and Wightman, 1992; Kulkarni and Colburn, 1998). The measured HRTFs can be saved in a database and used to synthesize sounds in virtual auditory space. Considering that HRTF measurement is time consuming and requires special equipment, an alternative method is to obtain HRTFs from an available database without measuring the actual individual HRTFs (i.e., using nonindividualized HRTFs). Both individualized and nonindividualized HRTFs have been applied in so-called virtual sound localization (VSL) experiments to synthesize virtual three-dimensional (3D) sound sources, which are delivered to the subjects via headphones (e.g., Wightman and Kistler, 1989b; Wenzel *et al.*, 1993, Qian and Eddins, 2006).

Compared to free-field sound localization experiments, where actual 3D sound sources are presented via speakers at specific locations in space, VSL has the advantage of allowing systematic manipulation of acoustic cues separately. A general problem in VSL is the relatively high rate of front–back confusions or difficulty differentiating front from back sound sources and vice versa (e.g., Wightman and Kistler, 1989b; Wenzel *et al.*, 1993; Qian and Eddins, 2006). However, ignoring front–back confusions by resolving them prior to data analysis reveals that the localization of virtual sources can be quite accurate and comparable to the corresponding free-field conditions, even when nonindividualized HRTFs are applied (e.g., Wenzel *et al.*, 1993; Qian and Eddins, 2006).

A series of VSL studies have focused on manipulating the phase and/or the magnitude spectra of the HRTFs (e.g., Kulkarni *et al.*, 1999; Langendijk and Bronkhorst, 2002; Wightman and Kistler, 1989b; Wenzel *et al.*, 1990; Kistler and Wightman, 1992). HRTFs are approximate minimum phase systems (Mehrgardt and Mellert, 1977; Kulkarni *et al.*, 1995), in which the phase spectrum of an HRTF can be uniquely specified by its magnitude spectrum. More specifically, with the property of minimum phase systems, a phase spectrum can be calculated based on the corresponding log-magnitude spectrum (Oppenheim and Schafer, 1989; Mehr-

[a)]Portions of this research were presented at the 29th Midwinter Meeting of the Association for Research in Otolaryngology, Baltimore, MD.
[b)]Author to whom correspondence should be addressed. Electronic mail: david_eddins@urmc.rochester.edu

gardt and Mellert, 1977). The minimum phase model of HRTFs has been tested in several studies (Wightman and Kistler, 1989b; Kistler and Wightman, 1992; Kulkarni *et al.*, 1999). Localization performance in the minimum phase condition is very similar to the performance in the free-field condition when individualized HRTFs are used in virtual 3D sound synthesis.

Compared to studies of the phase spectrum, researchers have focused more on specific features of the magnitude spectrum of HRTFs. Several previous studies have shown that the frequencies contributing to peaks and valleys in HRTFs provide the cues for elevation perception. Butler and Belendiuk (1977) reported that notches in HRTFs moved systematically with the sound source elevation and Bloom (1977) showed that the manipulation of spectral notch center frequency in the high-frequency region of the pinna transfer function evoked different elevation perceptions. Similarly, Langendijk and Bronkhorst (2002) studied the contribution of different frequency regions by selectively removing from the HRTF spectral components in $\frac{1}{2}$-, 1-, and 2-octave bands in the frequency range above 4 kHz. They determined that the up–down cues were determined mainly by the 1-octave band from 6 to 12 kHz and that the front–back cues were determined by the 1-octave band from 8 to 16 kHz.

Different from most studies discussed earlier, in which selected portions of HRTFs have been manipulated, Kulkarni and Colburn (1998) investigated the effect of spectral details on sound localization based on the Fourier transform of HRTFs. They smoothed the HRTFs to different degrees by omitting certain numbers of Fourier coefficients used in HRTF reconstruction. Their results showed that the performance of localization was largely unaffected by the HRTF smoothing, even when the number of Fourier coefficients used in reconstruction reduced from 512 to 16.

Since HRTFs are functions in the audio frequency domain, the Fourier transform of HRTFs can be referred to as functions in the spectral modulation frequency (SMF) or spectral envelope frequency domain, where SMF is in units of cycles per octave. Therefore, the HRTF smoothing is essentially a low-pass filtering of the HRTFs in the spectral modulation frequency domain. Furthermore, the minor effect on performance of sound localization using highly smoothed HRTFs suggests that important spectral cues in HRTFs might correspond to the relatively low SMF region, and that high SMFs computed from HRTFs might not be critical for cuing the sound source location.

Macpherson and Middlebrooks (2003) investigated the role of spectral cues in the SMF domain by determining the magnitude of interference a sinusoidal ripple superimposed on a flat-spectrum noise had on free-field sound localization. In separate experiments, the ripple density, depth, and phase of the interfering stimulus component were varied and sound localization was measured for azimuths between −30° and 30° and elevations from −60° to 60° in the front and rear hemifields. A sinusoidal ripple (on a logarithm audiofrequency axis) at a fixed density across the whole frequency region can be represented in the SMF domain by a single SMF component, and the depth and phase of a ripple correspond to the magnitude and phase of the related SMF component, respectively. Therefore, by imposing ripples on the flat noise spectrum, Macpherson and Middlebrooks essentially manipulated the stimulus spectrum in the SMF domain. They determined that the ripple densities between 0.5 and 2.0 ripples/octave (i.e., cycles/octave) induced the greatest error rates in vertical localization. Furthermore, SMF-induced interference was restricted to ripple depths $\geqslant$20 dB and was greatest for a depth of 40 dB. As noted by Macpherson and Middlebrooks, the fact that substantial interference only occurred for relatively high modulation depths (e.g., greater than typically seen in the peak-to-valley troughs of HRTFs), indicates that sound localization appears to be quite robust to SMF interference.

The idea of analyzing spectral cues in SMF domain is based on the growing body of evidence illustrating the tuning to SMF that may reflect "channels" tuned to SMFs. Both psychoacoustical (Eddins *et al.*, 2001; Eddins and Harwell, 2002; Saoji and Eddins, 2002; Eddins and Bero, 2007) and physiological studies (Shamma *et al.*, 1995; Versnel and Shamma, 1998) have suggested a possible auditory encoding procedure in the SMF domain (i.e., a spectrum in the audiofrequency domain can be represented by a bank of filters tuned to SMF). For example, Saoji and Eddins (2007) used a spectral modulation masking procedure to illustrate tuning to SMF while Eddins and Harwell (2002) used a selective adaptation procedure to show similar tuning to SMF. Several physiological studies (e.g., Versnel *et al.*, 1995; Versnel and Shamma, 1998) showed that units in the primary auditory cortex (AI) are tuned to different spectral ripples (i.e., SMFs) and that those units are topographically segregated with respect to the ripple responses.

In addition to the spectral investigation in the SMF domain, some advanced signal processing techniques have been applied to manipulate the spectral cues in HRTFs. One major example is principal component analysis (PCA; Martens, 1987; Kistler and Wightman, 1992; Chen *et al.*, 1995; Qian and Eddins, 2004, 2005). Principal component analysis is one type of eigenanalysis and it is an efficient way to represent the underlying structure in a highly variable data set (e.g., HRTFs). PCA represents the original data set as a smaller orthogonal set, known as principal directions (PDs). Linear combinations of PDs can account for most of the variance of the original set. The first PD can account for the largest portion of variance among the original data and successive PDs account for progressively less of the variance. PCA is essentially an analysis of the variance among the HRTFs in a data set. The HRTFs can be decomposed into a series of principal directions with their corresponding weights. These PDs and their weights can be used to reconstruct a new set of HRTFs later. The more PDs used in reconstruction, the more similar the reconstructed HRTFs are to the original set.

In our previous work (Qian and Eddins, 2004, 2005), the PCA technique was applied to study HRTFs and the high dimensional HRTF data set was represented using low dimensional basis functions or PDs. For example, the original 1250 HRTFs of an HRTF set can be very closely represented from linear combinations of only the first seven PDs by weighting them differently with the directional-dependent

weight values. The PCA was performed on each individual HRTF set and on the HRTFs for the left and right ears separately. Both the signal processing and virtual sound localization results indicated a relationship between several specific PDs and sound source direction. For example, the first PD carried azimuthal information, while the second and third PDs were associated with high and low elevations, respectively.

Principal component analysis can represent HRTFs more efficiently in the sense of reducing the data dimensions. However, it is generally difficult to associate a signal processing procedure with perceptual localization process in the auditory system. In other words, it might not be realistic to assume the auditory system is performing similar complex analyses or computations. It is possible, however, to gain some insight into the location-dependent features coded in HRTFs through PCA analysis. For example, given that the first three PDs can usually account for greater than 95% of the variance among a complete set of HRTFs (Qian and Eddins, 2004, 2005), if there are physiological-related channels which produce relatively consistent output corresponding to the information emphasized by the first three PDs, then these channels might be associated with or even critical to sound localization. Given the relationship between PD2 and high elevations, if there is a physiological channel which produces relatively consistent output corresponding to the information emphasized by PD2, then this channel might be related to high-elevation perception.

The possibility of a channel-based mechanism for analyzing spectral information (e.g., the spectral cues coded in HRTFs) is supported by both psychoacoustical and physiological studies as discussed earlier. These channels may be referred to as SMF channels and represent a domain related to the Fourier transform of the HRTFs. Since the PDs essentially extract the primary variance among HRTFs and emphasize the directional-dependent information of HRTFs in the audio frequency domain, the Fourier transform of the PDs is expected to reflect this variance and the important direction information in a series of putative SMF channels.

In the current study, we first analyzed PDs in the SMF domain, attempting to find the SMFs that might be related or critical to sound localization. Following this analysis, a virtual sound localization experiment was conducted to investigate the perceptual influences of manipulating the HRTFs in certain SMF regions. These manipulations build on several previous studies with a goal of better determining the relation between specific spectral features and vertical sound localization. Specifically, the SMF filtering conditions of the present experiment are similar to the smoothing operations performed in previous studies (e.g., Asano *et al.*, 1990; Kulkarni and Colburn, 1998; Langendijk and Bronkhorst, 2002), however, here we are able to directly relate this global filtering process to the relative importance of different regions in the SMF domain. Likewise, the SMF filtering procedure used here is similar in spirit to the experiments reported by Macpherson and Middlebrooks (2003). Although both sets of experiments consider sinusoidal spectral modulation, there are several important differences between the current study and the experiments reported by Macpherson and Middlebrooks (2003).

A primary goal of the current study is to determine whether or not certain SMF regions are related to specific sound source locations, as suggested by earlier PCA analyses of HRTFs (Qian and Eddins, 2004, 2005). While the results of Macpherson and Middlebrooks (2003) indicated that sinusoidal spectral modulation within the range of 0.5 and 2.0 cycles/octave (cyc/oct) and depths greater than 20 dB superimposed upon a broadband noise stimulus produce consistent localization errors, the design of their experiment did not permit an assessment of the potential relation between SMF regions and specific sound source locations. Furthermore, it is not known whether such discrete stimulus-induced interference leads to the same changes in sound localization performance that might occur as a result of introducing systematic modifications spanning a specific range in the SMF domain to the entire set of HRTFs in a virtual localization paradigm. To that end, the current study systematically investigates the potential effects of both filtering and enhancement of specific SMF regions on sound localization. It should also be noted that Macpherson and Middlebrooks (2003) used a free-field localization task, in which manipulation of the sound source spectra resulted in correlated changes in the magnitude and phase spectra of the stimuli as well as the ITDs cues. In the current study, we use VSL to systematically vary the spectral cues associated with the directional information provided by HRTFs. Since VSL allows one to separately alter different cues, and the primary goal of the current study is to explore the roles of spectral modulation cues in sound localization, the current study limits the spectral manipulation to the magnitude spectra of certain SMF components and preserves the phase spectra as well as the original ITDs. Thus the major goals of the current study are to test the hypothesis, based on the previous PCA analysis, that there is a specific relation between particular sound source locations and specific SMF regions and to test the hypothesis that attenuation in those (putative) regions degrades sound localization and/or enhancement in those regions improves sound localization.

II. METHODS

A. HRTF data source

The HRTFs used in the current study are available from the Tucker-Davis Technologies (TDT) HRTF database and include 26 sets of HRTF measured from 26 subjects. There are 360 HRTFs for each ear of each set corresponding to sound sources presented from 360 (36×10) different directions, including 36 azimuths (θ) and 10 elevations (ϕ). The azimuthal range is from −170° to 180° in steps of 10°, with 0° azimuth referring to the median plane of the subject. Those directions greater than 90° or less than −90° are behind the subject. The left side is negative azimuth and the right side is positive azimuth. Elevations range from −30° to 60° in steps of 10°, where negative directions are those be-

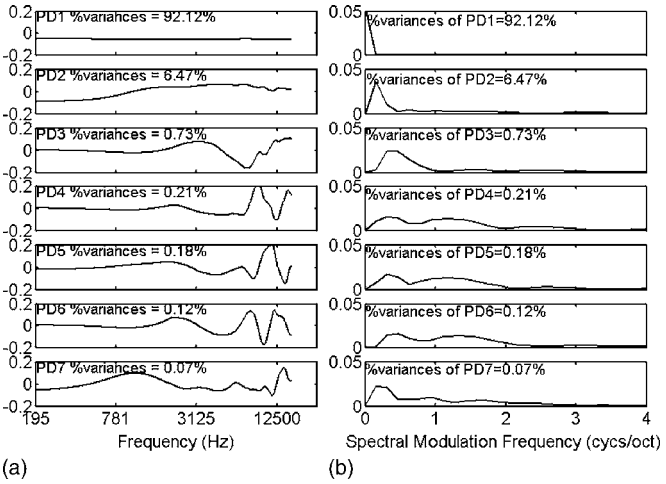


FIG. 1. (a) First seven PDs and (b) their corresponding FTPDs computed from left ear of HRTF set 11. The % of variance refers to the amount of variance that each PD can explain. Due to the lack of detail in the FTPD at high SMFs, only information below 4 cyc/oct is shown.

low the horizontal plane in front of the subject. The sample frequency of the HRTFs was 50 kHz and the duration of the impulse response was 10.24 ms.

B. Spectral modulation characteristics of HRTFs revealed by PCA

Principal directions derived from PCA are functions in the audio frequency domain, and they contain direction-specific information in this domain. Therefore, the Fourier transform of the PDs are expected to reveal the direction-specific information in the spectral modulation frequency domain. Based on this assumption, PCA was first applied to derive PDs for each ear of each HRTF set (including 360 HRTFs). PCA was only applied to the magnitude spectra, represented by components that were equally spaced on a logarithmic frequency scale. The phase spectra were unchanged and were not included in the PCA analysis. Next, the Fourier transforms of the derived PDs (FTPDs) were computed. Figure 1 shows an example of the first seven PDs and their corresponding FTPDs computed from the left ear of one HRTF set.

As shown in this example, the first seven PDs can account for approximately 99.9% of the total variance in 360 HRTFs. In other words, these PDs comprise most of the directional-dependent information in HRTFs. Since FTPDs can be considered as the representation of PDs in the SMF domain, the first seven FTPDs can represent most of the directional-dependent information in the SMF domain. Observation of the first seven FTPDs [see Fig. 1(b)] reveals that the major power of these FTPDs is located in the lower SMF region, while there is little power above 2 cyc/oct. Therefore, if the spectral cues coded in HRTFs are further processed by SMF channels in the auditory system, the directional-dependent information should coincide with the low SMF region (<2 cyc/oct). The lack of robust details in the FTPD above 2 cyc/oct was observed for all 26 HRTF sets. Our observation on FTPD pattern is consistent with Macpherson and Middlebrooks (2003), who reported that interference in the SMF region above 2 cyc/oct did not strongly influence processing of spectral cues for sound localization.

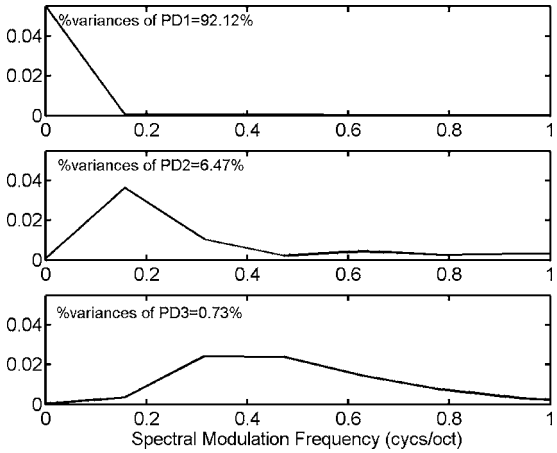


FIG. 2. The first three FTPDs computed from the left ear of HRTF set 11. The% of variance refers to the amount of variance explained by each PD.

Since our previous studies revealed specific relationships between the first three PDs and sound source locations, and the first three PDs generally can account for more than 98% of total variance in every set of HRTFs, the fine structure of the first three FTPDs warrant further study (see Fig. 2 for an example). In general, for each of all 26 HRTF sets, there are several common features: (1) The first three FTPDs have little power located in the SMF region above 1 cyc/oct; (2) the first FTPD has a consistent peak at the first SMF (i.e., DC); (3) the second FTPD has a consistent peak at the SMF region around 0.15 cyc/oct; and (4) the peak of the third FTPD varies from 0.25 to 0.6 cyc/oct across different HRTF sets. There are 13 out of 26 HRTF sets resulting in a peak of the third FTPD at around 0.3 cyc/oct.

If SMF channels exist and the spectral cues used in sound localization are further processed by SMF channels, the observations regarding the FTPDs for the 26 HRTF sets would suggest that: (1) The major spectral cues for sound localization may be processed in the SMF channels tuned to frequencies below 2 cyc/oct. Furthermore, the SMF channels below 1 cyc/oct may be relatively more important than the channels between 1 and 2 cyc/oct. (2) Specific SMF channels below 1 cyc/oct might be related to specific directions in sound localization. Regardless of the existence of SMF channels, observations based on the FTPDs provide insight into the nature and importance of various spectral features at a level of detail not provided in previous experiments. Based on these observations, we designed the virtual sound localization experiment to investigate the perceptual influences of manipulating spectral information in the SMF domain.

C. Customizing the HRTFs for individuals

In the virtual sound localization experiment, nonindividualized HRTFs were used to synthesize virtual stimuli. Considering the large individual differences among the different HRTF sets, an arbitrarily selected HRTF set may result in very different VSL performance for different subjects.

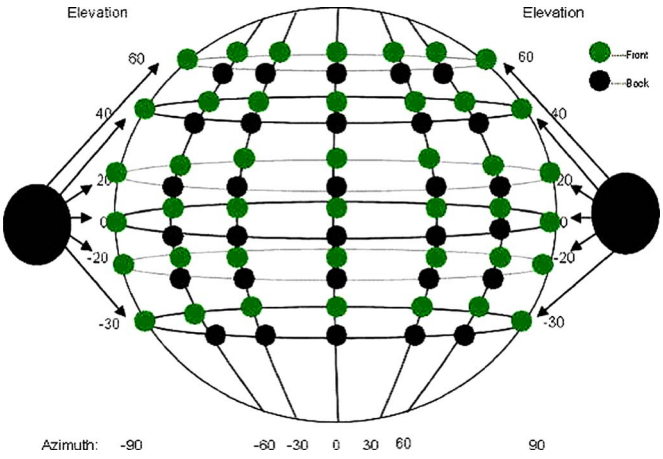


FIG. 3. (Color online) VSL subject interface, the 3D coordinate systems with 72 direction spots.

Therefore, an HRTF customization procedure has been applied for every subject to obtain individually customized HRTFs.

The HRTF customization procedure combined both subjective and objective psychoacoustical evaluations and included three phases: (1) Preselecting the 6 best matching HRTF sets from the 26 available sets, (2) determining the single best matching HRTF set from the 6 preselected sets, and (3) determining the proper factor to scale the selected HRTF set in audio frequency. This customization procedure is described in detail in the Appendix .

D. Virtual sound localization experiment

Six young adults served as paid volunteers. Subjects had normal hearing based on pure tone thresholds ($\leqslant$20 dB HL, ANSI, 1996), negative history of ear disease, and ranged in age from 23 to 27 years. A total of seven conditions were run for each subject, including a baseline condition and six HRTF-modified conditions. The stimulus was a 250-ms Gaussian noise gated on and off with a 10-ms linear rise–fall envelope, and then filtered by a specific HRTF from the individually customized HRTF set in the baseline condition or modified HRTFs in conditions 1 through 6 based on that set. The sound stimuli were generated digitally at a sampling rate of 48 828.125 Hz by TDT System3 hardware, and controlled by the TDT SykofizX software. Sounds were delivered to the subjects via Sennheiser HD265 headphones at an overall level of 70 dB SPL. The experiment was conducted in a standard double-walled sound booth.

Subjects identified the apparent source position of the synthesized stimuli via graphical user interface displayed on an LCD monitor (see Fig. 3) that represented the three-dimensional virtual space on the two-dimensional display using concentric ovals and color to simulate the third dimension. The stimuli were filtered by the individually customized HRTF set, simulating sound sources from 72 (12 azimuth $\times$ 6 elevations) different directions. The azimuth ranged from $-150°$ to 180°, in steps of 30°, and the elevation included $-30°$ and from $-20°$ to 60°, in steps of 20°. The VSL experiment consisted of two sessions: A training session and a test session. The training session was included to allow the subject to be familiar with the graphical interface and to learn to map the 3D physical space to the graphical representation of that space. It is unknown whether increased localization error is associated with mapping the physical space to the graphical response space used here, or whether such error would be greater or less than the error associated with verbal responses used to describe the 3D physical space. Nevertheless, the training procedure was included, in part, to reduce the likelihood of such response errors.

1. Subject training

In the training session, subjects were asked to listen to a sound presented via headphones and to choose the apparent direction of the sound stimulus from a three-dimensional coordinate system (as shown in Fig. 3). The same 72 directions were used in both the training and test sessions as described earlier. To choose the apparent direction, subjects clicked the mouse when the cursor overlapped a specific spot in the coordinate system. Before the training, subjects were instructed on responding in this coordinate system.

Following the subject's response, feedback was provided indicating the expected direction for each sound stimulus. There were 144 judgments (two judgments for each of the 72 directions) in each block of training, and subjects finished 6 blocks of training before starting the test session. All stimuli were presented in random order. The training session lasted approximately 2 h in total, conducted on two consecutive days, 1 h per day.

2. Applying filters to HRTFs in the SMF domain

The procedure in the test session was very similar to that in the training session, however, no feedback was provided and subjects made ten judgments for each direction. The test session consisted of seven conditions: The baseline and six other conditions (i.e., conditions 1 through 6). For the last six conditions, the Fourier transforms of the HRTFs (FTHRTFs) were first computed, followed by specific modifications of FTHRTFs in the SMF domain. In conditions 1 and 2, notch filters in the SMF domain were applied to the customized HRTF set, centered at 0.25 and 0.5 cyc/oct for conditions 1 and 2, respectively. The notch width for both conditions was 0.3 cyc/oct. In conditions 3 and 4, low pass filters in the SMF domain were applied to the customized HRTF set with cutoff frequencies of 1 and 2 cyc/oct for conditions 3 and 4, respectively. Figure 4 shows an example of the original and modified HRTFs in frequency domain (i.e., original and filtered HRTFs, see the first two rows) and in the SMF domain (i.e., original and filtered FTHRTFs, see the last two rows).

3. Spectral enhancement of HRTFs in SMF domain

In addition to SMF filtering, two additional conditions included enhancement to certain SMF components of the FTHRTFs. The enhanced regions for conditions 5 and 6 [see Figs. 4(e) and 4(f)] were 0.1–0.4 cyc/oct (centered at 0.25 cyc/oct, width of 0.3 cyc/oct) and 0.35–0.65 cyc/oct

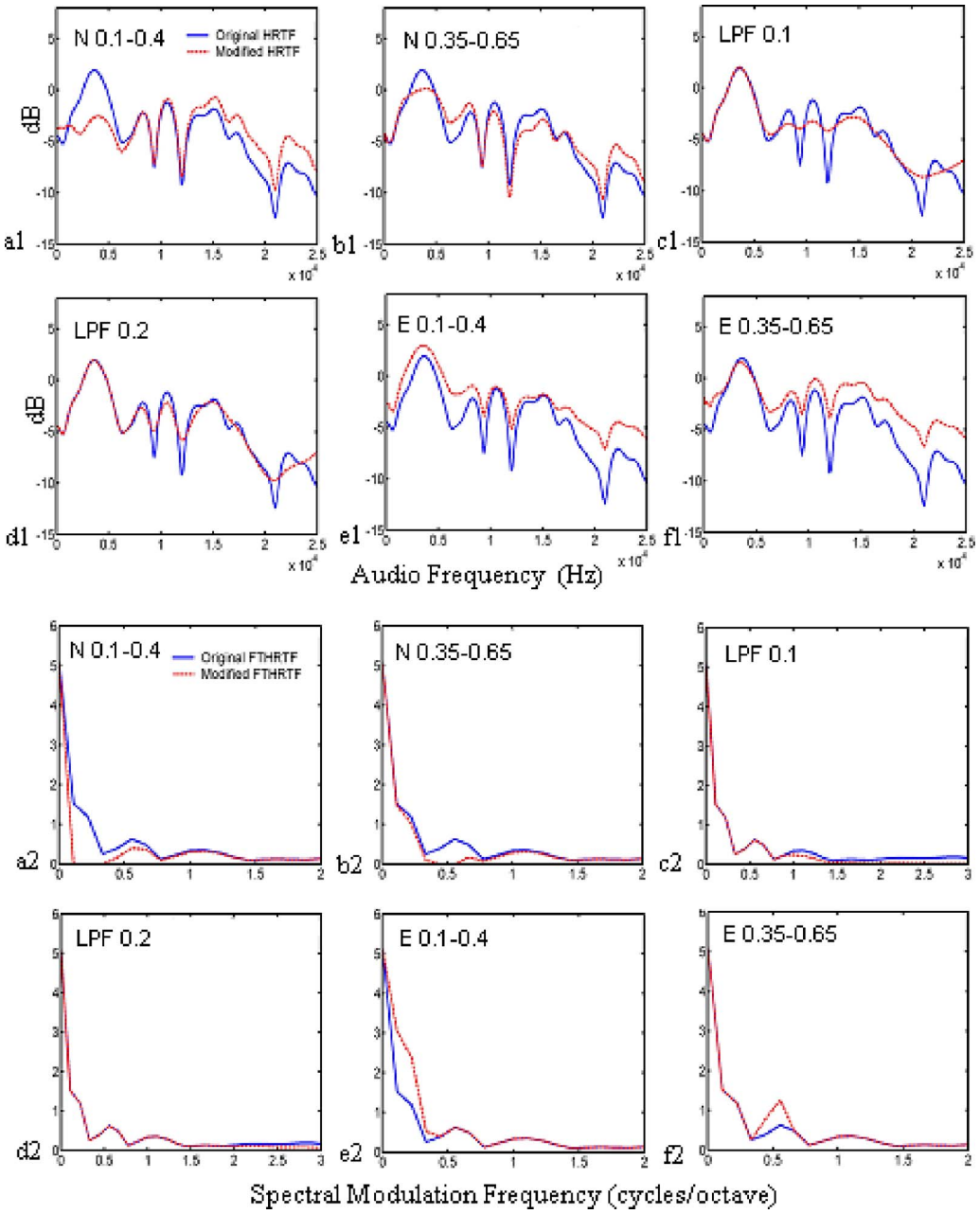


FIG. 4. (Color online) The original and modified HRTFs (azimuth=0°, elevation=0°) in the frequency domain [(a1)–(f1)] and in the SMF domain [(a2)–(f2)]. Data are from HRTF set 11 scaled by 0.917 (or −0.125 octave). Panels (a1) and (a2) were modified with a SMF notch filter between 0.1 and 0.4 cyc/oct; (b1) and (b2) were modified with a SMF notch filter between 0.35 and 0.65 cyc/oct; (c1) and (c2) were low pass filtered at 1 cyc/oct; (d1) and (d2) were low pass filtered at 2 cyc/oct; (e1) and (e2) were enhanced between 0.1 and 0.4 cyc/oct; (f1) and (f2) were enhanced between 0.35 and 0.65 cyc/oct.

(centered at 0.5 cyc/oct, width of 0.3 cyc/oct), respectively. The amount of enhancement for each SMF component was 3 dB or double the linear magnitude.

To synthesize HRTF filters for VSL, both the magnitude and the phase spectra are required. Since all HRTF modifications described previously are based on the magnitude of the HRTFs only, the phase spectrum still needs to be specified. In the present study, we preserved the original phases for all HRTF modifications, including the phases in audio frequency domain and the phases in the SMF domain. In addition, the modifications were limited to the spectral cues, so the ITDs were preserved as the original values from the HRTF database.

All stimuli were presented in a random order in each test condition. Subjects completed one more training block (approximately 15 min) before the test session of each day.

III. RESULTS

As detailed in Sec. II, VSL was performed in six conditions in which HRTFs were modified in the SMF domain and one baseline condition without modifications. Azimuth and elevation localization performance in these conditions is reported separately in the following to better illustrate the different effects of HRTF manipulations on localization in both azimuth and elevation.

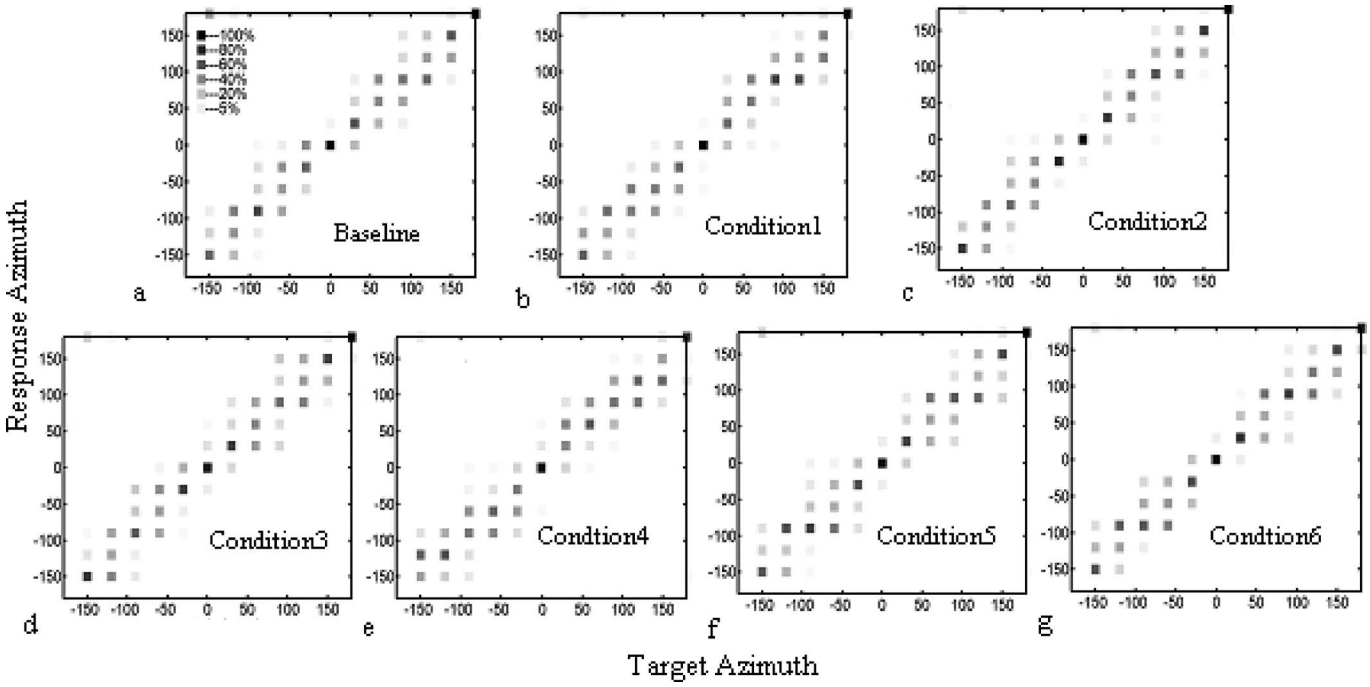


FIG. 5. The target-response plots for azimuth location (front–back confusion resolved) from subject S4 in the baseline (a) and six HRTF-manipulated conditions (b)–(g).

A. Azimuth localization

Consistent with previous VSL experiments (e.g., Wightman and Kistler, 1989b; Wenzel *et al.*, 1993; Qian and Eddins, 2006), front–back confusions were observed in the current study. The average front–back confusion rate across all subjects and conditions was 31.00% (s.d.=7.91%), consistent with the confusion rate reported by Wenzel *et al.* (1993) of 31%. An one-way repeated measures analysis of variance (ANOVA) on ranks indicated no significant differences in front–back error rate[1] across the seven conditions (H_6 =4.351, p=0.629). Due to the relatively high front–back confusion rates typically observed in VSL, the front–back confusions most often are resolved by replacing the response angle by the angle reflected about the vertical plane passing through the subject's ears when the angle between the target and response is greater than the angle between the target and the reflected angle (Wenzel *et al.*, 1993; Qian and Eddins, 2006). All azimuthal localization results reported here reflect such front–back resolution.

Figure 5 shows the target-response plots of the azimuth localization (front–back confusion resolved) from one subject (S4) in the baseline and six HRTF-manipulated conditions as an example. The squares use gray levels to represent different response percentages for particular locations with darkness proportional to the percentage of responses. The specific gray level is calculated by the number of responses at a particular azimuth for the given stimulus azimuth divided by the total number of stimuli with that given apparent azimuth times 100. As shown in this example, the azimuth localization is always quite accurate (this was true for all subjects and all conditions) when front–back confusions were resolved. A two-way repeated measures ANOVA with factors of azimuth and condition indicated a significant effect of azimuth ($F_{11,503}$=410.681, $p<0.001$) but no significant effect of condition ($F_{6,503}$=0.199, p=0.975), consistent with the interpretation that the spectral manipulations in conditions 1 through 6 did not significantly ($p>0.05$) affect azimuth responses.

B. Elevation localization

In general, localization accuracy at low elevations (−30° and −20°) was poorer in conditions 1 and 2 as compared to the baseline condition. No consistent changes were observed in localization performance at middle (0° and 20°) or high (40° and 60°) elevations across subjects and conditions. Figure 6 shows the target-response plots of the elevation localization from one subject (S4) in the baseline and six HRTF-manipulated conditions as an example. Similar to Fig. 5, the squares use gray levels to represent different response percentages for particular locations, with darkness proportional to the percentage of responses. Since the up–down confusion rates were relatively low (mean=10.80%, s.d.=4.01% across all subjects and all conditions), elevation localization is presented without resolving the up–down confusions.

To quantitatively compare the accuracy of elevation localization in the baseline and the other six conditions, the average elevation error was computed as a function of target elevation. Elevation error is defined as the unsigned elevation difference (in degrees) between each target elevation and the response elevation. This metric emphasizes the accuracy of elevation perception without including the accuracy of azimuth responses. The elevation error is calculated prior to resolving the up–down confusions and therefore includes errors due to up–down confusions. Regional averages of the elevation error in the baseline were also computed to evaluate whether particular elevation regions have different

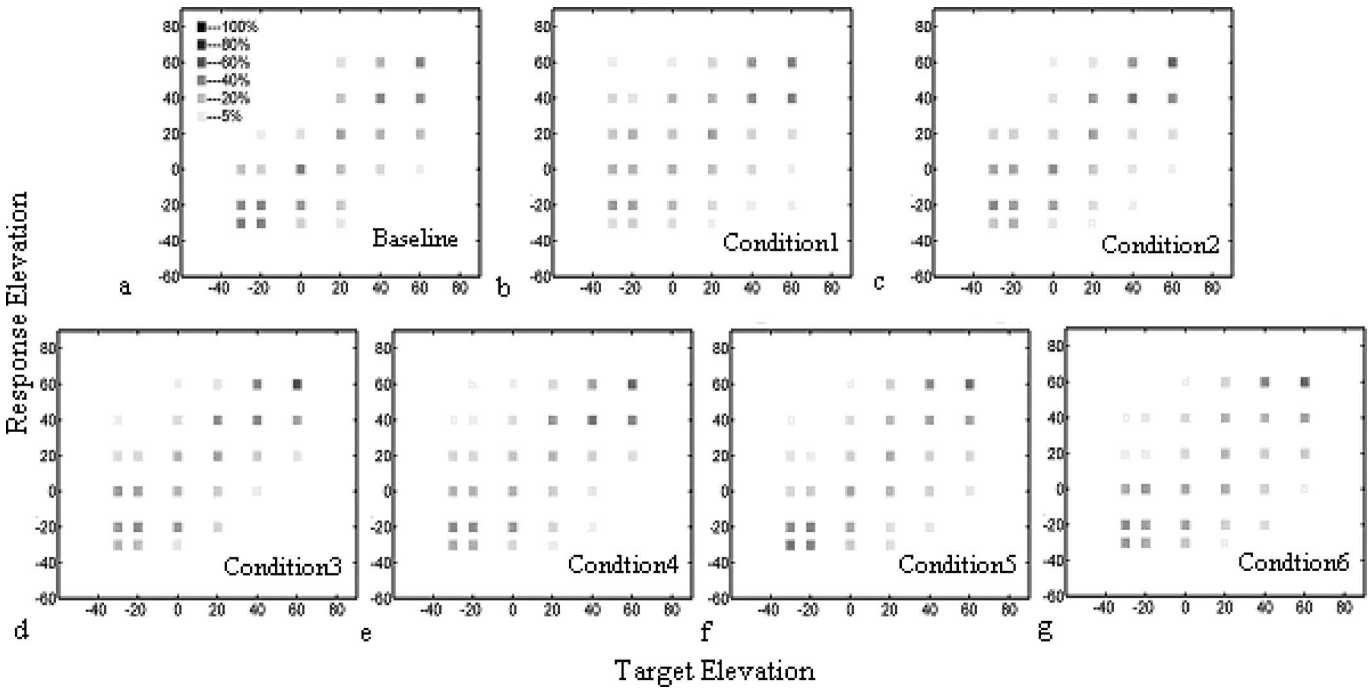


FIG. 6. The target-response plots for elevation location (up–down confusion not resolved) from subject S4 in the baseline (a) and six HRTF-manipulated conditions (b)–(g)

amounts of elevation error. There were a total of nine regions including low (−20° and −30°), middle (0° and 20°), and high (40° and 60°) elevations and front (−30°, 0°, and 30°), side (−120°, −90°, −60°, 60°, 90°, and 120°), and back (−150°, 150°, and 180°) azimuths. The results showed no significant difference in elevation error among these regions in the baseline condition ($F_{8,45}=1.43$), $p=0.209$).

Furthermore, to emphasize the effects of the specific HRTF manipulation, the differences between the average elevation error function in each of the HRTF-manipulated conditions (i.e., conditions 1 through 6) and that in the baseline condition are shown in Fig. 7.

Close inspection of the individual data in Fig. 7 reveals very different performance for subject 6 relative to the other subjects, particularly in conditions 1 and 5. As a result of this deviant data, the mean functions (in Fig. 8) reflect the average for the remaining five subjects, omitting subject 6.

To determine the influence of SMF filtering (conditions

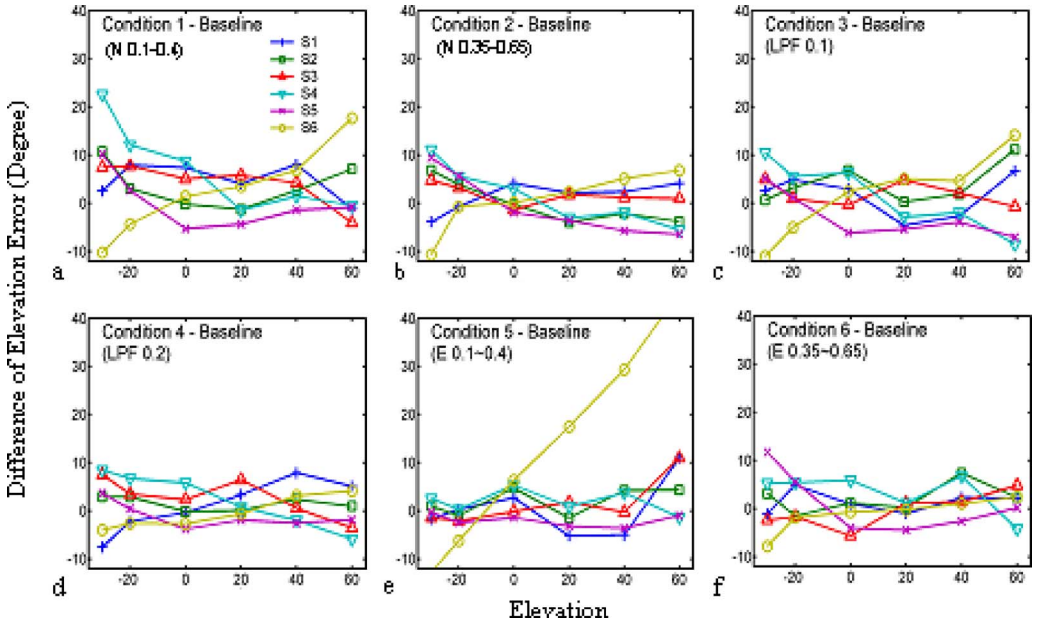


FIG. 7. (Color online) The differences of the average elevation error function in the baseline condition and in each of the HRTF-manipulated conditions for individual subjects.

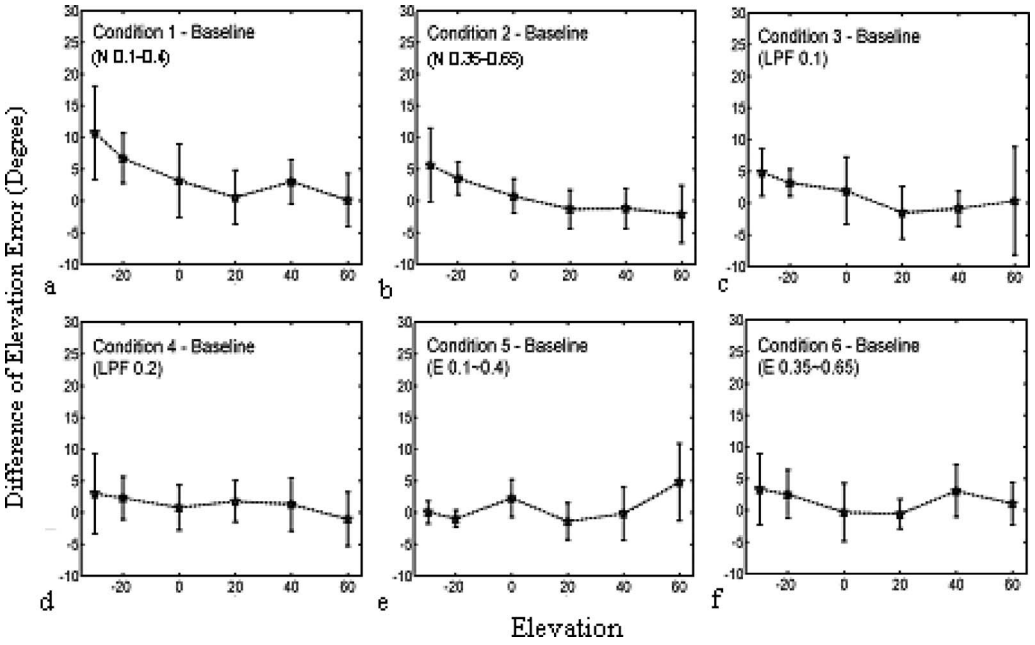


FIG. 8. Mean and standard deviations of differences of the average elevation error function in the baseline condition and in each of the HRTF-manipulated conditions. Data are computed across the first five individual subjects with S6's data excluded.

1 to 4), and SMF enhancement (conditions 5 and 6), statistical analyses were used to compare performance among conditions and to the baseline performance. A three-way ANOVA indicated a significant effect of subject ($F_{5,251}=2.919$, $p=0.015$), condition ($F_{6,251}=3.727$, $p=0.002$), and elevation ($F_{5,251}=5.364$, $p<0.001$). Subsequent pairwise multiple comparisons indicated that elevation errors were significantly greater for condition 1 than both the baseline and condition 2 while elevation errors among other conditions were not significantly different. Furthermore, elevation errors were significantly greater for −30° than for −20° or 40° and were greater for 20° than −20° or 40°. To explore possible differences among conditions within each elevation, separate one-way ANOVAs were computed for each elevation with condition as the factor. The average elevation error was significantly different across conditions for elevations of −30° ($F_{6,30}=2.60$, $p<0.05$) and −20° ($F_{6,30}=3.76$, $p<0.01$) but not for other elevations (all $p>0.05$). To explore individual differences in elevation errors across conditions, separate two-way ANOVAs were computed for each elevation with condition and subject as the factors. Consistent with the observation of the individual data in Fig. 7, there were significant individual differences [$F(5,30)=5.85$, $p<0.01$ at −30°, $F(5,30)=14.12$, $p<0.01$ at −20°]. Furthermore, pairwise comparisons were performed to compare the elevation error in the baseline condition to that in each of the HRTF-manipulated conditions. Considering the significant individual differences, the pairwise comparisons (t-tests) were performed on each individual's data separately (with the degree of freedom[2] of 239). In condition 1 (SMF notch filter between 0.1 and 0.4 cyc/oct), five out of six subjects showed significantly greater average elevation error ($p<0.01$) in the low elevation region (−30° and −20°) compared to the baseline condition. In condition 2 (SMF notch filter between 0.35 and 0.65 cyc/oct), four out of six subjects showed significantly greater average elevation error ($p<0.01$) in the low elevation region (−30° and −20°) compared to the baseline. In conditions 3–6 (SMF low pass filter and enhancement), there were no systematic differences from the baseline condition, with only 5 out of 24 possible comparisons reaching significance with a 0.01 criterion. The significant differences observed at low elevations (−30° and −20°) in condition 1 and condition 2 (notch filters in the SMF domain) are consistent with the PCA analysis reported in Sec. II. The PCA analysis revealed that the SMF region around 0.3 cyc/oct is related to low elevations. Conditions 1 and 2 have an overlap region at 0.3 cyc/oct. When the power in this region is reduced or removed, less accurate elevation response (or greater average elevation error) is observed at low elevations.

Also revealed in Fig. 7 is the similarity between the average elevation error function in the baseline condition and condition 4 in which the HRTFs were low pass filtered at 2 cyc/oct. While the previous analyses have focused on differences among conditions, a correlation analysis may be used to better gauge the similarity between specific conditions. The correlation between the functions relating change in elevation error to azimuth in the baseline and in each of the six HRTF-modified conditions were computed for each of the six listeners and are shown in Table I. Significant correlation coefficients highlight the similarity among conditions. The similarity between the error functions in the baseline and condition 4 (low pass at 2 cyc/oct) reached significance for three out of six subjects. This significant similarity suggests that the SMF region above 2 cyc/oct may not be important for sound localization, which is consistent with the conclusions drawn by Macpherson and Middlebrooks (2003). In addition, note that the similarity between the error functions in the baseline and condition 3 (low pass at 1 cyc/oct) reached significance for only one out of six sub-

TABLE I. Correlation coefficients between the baseline and each of the HRTF-manipulated conditions for all six subjects.

	Subject					
Condition	S1	S2	S3	S4	S5	S6
1	0.50	0.16	0.80	−0.64	−0.19	−0.21
2	0.54	0.64	0.95[a]	−0.29	0.09	−0.12
3	0.57	0.26	0.90[a]	−0.32	0.39	−0.15
4	0.04	0.89[a]	0.85[a]	0.08	0.92[a]	0.62
5	0.34	0.44	0.64	0.86[a]	0.97[a]	−0.29
6	0.79	0.35	0.69	0.52	−0.51	0.44

a: $p<0.05$.

jects, suggesting that the SMF region between 1 and 2 cyc/oct is important for VSL. Although relatively consistent data were observed for the correlation between the baseline and condition 4, there are large individual differences in the correlation coefficients between the baseline and the other manipulated conditions. This highlights the lack of similarity among most of the modified and the baseline conditions.

The spectral enhancement in conditions 5 (0.1–0.4 cyc/oct) and 6 (0.35–0.65 cyc/oct) did not seem to improve or degrade the elevation responses significantly across all subjects (based on the ANOVA and pairwise comparisons, all had p values greater than 0.05), although three out of six subjects showed significant differences between condition 5 and the baseline condition in localizing virtual stimuli at elevation of 60°. Only one subject (S5) showed a slight overall improvement in condition 5.

IV. DISCUSSION

A. Relationship among VSL, PCA, and SMF filtering

PCA has been applied to HRTF data by previous investigators (e.g., Kistler and Wightman, 1992; Martens, 1987; Qian and Eddins, 2004, 2005) to extract the important features of HRTFs. Those important features can be represented using principal directions (or basis functions) and their corresponding weights derived from the HRTFs. The relationship between certain PDs and some specific sound source locations has been suggested both by theoretical analyses and by empirical data from VSL experiments (Kistler and Wightman, 1992; Qian and Eddins, 2004, 2005).

Given the relationship between PDs and sound source locations, an important question is whether or not it is possible or reasonable to associate the PCA procedure with the perceptual localization process. Although it is difficult to imagine such a direct association, we may still gain some insight into the location-dependent features coded in HRTFs through PCA analysis. As detailed in Sec. I, the Fourier transform of PDs (i.e., FTPDs) are functions in the SMF domain and there is physiological evidence of systematic maps of spectral modulation coding in the auditory cortex (e.g., Shamma *et al.*, 1995; Versnel *et al.*, 1995; Versnel and Shamma, 1998). On the other hand, the PD-sound source location relationship could be transformed to a SMF region-sound source location relationship. For example, PD3 was found to be associated with low elevations (Qian and Eddins, 2004, 2005), while FTPD3 tended to emphasize the SMF region between 0.25 and 0.6 cyc/oct (with a peak near 0.3 cyc/oct). Therefore, this SMF region might be related to low elevation perception.

The relationship between SMF regions and sound localization derived from PCA analysis was tested here using a VSL experiment. The VSL results from the present study revealed:

(1) Azimuthal localization was not significantly affected by manipulations of the HRTFs in the SMF domain in conditions 1 through 6.
(2) Notch filters applied in the SMF domain between 0.1 and 0.4 cyc/oct and between 0.35 and 0.65 cyc/oct had the most pronounced effect on low-elevation localization accuracy.
(3) Low passing the HRTFs at 2 cyc/oct did not alter elevation localization.

Low passing the HRTFs at 2 cyc/oct had a negligible effect on localization performance, consistent with previous studies. Macpherson and Middlebrooks (2003) used a spectral modulation interference paradigm in which they imposed a single SMF component on a flat-spectrum stimulus and measured VSL performance. In separate conditions, they varied the SMF from 0.25 to 8 cyc/oct. Their results indicated that SMFs greater than 2 cyc/oct did not strongly influence spectral processing for sound localization. Kulkarni and Colburn (1998) performed HRTF smoothing in virtual stimulus synthesis by low passing the HRTFs in the SMF domain, and their VSL results showed that the highly smoothed HRTFs did not affect the perceived location of a sound. However, the significant differences between conditions 1 and 2 (notch filters between 0.1–0.4 and 0.35–0.65 cyc/oct) and the baseline condition contrast with Macpherson and Middlebrooks' (2003) study, where the SMF components less than 0.5 cyc/oct were reported to have no strong influence on sound localization. This may be related to the different SMF manipulation methods. Rather than superimposing spectral modulation at a single frequency onto a noise stimulus in an interference paradigm (i.e., Macpherson and Middlebrooks, 2003), the SMF filtering in conditions 1 and 2 of the present study involved the modification of an entire HRTF set over a specific SMF region that included a number of SMF components. In addition, the present study employed a VSL procedure using individually customized HRTFs while Macpherson and Middlebrooks (2003) conducted free-field sound localization experiment. The different types of localization tasks may also lead to differences in localization performance.

Overall, the filtering in conditions 1 and 2 had a significant effect on low-elevation perception, while no such effect was found for high-elevation perception for most of the six subjects. However, the elevation error analysis (see Figs. 7 and 8) did show significantly higher average elevation errors at high elevations (e.g., 60° and 40°) in conditions 1 and 3 compared to the baseline condition for two subjects (S2 and S6), especially for subject S6. Note that although the HRTF-manipulated conditions led to reduced accuracy for low elevations for most of our subjects, HRTF manipulations re-

sulted in reduced accuracy for high elevations for S6. In addition, the extent to which localization accuracy was influenced by HRTF manipulations was much greater for S6 (e.g., conditions 1 and 5) than for the other five subjects (in conditions 1 and 2). Perhaps S6 represents another group of listeners for whom HRTF-manipulated conditions influences high- rather than low-elevation perception. Therefore it may be helpful to recruit a larger group of subjects for further studies. Nevertheless, the present data do confirm the relationship observed in our previous (Qian and Eddins, 2004, 2005) and current PCA analyses indicating that low spectral modulation frequencies are associated with sound localization at low elevations. PCA analyses did not indicate systematic relationships between other SMF regions and sound locations, or did the behavioral measurements reported here.

The notion of relating SMFs to sound localization originally arose from the assumption of SMF channels in the auditory system. However, the present study establishes the influence of SMF filtering on sound localization regardless of the presence of SMF channels. Considering the importance of SMFs between 0.1 and 2 cyc/oct in elevation perception, deficits in the perception of spectral features within this SMF range may lead to poor elevation perception. A similar approach was adopted by Liu and Eddins (2004) to investigate the relative contribution of SMF region to vowel identification. Like the present results, they found that filtering in the SMF region below 2 cyc/oct significantly reduced vowel identification.

B. Relationship between VSL and SMF enhancement

The significant increase in elevation error observed in conditions 1 and 2 compared to the baseline condition suggests the importance of the SMF region between 0.1 and 0.65 cyc/oct for sound localization. However, the spectral enhancement in the SMF regions between 0.1 and 0.4 cyc/oct or between 0.35 and 0.65 cyc/oct did not produce a significant change in VSL performance. One possible explanation is based on the nonlinearity of the auditory system. If we consider the SMF intensity as the input and location perception as the corresponding output of the auditory system, when the SMF intensity is low, the location perception is unclear. As the SMF intensity increases, the location performance is improved until it reaches certain degree of accuracy, where the output saturates. If the Fourier transform of the original HRTFs already results in saturated localization accuracy, the spectral enhancement in the SMF domain will not significantly change the localization performance. Since no significant change in localization performance was observed when SMF enhancement between 0.1 and 0.65 cyc/oct was applied, the enhancement in this SMF region might be used in speech enhancement and noise reduction algorithms without impairing sound localization accuracy. Indeed, Eddins and Liu (2006) reported significant improvement in vowel identification in noise following SMF enhancement in the regions between 1.5 and 2.5 cyc/oct. Although the present study did not show a significant effect of SMF enhancement on VSL performance, varying the amount of enhancement might still result in different VSL performance.

C. Limitations and future work

The current study explored the effects of applying filtering and enhancement over several SMF regions to HRTFs using a VSL procedure. As detailed in Sec. I, the use of VSL allows one to separately manipulate ITDs and spectral cues. However, it is unclear the degree to which the current VSL results can be generalized to free-field localization conditions where the ITD and spectral cues are naturally related. In addition, the current study used nonindividualized HRTFs with a HRTF customization procedure in an effort to reduce the differences between an artificially selected and the individual HRTFs. Future work using individual HRTFs may provide better evaluation of the effects of filtering and enhancement over certain SMF regions on VSL localization performance.

ACKNOWLEDGMENTS

The authors appreciate the helpful comments of the associate editor and an anonymous reviewer. Portions of this work are reported in the Ph.D. thesis of J. Q. Work supported in part by Grant No. R01 DC04403 awarded by NIH/NIDCD to D. A. E.

APPENDIX: HRTF CUSTOMIZATION PROCEDURE

1. HRTF preselection

In the preselection phase, the sound stimuli consisted of Gaussian noise bursts filtered by a given HRTF set and presented via Sennheiser HD265 headphones at an overall level of 70 dB SPL. For each stimulus presentation, there was a moving virtual sound source presented at a known and fixed elevation with azimuth changing linearly (from −170° to 180°) over a duration of 10 000 ms. Ideally, the sound stimuli should form a horizontal circle at the fixed elevation in virtual auditory space, and there should be 2.5 circles over the duration of one presentation. However, due to individual differences in the subjects' own HRTFs and the HRTFs used to filter the Gaussian noise, the stimuli filtered by one HRTF set may sound like a horizontal circle at a given elevation for some subjects but not for others. Subjects were asked to evaluate the virtual sound stimuli based on the criteria listed in the following:

(1) The virtual stimuli result in externalization of auditory image (i.e., whether the virtual sources were perceived as being externalized or intracranial). Subjects evaluated this with "YES" or "NO" responses.
(2) Sound is clearly perceived in front, to the sides, and behind the listener (i.e., how well the sound is perceived as a horizontal circle in virtual auditory space). Subjects rated the circle criterion on a scale from 1 to 10, with 10 being most like a circle and 1 being least like a circle.
(3) Sound is perceived accurately at the desired and fixed elevation. Subjects rated the elevation criterion on a

scale from 1 to 10, with 10 representing the exact elevation and 1 representing far way from the desired elevation.

Three circle presentations were evaluated at the each of three elevations (60°, 0°, and −30°) for each of the 26 HRTF sets. Determination of the six best-matching HRTF sets was based on the total rating score summed over three presentations at three elevations (excluding sets resulting in "NO" responses for the externalization criterion). This procedure took approximately 35 min.

2. Selecting the best matching HRTF set

a. Paired comparisons of six best-matching HRTF sets

During the second phase, the best-matching HRTF set was selected from the six previously selected sets using paired comparisons. In this session, the sound stimulus was a 250-ms Gaussian noise filtered by a given set of HRTFs presented via headphones at 70 dB SPL. At each presentation, a pair of 250-ms stimuli filtered by two different HRTF sets was presented with a silent interval of 750 ms separating the two stimuli. The expected virtual sound source location was indicated by a red dot in a three-dimensional coordinate system (as in the VSL subject interface shown in Fig. 3) displayed on a computer monitor. Subjects were asked to compare the two stimuli in each pair and choose the one that sounds closer to or at the expected virtual location. Six virtual locations were tested (elevation, azimuth): (60°, 0°), (30°, 0°), (0°, 0°), (−30°, 0°), (0°, −90°), (0°, 90°). There were 15 possible paired combinations, and each possible pair was repeated 3 times so that each HRTF set was compared 15 times for every location tested. According to the selected percentage, the six HRTF sets were ranked from first to sixth, with the first HRTF set corresponding to the highest selected percentage.

b. Circle presentation for the six selected HRTF sets

Following the paired comparisons phase, the six preferred sets were reevaluated using circle presentations, but with a single criterion based on the general impression of virtual sources. Subjects rated on a 1–10 scale after presenting each circle.

The best matching HRTF set was determined by the results of the paired comparisons and the second circle presentation judgments. If the two sessions consistently resulted in the same HRTF set, then that set was chosen as the best match. If the paired comparisons and second circle presentation resulted in two different sets, then the one of these two sets with the higher score in the first circle presentation phase was chosen as the best match. The process of reducing the best six HRTF sets took approximately 18 min.

III. Scaling the selected HRTF set in frequency

In addition to selecting a particular HRTF set directly from a HRTF database, it has been shown that the individual differences in HRTFs could be reduced by scaling the given directional transfer functions (DTFs) in frequency by a certain value. DTFs are derived from HRTFs by removing their common component that includes the ear canal resonance and the microphone transfer function (Middlebrooks, 1999a). In addition, the VSL experiment (Middlebrooks, 1999b) showed an improved VSL performance using scaled DTFs compared to the original nonindividualized DTF sets. Middlebrooks *et al.* (2000) proposed a psychophysical procedure to obtain the appropriate scaling factor based on one given HRTF set. The procedure allowed listeners to compare the perceived sounds filtered by differently scaled DTFs at several given virtual sound source locations. The scaling factor was determined from individual preferences for those scaled DTFs. This procedure was adopted in the current study.

In the last phase of HRTF customization, the best matching HRTF set was scaled in frequency by a series of scaling factors and the proper scaling factor was determined from a paired comparison procedure. The scaling factor varied linearly on an octave frequency scale from −0.2 to 0.2 (corresponding to the range of 0.8706–1.1487 on a linear frequency scale) in 17 steps of 0.025 octaves. There were 136 possible pair combinations for 17 different scaled HRTF sets, however, to reduce the total number of comparisons and to avoid comparisons of neighboring scale factors the two scaling factors for each pair of HRTF sets differed by either 0.125, 0.175, or 0.225 octaves. As a result, there were a total of 60 paired combinations in which each scaled HRTF set was compared six, eight, or ten times (HRTF sets scaled by factors between −0.2 and 0.1 and between 0.1 and 0.2, including ±0.1 and ±0.2 octaves were compared six times; HRTF sets scaled by factors of 0, ±0.075, and ±0.05 octaves were compared eight times; HRTF sets scaled by factors of ±0.025 octaves were compared ten times). The best scale factor was chosen on the basis of the highest selection rate. The complete scaling session lasted for approximately 20 min. Thus, the final customized HRTF set was based on the best matching HRTF set scaled by the best scale factor. The complete HRTF customization procedure, reducing 26 possible HRTF sets down to one scaled HRTF set, lasted approximately 1 h and 10 min for each subject.

[1]A repeated measures ANOVA was performed to show the significance level of the difference among front–back confusion rates. Although arcsine transformations have usually been used to transform proportional data prior to computing an ANOVA, Studebaker (1985) has shown that for the proportions between 15% and 85%, the results from the untransformed data are similar to those from the transformed data. The front-back confusions in the current study are all greater than 14.5% (with only 3 out of 42 less than 15%). Therefore, an ANOVA was applied without the arcsine transformation.

[2]To reduce the number of possible pairs in pairwise comparisons, the average elevation errors in elevations at −30° and −20° were combined to form the average elevation errors in low elevation region. Therefore, there were 240 responses (120 for each elevation) in this elevation region for each subject in each condition.

American National Standards Institute. (**1996**). "Specification for audiometers," ANSI S3.6-1996, New York.

Asano, F., Suzuki, Y., and Sone, T. (**1990**). "Role of spectral cues in median plane localization," J. Acoust. Soc. Am. **88**, 159–168.

Blauert, J. (**1983**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT, Cambridge, MA).

Bloom, P. J. (**1977**). "Creating source elevation illusions by spectral manipulation," J. Audio Eng. Soc. **25**, 560–565.
Butler, R. A., and Belendiuk, K. (**1977**). "Spectral cues utilized in the localization of sound in the median sagittal plane," J. Acoust. Soc. Am. **61**, 1264–1269.
Chen, J., Van Veen, B. D., and Hecox, K. E. (**1995**). "A spatial feature extraction and regularization model for the head-related transfer function," J. Acoust. Soc. Am. **97**, 439–452.
Eddins, D. A., and Bero, E. M. (**2007**). "Spectral modulation detection as a function of modulation frequency, carrier bandwidth, and carrier frequency region," J. Acoust. Soc. Am. **121**, 363–372.
Eddins, D. A., and Harwell, R. M. (**2002**). "Spatial frequency channels in audition?," The 25th Meeting of the Association for Research in Otolaryngology, St. Petersburg, FL.
Eddins, D. A., Hoolihan, P. S., and Bero, E. M. (**2001**). "Just noticeable differences in spectral envelope frequency," The 25th Meeting of the Association for Research in Otolaryngology, St. Petersburg, FL.
Eddins, D. A., and Liu, C. (**2006**). "Enhancement of spectral modulation frequency in acoustics and identification of vowels," J. Acoust. Soc. Am. **119**, 3338.
Kistler, D. J., and Wightman, F. L. (**1992**). "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," J. Acoust. Soc. Am. **91**, 1637–1647.
Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **62**, 157–167.
Kulkarni, A., and Colburn, H. S. (**1998**). "Role of spectral detail in sound-source localization," Nature (London) **396**, 747–749.
Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (**1995**). "On the minimum-phase approximatioin of head-related transfer functions," in *Proceedings of the 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY pp. 84–87 (IEEE catalog No. 95TH8144).
Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (**1999**). "Sensitivity of human subjects of head-related transfer function phase spectra," J. Acoust. Soc. Am. **105**, 2821–2840.
Langendijk, E. H. A., and Bronkhorst, A. W. (**2002**). "Contribution of spectral cues to human sound localization," J. Acoust. Soc. Am. **112**, 1583–1596.
Liu, C., and Eddins, D. A. (**2004**). "Spectral frequency modulation in vowel identification," J. Acoust. Soc. Am. **115**, 2631.
Macpherson, E. A., and Middlebrooks, J. C. (**2003**). "Vertical-plane sound localization probed with ripple-spectrum noise," J. Acoust. Soc. Am. **114**, 430–445.
Martens, W. L. (**1987**). "Principal components analysis and resynthesis of spectral cues to perceived direction," in *Proceedings of the International Computer Music Conference*, edited by J. Beauchamp (International Computer Music Association, San Francisco, CA), pp. 274–281.
Mehrgardt, S., and Mellert, V. (**1977**). "Transformation characteristics of the external human ear," J. Acoust. Soc. Am. **61**, 1567–1576.
Middlebrooks, J. C. (**1999a**). "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Am. **106**, 1480–1492.
Middlebrooks, J. C. (**1999b**). "Virtual localization improved by scaling nonindividual external-ear transfer functions in frequency," J. Acoust. Soc. Am. **106**, 1493–1510.
Middlebrooks, J. C., and Green, D. M. (**1990**). "Directional dependence of interaural envelope delays," J. Acoust. Soc. Am. **87**, 2149–2162.
Middlebrooks, J. C., Macpherson, E. A., and Onsan, Z. A. (**2000**). "Psychological customization of directional transfer functions for virtual sound localization," J. Acoust. Soc. Am. **108**, 3088–3091.
Musicant, A. D., and Butler, R. A. (**1985**). "Influence of monaural spectral cues on binaural localization," J. Acoust. Soc. Am. **77**, 202–208.
Oppenheim, A. V., and Schafer, R. W. (**1989**). *Discrete-Time Signal Processing* (Prentice Hall, Englewood Cliffs, NJ).
Qian, J., and Eddins, D. A. (**2004**). "Principal component analysis of ear differences in head related transfer functions," The 27th Meeting of the Association for Research in Otolaryngology, Daytona Beach, FL.
Qian, J., and Eddins, D. A. (**2005**). "Virtual sound localization using PCA-modified, nonindividualized head related transfer functions," The 28th Meeting of the Association for Research in Otolaryngology, New Orleans, LA.
Qian, J., and Eddins, D. A. (**2006**). "Virtual sound localization using head related transfer functions modified in the spectral modulation frequency domain," The 29th Meeting of the Association for Research in Otolaryngology, Baltimore, MD.
Saoji, A. A., and Eddins, D. A. (**2007**). "Spectral modulation masking patterns reveal tuning to spectral envelope frequency," J. Acoust. Soc. Am., **122**, 1004–1013.
Shamma, S. A., Versnel, H., and Kowalski, N. (**1995**). "Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra," Aud. Neurosci. **1**, 233–254.
Strutt, J. W. (Lord Rayleigh, **1907**). "On our perception of sound direction," Philos. Mag. **13**, 214–232.
Studebaker, G. A. (**1985**). "A rationalized arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Versnel, H., Kowalski, N., and Shamma, S. A. (**1995**). "Ripple analysis in ferret primary auditory cortex. III. Topographic distribution of ripple response parameters," Aud. Neurosci. **1**, 271–286.
Versnel, H., and Shamma, S. A. (**1998**). "Spectral-ripple representation of steady-state vowels in primary auditory cortex," J. Acoust. Soc. Am. **103**, 2502–2514.
Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (**1993**). "Localization using nonindividualized head-related transfer functions," J. Acoust. Soc. Am. **94**, 111–123.
Wenzel, E. M., Stone, P. K., Fisher, S. S., and Foster, S. H. (**1990**). "A system for three-dimensional acoustic 'visualization' in a virtual environment workstation," in *Proceedings of Visualization '90*, San Francisco, CA, pp. 329–337.
Wightman, F. L., and Kistler, D. J. (**1989a**). "Headphone simulation of free-field listening. I. Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867.
Wightman, F. L., and Kistler, D. J. (**1989b**). "Headphone simulation of free-field listening. II. Psychological validation," J. Acoust. Soc. Am. **85**, 868–878.
Yost, W. A. (**1981**). "Lateralization position of sinusoids presented with interaural intensive and temporal differences," J. Acoust. Soc. Am. **70**, 397–409.

Comparison of adaptive psychometric procedures motivated by the Theory of Optimal Experiments: Simulated and experimental results

Jeremiah J. Remus and Leslie M. Collins[a)]
Department of Electrical and Computer Engineering, Duke University, Box 90291, Durham, North Carolina 27708-0291

(Received 27 October 2006; revised 19 October 2007; accepted 27 October 2007)

The wide use of psychometric assessments and the time necessary to conduct comprehensive psychometric tests has motivated significant research into the development of psychometric testing procedures that will provide accurate and efficient estimates of the parameters of interest. One potential framework for developing adaptive psychometric procedures is the Theory of Optimal Experiments. The Theory of Optimal Experiments provides several metrics for determining informative stimulus values based on a model of the psychometric function to be provided by the investigator. In this study, two methods based on a previous implementation of the Theory of Optimal Experiments are presented for comparison to two fixed step size staircase methods and also an existing adaptive method that utilizes a Bayesian framework. The psychometric procedures were used to measure detection thresholds and discrimination limens on two separate psychoacoustic tasks with normal-hearing subjects. Computer simulations were performed based on the outcomes of the experimental psychoacoustic detection task to analyze performance over a large sample size in the case of known truth. Results suggest that the proposed stimulus selection rules motivated by the Theory of Optimal Experiments perform better than previous techniques and also extend estimation to multiple parameters. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2816567]

PACS number(s): 43.66.Yw, 43.66.Fe, 43.64.Yp [JHG] Pages: 315–326

I. INTRODUCTION

Psychophysical variables are an important and useful tool for assigning a quantitative value to a subject's ability to perceive stimuli. In the specific field of cochlear implant research, psychophysical testing has been used to measure the threshold for perceiving a stimulus and the maximum level of comfortable stimulation, as well as the detection and discrimination of stimuli with differing parameters such as amplitude, pulse rate, and pulse duration. These variables are useful for understanding how information can be transmitted via sequences of electrical pulses, leading to the development of improved cochlear implant electrode arrays and speech processing algorithms. In cochlear implant research, the number of electrodes and the number of parameters available for investigation in an experiment make large-scale experiments time prohibitive in most circumstances, due the schedule and availability of study participants. Thus, researchers seek quick and efficient methods for measuring psychophysical variables with the least bias and lowest variance in the shortest possible number of trials.

One of the simplest and most straightforward adaptive techniques for measuring psychophysical variables is a fixed step size (FSS) staircase procedure. These techniques are appealing due to their basic and intuitive setup; after each trial or a set of trials with the same outcome, the stimulus parameters are changed by a fixed amount, either increased or decreased for either affirmative or negative responses, depending on the task. In cochlear implant psychophysics, the most frequently used FSS staircase method is the Levitt procedure (Levitt, 1971). Fixed step size staircase procedures are simple to implement and provide low-bias measurements of the psychophysical variables. However, the choice of method parameters such as step size may not be obvious; studies often use different stimulus selection rules (a survey of method parameters in different studies is provided in Garcia-Perez, 1998). Several studies provide guidelines for setting the fixed step size (Green *et al.*, 1989; Garcia-Perez, 1998), typically as a function of the estimated spread of the psychometric function. Alternatively, this study proposes that the difficulty of tuning fixed step size staircase procedures can be addressed using adaptive psychometric procedures to modify the step size in a more sophisticated manner that takes into account the previous trials. Fixed step size staircase procedures, e.g., Levitt, do not assume a form or model for the underlying psychometric function, which can be beneficial when the psychometric function is unknown or cannot be estimated accurately. However, in circumstances where a reasonable estimate of the psychometric function is available it can be used as an additional source of information to expedite psychophysical testing. Previous studies comparing fixed and adaptive step size procedures indicate improved performance with adaptive step size procedures (Pentland, 1980; Leek, 2001; Marvitt *et al.*, 2003). While more ad-

[a)]Author to whom correspondence should be addressed. Electronic mail: lcollins@ee.duke.edu

vanced psychometric techniques are available, they have not seen widespread use in the cochlear implant research community.

There are several promising adaptive psychometric techniques that have been developed in the field of vision research for measuring psychophysical thresholds that utilize a model of the psychometric function; this study will focus on a subset of techniques that incorporate a Bayesian framework. In the basic Bayesian framework, a prior probability distribution reflecting current knowledge or "prior information" about the parameters is updated based on the outcome of a trial to produce a posterior probability distribution. The updated posterior probability distribution incorporates both the prior information and the "evidence" or outcome of the trial. The updated parameter probability distribution is used to determine the next stimulus value; hence the process is adaptive. Bayesian parameter estimation is an appropriate technique for parameter estimation via sequential test trials, and is both robust and efficient. One advantage to Bayesian parameter estimation over alternative parameter estimation techniques that use curve fitting of the psychometric function model to the data is the fact that the individual trials in psychophysical testing have binary outcomes (e.g., "correct" or "incorrect"). The psychometric function is not observed directly, but instead measured with quantization error that is dependent on the number of observations at each point and can only be reduced by further trials. This can result in an observed set of data that is poorly fit by the model of the psychometric function, producing inaccurate estimates of the parameters that will also lead to the selection of poor, uninformative stimulus values. This phenomenon has been observed in Remus and Collins (2007).

Several examples of Bayesian adaptive psychometric procedures exist in the literature. The quick estimation by sequential testing (QUEST) method developed by Watson and Pelli (1983) uses Bayesian techniques to place the next stimulus at the current best estimate of the threshold. Their analysis suggests a significant improvement in asymptotic efficiency over the parameter estimation by sequential testing (PEST) method. King-Smith *et al.* (1994) compared the QUEST method to two techniques using modified definitions of threshold estimate, resulting in the development of the zippy estimation by sequential testing (ZEST) method, which also places the stimuli at the estimated threshold value. Whereas the QUEST method uses the maximum likelihood of the parameter probability distribution to estimate the threshold, the ZEST method uses the parameter probability distribution to calculate the expected value of the threshold, i.e., mean likelihood. The Ψ method, developed and implemented by Kontsevich and Tyler (1999), utilizes Bayes' rule to update the two-dimensional probability distribution over the slope and threshold parameter values after each trial. Pentland (1980) modified the rules in PEST to create bestPEST, which starts with a binary search and then uses logic similar to the QUEST method following the first reversal by using the maximum likelihood estimate of the threshold as the next stimulus level.

The ZEST method has been shown in several studies (Phipps *et al.*, 2001; Turpin *et al.*, 2002; Marvitt *et al.*, 2003; Anderson and Johnson, 2006) to quickly and efficiently estimate psychophysical thresholds through sequential trials. Additionally, it has the desirable characteristics of being computationally efficient and simple to implement. However, one potential limitation of the ZEST method is the assumption of a fixed, pre-defined value for the slope of the psychometric function. Green (1990) concluded that threshold estimates are relatively insensitive to minor mismatch between the assumed and true psychometric function, suggesting that the effects of assuming an incorrect slope value may be minimal. Additionally, the slope parameter may have originally been considered fixed since it was not considered a parameter of interest. However, the assumptions in the ZEST method could have two effects: it restricts the use of the ZEST method to the estimation of a single parameter of the psychometric function when multiple parameters of interest may exist, and it imposes a form of the psychometric function that could negatively impact performance if incorrect. Amitay *et al.* (2006) performed a comparison of psychometric methods using both maximum likelihood and staircase procedures. In their study they observed that one of the shortcomings when using maximum likelihood techniques versus fixed step size staircase procedures was the requirement for prior knowledge about the psychometric function, both the assumed functional form (e.g., logistic cumulative distribution function (cdf), Weibull cdf, hyperbolic tangent cdf) and the parameters (psychometric function slope, false negative rate). Additionally, Klein (2001) outlined several arguments in favor of measuring the slope of the psychometric function. Thus it is desirable to find an adaptive psychometric procedure that is easily extended to estimating multiple model parameters without any drawbacks of significant concern. In the context of the current study, it may be advantageous to consider the slope parameter as a variable to allow flexibility but to do so in a manner that does not impede the convergence of the Bayesian parameter estimation procedure.

Adaptive psychometric procedures determine the stimulus value for the next trial based on the outcomes of previous trials; however, there are several potential rules presented in the literature for selecting the next stimulus value in Bayesian adaptive methods. The QUEST and ZEST methods place the next stimulus value at the current estimate of the threshold parameter. Since these methods do not seek estimates of the slope value and only operate on a single unknown parameter, this rule is an appropriate solution. In the Ψ method (Kontsevich and Tyler, 1999) and multiple estimation by sequential testing (MUEST) method (Snoeren and Puts, 1997), multiple parameters are estimated, which complicates placement of the next stimulus values since each parameter may have a unique optimal stimulus value, and the next trial can only occur at a single point. In the Ψ method, the stimulus values were selected to minimize the entropy of the parameter probability distribution (represented as a probability mass function with probabilities assigned to discretely sampled points in parameter space), essentially seeking the stimulus value that, given either a correct or incorrect outcome, would most increase the certainty of the parameter probabilities. However, minimizing the entropy of the pa-

rameter probability distribution in not always desirable, i.e., when the true parameter value is between two sample points in the parameter space, and can result in parameter estimates clustered at the sample points in the parameter probability distribution (Remus and Collins, 2007). In the MUEST method, a two-dimensional extension of Taylor's sweat factor (Taylor and Creelman, 1967; Taylor, 1971) was developed and used to calculate ideal points for each of the two parameters. Trials were assigned one of the two stimulus values according to pre-determined probabilities (Snoeren and Puts, 1997). While the selected sample points are ideal under Taylor's sweat factor, a more principled framework may exist for determining which of the two stimulus values should be used in the next trial.

One potential framework for developing such a method is the Theory of Optimal Experiments. The Theory of Optimal Experiments is described in detail by Federov (1972) and Chernoff (1972), and further developed and applied to a variety of problems by others (Tsutakawa, 1972; Chaloner and Larntz, 1989; El-Gamal, 1991; MacKay, 1992; Chaloner and Verdinelli, 1995; Whaite and Ferrie, 1997; Liao and Carin, 2004). The Theory of Optimal Experiments can be utilized to specify a sequential testing procedure given a reasonably accurate model of the process under investigation. Psychometric functions that model the probabilities of responses in psychophysical testing exist and are supported by substantial amounts of experimental data. Having a specified form of the psychometric function allows calculation of the Fisher information (Cover and Thomas, 1991), which quantifies the amount of information an observation provides about the underlying parameters of the observation likelihood function. As outlined in Fedorov (1972), stimulus selection rules can be based on maximizing various quantities calculated from the Fisher information matrix, such as the determinant or trace. Experimental designs using the Fisher information matrix and Theory of Optimal Experiments have been proposed in other fields such as bioassay (Tsutakawa, 1972), but have not been previously considered for psychophysical testing. By dividing the experiment design into two stages, first estimating the model parameters using the available data and then using those parameter values to select the next stimulus location that will provide the most information about the estimated model parameters, the resulting procedure is locally optimal.

The Theory of Optimal Experiments was considered previously (Remus and Collins, 2007) as a framework for adaptive psychometric procedures intended for psychophysical testing in cochlear implant subjects. A performance comparison was made using computer simulations of a psychophysical task; results favored the use of Bayesian parameter estimation over curve-fitting techniques. Results also indicated that stimulus selection rules motivated by previous implementations of the Theory of Optimal Experiments (e.g., Liao and Carin, 2004) select informative stimulus values. However, it is necessary to test the performance of parameter estimation techniques and stimulus selection rules on human subjects, since variability in subject parameters and characteristics as well as assumptions made in the modeled psychometric functions may mitigate any benefits observed in computer simulations.

In this study, five adaptive psychometric procedures were evaluated via simulations and experiments with normal-hearing subjects. The psychometric procedures considered in this study were two proposed Bayesian adaptive psychometric procedures based on the Theory of Optimal Experiments, the previously mentioned ZEST method, the Levitt procedure, and a fixed step size staircase procedure proposed by Kaernbach (1991). Section II describes the psychometric procedures considered in this study as well as the psychoacoustic experiment and computer simulation methods used to evaluate and compare the different procedures. Section III presents the results for the psychoacoustic study using normal-hearing subjects and computer simulations with larger sample sizes and known truth. A discussion of the results and outcomes of this study is presented in Sec. IV, along with suggestions for further investigation.

II. METHODS

Three Bayesian adaptive sequential testing procedures for measuring psychophysical variables were compared to two fixed step size staircase procedures in this study. Each of the Bayesian adaptive methods utilizes a two-stage framework: the psychometric function parameters are estimated after each trial, and then the stimulus value for the next trial is calculated using the updated estimates of the psychometric function parameter values. The three Bayesian adaptive methods considered in this study include one existing technique (ZEST) and two proposed techniques (Bayes Fisher information gain (FIG) and Bayes Greedy), utilizing different implementations of the Bayesian parameter estimation routine and different rules for calculating the stimulus value for the next trial.

In the first stage of each of the three Bayesian adaptive psychometric methods, the parameters are estimated using the outcomes of the previous trials and any prior information about the parameter values. The ZEST, Bayes FIG, and Bayes Greedy methods all utilize a Bayesian update framework for estimating the parameter values. The ZEST method assumes a fixed slope value for the psychometric function and only updates a probability distribution over the threshold parameter. The Bayes FIG and Bayes Greedy methods perform a two-parameter estimation of both the slope and threshold values. The details of the Bayesian parameter estimation routine are provided below. The second stage is the selection of the stimulus value for the next trial. Two rules for stimulus selection were investigated in the Bayesian adaptive procedures; these will also be described below.

To clarify the following presentation of adaptive psychometric procedures and measurement of psychometric parameters, some standard notation will first be defined. The psychometric function $\Psi(x;s,m)$ is modeled in this study using the logistic cumulative distribution function (cdf), which has the form shown in Eq. (1). The logistic function is a popular model for the psychometric function; other frequently used

models include the Weibull cdf, hyperbolic tangent function, and the Gaussian cdf (Macmillan and Creelman, 1991; Strasburger, 2001).

$$\Psi(x;s,m) = \frac{1}{1+e^{-s(x-m)}}. \quad (1)$$

In Eq. (1), x is the stimulus value and the parameters of the psychometric function are the "threshold," or median m of the underlying logistic probability distribution function, and s, the parameter controlling the slope of the psychometric function. For convenience and generalization, the psychometric function will often be referenced as $\Psi(x;\lambda)$, with λ representing the parameters of the psychometric function. Estimates of the psychometric function parameters are denoted as $\hat{\lambda}$. Each trial of a psychometric experiment produces either a correct or incorrect response. The binary-valued discrete random variable r is the response from the subject, either correct ($r=1$) or incorrect ($r=0$). Let the set of all possible stimulus values be defined as $X=[x_1,x_2,\ldots,x_n]$ (within the scope of this study the stimulus values will always be discrete). The set of stimulus values presented in the previous t trials is the vector $X_t^{\text{obs}}=[x_1^{\text{obs}},x_2^{\text{obs}},\ldots,x_t^{\text{obs}}]$ with responses $R_t=[r_1,r_2,\ldots,r_t]$.

A. Parameter estimation—Bayesian update equation

The Bayesian adaptive procedures utilize a nonparametric probability distribution $p(\lambda)$ constructed from probabilities assigned to parameter values at $\tilde{\lambda}$, where $\tilde{\lambda}$ generally consists of $\Pi_{i=1}^{D}K_i$ discretely sampled points in the parameter space (D is equal to the number of psychometric function parameters and K_i equals the number of discrete sample points for the ith parameter). Thus, $p(\lambda)$ is a probability mass function that only takes values at the points in $\tilde{\lambda}$ and can be represented as a D-dimensional matrix with a total of $\Pi_{i=1}^{D}K_i$ entries. The value for the $(k_1k_2\ldots k_D)$th entry in $p(\lambda)$ is denoted as $p(\tilde{\lambda}^{k_1k_2\ldots k_D})$ and indicates the probability that the true psychometric function parameters equal $\tilde{\lambda}^{k_1k_2\ldots k_D}$. The formulation for the Bayesian update of the probability distribution over the parameter values is shown in Eq. (2)

$$p(\tilde{\lambda}^{k_1k_2\ldots k_D}|X_t^{\text{obs}},R_t) = \frac{p(r_t|x_t^{\text{obs}},\tilde{\lambda}^{k_1k_2\ldots k_D})p(\tilde{\lambda}^{k_1k_2\ldots k_D}|X_t^{\text{obs}},R_{t-1})}{\sum_{k_1'=1}^{K_1}\sum_{k_2'=1}^{K_2}\cdots\sum_{k_D'=1}^{K_D}p(r_t|x_t^{\text{obs}},\tilde{\lambda}^{k_1'k_2'\ldots k_D'})p(\tilde{\lambda}^{k_1'k_2'\ldots k_D'}|X_t^{\text{obs}},R_{t-1})}, \quad (2)$$

where x_t^{obs} is the stimulus value for the tth trial, r_t is the outcome of the tth trial (correct or incorrect), and the primed values in the denominator are summation indices. The likelihood of a correct or affirmative response $p(r=1|x,\lambda)$ can be calculated from Eq. (3), using the model of the psychometric function defined in this study by the logistic cumulative distribution function shown in Eq. (1)

$$p(r=1|x,\lambda;\gamma,\delta) = \min\left[\frac{1}{\delta}+\left(1-\frac{1}{\delta}\right)\Psi(x;\lambda),1-\gamma\right]. \quad (3)$$

In Eq. (3), the probability of observing a correct response is dependent on two additional parameters, γ and δ. The parameter γ is the probability of an incorrect response to a suprathreshold stimulus, i.e., false negative rate. It is also referred to as the lapsing rate (e.g., Wichmann and Hill, 2001). This implementation of the psychometric function, specifically the handling of the false negative parameter γ, has been used previously (e.g., Watson and Pelli, 1983). It is an alternative to incorporating γ into the psychometric function to scale the maximum probability of a correct response, which can bias the slope parameter as well as the threshold parameter when modeling a two-interval forced choice (2IFC) task (Wichmann and Hill, 2001). It may be better to represent the occurrence of false negative responses as a separate process not described by the psychometric function but rather superimposed over it. Thus, a ceiling on the probability of correct response, set equal to $1-\gamma$, was added to the overall model of subject responses. The parameter δ is the number of intervals in the test (i.e., $\delta=2$ for a two-interval forced-choice test) and is responsible for scaling the probability of a correct response to account for the chance probability of correct guessing. This study will focus on the two-interval forced-choice test and the logistic function model of the psychometric metric. Thus Eq. (3) can be rewritten as Eq. (4) and the probability of an incorrect response is simply unity minus this equation. For convenience, the γ and δ parameters will not be stated in future references to the probability of a correct response but are still implied.

$$p(r=1|x,s,m;\gamma,\delta) = \min\left[\frac{1}{2}+\frac{1}{2+2e^{-s(x-m)}},1-\gamma\right]. \quad (4)$$

The ZEST method assumes a fixed, known value for the slope parameter and only applies the Bayesian update in Eq. (2) to the threshold parameter, whereas the Bayes FIG and Bayes Greedy methods use Eq. (2) to calculate probabilities for both the slope and threshold parameters in Eq. (4). Thus, in the ZEST method $\tilde{\lambda}$ is a one-dimensional vector over the threshold parameter whereas the Bayes FIG and Bayes Greedy methods use a two-dimensional matrix with size $[K_1,K_2]$ over both the threshold and slope parameters. After each trial, given the outcome (correct or incorrect response), the parameter probabilities $p(\lambda|X_t^{\text{obs}},R_t)$ are updated and normalized, resulting in a new set of probabilities for the psy-

chometric function parameters that include the outcome of the latest trial. The expected value of the ith psychometric function parameter is denoted as $\hat{\lambda}_t^{\langle i \rangle}$ and is used as the estimate of the ith parameter after each trial (Eq. (5)). The implementations in this study were used to estimate either one (in ZEST) or two (in Bayes FIG and Bayes Greedy) parameters, resulting in values of D in Eqs. (2) and (5) equaling one or two, respectively. However, the methods were developed and shown here for the general case of estimating D unknown parameters

$$\hat{\lambda}_t^{\langle i \rangle} = E\{\tilde{\lambda}^{\langle i \rangle}\} = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \cdots \sum_{k_D=1}^{K_D} \tilde{\lambda}^{k_i} p(\tilde{\lambda}^{k_1 k_2 \ldots k_D} | X_t^{\text{obs}}, R_t). \tag{5}$$

B. Stimulus selection

The stimulus placement rules considered in this study were based on an implementation of the Theory of Optimal Experiments applied to subsurface sensing by Liao and Carin (2004). Liao and Carin used the model of the process under investigation and the stimulus values for the previous trials to calculate the gain in Fisher information (Cover and Thomas, 1991) for each possible stimulus value. The stimulus value for the next trial is then selected to maximize the Fisher information gain (FIG). Using the notation presented previously, the gain in Fisher information for each possible stimulus value x_j for the next trial can be calculated using Eqs. (6)–(8) (Liao and Carin, 2004; Remus and Collins, 2007).

Equation (6) calculates the two-element vector B consisting of the partial derivatives of the psychometric function with respect to each of the two parameters. The vector B is a function of the stimulus value x and is utilized in Eqs. (7) and (8) where it is evaluated at the estimates of the psychometric function parameters (calculated using the Bayesian parameter estimation in Eq. (5)). The estimate of the threshold $\hat{m}$ is defined as $\hat{\lambda}_t^{\langle 1 \rangle}$ in Eq. (5) and the estimate of the slope $\hat{s}$ is defined as $\hat{\lambda}_t^{\langle 2 \rangle}$. The first term in B, the partial derivative with respect to the threshold parameter m, is maximized when x equals the estimate of m. The second term, the partial derivative with respect to the slope parameter, has maximum magnitude at the stimulus values x that correspond to 5% and 95% probability of correct response on the psychometric function. These are similar to the values provided in Levitt (1971) for minimizing the expected error variance of the estimates of μ and σ in an integrated normal distribution model of the psychometric function. Equation (7) computes a two-by-two matrix A equal to the approximation of the Fisher information matrix that is dependent on the previously selected stimulus values in the set X^{obs}. Equation (8) calculates FIG, the n-element vector where $\text{FIG}(x_j)$ is the gain in Fisher information for a trial using stimulus value x_j.

$$B(x) = \left[\frac{\partial p(r=1|x=x,s,m)}{\partial m} \frac{\partial p(r=1|x=x,s,m)}{\partial s} \right] \Bigg|_{\substack{m=\hat{m}_t \\ s=\hat{s}_t}}, \tag{6}$$

$$A = \sum_{i=1}^{t} B(x_i^{\text{obs}})^{\text{T}} B(x_i^{\text{obs}}), \tag{7}$$

$$\text{FIG}(x_j) = \log|1 + B(x_j) A^{-1} B(x_j)^{\text{T}}|. \tag{8}$$

The gain in Fisher information is used to select the stimulus value for the next trial via Eq. (9).

$$x_{t+1}^{\text{obs}} = \arg\max(\text{FIG}). \tag{9}$$

This formulation of the gain in Fisher information implemented by Liao and Carin (2004) assumes that the outcome of each trial is an observation of the true model corrupted by white Gaussian noise, an assumption that does not hold in psychometric testing since outcomes are binary (correct or incorrect). However, this concern was investigated by Remus and Collins (2007) and it was found that the Liao and Carin solution provides an appropriate approximation of the Fisher information gain.

The gain in Fisher information is related to the Cramer-Rao bound (Cover and Thomas, 1991), which specifies the lower limit on the variance of an unbiased estimator. Selecting stimulus values that seek to maximize the gain in Fisher information corresponds to a goal of achieving the Cramer-Rao lower bound. As shown in Eqs. (6)–(8), the gain in Fisher information is dependent on the estimated psychometric function (the model using the estimated parameters) and the stimulus values selected in the previous t trials. Previous work (e.g., Levitt, 1971) has shown that different stimulus values are required to minimize the expected error variance for either the threshold or slope parameter of the psychometric function. Similar conclusions are implied by the above discussion about the stimulus values that maximize the terms of Eq. (6). Whether the next stimulus value should provide more information about the threshold parameter or slope parameter depends on the information provided about each parameter in the previous t trials; this is determined by the Fisher information measure. The number of times a stimulus value has been sampled provides an estimate of the variance of the observed psychometric function at that point. The measure of information outlined by Eqs. (6)–(8) combines the estimated variance of each parameter and the information provided by each possible stimulus value about the parameters to find the best stimulus value for the next trial. This stimulus selection rule (Eqs. (6)–(9)) was used in the Bayes FIG procedure.

An alternative stimulus selection rule used in this study and previous psychometric methods is placing the next trial at the stimulus value nearest the estimated value of the threshold. It can be shown that when the slope parameter is ignored this stimulus selection rule will find the most informative sample points for the threshold value, according to the gain in Fisher information metric. Inserting Eq. (4) into Eq. (6) yields the derivative of the psychometric function. It is permissible to ignore the min operator in Eq. (4) and assume that the stimulus values of interest are below the point where $\frac{1}{2} + \frac{1}{2+2e^{-s(x-m)}}$ exceeds $1-\gamma$. Keeping only the term for the parameter of interest (the threshold parameter m) yields Eq. (10)

$$B(x) = \frac{\partial p(r|x=x,s,m)}{\partial m}\bigg|_{\substack{m=\hat{m}_t \\ s=\hat{s}_t}} = \frac{-2\hat{s}_t e^{-\hat{s}_t(x-\hat{m}_t)}}{(2+2e^{-\hat{s}_t(x-\hat{m}_t)})^2}. \quad (10)$$

When considering only the threshold parameter, the approximation of the Fisher information matrix A in Eqs. (7) and (8) will be a one-by-one matrix (i.e., scalar) that is independent of the possible stimulus value x_j. Therefore, it can be represented as a constant α, and the maximum gain in Fisher information will occur where the following is maximized (using B as defined in Eq. (10) and the value of A represented as α)

$$B(x_j)A^{-1}B(x_j) = \frac{4\hat{s}_t^2 e^{-2\hat{s}_t(x_j-\hat{m}_t)}}{\alpha(2+2e^{-\hat{s}_t(x_j-\hat{m}_t)})^4} \quad (11)$$

A simple modification puts Eq. (11) in the form

$$B(x_j)B(x_j) = \frac{1}{\alpha}\left(\frac{\hat{s}_t^2}{4e^{2\hat{s}_t(x_j-\hat{m}_t)} + 16e^{\hat{s}_t(x_j-\hat{m}_t)} + 24 + 16e^{-\hat{s}_t(x_j-\hat{m}_t)} + 4e^{-2\hat{s}_t(x_j-\hat{m}_t)}}\right), \quad (12)$$

which is maximized when $x_j=\hat{m}_t$, the estimate of the threshold. The decision to ignore the min operator in Eq. (4) during calculation of the ideal stimulus value is valid as long as $\gamma < 25\%$ when using a threshold value at probability of a correct response equals 75%, such that the logistic cdf form of the psychometric function still applies at the threshold value. This stimulus selection rule is used frequently in the literature (e.g., QUEST and ZEST methods) and will be investigated in this study in combination with the two-parameter Bayesian parameter estimation in a method termed Bayes Greedy. This technique, in combination with the Bayes FIG technique, will be investigated in this study as two new adaptive psychometric procedures.

C. Psychoacoustic experiment setup

To evaluate and compare the performance of the three Bayesian adaptive psychometric procedures, a set of psychoacoustic studies were conducted using normal-hearing subjects. A fixed step size procedure was included as a base line measure for a total of four psychometric methods (the fifth psychometric method, a second fixed step size procedure, was added after the human subject study and evaluated only via computer simulation). Sixteen normal-hearing subjects were recruited from the population of students and staff at Duke University for two different psychoacoustic tasks. Eight subjects performed each task. The psychoacoustic tasks did not have any specific audiometric or language requirements. Subjects were compensated for their participation. All procedures involving human participants were approved by the Duke University IRB.

Two separate psychoacoustic tasks (detection and discrimination) were investigated to evaluate any task-specific performance factors. The detection task asked subjects to identify the presence of a sinusoid in background noise. The discrimination task used two sinusoids of different intensity to measure intensity discrimination limens. Thus, the four psychometric methods could be evaluated on two different types of tasks (detection and discrimination) of interest in psychophysical testing.

Each task consisted of two stimuli: a target or anomalous stimulus (H_1) and a reference stimulus (H_0). For the detection of a pure tone in noise, the H_1 and H_0 stimuli were specified as

$$H_1: y = N(0,1) + 10^X \sin(2\pi f_{\mathrm{Hz}} t)$$

$$H_0: y = N(0,1) \quad (13)$$

where the sinusoid frequency f_{Hz} was set equal to 500 Hz and the variable X controlling the amplitude of the sinusoid is the variable of interest (i.e., the stimulus values calculated for each trial). The background noise N(0,1) is zero-mean white Gaussian noise with unit variance, which has a uniform spectrum level over the full signal bandwidth (0–11 kHz). For the discrimination between two sinusoids of different intensity, the stimuli were

$$H_1: y = (1 - 10^X)\sin(2\pi f_{\mathrm{Hz}} t)$$

$$H_0: y = \sin(2\pi f_{\mathrm{Hz}} t), \quad (14)$$

where again the sinusoid frequency f_{Hz} was set equal to 500 Hz and the variable X controls the intensity of the target/anomaly stimulus (the H_0 reference stimulus has a constant intensity whereas the H_1 target stimulus is to be identified based on its smaller magnitude). All stimuli were produced using a sampling rate of 22 kHz.

The psychoacoustics tasks were conducted using a two-interval forced-choice (2IFC) paradigm and a graphical user interface implemented in MATLAB®. The user interface contains two buttons, one for each of the stimulus intervals (either reference or target/anomaly), and subjects responded after each pair of stimuli presentations by selecting the button corresponding to the interval containing the H_1 target/anomaly pulse. The interval containing the target/anomaly stimulus was randomly assigned for each trial and no feedback was provided. The stimulus duration and inter-stimulus interval were 500 ms. The experimental interface was configured for three interleaved runs of a single psychometric method, each 60 trials in length, for a total of 180 uninterrupted trials that provided three measurements of the psychometric function parameters. Subjects were offered a short

break, and then repeated the same process with another set of 180 trials to provide an additional three measurements of the threshold. This concluded a single test session. All measurements in a single test session used the same psychometric procedure (i.e., ZEST, Bayes FIG, etc.). Each subject participated in four separate sessions, one for each psychometric procedure: the previously described Bayesian adaptive techniques (ZEST, Bayes FIG, Bayes Greedy) as well as the standard transformed Levitt procedure (Levitt, 1971) using a two-down one-up rule as a base line performance measure. The orders of the psychometric methods were randomized across subjects to reduce bias resulting from learning effects or study fatigue. Tests were performed in a double-walled soundbooth with stimuli presented via Sony MDR-V600 headphones.

The task measuring detection of a pure tone in noise used the following set of parameters. All four psychometric methods started at the same initial stimulus value, $X_1=0.2$, a level at which the pure tone is clearly audible against the background noise (+1 dB signal-to-noise ratio (SNR)). The set of possible stimulus values X for the three Bayesian adaptive procedures were 100 linearly spaced values between −2.5 and 0.5, corresponding to signal-to-noise ratios ranging from −50 dB SNR to +7 dB SNR. For the Levitt procedure, the stimulus value changed by 0.5 (change of 10 dB SNR) at each transition point until three reversals were observed, after which the step size decreased to 0.075 (change of 1.5 dB SNR). All stimuli in the sinusoid in noise detection task were scaled to within the range of $[-1,+1]$ to prevent clipping during playback. The volume control on the computer was set to 70 dB sound pressure level through the headphones using a 1 kHz calibration tone.

The Levitt procedure averaged the last eight reversals to estimate the parameter (detection threshold or discrimination limen) and excluded reversals observed prior to decreasing the step size. The Bayesian adaptive procedure assumed a uniform initial parameter probability distribution. For Bayes FIG and Bayes Greedy, the parameter probability distribution is a 75-by-18 matrix, linearly sampled in the parameter space with 75 threshold parameter sample points from −2.5 to 0.5 and 18 slope parameter sample points ranging from 1–18. In terms of the variables introduced for the Bayesian parameter estimation calculations, $\tilde{\lambda}$ is a 75-by-18 matrix ($K_1=75$, $K_2=18$), with $\tilde{\lambda}^{k_1}$ equaling 75 linearly spaced points from −2.5 to 0.5 and $\tilde{\lambda}^{k_2}$ equaling [1,2,…,18]. The ZEST method assumed a fixed, constant value of 15 for the slope parameter and the threshold probability distribution was one-dimensional with the same 75 sample points as Bayes FIG and Bayes Greedy.

For the task measuring intensity discrimination using sinusoids, the initial stimulus value for all four methods was −0.3, which results in a signal difference of 13 dB. The 100 possible stimulus values for the Bayesian procedures were linearly spaced from −3 to 0 (an intensity difference ranging from 0.02 dB to ∞ dB), and the threshold parameter probability distributions were spaced over this range ($x=-3$ to $x=0$) using the same number of sample points as in the detection study. The stimuli in the intensity discrimination task were not scaled, preserving the difference in amplitude between the H_1 and H_0 stimuli. All other parameters of the psychometric procedures were the same as for the detection task.

D. Computer simulation setup

Following the psychoacoustic studies, computer simulations using similar parameter values to those observed in the psychoacoustic study were conducted to provide results with a large number of measurements and known truth. The psychoacoustic tasks are an important tool for analyzing performance since there are many factors to consider when testing human subjects, such as fatigue and subject variability. However, computer simulations are also useful due to two benefits: (1) a large sample size can be simulated to finely resolve asymptotic performance and (2) truth is known. Establishing truth with human subjects can be difficult and time consuming. For these reasons, computer simulations of psychometrics tests were conducted to extend the comparison of the different adaptive and fixed step size (FSS) staircase procedures.

Computer simulations were set up to mimic the sinusoid-in-noise detection task. In the simulations, responses were generated using simulated psychometric functions with parameter values reflecting the results of the psychoacoustic testing with human subjects. The psychometric function has two parameters: the threshold parameter m, which was uniformly distributed over the range $[-1.25,-1]$, and the slope parameter s, uniformly distributed over the range [5, 15]. The probability of an incorrect response to a suprathreshold stimulus value, γ, was set to 2%.

The computer simulations included the three Bayesian adaptive procedures, the Levitt procedure with parameter estimates calculated using averaged reversals and psychometric curve fitting, and a FSS staircase procedure proposed by Kaernbach (1991) that was added to the computer simulations as a second example of FSS staircase methods. The Kaernbach procedure changes the stimulus value after every trial, but can converge to the 75% point on the psychometric function by making the down step size after a correct response equal to one-third the up step size after an incorrect response. Two sets of computer simulations were performed. In the first set of simulations, the ZEST, Levitt, Bayes Greedy, and Bayes FIG procedures used the same parameter values (e.g., possible stimulus values, parameter probability distributions) as in the psychoacoustic tests. The slope of the psychometric function assumed by the ZEST method (slope=15) equals the upper limit of the true values of the simulations. This assumed slope value provides an appropriate approximation to the accuracy that may be expected in psychometric testing of tasks where there is little prior information about the true range of slope values; also, it is consistent with the experiment setup used in the human subject study and illustrates the potential drawback of using fixed values for unknown parameters. The γ parameter for all Bayesian adaptive methods was set equal to 1%, the same value set by Watson and Pelli (1983) for the QUEST method and by King-Smith *et al.* (1994) for ZEST. In the second set

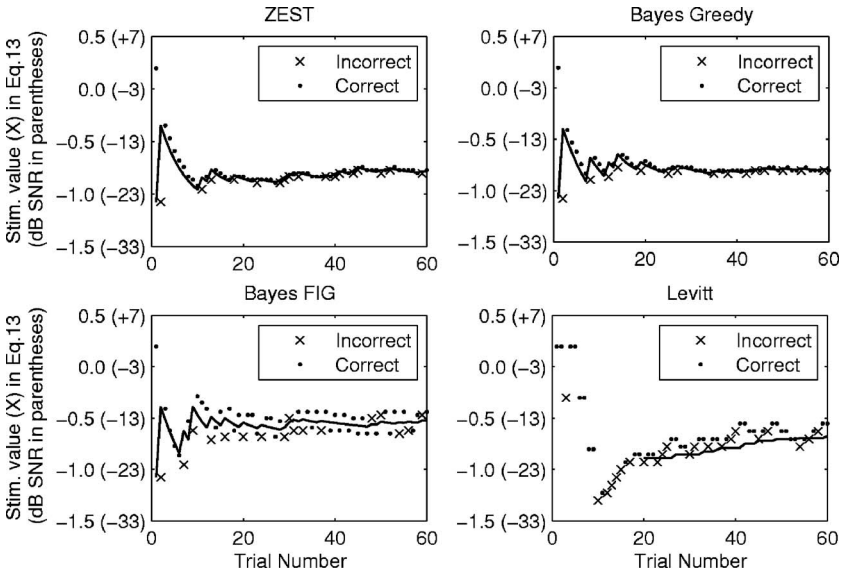


FIG. 1. Trial-by-trial data for a single run of each psychometric procedure (ZEST, Bayes Greedy, Bayes FIG, Levitt) for Subject S106 performing the sinusoid in noise detection task. Each subplot includes the stimulus value for each trial and whether the response was correct (·) or incorrect (x), as well as the estimated threshold value (solid line). The ordinate axis is in units of the stimulus value X defined in Eq. (13), as well as the signal-to-noise ratio (dB SNR).

of simulations, the Levitt and Kaernback procedures were run using one of three initial step sizes ranging from 1.2σ to 0.6σ until observing three reversals and then switching to one of three final step sizes ranging from 0.3σ to 0.06σ (where σ is the range of the psychometric function (Green *et al.*, 1989)). Considering all combinations of initial and final step size resulted in nine implementations of each FSS staircase procedure. In both sets of simulations, the number of trials in each run was increased to 100 to observe effects that occur with higher numbers of trials and to see if asymptotic limits were reached. The simulations consisted of 5000 runs, which should be a sufficient number of runs to provide a dense sampling from the distributions of psychometric function parameter values as well as to have enough repeated measurements to calculate performance statistics such as estimate error bias and variance for each psychometric procedure.

III. RESULTS

A. Psychoacoustic experiments

The results of a single run with each of the four psychometric procedures are shown in Fig. 1 for a single subject who performed the sinusoid in noise detection task. Plotted for each technique are the stimulus values for each trial and whether the subject's response was correct (·) or incorrect (x). The estimated threshold value at each trial is also shown as a solid line. This trial-by-trial plot illustrates some basic differences between the four procedures used in the psychoacoustic study. The ZEST and Bayes Greedy procedures are displayed in the top two subplots. The similarity of the curves reflects the similarity of these two methods. Both ZEST and Bayes Greedy select the next stimulus value to be at the most recent estimate of the threshold value. The Bayes FIG procedure, shown in the lower left subplot, considers both the threshold and slope parameters of the psychometric function when selecting the most informative stimulus values. This results in stimulus values further away from the estimated threshold value as more information about the psychometric function parameters is collected. The Levitt procedure exhibits the expected behavior of fixed step size staircase procedures. Since the step size cannot adapt based on the previous trials, it requires additional trials to find the threshold level.

Each subject was tested with each of the four psychometric methods repeated six times to produce repeated measurements of the parameters. The repeated measurements with each procedure were used to calculate the test/re-test reliability, which has been used previously (Turpin *et al.*, 2002; Amitay *et al.*, 2006) in evaluating psychometric procedures using normal-hearing subjects and is particularly appropriate for analyzing the results of human subject studies since determining truth requires an excessive number of trials.

The test/re-test reliability for the detection task is shown in Fig. 2 *left*. Results are shown for each of the four psychometric procedures considered in the psychoacoustic study (Levitt, ZEST, Bayes FIG, Bayes Greedy). Each point plotted in the figure represents the variance of the six threshold measurements with a single psychometric procedure for one subject. The ordering of the individual points is consistent across procedures, allowing comparison of the threshold variability across procedures for each subject. The line through each method shows the geometric mean of the variances across all subjects. The test/re-test reliability results for the detection task show a substantial reduction in variance, i.e., improvement in consistency of the threshold estimates, using the Bayesian adaptive procedures. However, there appear to be only slight differences between the average performances of the three Bayesian adaptive procedures.

Figure 2 *right* shows the test/re-test reliability for the second psychoacoustic task, intensity discrimination for two sinusoids. The variances for each procedure are slightly higher than in the detection task. Comments from the study participants and personal observations by the authors suggest that this task is more tedious than the detection of the sinusoid in background noise task. The additional difficulty of the intensity discrimination task most likely resulted in higher variance in the estimated limens. The trends in the results are similar to those seen in the detection task; the

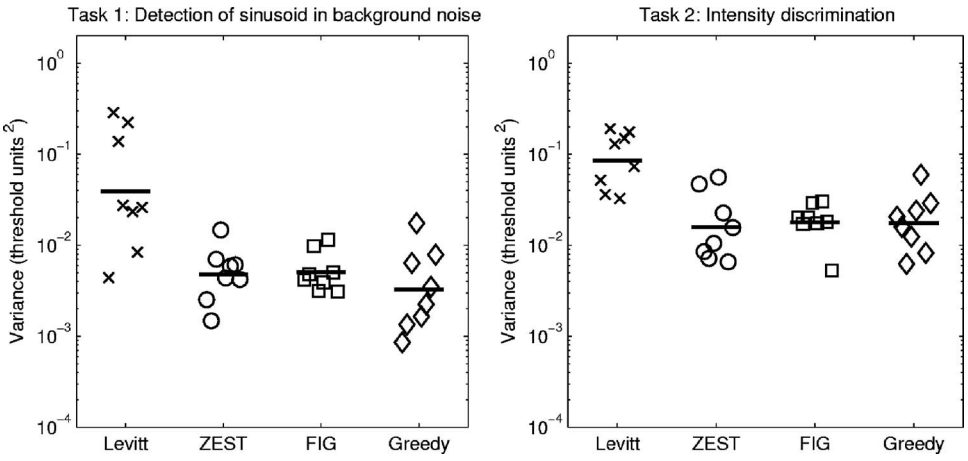


FIG. 2. Test/re-test reliability of the (left) thresholds for detection of a pure tone in background noise and (right) intensity discrimination limens using pure tones with each psychometric method (Levitt, ZEST, Bayes FIG, Bayes Greedy). Each point is the variance of the threshold estimates for one subject. Lines indicate the geometric average of the variances for each method.

Bayesian adaptive procedures outperform the Levitt procedure, but only small differences are observed between the Bayesian adaptive techniques.

To further analyze the performance of the four psychometric procedures, the average variance versus trial number is plotted to illustrate how quickly the threshold estimates from multiple runs converge (i.e., in number of trials). Figure 3 shows the average variance for each psychometric procedure as a function of trial number for the two psychoacoustic tasks performed in this study. Variances were calculated for the six runs with each psychometric procedure for each subject, and then averaged across all subjects. Average variance for the Levitt procedure cannot be calculated until a sufficient number of reversals (five reversals, i.e., two reversals at the smaller step size) have been observed to produce threshold estimates for all runs with all subjects. In Fig. 3, the difference in average variance between the two tasks is clearly visible, again illustrating the higher variance in the measurements of the intensity discrimination limens. The Bayesian adaptive procedures exhibit similar performance on both tasks. For the intensity discrimination task it may be that the Bayesian FIG technique is approaching a lower bound on the variance after approximately 40 trials. Due to the tedious nature and/or difficulty of this task, the measured discrimination limens may include a certain amount of testing "noise" which would set a lower limit on the consistency achievable through repeated measurements of the discrimination limens.

B. Computer simulations

The results of the computer simulations of the sinusoid in noise detection task are shown in Fig. 4 for the four psychometric procedures considered in the psychoacoustic study. The mean and variance of the threshold estimate error across all 5000 simulated runs is plotted versus trial number. The error for each simulated run was normalized by the true threshold value for that run prior to calculating the mean and variance. The ZEST, Bayes FIG, and Bayes Greedy methods converge to the 75% correct point on the psychometric function whereas the Levitt procedure using a two-down one-up rule converges to the 70.7% point, which has been accounted for in the error calculations. The results for these simulations show low average error (<0.05) with the three Bayesian adaptive procedures after 17 trials, and almost no average error (<0.01) with the Bayes FIG and Bayes Greedy methods after 40 trials. The error variances are very similar for the three Bayesian adaptive procedures, with the ZEST method having lower variance after 40 trials. The ZEST method used an assumed slope value equal to 15, which is the upper limit of the slope values of the simulated psychometric functions (distributed over the range [5, 15]) and may have contributed to higher bias. However, as stated previously, this was the best available estimate of the slope of the psychometric function for use in the psychoacoustic study, and is representative of the mismatch that may occur when testing on tasks where little slope data have been collected

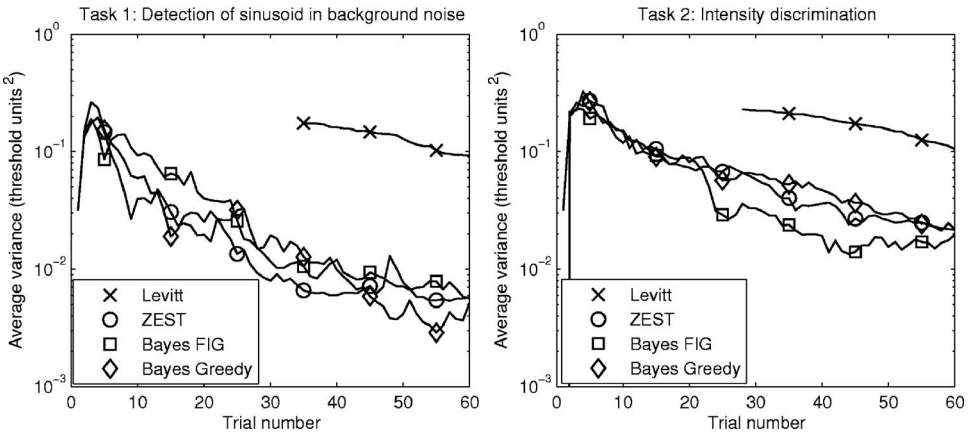


FIG. 3. Test/re-test reliability for the detection task (left) and intensity discrimination task (right) as a function of trial number, averaged across subjects. Performance at trial 60 is reported in the results shown in Fig. 2. Note: threshold, and consequently average variance, cannot be computed for the Levitt procedure until a sufficient number of reversals have been observed.

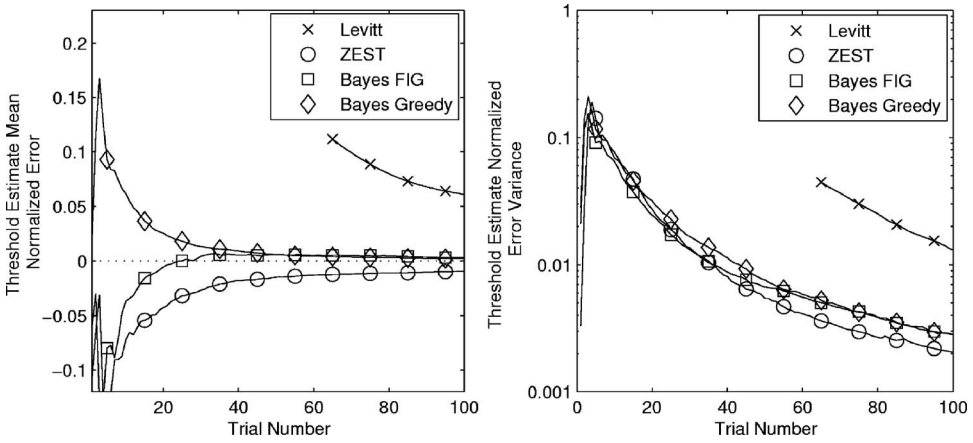


FIG. 4. Mean (left) and variance (right) of the error for the threshold parameter estimates versus trial number for each of the four psychometric procedures used in the psychoacoustic study. The *y*-axis units are in normalized error, calculated using the true threshold value for each of the 5000 simulated runs. The dashed line in the left plot is at zero, indicating zero bias.

previously. Analysis of statistical significance was performed on the error means using a z test to compare the difference between two population means and on the error variances using the F distribution to test the ratio of the sample variances (Devore, 1995). The differences between the mean errors for each method were statistically significant ($p<0.05$) with the exception of the difference between Bayes FIG and Bayes Greedy after 30 trials. The differences between the error variances for ZEST and both Bayes FIG and Bayes Greedy are statistically significant after 40 trials; the differences between Bayes FIG and Bayes Greedy converge and are not statistically significant after 50 trials.

By using a two-dimensional parameter probability distribution over both the threshold and slope of the psychometric function, the Bayes FIG and Bayes Greedy methods are gaining information about the slope parameter from the outcome of each trial. This is true even for the Bayes Greedy procedure; while it is attempting to place the stimulus value for each trial at the threshold, a point on the psychometric function that does not provide information about the slope since psychometric functions with all slope values pass through that point, some stimulus values will be offset from the true threshold value (either due to quantization of the stimulus values not allowing the true threshold value to be sampled or imperfect estimates of the threshold value) and information about the slope parameter will be provided. However, collecting substantive information about the slope parameter typically requires a greater number of trials than were presented in the psychoacoustic study or in the simulations (Macmillan and Creelman, 1991; Remus and Collins, 2007). Thus, the main function of the variable slope parameter in the Bayes FIG and Bayes Greedy methods in this study was to allow flexibility in fitting the psychometric function. However, the framework exists for extending estimation to multiple parameters without modification of the parameter estimation technique or stimulus selection rule.

The results of the computer simulations of the sinusoid in noise detection task using the Levitt and Kaernbach procedures with a variety of parameterizations are shown in Fig. 5. Results are also shown for the Bayes FIG procedure for comparison. The mean and variance of the threshold estimate error over the 5000 runs are again plotted versus trial number. Figure 5 illustrates the variability in bias and variance that can result from the different choice of step size. The Bayes FIG procedure provides lower bias, lower variance estimates of the threshold parameter than the Levitt or

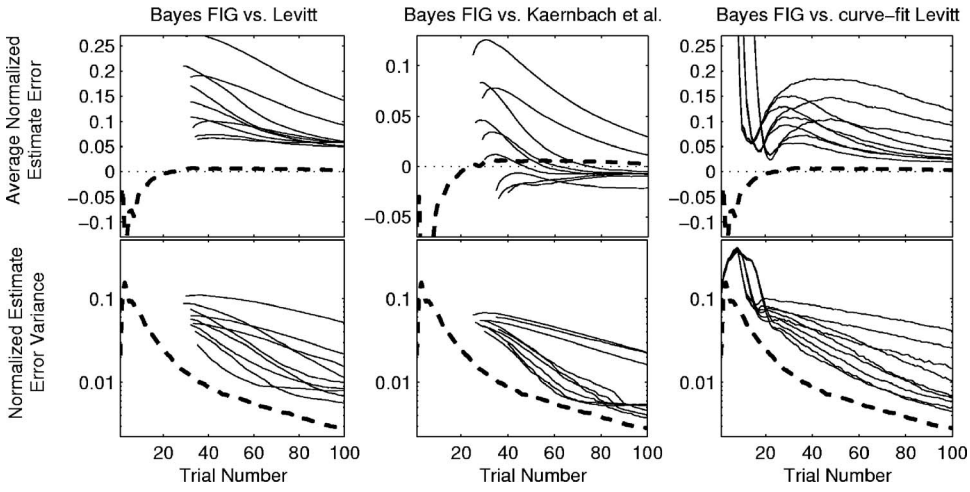


FIG. 5. Mean (top row) and variance (bottom row) of the error for the threshold parameter estimates versus trial number for the Bayes FIG, Levitt, and Kaernbach procedures. Results are shown for the Levitt procedure using averaged reversals (left column), the Kaernbach procedure (center column), and the Levitt procedure using curve-fitting for parameter estimation (right column). Bayes FIG results are the same as Fig. 4 and are shown as thick dashed lines in each subplot. Each subplot contains results for nine different combinations of initial and final step size. The *y*-axis units are in normalized error, calculated using the true threshold value for each of the 5000 simulated runs. The dashed line in the top row of subplots is at zero, indicating zero bias.

Kaernbach implementations, despite a sampling of step sizes from the suggested optimal range (Green *et al.*, 1989; Garcia-Perez, 1998). There are parameterizations for which the fixed step size procedures perform well, approaching the threshold estimate error levels achieved with the Bayes FIG procedure. However, there is no straightforward way to determine the "optimal" step sizes in a fixed step size procedure prior to measuring the subject's actual psychometric function since the optimal step sizes are dependent on the spread of the psychometric function.

IV. DISCUSSION

This study presented for comparison two fixed step size (FSS) staircase psychometric procedures and three Bayesian adaptive psychometric procedures, including two proposed procedures that were developed based on the Theory of Optimal Experiments. Results were presented for psychoacoustic studies conducted for two different tasks: detection of a pure tone in background noise and intensity discrimination using pure tones. Additional analyses were carried out through computer simulations using parameter values observed in the detection task. The outcomes of both the psychoacoustic studies as well as computer simulations suggest that the Bayesian adaptive procedures outperform the standard FSS staircase procedure, estimating parameters with lower error using fewer trials. This is in agreement with previous studies (Pentland, 1980; Leek, 2001; Marvitt *et al.*, 2003) investigating alternatives to FSS staircase procedures for psychometric evaluation.

Two psychometric procedures proposed in this study utilize a probability distribution over both the threshold and slope parameter. This may partially address the concern expressed in Amitay *et al.* (2006) that a limitation in many recently proposed psychometric procedures is the assumption of some form for the underlying psychometric function. A variable slope parameter provides an additional degree of freedom for fitting the observed data. The findings of Green (1990, 1993) may suggest that the slope parameter is not critical for psychometric procedures. The performance of the ZEST method in this study supports those conclusions, given the existence of a minimal amount of prior information such that the assumed slope value is not unreasonable. However, the performance shown by Bayes FIG and Bayes Greedy does not suggest any penalty when using a variable slope parameter. Computation time using the two-dimensional parameter estimation does not increase to the extent of making these techniques unfeasible for human subject testing (on a 866 MHz Pentium III, calculating the next stimulus value takes 7 ms with ZEST, 165 ms with Bayes FIG, and 150 ms with Bayes Greedy) and scales accordingly with increases to the number of sample points in $p(\lambda)$. Thus, the proposed methods appear to provide an advancement that addresses the concerns expressed by both Amitay *et al.* and Green.

The existing psychometric procedures considered in this study (Levitt and ZEST) are only two examples of the numerous techniques available for adaptive psychometric testing. The scope of this study was restricted to Bayesian adaptive procedures that estimate parameters using expected values calculated from the parameter probability distribution. The ZEST method was one of the original techniques to consider mean likelihood estimates rather than maximum likelihood estimates, and is thus the most appropriate predecessor to the techniques proposed in this study. Many studies have evaluated different maximum likelihood adaptive psychometric procedures, e.g., the method of maximum likelihood by Green (1993), and have shown performance gains over standard fixed step size staircase procedures as well.

The Bayesian update equation (Eq. (2)) used to estimate the psychometric function parameters requires an assumed value for γ, the probability of observing an incorrect response to a suprathreshold stimulus. Preliminary investigations to determine the effect of γ in simulated experiments suggests that moderate changes in the estimated or true values of γ can change the observed trends. This presents an interesting challenge since γ is a difficult parameter to understand intuitively, and is often handled cautiously or conservatively estimated. Several studies (e.g., Watson and Pelli, 1983; Kontsevich and Tyler, 1999) assume the same estimated and true value of γ, which is likely a best-case scenario since mismatch in γ can affect performance (Kontsevich and Tyler, 1999). Thus it seems important to consider a variety of cases to ensure that trends generalize over most reasonable conditions. One observed concern is the use of a constant γ parameter in most simulation studies. Based on subject feedback and personal observations during testing, the probability of a false incorrect response seems to be a function of subject fatigue, increasing as the number of trials increases, or after a string of subthreshold trials that decrease subject concentration. However, after a moderate number of trials, providing at least minimal information about the parameter values, the effect of γ appears to decrease dramatically such that there may no longer be a benefit from precisely modeling γ.

The results of the psychoacoustic study show that the intensity discrimination task had a higher average variance than the pure-tone detection task, which directly illustrates the relationship between the selected task and performance. The results of this study hint that when taking repeated measurements of a subject's ability to perform a perceptual task, there is some base line variability in the subject's performance that, at least in the current study, appears to be task dependent. An examination of the raw psychoacoustic data collected in this study suggests that task variability resulting from subject fatigue and lapses in concentration is greater than the parameter estimate variability measured in the computer simulations. Thus, there may be some lower performance limit on the variance of parameter estimates below which performance does not improve due to subject variability. This subject variability can be considered testing noise that models a subject's ability to repeat a task with consistency, and is likely to be dependent on the subject's focus, fatigue level, and task difficulty. Once that minimum is achieved, psychophysical testing will not benefit from lower variance procedures. However, there may still be improvements in performance to be achieved by focusing on reaching the lower limits in fewer trials.

The current study, in addition to several previous studies, indicate benefits to using Bayesian adaptive procedures, or similar psychometric procedures that model a functional form of the psychometric function, over the standard fixed step size staircase techniques. Fixed step size staircase procedures operate using a very intuitive stimulus selection rule: determine in which direction the threshold lies and select a new stimulus value in that direction. More sophisticated psychometric procedures, such as the Bayesian adaptive procedures considered in this study, have the appealing feature of not using a fixed, predetermined step size. Rather, they are able to use the information from all previous trials and the current estimate of the threshold parameter to select the best stimulus value. Bayesian parameter estimation, when paired with stimulus values placed at the estimated threshold, provides quick and accurate estimates of the thresholds for use in the next stimulus selection stage. This study suggests substantial benefit from using more sophisticated psychometric procedures and provides results for both psychoacoustic studies and computer simulations to verify the benefits of using more complex stimulus selection rules. Any study involving a substantial amount of psychophysical testing should consider alternatives to the fixed step size staircase procedures for significant increases in efficiency and performance.

ACKNOWLEDGMENTS

The authors would like to thank the listening study participants for their patience and cooperation. The authors are grateful to the anonymous reviewers for their valuable suggestions. This research was supported by NIH Grant No. 1-R01-DC007994-01.

Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., and Moore, D. R. (**2006**). "A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners," J. Acoust. Soc. Am. **119**, 1616–1625.

Anderson, A. J., and Johnson, C. A. (**2006**). "Comparison of the ASA, MOBS, and ZEST threshold methods," Vision Res. **46**, 2403–2411.

Chaloner, K., and Larntz, K. (**1989**). "Optimal Bayesian design applied to logistic regression experiments," J. Stat. Plan. Infer. **21**, 191–208.

Chaloner, K., and Verdinelli, I. (**1995**). "Bayesian experimental design: A review," Stat. Sci. **10**, 273–304.

Chernoff, H. (**1972**). "Sequential analysis and optimal design," in *Regional Conference Series in Applied Mathematics* (Society for Industrial and Applied Mathematics, Philadelphia).

Cover, T. M., and Thomas, J. A. (**1991**). *Elements of Information Theory* (Wiley-Interscience, New York).

Devore, J. L. (**1995**). *Probability and Statistics for Engineering and the Sciences* (Duxbury, Belmont, CA).

El-Gamal, M. A. (**1991**). "The role of priors in active Bayesian learning in the sequential statistical decision framework," in *Maximum Entropy and Bayesian Methods*, edited by W. Grandy, and L. Schick (Kluwer, Dordrecht).

Fedorov, V. V. (**1972**). *Theory of Optimal Experiments* (Academic, New York).

Garcia-Perez, M. A. (**1998**). "Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties," Vision Res. **38**, 1861–1881.

Green, D. M. (**1990**). "Stimulus selection in adaptive psychophysical procedures," J. Acoust. Soc. Am. **87**, 2662–2674.

Green, D. M. (**1993**). "A maximum-likelihood method for estimating thresholds in a yes-no task," J. Acoust. Soc. Am. **93**, 2096–2105.

Green, D. M., Richards, V. M., and Forrest, T. G. (**1989**). "Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics," J. Acoust. Soc. Am. **86**, 629–636.

Kaernbach, C. (**1991**). "Simple adaptive testing with the weighted up-down method," Percept. Psychophys. **49**, 227–229.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., and Supowit, A. (**1994**). "Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation," Vision Res. **34**, 885–912.

Klein, S. A. (**2001**). "Measuring, estimating, and understanding the psychometric function: A commentary," Percept. Psychophys. **63**, 1421–1455.

Kontsevich, L. L., and Tyler, C. W. (**1999**). "Bayesian adaptive estimation of psychometric slope and threshold," Vision Res. **39**, 2729–2737.

Leek, M. R. (**2001**). "Adaptive procedures in psychophysical research," Percept. Psychophys. **63**, 1279–1292.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Liao, X., and Carin, L. (**2004**). "Application of the theory of optimal experiments to adaptive electromagnetic-induction sensing of buried targets," IEEE Trans. Pattern Anal. Mach. Intell. **26**, 961–972.

MacKay, D. J. C. (**1992**). "Information-based objective functions for active data selection," Neural Comput. **4**, 589–603.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, New York).

Marvitt, P., Florentine, M., and Buus, S. (**2003**). "A comparison of psychophysical procedures for level-discrimination thresholds," J. Acoust. Soc. Am.

Pentland, A. (**1980**). "Maximum likelihood estimation: The best PEST," Percept. Psychophys. **28**, 377–379.

Phipps, J. A., Zele, A. J., Dang, T., and Vingrys, A. J. (**2001**). "Fast psychophysical procedures for clinical testing," Clin. Exp. Optom. **84**(5), 264–269.

Remus, J. J., and Collins, L. M. (**2007**). "A comparison of adaptive psychometric procedures based on the Theory of Optimal Experiments and Bayesian techniques: Implications for cochlear implant testing," Percept. Psychophys. **69**, 311–323.

Snoeren, P. R., and Puts, M. J. H. (**1997**). "Multiple parameter estimation in an adaptive psychometric method: MUEST, an extension of the QUEST method," J. Math. Psychol. **41**, 431–439.

Strasburger, H. (**2001**). "Converting between measures of slope of the psychometric function," Percept. Psychophys. **63**, 1348–1355.

Taylor, M. M. (**1971**). "On the efficiency of psychophysical measurement," J. Acoust. Soc. Am. **49**, 505–508.

Taylor, M. M., and Creelman, C. D. (**1967**). "PEST: Efficient estimates on probability functions," J. Acoust. Soc. Am. **41**, 782–787.

Tsutakawa, R. K. (**1972**). "Design of experiment for bioassay," J. Am. Stat. Assoc. **67**, 584–590.

Turpin, A., M., M. A., Johnson, C. A., and Vingrys, A. J. (**2002**). "Performance of efficient test procedures for frequency-doubling technology perimetry in normal and glaucomatous eyes," Invest. Ophthalmol. Visual Sci. **43**, 709–715.

Watson, A. B., and Pelli, D. G. (**1983**). "QUEST: A Bayesian adaptive psychometric method," Percept. Psychophys. **33**, 113–120.

Whaite, P., and Ferrie, F. P. (**1997**). "Autonomous exploration: Driven by uncertainty," IEEE Trans. Pattern Anal. Mach. Intell. **19**, 193–205.

Wichmann, F. A., and Hill, N. J. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.

Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002

Brad H. Story[a)]
Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

(Received 7 March 2007; revised 12 October 2007; accepted 16 October 2007)

A new set of area functions for vowels has been obtained with magnetic resonance imaging from the same speaker as that previously reported in 1996 [Story *et al.*, J. Acoust. Soc. Am. **100**, 537–554 (1996)]. The new area functions were derived from image data collected in 2002, whereas the previously reported area functions were based on magnetic resonance images obtained in 1994. When compared, the new area function sets indicated a tendency toward a constricted pharyngeal region and expanded oral cavity relative to the previous set. Based on calculated formant frequencies and sensitivity functions, these morphological differences were shown to have the primary acoustic effect of systematically shifting the second formant (F2) downward in frequency. Multiple instances of target vocal tract shapes from a specific speaker provide additional sampling of the possible area functions that may be produced during speech production. This may be of benefit for understanding intraspeaker variability in vowel production and for further development of speech synthesizers and speech models that utilize area function information.
 [DOI: 10.1121/1.2805683]

PACS number(s): 43.70.Bk, 43.70.Aj [CS] Pages: 327–335

I. INTRODUCTION

Collections of vocal tract area functions are useful for the development and testing of many types of speech production models and speech synthesizers. Since the early 1990s, magnetic resonance imaging (MRI) has been used extensively to acquire volumetric image sets of the head and neck from which area functions can be directly measured (e.g., Lakshminarayanan *et al.*, 1991; Baer *et al.*, 1991; Yang and Kasuya, 1994; Dang *et al.*, 1994; Dang and Honda, 1997; Story *et al.*, 1996, 1998; Narayanan *et al.*, 1995, 1997; Alwan *et al.*, 1997; Narayanan *et al.*, 1997; Story, 2005b). These area functions are typically obtained for static vocal tract shapes and are assumed to be representative of a particular speaker's "normal" production of specific vowels or consonants.

It has been shown, however, that a particular speaker can produce a range of different vocal tract shapes for the same target vowel. Story *et al.* (2001) reported area functions obtained from one male and one female speaker who were each asked to deliberately produce the vowels [i ɑ æ u] with three different qualities: normal, twang, and yawny. Production of each quality was hypothesized to require a different overall "setting" of the vocal tract shape (Laver, 1980). The results for each speaker indicated quite different area functions for each vowel across the three quality conditions, but production of each quality involved fairly systematic changes to the vocal tract shape across the vowels. Specifically, the yawny quality was characterized by a general widening of the oral cavity and lengthening of the vocal tract, whereas the twang was produced with a widened lip opening, constricted oral cavity, and shortened tract length. Although these speakers were trained vocal performers and the twang and yawny qualities imposed rather extreme changes on the vocal tract shape, the results suggest that a typical speaker would have the ability to generate a variety of tract shapes for a given target vowel depending on speaking conditions (e.g., Lindblom, 1996).

Takemoto *et al.* (2006) has also demonstrated differences in area functions of the same target vowel obtained from the same person, albeit one for dynamic speech and the other for static. They used MRI techniques that allow a three-dimensional (3D) volume of the head and neck to be acquired over the time course of an utterance (Takemoto *et al.*, 2006). Postacquisition image analysis was then used to generate a time-varying area function. Area functions taken from specific points in time within the vowel sequence [ɑeuio] were compared with those, of the same five vowels, obtained from the same speaker with a more conventional static imaging procedure. The area functions for each vowel were quite similar in overall shape, but there were localized differences in the magnitude of the cross-sectional area in both the anterior and posterior portions of the area function, depending on the particular vowel.

Information concerning multiple instances of target vocal tract shapes of a specific speaker is useful from the point of view that it provides a better sampling of the possible area functions that may be produced during speech production. This may be of particular benefit for understanding intraspeaker variability as well as for further development of vocal tract models based on statistical analyses of area functions

[a)]Electronic mail: bstory@u.arizona.edu

(e.g., Meyer *et al.*, 1989; Yehia *et al.*, 1996; Story and Titze, 1998; Story, 2005a, b; Mokhtari *et al.*, 2006).

The purpose of this article is to report a new set of vocal tract area functions for vowels obtained from the same speaker as that reported in Story *et al.* (1996). Between 2001 and 2003, image sets were collected from six speakers whose area functions were reported in Story (2005b). During this time, the opportunity existed to also obtain image sets of vowels from the speaker in Story *et al.* (1996). The aim in collecting these images was simply for purposes of comparison to the previous set, and to supply an additional instance of each vowel produced by the same person. Reporting similar data collected from the same speaker twice is perhaps a bit unusual, and, if such data were identical, would not be particularly interesting. The results, however, indicate that the new set of area functions support a downward shift in the second formant (F2), relative to the Story *et al.* (1996) set, over almost the entire [F1, F2] vowel space. This suggests that a different vocal tract setting, as described previously, was used for production of these vowels. The specific aim of this article is to report the area functions in numerical form, compare them graphically and acoustically with those from Story *et al.* (1996), and provide some explanation of how the differences observed between the two area function sets support a downward shift in F2.

II. METHOD

A. Speaker

The speaker was the same person who was scanned with MRI for the Story *et al.* (1996) study (henceforth referred to as STH96). It is acknowledged that he was the first author of STH96 and is the sole author of the present study. All new image sets were collected on 22 May 2002 at the University of Arizona Medical Center in Tucson, AZ. At this time the speaker was 37 years old. The previous data (reported in 1996) had been collected in June 1994. In the intervening years the speaker moved from Iowa City, IA to Denver, CO where he lived until 2000; he then relocated to Tucson, AZ. The speaker's height and weight were identical to those stated in STH96 (ht.=5 ft 7 in. wt.=145 lbs).

B. Image collection

MRI was used to obtain volumetric image sets of vocal tract shapes that corresponded to the speaker's production of the American English vowels [i ɪ e ɛ æ ʌ ɑ ɔ o ʊ u]. The MR images were acquired with a General Electric Sigma 1.5 Tesla scanner. The data acquisition mode was fast spin echo and the scanning parameters were set to TE=13 ms, TR=4000 ms, ETL=16 ms, and NEX=2. During the speaker's production of a particular vocal tract shape, a 28 slice series was collected with an interleaved acquisition sequence. Each image set consisted of contiguous, parallel, axial sections (slices) extending from a location just superior of the hard palate to an inferior location near the first tracheal ring. The field of view (slice dimensions) for each slice was 24 cm×24 cm which, with a 256×256 pixel matrix, provided an in-plane spatial resolution of 0.938 mm/pixel. Although the STH96 images were obtained with a rigid anterior/posterior neck coil, at the time these new images were collected only a flexible anterior neck coil was available. The scanning parameters were set to allow an image slice thickness of 4 mm for all image sets. In STH96 the slice thickness was 5 mm.

The procedure for acquiring the image sets was nearly identical to that described in Story (2005b). Each vowel was produced as if it were to be spoken in an hVd syllable, but was instead sustained. Thus, any vowels typically spoken as diphthongs (e.g., [e] and [o]) would be represented in the image set as the onset vowel of that diphthong. Shortly after the speaker began phonation for a particular target vowel, the MR technologist initiated the scan. After 8 s the scan was paused to allow time for the speaker to breathe. The scanning was continued when the speaker resumed phonation. The scanning time required for each image set (i.e., for one complete tract shape) was 4 min and 32 s as compared to 4 min and 16 s in STH96. This required approximately 30 repetitions of each target vowel. With pauses for respiration between repetitions, each image set was completed in about 10–15 min. The speaker's goal was a "normal" production of each vowel, with a strong focus on maintaining a consistent shape.

C. Image analysis and area functions

The image analysis was the same as that described in Story (2005b) and Story *et al.* (1996, 1998, 2001). In brief, for each vocal tract shape the procedure included segmentation of the airspace from the surrounding tissue, shape-based interpolation to generate a 3D reconstruction of the airspace, and cross-sectional area analysis of the airspace. The collection of areas obtained, which extends from just above the glottis to the lips, along with the distance of each cross section from the glottis, comprise the area function. Each area function was subsequently resampled with a cubic spline from which 44 area sections were obtained at equal length increments. A smoothing filter was subsequently applied to remove small discontinuities assumed to be imaging artifacts (see Story *et al.*, 2001, p. 1653). The piriform sinuses were segmented in each image set but were not included in the cross-sectional area analysis. Hence, information about them will not be reported here.

D. Audio recording and formant analysis

On the day immediately following image collection, the speaker's production of all 11 vowel sounds was recorded. For this session the speaker inserted earplugs and lay supine on a cushioned table in a sound-treated booth. At least three repetitions of each vowel were produced as long sustained productions with approximate durations of 4–8 s, similar to the duration of a single vowel repetition in the MR scanner. The speaker attempted to produce the vowels as similarly as possible, in both quality and loudness (self-perception), to those produced the day before in the MR scanner. The audio signal was transduced with an AKG CK92 microphone positioned 30 cm from the speaker and off-axis at 45°. The signal was recorded on digital audio tape at a sampling frequency of 44.1 kHz and later transferred to separate digital

computer files for each vowel. In addition, audio recordings obtained in 1994 that coincide with STH96 were transferred to digital files so that they could be analyzed with the same methods used for the 2002 recordings. These were also long sustained vowels recorded in the supine position with earplugs.

For the 2002 recordings formant frequencies were estimated over the time course of three repetitions of each vowel with PRAAT's formant analysis module (Boersma and Weenink, 2007). Depending on the particular vowel, formant analysis parameters were manually adjusted so that the formant contours of F1, F2, and F3 were aligned with the centers of their respective formant bands in a simultaneously displayed wide-band spectrogram. All time-dependent formant values for each vowel were transferred to MATLAB (Mathworks, 2006) where means and standard deviations were computed. Depending on the vowel, the three repetitions provided 15–25 s of recorded signal over which the analysis was performed. The 1994 recordings were analyzed in exactly the same manner but for some vowels there were only two repetitions available. Nonetheless, this still provided 10 or more seconds of recorded signal for analysis.

E. Calculation of formant frequencies

Frequency response functions were calculated for each newly obtained area function as well as for the ten vowels of STH96. This was performed with a frequency-domain technique (Sondhi and Schroeter, 1987; but specifically as presented in Story *et al.*, 2000) that included energy losses due to yielding walls, viscosity, heat conduction, and radiation. Prior to the calculations, the STH96 area functions were smoothed with the same process applied to those of the present study (see Sec. II C). Formant frequencies were determined by finding the peaks in the frequency response functions with parabolic interpolation (Titze *et al.*, 1987).

Calculation of the formant frequencies reported in STH96 was performed differently than in the present study. These differences do have an effect on the actual formant values, hence, they will be briefly summarized. For STH96 a wave-reflection algorithm (Liljencrants, 1985; Story, 1995) was used to generate a time-domain simulation of a particular vowel sound. The algorithm used a glottal flow pulse signal as the voice source and included losses due to yielding walls, viscosity, heat conduction, and radiation. This particular simulation required that each section (tubelet) of the area function have a length of $c/(2F_s)$, where c is the speed of sound and F_s is the sampling frequency. Accordingly, the area functions were reported as a variable number of sections, each with the same section length (STH96, p. 547). In the present study, all area functions contain 44 sections but the section length can vary across vowels. Another important difference is that a side-branch representing the piriform sinuses was coupled to the main vocal tract at a point 2.4 cm from the glottis for all vowels. The cross-sectional areas for the piriform sinuses were reported in Story (1995). Each simulated vowel was then subjected to exactly the same LPC analysis that was used to measured formants from recorded speech. This involved an initial preemphasis prior to the LPC analysis.

Each of these differences can potentially contribute to a slightly different calculation of the formants than would be given by the frequency-domain approach used in the present study. As can be discerned from Sec. III, the formants calculated for the STH96 area functions with the present frequency-domain approach are not the same as those originally reported. A piriform sinus branch, however, was added to the frequency-domain model and used to recalculate the formants of the STH96 area functions. In this case, the formants were well-matched to the originals, suggesting that the inclusion of the piriform sinus is the largest contributor to the difference in the calculated formants. Although the effects of these sinuses on vowel formants are an important area for further study (e.g., Dang and Honda, 1997), they were considered to be outside the scope of the present report. Hence, the calculated formants for both sets of area functions are reported in Sec. III for the condition *without* a piriform sinus branch.

III. RESULTS

The area functions are presented numerically in Table I. Each column contains 44 cross-sectional areas that extend from the glottis (section 1) to the lips (section 44). The bottom two rows contain the section length and total tract length of each area function, respectively. The measured and calculated acoustic characteristics are shown in Table II. In the top row are the measured fundamental frequencies (F0) that range from 149 Hz for [e] to 159 Hz for [i]. These are somewhat higher than the speaker's typical F0 but were thought, by the speaker, to be representative of those produced during MR scanning. Measured and calculated formant frequencies are shown in the middle portions of Table II. The lower three rows indicate the percent error of the computed formants relative to the mean value of the natural speech formants. These range from a low of 1.6% for the second formant of [e] to high of 35.7% for the second formant of [ʌ].

Each new area function is plotted along with its STH96 counterpart in Fig. 1. To accurately depict the time period in which the respective image sets were acquired, the legends refer to the area functions from the present study as "2002" and those from STH96 as "1994." An [e] vowel was not reported in STH96, hence, the subplot for it contains only the 2002 version. There are both similarities and differences between the two area function sets. For example, the variation in the cross-sectional area along the initial 2–4 cm of VT length is approximately the same for both versions of the vowels [i ɪ ɛ æ o ʊ u]. For [ʌ], [ɑ], and [ɔ], however, an increase in area occurs closer to the glottis in the new versions than in the old. This is apparently a consequence of lengthening the pharyngeal section, perhaps by larynx lowering, which extends from about 3.5 to 9 cm above the glottis in the 1994 area functions and from 2.5 to 9 cm in the new ones. The new versions of these same three vowels, along with [o] and [ʊ], also exhibit larger areas within the oral cavity but have nearly the same lip termination area as

TABLE I. Area vectors for each vocal tract shape. Each original area function has been resampled to consist of 44 area sections given in cm^2; the length of each section is given by Δ in cm. The glottal end of each area vector is at section 1 and the lip end at section 44. The total vocal tract length (VTL) in cm is computed as 44Δ.

Section	i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u
1	0.51	0.28	0.29	0.37	0.31	0.23	0.56	0.27	0.38	0.37	0.54
2	0.59	0.21	0.26	0.30	0.21	0.34	0.62	0.43	0.45	0.38	0.61
3	0.62	0.21	0.30	0.24	0.18	0.47	0.66	0.54	0.57	0.49	0.66
4	0.72	0.30	0.40	0.23	0.23	0.60	0.78	0.67	0.77	0.62	0.75
5	1.24	0.47	0.55	0.29	0.33	0.77	0.97	0.83	1.31	0.85	1.13
6	2.30	0.71	0.74	0.41	0.50	1.06	1.16	0.92	1.92	1.28	1.99
7	3.30	1.12	0.99	0.58	0.78	1.26	1.12	0.89	1.74	1.62	2.83
8	3.59	1.48	1.09	0.82	0.96	1.09	0.82	0.73	1.11	1.47	2.90
9	3.22	1.35	0.90	0.97	0.85	0.80	0.55	0.55	0.75	1.04	2.52
10	2.86	1.05	0.69	0.82	0.63	0.65	0.45	0.44	0.59	0.81	2.40
11	3.00	0.92	0.77	0.62	0.46	0.56	0.37	0.37	0.57	1.03	2.83
12	3.61	0.92	1.31	0.54	0.36	0.47	0.29	0.28	0.68	1.44	3.56
13	4.39	1.19	2.13	0.54	0.33	0.37	0.21	0.18	0.73	1.49	3.99
14	4.95	1.94	2.74	0.68	0.46	0.24	0.15	0.14	0.67	1.28	3.89
15	5.17	2.83	3.03	1.09	0.73	0.17	0.16	0.15	0.58	1.06	3.50
16	5.16	3.31	3.23	1.62	1.00	0.18	0.25	0.14	0.49	0.85	3.04
17	5.18	3.48	3.33	2.03	1.30	0.23	0.34	0.13	0.44	0.69	2.64
18	5.26	3.60	3.27	2.35	1.66	0.27	0.43	0.14	0.42	0.56	2.44
19	5.20	3.64	3.09	2.51	1.97	0.28	0.54	0.18	0.49	0.42	2.31
20	5.02	3.49	2.84	2.39	2.06	0.29	0.61	0.20	0.53	0.27	2.07
21	4.71	3.20	2.66	2.22	2.03	0.33	0.67	0.23	0.38	0.27	1.80
22	4.13	2.90	2.46	2.13	2.01	0.52	0.98	0.49	0.30	0.41	1.52
23	3.43	2.59	2.14	2.00	1.89	0.97	1.76	1.03	0.45	0.51	1.14
24	2.83	2.21	1.79	1.78	1.66	1.50	2.75	1.58	0.61	0.49	0.74
25	2.32	1.87	1.44	1.58	1.49	1.91	3.52	2.06	0.71	0.47	0.42
26	1.83	1.54	1.17	1.43	1.42	2.23	4.08	2.61	0.79	0.50	0.22
27	1.46	1.20	1.00	1.31	1.37	2.65	4.74	3.35	0.86	0.58	0.14
28	1.23	0.92	0.88	1.23	1.34	3.29	5.61	4.34	1.01	0.80	0.20
29	1.08	0.74	0.80	1.24	1.41	4.13	6.60	5.51	1.41	1.19	0.47
30	0.94	0.59	0.81	1.38	1.58	5.00	7.61	6.70	2.09	1.62	0.89
31	0.80	0.52	0.85	1.61	1.82	5.77	8.48	7.75	3.00	2.27	1.15
32	0.67	0.54	0.84	1.82	2.19	6.33	9.06	8.63	4.10	3.24	1.42
33	0.55	0.59	0.86	1.96	2.63	6.61	9.29	9.29	5.16	4.16	2.17
34	0.46	0.65	0.96	2.01	2.97	6.63	9.26	9.59	6.22	5.00	3.04
35	0.40	0.71	1.18	2.00	3.17	6.45	9.06	9.42	7.34	5.70	3.69
36	0.36	0.67	1.35	1.95	3.40	6.04	8.64	8.78	8.15	6.11	4.70
37	0.35	0.61	1.48	1.77	3.56	5.39	7.91	7.82	8.61	6.21	5.74
38	0.35	0.57	1.62	1.48	3.57	4.42	6.98	6.50	8.37	6.29	5.41
39	0.38	0.50	1.49	1.30	3.58	3.29	6.02	4.95	6.76	6.24	3.82
40	0.51	0.48	1.29	1.21	3.44	2.37	5.13	3.47	4.37	4.91	2.34
41	0.74	0.54	1.24	1.10	3.15	1.74	4.55	2.15	2.30	2.61	1.35
42	0.92	0.73	1.17	0.99	3.38	1.36	4.52	1.38	1.06	1.09	0.65
43	0.96	0.93	1.04	0.91	3.99	1.17	4.71	1.11	0.58	0.63	0.29
44	0.91	0.82	0.95	0.78	4.17	0.99	4.72	0.90	0.47	0.59	0.16
Δ (cm)	0.384	0.376	0.386	0.393	0.366	0.390	0.388	0.395	0.417	0.440	0.445
VTL (cm)	16.90	16.55	16.98	17.30	16.11	17.14	17.09	17.40	18.33	19.34	19.59

they did previously. For [ʌ ɔ o ʊ], the speaker maintained a more constricted pharyngeal region in the new versus the old area functions that, with the expansion of the oral cavity, would suggest they were produced with an increased degree of the "back" dimension. The overall shape for [ɪ], [ɛ], and [æ] is similar across the 1994 and 2002 sets even though the magnitude of the areas is different. This is especially prominent for [æ] where peaks occur near 8 and 14 cm from the glottis, respectively, in both versions, but the more recent area function is on the order of 1 to 2 cm^2 smaller within this region.

For nearly all vowels, the vocal tract length in the new set of area functions was greater or equal to those from 1994. The exception is the [ɑ] vowel which is slightly shorter. The [ɛ], [o], [ʊ], and [u] were longer by more than 1 cm. In the case of [ɛ], the entire length axis of the area function appears

TABLE II. Fundamental frequencies F_0, and measured and calculated formants for the 11 vowels produced by the speaker. Each measured formant (denoted by superscript "N") is the mean across several seconds of recording and s.d. is the standard deviation. The calculated formant values are denoted by "C." The Δ's represent the percent error of the computed formants relative to the mean value of the natural speech formants (e.g., $\Delta 1 = 100(F1^C - F1^N)/F1^N$). All values have units of Hertz except the Δ's which are percentages.

	i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u
F_0	159	154	149	154	153	155	154	154	155	156	155
$F1^N$	295	478	509	659	765	752	715	665	563	518	430
s.d.	±17	±9	±10	±15	±15	±13	±9	±19	±9	±8	±5
$F2^N$	1923	1965	1917	1740	1724	1300	1180	932	835	947	917
s.d.	±28	±11	±10	±26	±22	±23	±31	±12	±30	±13	±14
$F3^N$	2797	2688	2617	2485	2501	2698	2691	2814	2637	2355	2087
s.d.	±60	±27	±42	±69	±59	±33	±21	±54	±26	±30	±14
$F1^C$	325	413	484	565	810	593	694	557	499	477	314
$F2^C$	2139	2103	1887	1595	1655	836	942	709	768	789	702
$F3^C$	2976	2626	2404	2200	2310	3099	2948	3176	2421	2499	2298
$\Delta 1$	10.0	−13.6	−4.9	−14.3	5.8	−21.1	−3.0	−16.2	−11.3	−7.9	−27.1
$\Delta 2$	11.2	7.0	−1.6	−8.3	−4.0	−35.7	−20.2	−23.9	−8.0	−16.7	−23.5
$\Delta 3$	6.4	−2.3	−8.1	−11.5	−7.6	14.9	9.6	12.9	−8.2	6.1	10.1

to be stretched, whereas, for the latter three vowels, the length increase could be attributed to more extreme lip rounding.

The F1 and F2 formant frequencies calculated for both the 1994 and 2002 versions of each vowel are plotted against each other in Fig. 2(a). A data point for [e] is only present for the 2002 vowels. Relative to the 1994 vowels, a predominant feature is that the entire 2002 vowel space is shifted downward along the F2 dimension, except for [ɪ]. Such a global change in F2 suggests a systematic, rather than random, difference in the vocal tract shapes of the 2002 vowels. To investigate this difference, acoustic sensitivity functions (Schroeder, 1967; Fant and Pauli, 1975) corresponding to F2 were calculated for each of the 2002 vowels using the method described in Story (2006, p. 715). For each vowel, regions along the vocal tract length were identified from the F2 sensitivity function where an increase or decrease in cross-sectional area would increase the frequency of F2. That is, what change in the area function would move F2 toward the value calculated for the corresponding 1994 vowel. Two of these regions are indicated by the bold portion of each 2002 area function shown in Fig. 1; the solid dots and vertical lines denote the division of the regions. The arrows above the area function, within each region, show the direction of cross-sectional area change that would perturb F2 upward in frequency. In all cases, a synergistic expansion and constriction of the marked posterior and anterior regions, respectively, would increase the frequency of F2. For all of the vowels, these prescribed changes in area (to increase F2), in at least one of the regions, would account for differences relative to the 1994 vowels. For instance, a decrease in the area of the 2002 [i] vowel between 8.5 and 14 cm from the glottis would bring it more in line with the cross-sectional area of the 1994 [i], and consequently increase F2. Increases in area for at least some portion of the posterior regions of the vowels [ɛ æ ʌ ɑ ɔ o ʊ u] and decreases in the area of the anterior regions of [ʌ ɑ ɔ o ʊ u], all of which would increase F2, are also in the direction needed to more closely match the 1994 versions of these same vowels.

Certainly other characteristics of the area function, such as tract length and cross-sectional area near the lips, may contribute to the differences in F2. But the dominant structural difference between the two sets of area functions that corresponds to their acoustic differences is a tendency toward a constricted pharynx and expanded oral cavity in the 2002 vowels. This is exemplified by the mean area functions calculated for both sets, that are plotted in the lower right panel of Fig. 1. Relative to the 1994 version, the 2002 mean area function is similar in the initial 2–4 cm of VT length, slightly constricted in the pharyngeal region (between 4 and 11 cm from glottis), slightly expanded in the oral cavity (between 11 and 16 cm from glottis), and again similar near the lip termination. These differences result in an F2 that is about 200 Hz lower than that calculated for the 1994 mean.

The F1 and F2 formant frequencies measured from the recorded speech are shown in Fig. 2(b). The center of each black ellipse represents the mean [F1,F2] value over the time course of the 15–25 s duration of each of the 2002 vowels. The horizontal and vertical extent of each ellipse indicates ±1 s.d. The 1994 formants are similarly plotted with gray ellipses. The standard deviations are small enough that there is no overlap of the corresponding vowels in the two sets except for and [ɔ] and nearly for [ɪ] (the ellipses for the 1994 [ʌ] and the 2002 [ɑ] do, however, overlap). There is a downward trend in the second formant (F2) of some of the 2002 vowels relative to those of 1994, although less prominent than for the calculated formant values. Specifically, the second formants of [i], [æ], and [ɔ] decreased by 115 Hz or more, whereas, [ɛ], [ʊ], and [u] also showed decreases in F2 but to a lesser degree. The remaining three vowels [ɪ], [ʌ], and [ɔ], had nearly identical F2 values as those of the 1994 vowels (their F1 values were different enough to prevent

them from overlapping in the [F1,F2] space). There is also a notable upward shift in the first formant (F1) of [æ], similar to that observed for the calculated formants.

IV. DISCUSSION

A new set of area functions for 11 vowels has been obtained from the same speaker who provided those reported in STH96. Eight years separated the collection of these two data sets. Comparing them, differences were observed in the cross-sectional area variation along the vocal tract axis as well as differences in vocal tract length. In a general sense, the 2002 vowels tended to be slightly more constricted in the pharyngeal region and slightly more expanded in the oral cavity than their 1994 counterparts, although there are exceptions for a few vowels. Based on calculated sensitivity functions, these morphological differences were shown to have a primary acoustic effect of systematically moving F2 downward in frequency. Although the results have indicated how the observed acoustic differences are related to the area function differences, two questions remain unanswered. First, why are the 1994 and 2002 area functions sets different? Second, why are the differences in calculated formant frequencies for each of the two sets, especially F2, greater in magnitude than the differences observed from the measured formants?

One obvious potential source of area function differences is the 8 year separation between collection of the two image sets. During this time the speaker aged from 29 to 37 years, and it is known that structural changes to the craniofacial complex and pharynx can occur in adulthood (Kollias and Krogstad, 1999; West and McNamara, 1999). Anatomical changes seem somewhat unlikely, however, because of the way in which many of the peaks and valleys in the two sets of area functions are fairly well aligned, even though cross-sectional areas are different. This was noted in particular for the [æ] but can also be observed for most of the other vowels. A detailed anatomical analysis based on the original MR image sets could potentially reveal whether structural changes did occur, but this was not carried out for the present study.

A second possible source of change is the speaker's relocation during the 8 years between image collections. As stated in Sec. II, he first moved from Iowa City, IA to Denver, CO, and then to Tucson, AZ. There are some dialectal differences relative to these three cities. Particularly notable is the merging of /ɔ/ with /ɑ/ in the western states (Labov, 1996). There is little evidence of these two vowels merging in the 2002 area function set, however, as the tract shapes of [ɔ] and [ɑ] are quite different near the lip termination and their corresponding [F1, F2] values are well separated in the vowel space plot. This does not rule out the possibility that other dialectal changes could have affected the speaker's vowel production, but does show that a well-known vowel change relative to these three cities was not observed in the present data.

A third possible source of area function change could be slight differences in the imaging procedure. The new image sets were collected using a flexible neck coil and 4 mm axial slices thickness, whereas a rigid anterior/posterior neck coil and 5 mm axial slices were used for acquiring the 1994 image data. Although different coils could affect image quality and slice thickness could affect the accuracy of the 3D vocal tract reconstruction, especially in the oral cavity because of exclusive use of axial images, one would expect such differences to be rigidly systematic with respect to particular regions of the vocal tract. For example, if slice thickness were the cause of the area function differences, the oral cavity portion might be expected to be consistently under-or overestimated in area across all the vowels. But this is not what was observed: Six of the new vowels have larger areas in the oral cavity than the 1994 versions and three are clearly smaller.

A more likely reason for the differences in the two area function sets is that the speaker had, over time, acquired a slightly different habitual "setting" of the vocal tract shape. Laver (1980) referred to such settings as "tendencies to maintain a particular constrictive (or expansive) effect" within some region of the vocal tract that biases the resulting formant frequency patterns toward a particular type of global timbre. The downward shift in F2 supported by the 2002 area functions indicates such a bias. In addition, the overall differences between area functions of the 1994 and 2002 data sets coincides with Laver's (1980) description of a "pharyngealized" quality where a constrictive effect is imposed on the middle pharynx which also results in some expansion of the oral cavity. Area functions and formant frequencies reported for a "yawny" voice quality (Story *et al.*, 2001) also coincide somewhat with the overall shape changes and the downward shift in F2 observed in the present study. That a change in habitual setting occurred during the 8 years between the data collection is, perhaps, not surprising. During this time, in addition to relocating, the speaker became familiar with many voice therapy techniques as well as methods for coaching and enhancing the professional voice. He also studied various voice qualities that involved subtle changes to the vocal tract shape (e.g., Story *et al.*, 2001; Story and Titze, 2002). Through these investigations he became well-practiced in producing many of the various qualities described by Laver (1980). Taken together these experiences likely had an effect on his typical speech production pattern that is perhaps exemplified in the differences observed between the two sets of area functions.

Also supporting the notion that a vocal tract "setting" could account for differences in the area functions is the fact that the formant frequencies of the recorded vowels from 1994 and 2002 were in closer proximity than the calculated formants based on the two area function sets. This suggests that the speaker was not obligated by anatomical, dialectal, or even habitual influences to produce the vowels with the same degree of decrease in F2 as generated by calculation using the 2002 area functions. For whatever reason, the speaker appears to have reduced the degree of a pharyngeal/yawny-type setting during the audio recording on the day following image collection. Alternatively, it might be concluded that the speaker cannot actually produce vowels with formant frequencies like those resulting from the calculations, and that the area functions are simply in error.

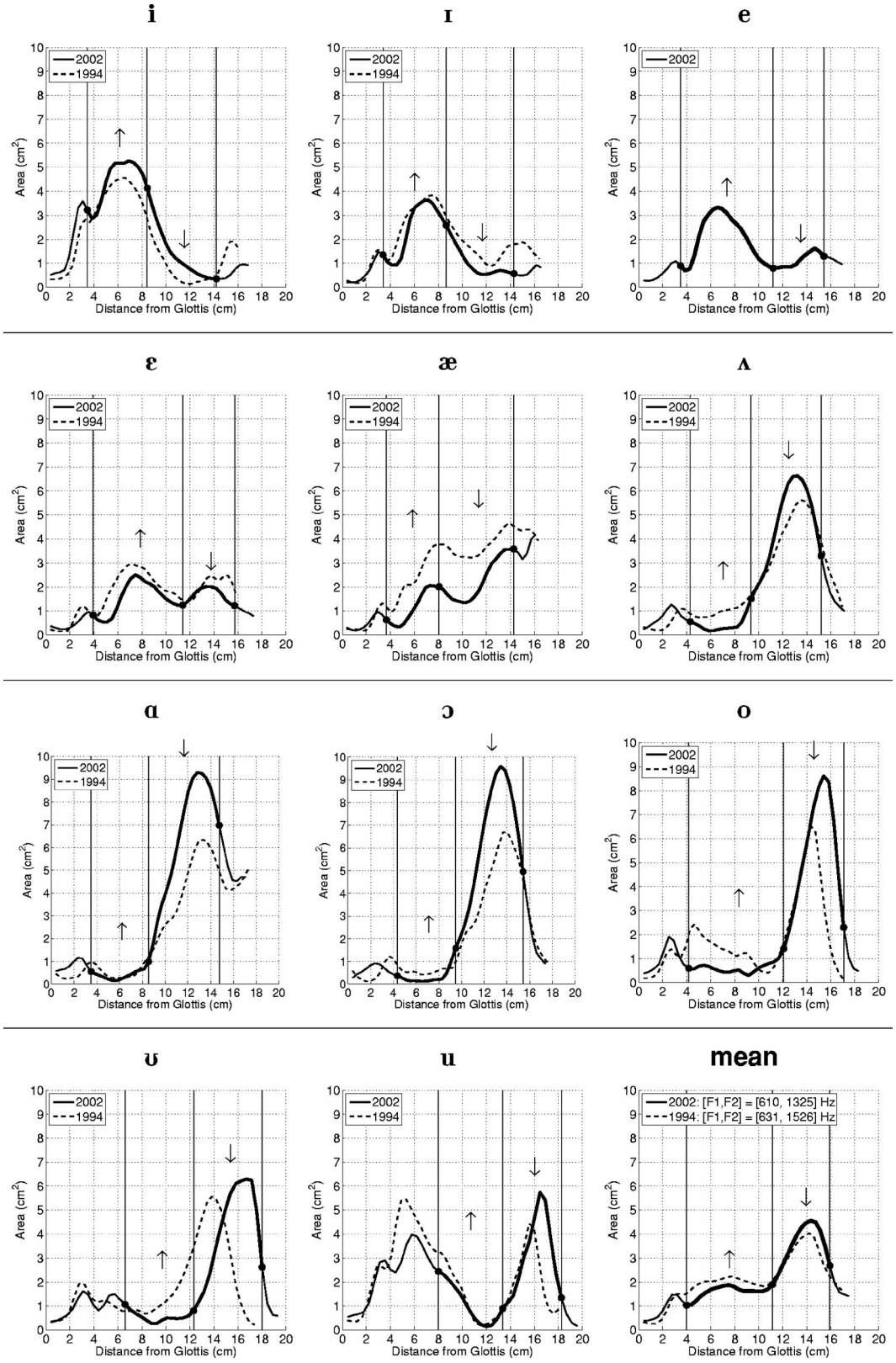
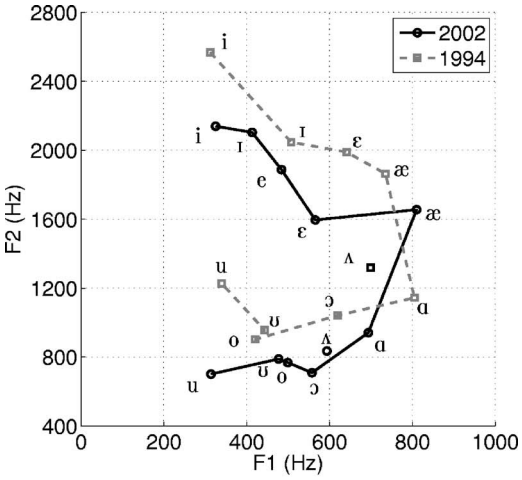
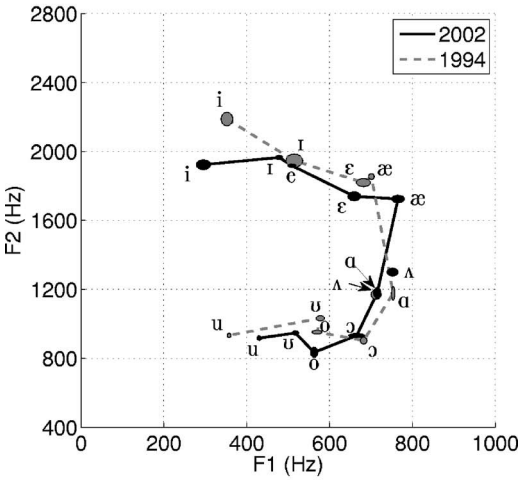


FIG. 1. Two sets of vocal tract area functions obtained from the same speaker. The solid lines represent area functions from the present study and the dashed lines are those area functions reported in STH96. The lower right subplot contains the mean area functions from each set. The legend on each subplot designates each set according to the year (1994 or 2002) in which the images were collected with MRI. The solid dots on the 2002 area functions along with the vertical lines denote regions within which a change of cross-sectional area, in the direction of the arrows, would increase F2.

The alternative conclusion begs the question of whether the speaker can actually produce vowels with the same downward shift in F2 as produced by the calculations. To test this, an audio recording was recently (September 2007) made of the same speaker producing two different series of 11 hVd syllables (a vowel embedded between an initial /h/ and final



(a) Calculated formant frequencies



(b) Measured formant frequencies

FIG. 2. Vowel space plot of F1 and F2 frequencies calculated and measured from the 1994 and 2002 versions of each vowel (the [e] exists only for the new set). The data points are connected by solid or dashed lines to clarify the set to which they belong and to provide a rough outline of the possible vowel space. Because of their positions in the F2 vs F1 plane, the two [ʌ] vowels are not connected to the other vowels within their respective sets. (a) Calculated formant frequencies. (b) Measured formant frequencies. Here the data points are indicated by ellipses in which the horizontal and vertical extents denote ±1 s.d. in the F1 and F2 dimensions, respectively; the mean value is located in the center of each ellipse.

/d/) containing the same 11 vowels as represented by the 2002 area function set. In the first series, the syllables were spoken in a pharyngealized/yawny quality. This was done based on the speaker's knowledge that the pharynx should tend toward a constrictive configuration to the degree allowed by a given vowel. For contrast, a second series of hVds was produced by releasing the constrictive effect in the pharynx and imposing somewhat of a constrictive tendency in the palatal region. The speaker, who was in the supine position with ear plugs inserted, was recorded while producing three repetitions of each hVd. The audio signals were saved in digital form directly to a computer disk. The speaker received no feedback concerning formant frequency locations during the recording. Formant frequencies were extracted from the vowel portion of each hVd (three repetitions of each) with an LPC technique programmed in MATLAB

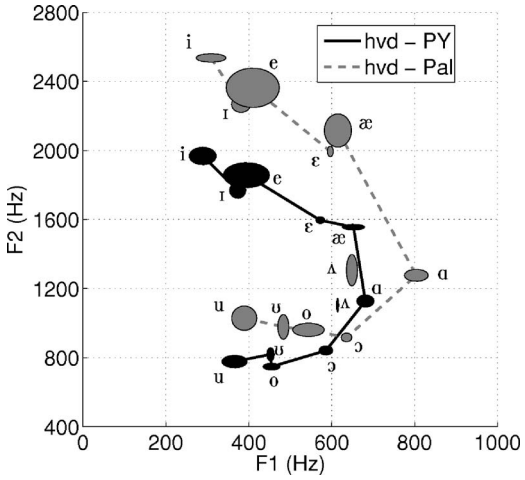


FIG. 3. Vowel space plot of F1 and F2 frequencies measured for the vowels in two series of hVd syllables. The data points are indicated by ellipses as they were in Fig. 2. The vowel formants denoted by solid black points are for a pharyngealized/yawny quality (PY), whereas the gray points indicate formants for the palatalized quality (Pal).

(these were verified to be essentially the same as would have been given by the PRAAT analysis algorithm). The results are shown in Fig. 3. The vowels of the pharyngealized/yawny quality, indicated by the black ellipses (horizontal and vertical extent represents ±1 s.d.), have a range of F2 values similar to those produced by the 2002 area functions and are shifted downward relative to the platalized vowels (gray ellipses). The standard deviations are larger than for the sustained vowels shown in Fig. 2(b) because of the nearly continuous movement of the vocal tract required to produce an hVd syllable.

That the speaker was indeed capable of producing a set of vowels with second formant frequencies as low as those calculated from the 2002 area functions, and at the same time could also produce a set of vowels with much higher values of F2 (more like those of the 1994 set), suggests that the observed differences in area function shapes could have resulted from different vocal tract "settings" rather than anatomical or dialectal change, or from imaging method differences. The reason that differences in F2 were less extreme between the measured formants from 1994 and 2002 [Fig. 2(b)] compared to the F2 differences of the corresponding calculated formants is apparently because the speaker utilized a more extreme version of the pharyngealized/yawny setting during image collection than in the subsequent audio recording. Perhaps this setting allows for production of vocal tract shapes that are easier to maintain over many repetitions in the noisy environment of a MR scanner.

The second set of area functions reported in the present study provides additional instances of target vocal tract shapes produced by one specific speaker. These data show that the vocal tract shape may be highly variable for the same target vowel depending on the particular setting used by the speaker. Such multiple instances of target vocal tract shapes may be useful for understanding intraspeaker variability and for purposes of speech synthesis and speech production modeling.

ACKNOWLEDGMENTS

The author would like to thank Ted Trouard for consulting on image acquisition and Jennifer Johnson for operating the MR scanner. This research was supported by NIH Grant No. R01-DC04789.

Alwan, A. A., Narayanan, S. S., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. 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.

Boersma, P., and Weenink, D. (**2007**). "PRAAT, Version 4.6.09," www.praat.org. (last viewed on 8 August, 2007).

Dang, J., and Honda, K. (**1997**). "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101**, 456–465.

Dang, J., Honda, K., and Suzuki, H. (**1994**). "Morphological and acoustical analysis of the nasal and the paranasal cavities," J. Acoust. Soc. Am. **96**, 2088–2100.

Fant, G., and Pauli, S. (**1975**). "Spatial characteristics of vocal tract resonance modes," in Proceedings of the Speech Communications Seminar, Vol. **74**, Stockholm, Sweden, 1–3 August, pp. 121–132.

Kollias, I., and Krogstad, O. (**1999**). "Adult craniocervical and pharyngeal changes—a longitudinal cephalometric study between 22 and 42 years of age. I. Morphological craniocervical and hyoid bone changes," Eur. J. Orthod. **21**, 333–344.

Labov, W. (**1996**). "The organization of dialectic diversity in North America," Presented at the Fourth International Conference on Spoken Language Proceedings, Philadelphia, 6 October. Available online at www.ling.upenn.edu/phono_atlas/ICSLP4.html. (last viewed on 6 August 2007).

Lakshminarayanan, A. V., Lee, S., and McCutcheon, M. J. (**1991**). "MR imaging of the vocal tract during vowel production," J. Magn. Reson Imaging **1**, 71–76.

Laver, J. (**1980**). "The Phonetic Description of Voice Quality," Cambridge University Press, Cambridge, UK.

Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type line analog," DS dissertation, Dept. of Speech Communication and Music Acoustica, Royal Institute of Technology, Stockholm, Sweden.

Lindblom, B. (**1996**). "Role of articulation in speech perception: Clues from production," J. Acoust. Soc. Am. **99**, 1683–1692.

The Mathworks (**2007**). "MATLAB, Version 7.4.0.287," R2007a.

Meyer, P., Wilhelms, R., and Strube, H. W. (**1989**). "A quasiarticulatory speech synthesizer for German language running in real time," J. Acoust. Soc. Am. **86**, 523–539.

Mokhtari, P., Kitamura, T., Takemoto, H., and Honda, K. (**2006**). "Principal components of vocal tract area functions and inversion of vowels by linear regression of cepstrum coefficients," J. Phonetics **35**, 20–39.

Narayanan, S. S., Alwan, A. A., and Haker, K. (**1995**). "An articulatory study of fricative consonants using magnetic resonance imaging," J. Acoust. Soc. Am. **98**, 1325–1347.

Narayanan, S. S., Alwan, A. A., and Haker, K. (**1997**). "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. I. The laterals," J. Acoust. Soc. Am. **101**, 1064–1077.

Narayanan, S. S., Alwan, A. A., and Song, Y. (**1997**). "New results in vowel production: MRI, EPG, and acoustic data," Proceedings of the 1997 European Speech Proceedings Conference, Rhodes, Greece, Vol. **2**, pp. 1007–1009.

Schroeder, M. R. (**1967**). "Determination of the geometry of the human vocal tract by acoustic measurements," J. Acoust. Soc. Am. **41**, 1002–1010.

Sondhi, M. M., and Schroeter, J. (**1987**). "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-35**, 955–967.

Story, B. H. (**1995**). "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, University of Iowa, Ames, IA.

Story, B. H. (**2005a**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am., **117**, 3231–3254.

Story, B. H. (**2005b**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.

Story, B. H. (**2006**).

Story, B. H., Laukkanen, A. M., and Titze, I. R. (**2000**). "Acoustic impedance of an artificially lengthened and constricted vocal tract," J. Voice **14**, 455–469.

Story, B. H., and Titze, I. R. (**1998**). "Parameterization of vocal tract area functions by empirical orthogonal modes," J. Phonetics **26**, 223–260.

Story, B. H., and Titze, I. R. (**2002**). "A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function," J. Phonetics **30**, 485–509.

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.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1998**). "Vocal tract area functions for an adult female speaker based on volumetric imaging," J. Acoust. Soc. Am. **104**, 471–487.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**2001**). "The relationship of vocal tract shape to three voice qualities," J. Acoust. Soc. Am. **109**, 1651–1667.

Takemoto, H., Honda, K., Masaki, S., Shimada, Y., and Fujimoto, I. (**2006**). "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," J. Acoust. Soc. Am. **119**, 1037–1049.

Titze, I. R., Horii, Y., and Scherer, R. C. (**1987**). "Some technical considerations in voice perturbation measurements," J. Speech Hear. Res. **30**, 252–260.

West, K. S., and McNamara, J. A. (**1999**). "Changes in the craniofacial complex from adolescence to midadulthood: A cephalometric study," Am. J. Orthod. Dentofacial Orthop. **115**, 521–532.

Yang, C. S., and Kasuya, H. (**1994**). "Accurate measurement of vocal tract shapes from magnetic resonance images of child, female, and male subjects," Proceedings of ICSLP 94, Yokohama, Japan, pp. 623–626.

Yehia, H. C., Takeda, K., and Itakura, F. (**1996**). "An acoustically oriented vocal-tract model," IEICE Trans. Inf. Syst. **E79-D**, 1198–1208.

Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in Swedish vowels: Statistical considerations

Michel T.-T. Jackson[a] and Richard S. McGowan
CReSS LLC, 1 Seaborn Place, Lexington, Massachusetts 02420

(Received 28 June 2007; revised 1 November 2007; accepted 1 November 2007)

In a re-analysis of x rays of speakers producing Swedish vowels, midsagittal pharyngeal dimensions were predicted from anterior tongue positions using procedures based on estimated tongue pellet positions. Principal component analysis was used to reduce the number of pellet degrees of freedom from eight to three prior to applying linear regression from these three independent variables to dependent vocal tract midsagittal cross distances. Except for the regions around the laryngopharynx and uvula, the pharynx dimensions were predictable from linear regressions and were significant at the $p<0.05$ level. Numerical experiments show that it is crucial to reduce the number of independent variables in tests of statistical significance. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816579]

PACS number(s): 43.70.Bk, 43.70.Kv [BHS] Pages: 336–346

I. INTRODUCTION

One goal of phonetics is to quantitatively understand articulation in human speech production. As part of this goal, the authors are interested in the extent to which some vocal tract dimensions—specifically, midsagittal pharyngeal cross distances—are predictable from more readily observed articulatory measures. The predictability of midsagittal pharyngeal cross distances is of interest for two reasons.

First, a number of modern articulatory research techniques yield high-quality data on the position and movement of the anterior articulators (including the anterior portion of the tongue, the lips, and the upper and lower incisors) during speech, but cannot be extended to the posterior portion of the oral cavity and pharyngeal articulation. These techniques include x-ray microbeam pellet tracking (e.g., Kiritani, 1986; Westbury, 1994), electromagnetic midsagittal articulometry (e.g., Perkell *et al.*, 1992), and optical tracking (e.g., Ramsay *et al.*, 1996). To the extent that pharyngeal dimensions can be predicted or estimated from the positions of the other articulators, including the anterior of the tongue, a "missing" portion of the articulation can be recovered from these data sets. For instance, Kaburagi and Honda (1994) showed that much of the anterior portion of the midsagittal tongue shape, including regions between coils and just posterior to the rearmost coil, can be predicted from electromagnetic midsagittal articulometry data for one speaker of Japanese CVCs, where the Cs had labial, alveolar, and velar places of articulation.

Second, the degree to which pharyngeal dimensions are predictable from other articulatory information is important in attempting to recover articulatory positions from speech acoustics. There are some instances in which the position of a few points on the tongue has been shown to be highly correlated with the acoustic output or other articulatory variables. For instance, Nguyen *et al.* (1994) showed that coordinates of three points on the tongue predict the acoustic output of the vocal tract in [s] and [ʃ] in French, and Lindau-Webb and Ladefoged (1989) showed that as few as two fleshpoint positions suffice to determine the entire tongue shape in static American English vowels. Whalen *et al.*, (1999), using the coordinates of four manually estimated fleshpoints from magnetic resonance (MR) images of 11 American English vowels, showed high correlations between these eight variables (x and y coordinates of each fleshpoint) and midsagittal vocal tract cross distances measured along gridlines.

However, there are two reasons to treat these results with caution: First, there are languages in which pharyngeal dimensions—and therefore the acoustic output of the vocal tract—cannot be predicted from anterior tongue position. Tiede (1996) showed that the phonological advanced tongue root (ATR) contrast introduces an articulatory degree of freedom in the pharynx that makes the pharyngeal dimensions vary independently of the anterior portion of the tongue's position. Tiede (1996) also showed that this pattern of variation is different from the one observed, as for example, in American English. Thus, the results presented above have limited interpretive power with regard to phonological considerations. In particular, it would not be expected that pharyngeal cross distance can be predicted from anterior tongue position for languages with ATR contrast.

Second, there are several statistical reasons for treating these results with caution. In these studies, the number of observations on which t tests and F tests are based is small. But most statistical tests, including t and F tests, are based on a theory that requires large numbers of observations. Furthermore, many statisticians (see, for example, Hill and Lewicki, 2006). recommend that multiple regressions have more than ten times as many observations of the dependent variables as there are independent variables. Having a number of independent variables too close to the number of ob-

[a]Author to whom correspondence should be addressed. Electronic mail: ladmttj@ix.netcom.com

 0001-4966/2008/123(1)/336/11/$23.00

servations in a regression affects the numerical stability of the regression, may lead to overfitting, and increases the likelihood of capitalizing on chance in finding a high correlation in the results.

Similarly, parametric statistical tests, again including *t* tests and *F* tests, depend on some assumption of normality in the data sets. By definition, the normal distribution extends to positive and negative infinity. Since the midsagittal cross distance has a minimum possible value of zero and a maximum around 30 or 40 mm, the distribution of the dependent variable in the regressions discussed above *cannot* be normal. Common transformations, such as converting to *z* scores, do not improve the normality of the distribution of these data as measured by skewness, kurtosis, and higher-moment measures of a distribution's normality (Cohen *et al.*, 2003). Thus, the significance of correlations reported in some speech production studies is open to debate due to the small number of observations and the non-normality of the data. For instance, studies such as Jackson (1988) with 16 observations (vowels) per speaker, and Whalen *et al.* (1999) with 11 observations, typically report high correlations, over 0.95. However, studies using more data, such as Beaudoin and McGowan (2000) with over 3000 observations per speaker, report correlations of around 0.9.

Instead of relying solely on *t* and *F* tests, a series of numerical experiments was designed to examine the likelihood of finding our observed results under suitable null hypotheses. These numerical experiments calculate distributions of *R* in multiple regressions predicting midsagittal pharyngeal cross distance for various numbers of uniformly distributed, randomly generated independent variables. *R* is the correlation between a dependent variable and an optimally weighted combination of two or more independent variables (Cohen *et al.* 2003). These distributions can be used to determine significance levels for various values of *R*.

Finally, caution is necessary because the measured or estimated fleshpoint positions, which are the independent variables, are strongly correlated with each other. This produces multicollinearity, which is a well-known problem in multiple regression. Multicollinearity is the situation in which two or more independent variables are strongly correlated with each other. In this situation, regression coefficients are not reliable (Cohen *et al.* 2003).

These statistical concerns were addressed using principal components regression (Cohen *et al.* 2003, p. 428), which uses principal component analysis (PCA) to produce a reduced number of orthogonal (uncorrelated) independent variables from the original independent variables (pellet coordinates). By orthogonalizing the independent variables, multicollinearity is eliminated. Further, reducing the number of independent variables, relative to the number of observations, increases the power of statistical tests. Multiple regression using PCA components as independent variables is known as principal components regression. Previous studies (Shirai and Honda, 1978; Sekimoto *et al.*, 1978; Maeda, 1978, 1990; Beaudoin and McGowan, 2000; Story, 2007) have also used PCA and combinations of PCA with other forms of factor analysis (Hoole, 1999) to study midsagittal tongue shape, particularly in vowel production. Shirai and Honda (1978) used 50 vowel tokens from ten different speakers and Sekimoto *et al.* (1978) used over 1000 vowel tokens from two different speakers. The other studies segregated the observations by speaker, with Maeda (1978) using 77 observations from frames of an x-ray movie, Maeda (1990) using about 1000 observations of vowel production in sentences per speaker, Beaudoin and McGowan (2000) using over 3000 observations per speaker including consonant-vowel transitions, and Story (2007) using data from the centers of 11 vowels and the frames from six vowel pairs per speaker. In his hybrid approach, Hoole (1999) appears to have used 90 observations per speaker. These studies have shown that the number of degrees of freedom for the midsagittal tongue can be reduced to two or three, accounting for between 80% and 97% of the variance in tongue fleshpoint coordinates.

An overview of the current work is as follows. In order to investigate the predictability of midsagittal pharyngeal dimensions, midsagittal vocal tract cross distances were measured along gridlines and fleshpoint positions were estimated on the anterior of the tongue from x-ray tracings. The x-ray tracings come from previous studies (Sundberg 1969, Fant 1965). Principal components regression was used to predict the dependent midsagittal vocal tract cross distances. Numerical experiments were used to produce distributions of *R* under suitable null hypotheses. The *R* values from the regressions were compared with these distributions and were examined in order to determine significance levels.

II. PROCEDURE

A. Image acquisition, processing, and coordinate system

The data are tracings of static x-ray images of spoken vowels produced by four speakers of Swedish. The first three speakers are described in Sundberg (1969); photocopies of the original x-ray tracings were re-analyzed in this study. The fourth speaker is described in Fant (1965); photocopies of the x-ray tracing figures as prepared for publication were re-analyzed. The vowel tracings analyzed in this study are summarized in Table I. Each tracing was digitized on a flatbed scanner at a resolution of 300 dpi to produce a digital image that was subsequently contrast enhanced using the MATLAB Image Processing Toolkit.

A notable feature of the tracings of the first three speakers (from Sundberg, 1969) is that they include registration marks that provide accurate image alignment. For each speaker, tracings were translated, rotated, and scaled to bring the registration marks into coincidence using the MATLAB Image Processing Toolkit. The degree of scaling was manually checked by examining the coefficients of the MATLAB image transformation matrix. In all cases, the degree of scaling was less than 5%, indicating that the image acquisition chain, including photocopying original tracings, scanning and digitization, did not introduce notable image distortions. The accuracy of image alignment was verified by checking that the tracings of the static hard structures of the vocal tract, especially the upper incisors and hard palate, overlapped. The first two rows of Table II summarize the results

TABLE I. Vowels analyzed in this study. "√" indicate that the vowel was produced by the given speaker; "n/a" indicates that it was not; a phonetic transcription indicates that a variant with different phonemic length (e.g., [ø] rather than [ø:]) was produced.

	[i:]	[e:]	[æ:]	[y:]	[ø:]	[œ:]	[ʉ:]	[u:]	[o:]	[ɑ:]	[ɵ]	[a]	[ɔ]
BE	√	√	√	√	[ø]	n/a	√	√	√	√	n/a	n/a	n/a
JS	√	√	n/a	√	√	n/a	√	√	n/a	√	n/a	n/a	n/a
RL	√	√	√	√	√	n/a	√	√	√	√	n/a	n/a	n/a
F	√	√	√	√	√	√	[ʉ]	√	√	√	√	√	√

of comparing the positions of the tips of the upper incisors outline in these images. Since the checkpoint, the tips of the upper incisors, was a considerable distance—approximately 100 mm—from the registration marks that were used for alignment, this was considered a stringent check on the accuracy of the overall image alignment.

After satisfactory image alignment was obtained, an absolute scale in pixels per cm was obtained for the first three speakers from the original tracings″ measurements of the length of the midline of the vocal tract. The mean scale of all the images for the first three speakers was 4.6 pixels/mm; the standard deviations were all reported less than 0.6 pixels/mm. The small standard deviations reflect the accuracy of the original tracings and measurements.

For the fourth speaker (Fant, 1965) the tracings do not include registration marks. The scanned images of these tracings were translated to bring the tip of the upper incisor on each image into coincidence. Then, between five and ten points on the superior surface of the maxilla on each image were selected, and the best-fit line through these points (equivalent to the major axis of the distribution of the points) was determined. Each image was then rotated to bring this line, which approximated the upper surface of the maxilla, to horizontal. The images were not scaled relative to one another; the scale was stipulated to be 4.5 pixels/mm to make the measurements on these images comparable to the measurements on the images from the first three speakers. Since there were no lengths measured on the original tracings an absolute scale could not be determined, but all the images were scaled so as to make the range of measurements comparable with the range of measurements found among the other speakers.

From each image, more than 300 outline points on the upper and lower lip surfaces, hard palate and upper incisor outline, soft palate outline, tongue, laryngeal structures, and dorsal wall of the pharynx were selected from digitized images. Points on the outlines of the other upper teeth, especially the molars, were also selected where available.

Adopting Westbury's (1994) coordinate system, the maxillary occlusal plane (MaxOP) was estimated as the horizontal axis and the tip of the upper incisors as the origin. For the first three speakers (Sundberg, 1969), the MaxOP was determined from the x-ray tracing that had the most extensive tracing of the upper incisors and other upper teeth. The MaxOP was determined by finding the line through the tip of the upper incisors that was tangent to the outline of the other upper teeth. For the fourth speaker (Fant, 1965), because the outline of the upper teeth was not traced, the MaxOP was estimated by rotating the images to make the upper surface of the maxilla horizontal and passing the horizontal axis through the tip of the upper incisor.

TABLE II. Standard deviations (mm) in the *x* and *y* coordinates of reference points across images. P1 is on the alveolar ridge; P2 is the most superior point of the hard palate; P3 is on the dorsal wall of the pharynx at the level of the anterior tubercle of the atlas; P4 is on the dorsal wall of the pharynx at the level of the bottom of the vallecular sinus. Coordinates are relative to the tip of the upper incisor.

Speaker:	BE	JS	RL	F
Upper incisor-*x*	0.3	0.5	0.3	N/A
Upper incisor-*y*	0.3	0.5	0.3	N/A
P1-*x*	0.7	0.9	0.3	0.9
P1-*y*	0.4	0.5	0.1	0.8
P2-*x*	1.4	0.8	1.3	5.2
P2-*y*	0.7	1.0	0.7	0.8
P3-*x*	1.2	0.6	1.2	1.8
P3-*y*	1.5	2.8	1.3	1.2
P4-*x*	1.8	1.0	1.9	2.8
P4-*y*	3.3	4.3	1.6	3.5

B. Reference points and measurement grids

A measurement grid was constructed for each speaker. Each measurement grid was based on four fixed reference points, P1 through P4, similar to the ones used in Whalen *et al.* (1999). However, Whalen *et al.* (1999) analyzed MR images, whereas this study analyzes tracings of x-ray images, leading to differences in choosing anatomical landmarks.

Teeth do not image well in MR, because conventional MR imaging relies on the magnetic resonance of hydrogen nuclei, and the density of hydrogen atoms in the enamel and dentin of teeth is low. However, bone does image well and thus the bony alveolar ridge can be imaged. In x rays, both teeth and bone project shadows, and the shadow of the roots of the teeth may obscure the shadow of the surrounding bony alveolar socket and vice versa. In order to pick P1 to be on the alveolar ridge, an inflection point on each tracing at the juncture between the incisor and the alveolar socket in the maxilla was identified visually (see Fig. 1). The standard deviations of the coordinates of this point over all the vowels for each speaker are reported in rows three and four of Table II. The standard deviations are generally less than 1 mm. The mean position of these points for each speaker was used as reference point P1.

The most superior point of the hard palate, P2, is well

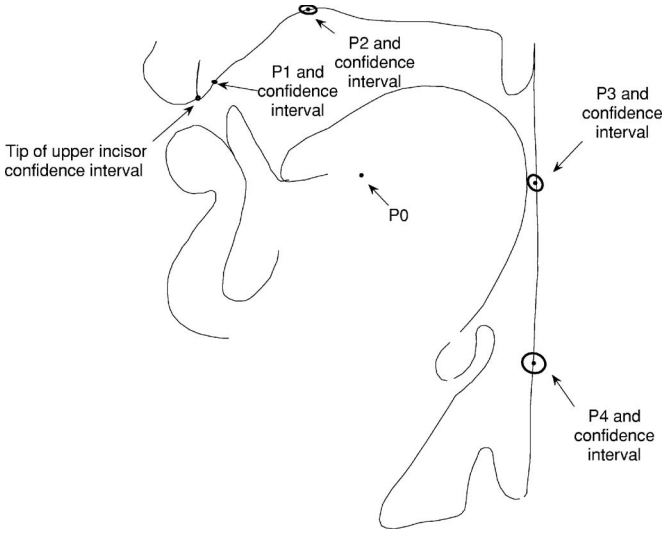


FIG. 1. Reference points for speaker BE. P1—mean position of the alveolar ridge reference point on all nine digitized vowel tracings; P2—mean position of the highest point of the palate; P3—mean position of the superior point on the rear pharyngeal wall; P4—mean position of the inferior point on the rear pharyngeal wall; P0—center of circular arc through P1, P2, and P3.

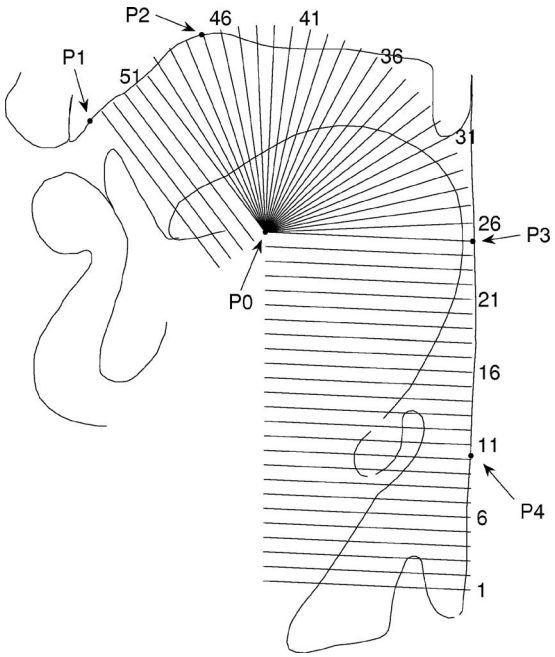


FIG. 2. Gridlines for speaker BE constructed based on the reference points P1 through P4.

defined because of the definition of vertical in the coordinate system based on the MaxOP. The highest point on the bicubic spline tracing the outline of the hard palate was determined automatically for each image. Although it is possible that the hard palate could contain a small region that appears to be horizontal, in fact none were found, and the highest point on the hard palate was unique on each tracing. The standard deviations of the coordinates of this point over all the vowels for each speaker are reported in rows five and six of Table II. The mean position of these points for each speaker was used as reference point P2.

The definitions of P3 and P4 used landmarks that could be found in the x rays. P3 is the point on the rear pharyngeal wall that is at the same height as the most anterior point of the anterior tubercle of the atlas. A point corresponding to P3 was identified on each image for each speaker, and the mean position (within speaker) was used as the fixed location of P3 in subsequent analysis. Following Tiede (1996), who described an earlier version of the image-measurement methods used in Whalen *et al.* (1999), the point on the rear pharyngeal wall at the same height as the point at the bottom of the vallecular sinus was adopted to be P4. Table II reports the standard deviations of the x and y positions of P3 in rows seven and eight; P4 in rows nine and ten.

Figure 1 shows the results of applying these procedures to pick reference points for one speaker (BE). The figure shows the joint (x,y)-distribution of these points in all the vowels produced by this speaker by plotting the ellipses where the semimajor and semiminor axes are one standard deviation. The center of the circle passing through P1, P2, and P3 is denoted P0, which is an auxiliary point required to construct the set of gridlines for each speaker.

C. Gridlines

There are three groups of gridlines in this procedure: gridlines in the lower pharynx, which are parallel to each other; gridlines in the upper pharynx and posterior oral cavity, which are radii of the circle centered on P0, and gridlines in the anterior oral cavity, which are also parallel.

1. Lower pharynx

The gridlines of the lower pharynx were specified to be perpendicular to the P3–P4 line, 3 mm apart (Fig. 2). They are constructed so that the topmost gridline passed through P0 (i.e., it is a radius of the circle passing through P1, P2, and P3) and the lowest gridline was specified to be less than 3 mm above the position of the vocal folds in the vowel that had the highest vocal fold position. In this way, there was no missing data on any of the lower gridlines due to the vocal folds having risen above them. These gridlines are approximately perpendicular to the axis of the pharynx and measure the cross distance from the posterior surface of the tongue to the dorsal wall of the pharynx.

Where there appeared to be an acoustically relevant airspace posterior to the opening of the larynx (e.g., posterior to the arytenoid cartilages) in the x-ray tracing, the midsagittal cross distance of that airspace was included in the measured gridline cross distance, but the thickness of the tissue separating the larynx from the posterior airspace was excluded. Similarly, when the vallecular sinus, between the tongue and the epiglottis, was open, the width of the airspace between the tongue and the epiglottis was included but the thickness of the epiglottis itself was excluded.

2. Upper pharynx and posterior oral cavity

The second set of gridlines was constructed using the topmost of the previous gridlines as a starting point. From that gridline, which passes through P0, radial gridlines centered at P0 were constructed every 5° until the gridlines were within 5° of being perpendicular to the P1–P2 line segment,

i.e., until the gridlines (nearly) coincided with the perpendicular bisector of the P1–P2 line segment. These gridlines cover the region of the oral cavity—pharyngeal cavity junction. As in the lower pharynx, the width of any airspace posterior to the uvula was included in the cross distance measured along the gridline, but the thickness of the uvula itself along the gridline was excluded.

3. Anterior oral cavity

From the last radial gridline, which was within 5° of being perpendicular to the P1–P2 line, gridlines perpendicular to the P1–P2 line were placed every 3 mm as far as P1. These gridlines measure the cross distance from the anterior portion of the tongue to the anterior portion of the hard palate.

Because of variations in overall vocal tract length, a different number of gridlines was produced for each speaker. In addition, varying proportions resulted in a different number of gridlines in the pharynx and oral cavity for each speaker. The gridlines for the same speaker (BE) illustrated in Fig. 1 are shown in Fig. 2.

D. Estimated fleshpoint locations and principal components regression

The procedure in Nix *et al.* (1996) estimates the position (x and y coordinates) of 13 equally spaced points along each tongue curve from traced x-ray images similar to the ones analyzed in this paper. The result is 13 estimated fleshpoint locations, for a total of 26 coordinates. In the present work, their procedure was used to estimate the length of the tongue surface in the midsagittal plane, and then to estimate the location of four fleshpoints.

Numerical integration was used to estimate the length of the tongue surface from the bicubic spline interpolating the selected outline points described in Sec. II A (not points on gridlines) from tongue tip to the point at the bottom of the vallecular sinus. This distance was then divided by 13, effectively dividing the portion of the tongue between the tip and the root of the epiglottis into 13 sections. Estimated fleshpoint locations were then calculated at 1.5, 3.0, 4.5, and 6.0 times this section length. These values were taken in an attempt to roughly simulate typical fleshpoint locations on the anterior of the tongue. The x and y coordinates of each fleshpoint were recorded for each vowel and used as the raw independent variables for the data analysis.

PCA of the fleshpoint coordinates was performed for each subject. The PCA components that accounted for the largest amount of the variance were employed as independent variables in predicting all the midsagittal cross distances in a principal components regression. The principal components regressions reduced the number of independent variables from the eight Cartesian tongue pellet coordinates, as well as alleviating any multicollinearity. These regressions were performed using the SPSS statistical package.

One of the assumptions underlying regression models and significance tests is that the values of the independent variable in the regression are under experimental control and known precisely. However, in this study, the quantities used as independent variables—the x- and y-fleshpoint coordinates—are estimates and not known precisely.

There are two notable sources of errors in the estimation of fleshpoint coordinates. First, of course, there are possible errors in the original x-ray tracing and its duplication (photocopying, scanning, and digitization). These errors are intrinsic to the data and there seems to be no way of estimating their magnitude or effect without independent instrumental control. Another kind of intrinsic uncertainty in the x-ray tracings is overall scaling; however, the presence of registration marks in some of the original tracings and the degree of overlap of the hard structures of the vocal tract from tracing to tracing provides a relatively strong control on this kind of uncertainty.

A second kind of error would occur if the tongue has substantial nonuniform longitudinal dilation or contraction during vowel production. If, e.g., the surface of the blade of the tongue were to stretch while the surface of the tip of the tongue contracted, the true fleshpoint positions would change in a way that could not be estimated by our algorithm. However, we find no literature suggesting that stretching of the anterior tongue surface occurs in speech. This evidently remains a topic for further investigation.

Despite these possibilities, we believe that the error in the estimates of fleshpoint positions is small. This is mainly because the PCA has residual errors that are comparable to residual errors found in other studies.

E. Numerical experiments

The goal of these numerical experiments was to empirically determine the distribution of R given the measured dependent variables, the midsagittal cross distances along the gridlines, under the hypothesis that the values of the independent variables, the PCA components, are drawn from a uniform random distribution. This distribution of R is, in effect, the expected distribution of R under a null hypothesis of no underlying relationship between the independent and dependent variables. These distributions provide an empirically derived relation between the value of R and the probability of obtaining that value of R by chance, since with random values of the independent variables, there is no relation between the independent variables and the midsagittal cross distances. A particular interest was the effect of the number of independent variables on the statistical significance of the multiple regressions.

A computer program was written to perform this experiment by generating different numbers of uniformly distributed random variates to use as independent variables in a regression model for the midsagittal pharyngeal cross distances for each speaker. The program then calculated the fraction of the total variance accounted for by the regression model (i.e., the R^2), and from that, the R value that would have been obtained in multiple regression. This program performed regressions 10 000 times for each gridline, and sorted the resulting R values into 100 bins, that is, into percentile distributions. The statistical significance (i.e., the probability of incorrectly rejecting the null hypothesis) of R values can be determined from these distributions. For instance, if 5%

TABLE III. Numerical experiment parameters. Although a potential maximum of eight independent variables (an *x*- and a- and a *y*-coordinate for each of four estimated fleshpoints) could be simulated, the small number of observations (vowels) and / or numerical degeneracy limits the actual maximum number of independent variables that can be used in each numerical experiment.

Speaker:	BE	JS	RL	F
Number of vowels	10	7	9	13
Potential maximum number of independent variables	8	6	8	8
Final maximum number of independent variables	6	3	5	8

of the simulated regressions possess an R value greater than 0.92, then the null hypothesis can be rejected when $R > 0.92$ with a 5% probability that, in fact, the null hypothesis is true.

For each speaker, a range of numbers of independent variables was simulated. Simulating the x and y coordinates of the four estimated fleshpoints would yield eight independent variables. However, there cannot be more independent variables than one less the number of observations. Since some of the speakers in this study produced fewer than nine vowels (see Table I), eight independent variables could not be simulated for those speakers (see row two of Table III). Further, numerical degeneracy also arose in some cases when the number of independent variables was two or three fewer than number of vowels, further reducing the maximum number of independent variables that could be simulated. Finally, it was stipulated that the a significance level of $0.01 < p$ should be attained with $R < 0.99$. With these constraints, the potential maximum number of independent variables simulated for each speaker is shown in the third row of Table III. The minimum number of independent variables simulated for each speaker was three.

III. RESULTS

A. PCA components

The cumulative variance accounted for by the principal components is summarized in Table IV for each speaker. Table IV shows that the first three PCA components account for more than 99% of the variance in the estimated fleshpoint positions. The remaining components clearly account for negligible amounts of the variance. Therefore only the first three components derived by PCA were retained for the remaining analysis. With the small number of observations (vowels) per subject it would have been possible to proceed with two principal components accounting for at least 97% of the variance. However, in other circumstances with small numbers of observations, three components would need to be retained. This could occur if a set of observations contains consonant-vowel transitions and rhoticized articulations (cf. Hoole, 1999). It will be shown that even with a small number of observations (between 7 and 13), a reduction to three PCA components is sufficient for statistically meaningful results.

The "shapes" of the first three PCA components, in terms of the pellet displacements from mean position associated with variation in each component, are shown in Fig. 3 for the same speaker (BE) as illustrated in Fig. 1. The component shapes are plotted together with the mean position averaged across all vowel tracings. Although we are not focusing on the interpretations of these components in this study, the first component resembles the "front raising" component of Harshman *et al.* (1977). This has been found to be true in a number of other studies, e.g., Shirai and Honda (1978), Sekimoto *et al.* (1978), Maeda (1978), Zerling (1979), Jackson (1988), and Beaudoin and McGowan (2000). The second and third components are more variable. However, there is evidence of a "back raising" component and either a "bunching" or a "tongue blade raising" component.

B. Principal component regressions predicting midsagittal vocal tract cross distances

Multiple regression models for each speaker were constructed with the first three PCA components as independent variables and midsagittal vocal tract cross distances measured along the gridlines as dependent variables using SPSS. The model sum of squares, error sum of squares, R values, and the R threshold corresponding to the $p < 0.05$ significance level according to the F test for each gridline are plotted in Figs. 4–7 for the four subjects. In general the R values are over 0.9 except in the regions of the laryngopharynx and uvula. In particular, the midsagittal pharyngeal cross distances that are the focus of this study generally exhibit high values of R that are significant according to an F test.

The gridline construction procedure distributes different numbers of gridlines in different parts of the vocal tract for each speaker (see Figs. 4–7). For simplicity, Table V reports the mean R for the gridlines listed as being in the pharyngeal region of each speaker in Figs. 4–7. All of these values are over 0.9. The predictability of the vocal tract cross distances measured along the gridlines in the oral cavity is extremely high, which is not surprising since the estimated fleshpoints are also in the oral cavity.

Closer examination of the results reveals other trends. First, the variance of the midsagittal cross distance, as mea-

TABLE IV. Cumulative variance accounted for by PCA components of eight pseudopellet variables (mm^2; %)

Speaker:	BE	JS	RL	F
Total variance	402.56 (100)	218.40 (100)	143.91 (100)	214.51 (100)
Component 1	365.17 (90.71)	186.96 (85.60)	104.78 (72.81)	162.58 (75.79)
Components 1–2	395.93 (98.35)	213.46 (97.73)	139.60 (97.01)	209.38 (97.61)
Components 1–3	400.6 (99.51)	216.99 (99.35)	143.17 (99.49)	213.19 (99.39)
Residual components 4–8	1.96 (0.49)	1.41 (0.65)	0.74 (0.51)	1.32 (0.61)

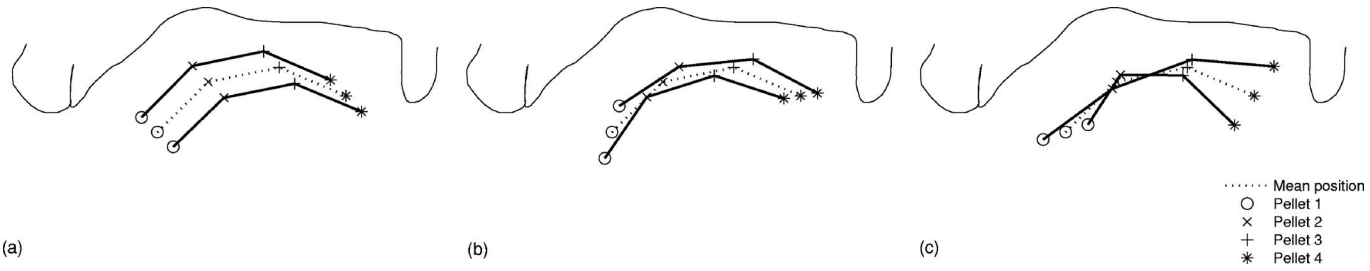


FIG. 3. PCA component shapes for speaker BE. Pellet displacements from mean position associated with variation in each component: (a) first principal component; (b) second principal component; (c) third principal component.

sured along the gridlines, is lowest in the laryngopharyngeal and uvular regions for all four speakers. These are also areas in which the F test suggests the regression model to be nonsignificant. Another trend worth noting is that even when R is fairly high, the regression model can be nonsignificant according to the relevant F test. For example, Fig. 5 shows that gridline 9 for speaker JS has an R over 0.92 but nonetheless, fails to achieve significance at the $p<0.05$ level.

C. Numerical experiments

Figure 8 plots the distribution of R resulting from the numerical experiments for each gridline together with the $p<0.05$ (95th percentile) and $p<0.01$ (99th percentile) levels, and the thresholds for the same significance levels according to F tests (calculated according to Hays, 1981). For speaker BE (Fig. 8(a)) with six independent variables, the $p<0.05$ significance level according to the F test is at $R\approx0.97$. But the empirically derived 95th percentile of the R distribution—the dark gray band in the figure—requires a larger $R(\approx0.98)$. Even for the speaker with the most vowels (observations), F, the 95th and 99th percentile levels derived from the numerical experiments require a larger R than the $p<0.05$ and $p<0.01$ significance levels according to the F test. Figure 8(d) shows the results with eight independent variables. The $p<0.05$ significance level according to the F test is at $R\approx0.96$; the empirically derived 95th percentile of the R distribution is ≈0.98. Similarly, the $p<0.01$ significance level according to the F test is $R\approx0.985$, but the empirically derived 99th percentile level is $R\approx0.99$. Similar results obtain for speakers JS (Fig. 8(b)) and RL (Fig. 8(c)).

On the other hand, when the number of independent variables is limited to only three, the empirically derived 95th and 99th percentile levels differ from the thresholds according to F test by less than 0.001. As the number of independent variables decreases, the empirical thresholds

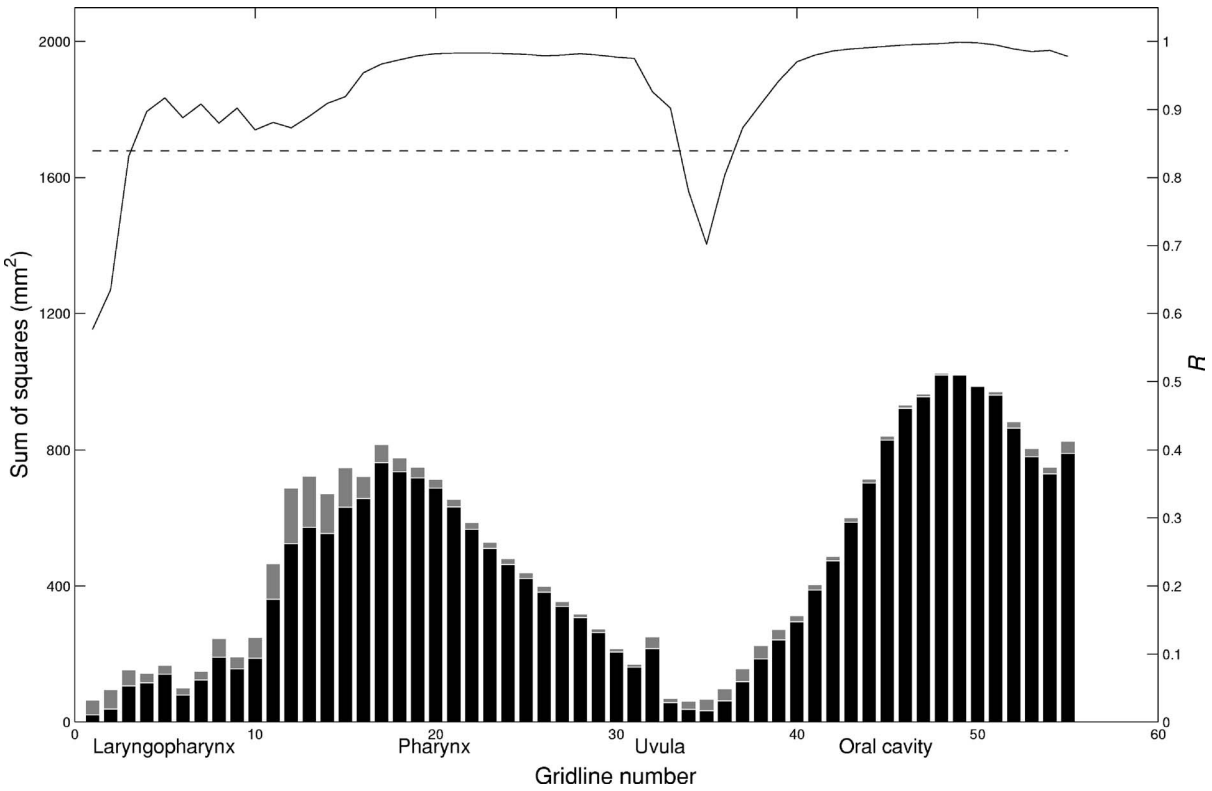


FIG. 4. Regression results for speaker BE. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R; dashed line shows the level at which R is significant at the $p<0.05$ level according to the F test.

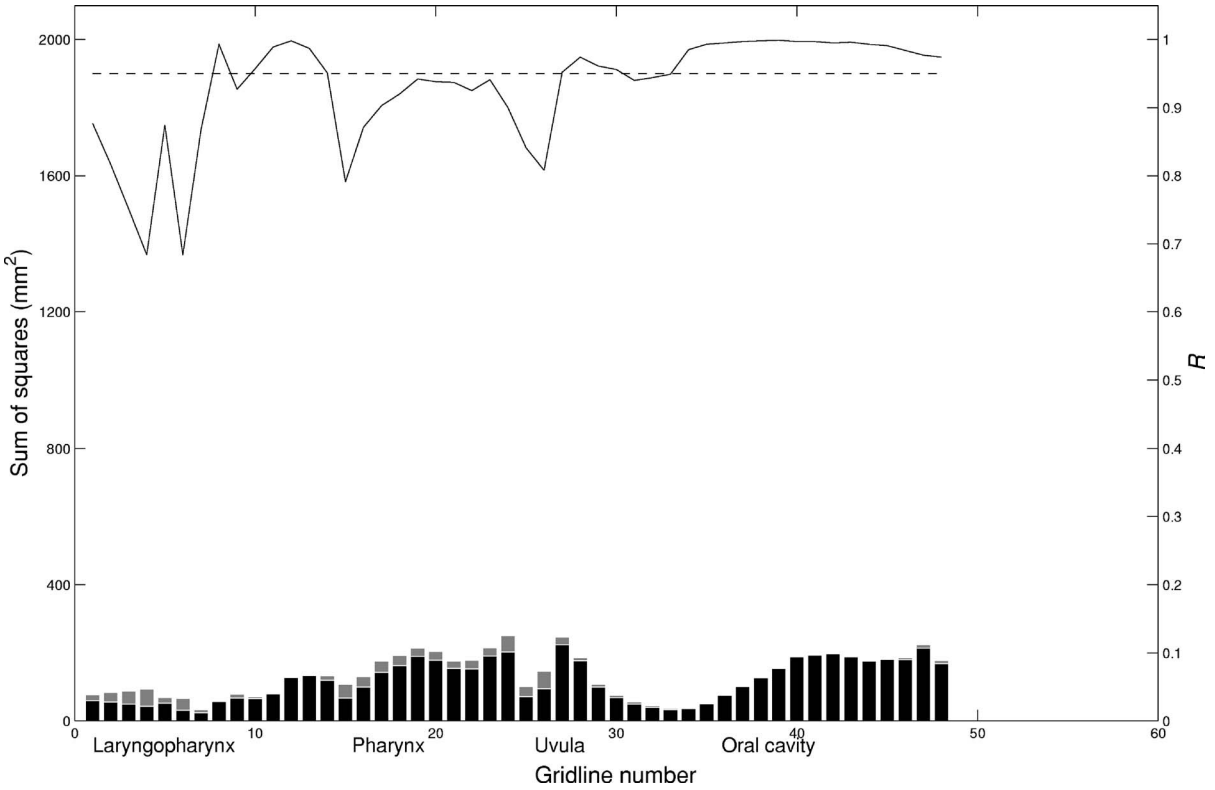


FIG. 5. Regression results for speaker JS. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R; dashed line shows the level at which R is significant at the $p<0.05$ level according to the F test.

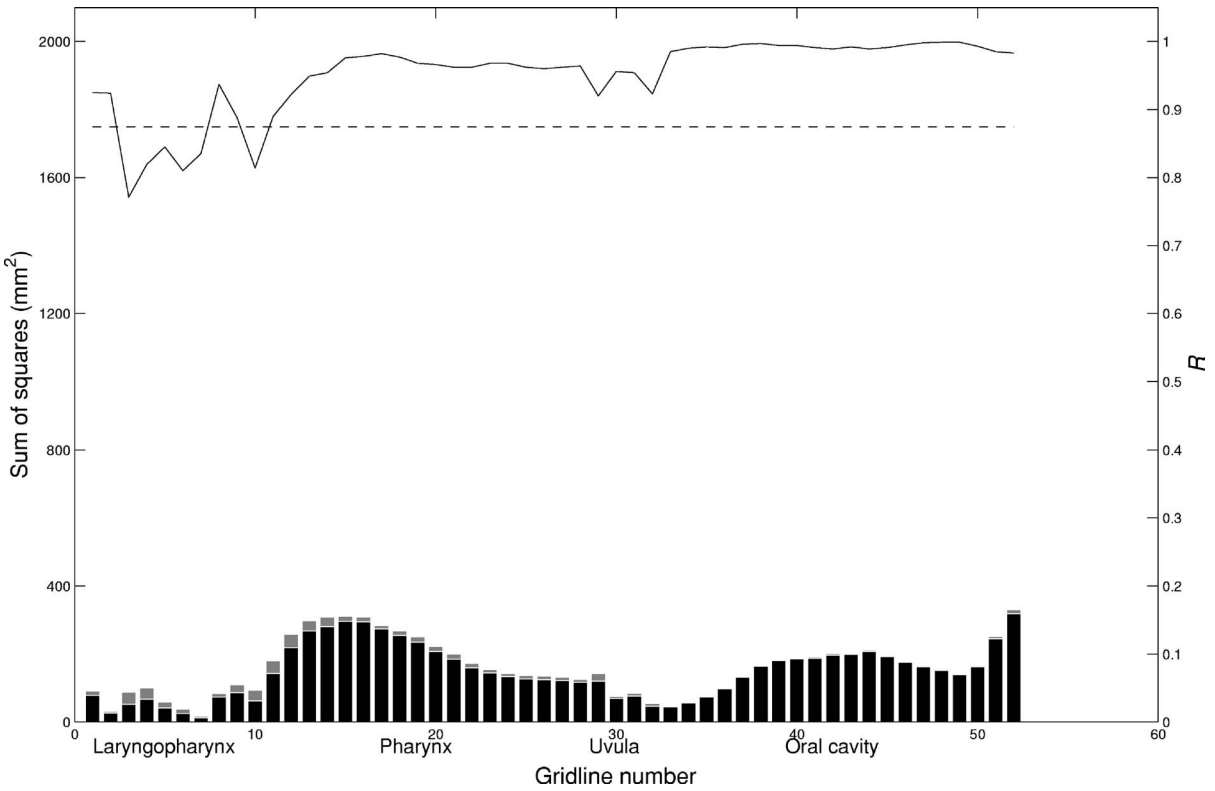


FIG. 6. Regression results for speaker RL. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R; dashed line shows the level at which R is significant at the $p<0.05$ level according to the F test.

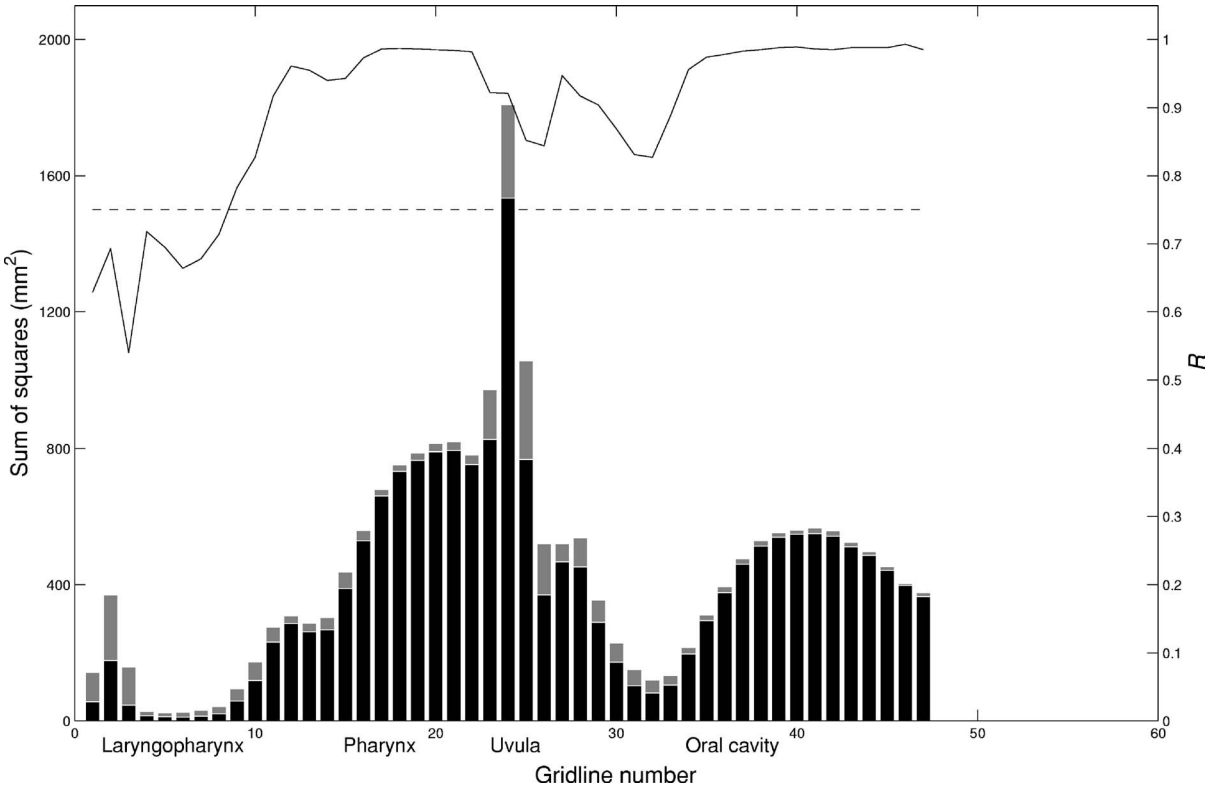


FIG. 7. Regression results for speaker F. For each gridline, black bars show the model sum of squares for the regression; gray bars show the error sum of squares; the total sum of squares is shown by the total height of the black and gray bars; solid line shows the regression R; dashed line shows the level at which R is significant at the $p<0.05$ level according to the F test.

and the F-test thresholds converge. Thus, in Figs. 4–7, the dashed lines denoting the R level for $p<0.05$ significance remain virtually unchanged as a result of the numerical experiments.

IV. DISCUSSION

This study confirms that three, or even two, degrees of freedom can describe anterior tongue position during vowel production in Swedish. Three principal components suffice to describe more than 99%, and two principal components more than 97%, of the variance in the anterior tongue position (as measured by pellet positions) in these vowels for each speaker. In addition, it has been shown that these measures of the anterior tongue position can be sufficient to describe tongue position in the pharynx, excluding the regions of the laryngopharynx and uvula. At least for the Swedish vowels examined here, restricting the data to anterior tongue pellet positions would not lose any essential part of the pattern of variation of tongue position in the mid-pharynx; there is no component or "gesture" that affects the posterior of the tongue to the exclusion of the anterior. This is shown by the fact that the cross distance of the posterior portion of the tongue from the rear pharyngeal wall can be predicted from the anterior pellet positions, via principal components analysis and regression equations.

TABLE V. Mean R for pharyngeal region gridlines. The pharyngeal regions include gridlines 11 through 30 for speaker BE; 11–24 for speaker JS; 11–28 for RL; and 12–24 for F. (See Figs. 4–7 for comparison.)

Speaker:	BE	JS	RL	F
R in pharyngeal region	0.96	0.93	0.96	0.97

This study does not attempt to formulate the components of tongue position exactly in Swedish vowel production. We do not know, in fact, whether or not different speakers exploit similar or different articulatory possibilities to produce roughly similar acoustic output. Furthermore, in languages which have pharyngealized or ATR vowels (cf. Jacobsen 1978; Lindau 1979; Tiede 1996), there will be different degrees of freedom or components, but there seem likely to be at least three. In consonants there may also be additional degrees of freedom (cf. Zerling 1979; Hoole 1999).

The results of the numerical experiments show that it is somewhat more difficult to demonstrate significance with this data than it might appear. First of all, it is not possible to construct numerically valid regression models with eight independent variables for all speakers. Second, the numerical experiments with the maximum number of independent variables result in thresholds for the $p<0.05$ significance level ranging from $R>0.96$ to $R>0.98$. But the threshold R value according to the F test can be calculated, and in each case, the F test gives a more liberal, and potentially erroneous, threshold for significance with these data than the numerical experiment. Multiple regression models relating four pellet positions with eight independent variables to midsagittal pharyngeal cross distances may be valid, but simple F tests are not appropriate for determining significance levels because the data do not obey the assumption of multivariate normal distribution, and the number of observations is small.

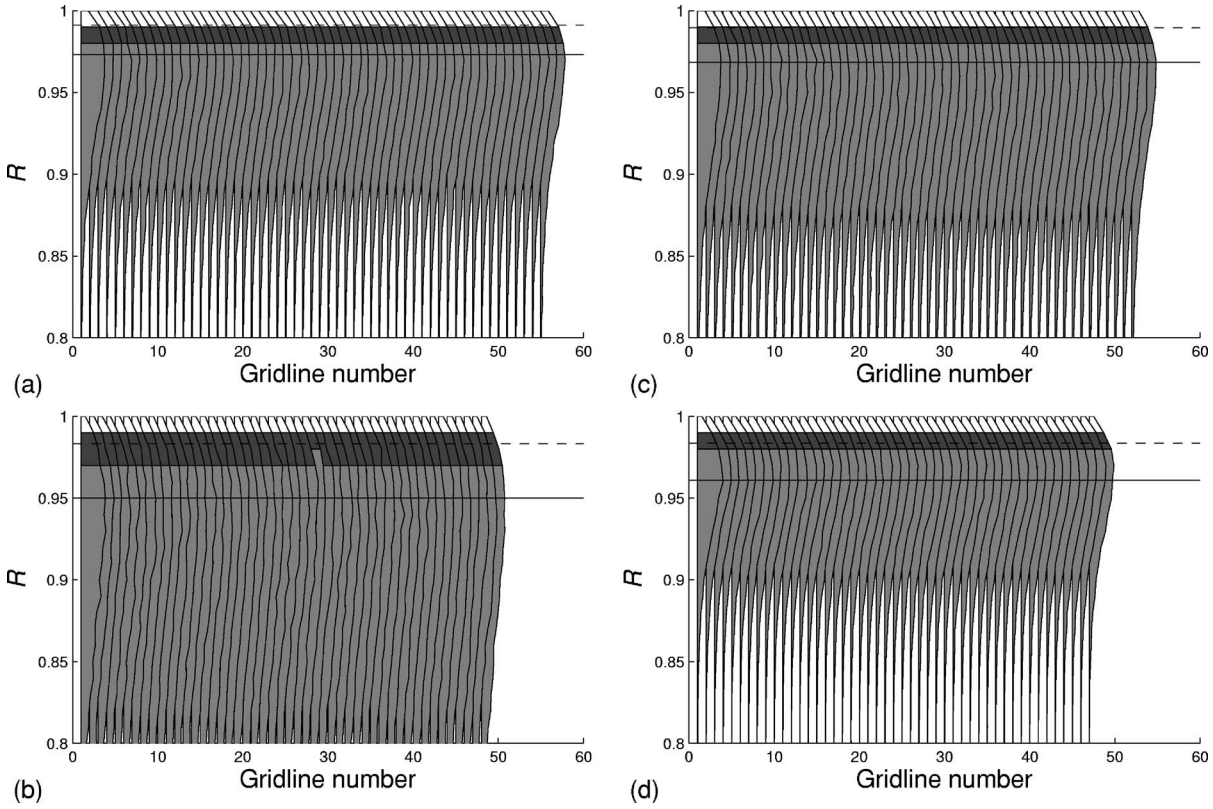


FIG. 8. The distribution of R from multiple regressions with the final maximum number random, uniformly distributed independent variables and midsagittal cross distances at each gridline as dependent variables. The light gray region of each distribution shows the fraction of regressions below the 95th percentile of each gridline's R distribution, the dark gray shows fraction between the 99th and 99th percentiles, and the white shows the regressions above the 99th percentile. The solid horizontal line denotes the minimum R value that would need to be attained for $p<0.05$ according to an F test, and the dashed line the minimum R values that would need to be attained for $p<0.01$ according to an F test. Subject BE is shown in (a); subject JS in (b); subject RL in (c); and subject F in (d).

For multiple regression models with only three independent variables derived via PCA, the values of R accorded significance by numerical experiment are virtually the same as the values of R judged significant by the F tests. Thus, the approach to determining significance levels by numerical experiment is more conservative than the F test, but they converge as the ratio of number of observations-to-independent variables increases.

This study shows that it is methodologically difficult to justify using raw x and y coordinates from fleshpoint positions (x-ray microbeam pellets, electromagnetic midsagittal articulometry sensor coils) as independent variables in statistical analysis of static vowels. The problems induced by the high correlations among the independent variables and the small number of observations relative to the number of independent variables make the true threshold for significant results substantially higher than routine application of statistical methods suggests. Therefore it is imperative to reduce the number of independent variables in this kind of multiple regression, and to use numerical experiments to test statistical significance.

This work shows that a large portion of the posterior position of the tongue can be deduced—quantitatively—from the anterior position in these vowels with a high degree of statistical confidence. This in turn implies that if the position of the anterior portion of the tongue can be deduced from its acoustic output, then much of the entire tongue position is, in turn, recoverable from the acoustic description of a vowel, via pellet positions on the anterior portion of the tongue. In fact, the recent work of Story (2007) has shown that two PCA factors of cross distance in the oral cavity can be predicted from the first two formant frequencies. Because the formant frequencies are shaped by the pharyngeal, as well as oral, cavity, Story's work in English indirectly corroborates the findings here. This opens the way to both a compact representation of the vowel (anterior tongue position only) as a basis for articulatory speech synthesis, and to the inverse task, namely reconstruction of the articulatory configuration in a given vowel from its acoustic representation.

For speech production research in general it should be borne in mind that statistical significance cannot always be quantified using parametric statistics. There can be a number of reasons for this, including a small number of observations with respect to number of independent variables (e.g., Whalen *et al.*, 1999), high variability in the data, as is often the case with children (e.g., McGowan *et al.*, 2004), the complexity of algorithms, such as PARAFAC (parallel factors, e.g., Jackson, 1988), or because the data are not normally distributed. In these cases numerical experiments and/or nonparametric statistics provide a means for enabling us to make statistically meaningful statements.

ACKNOWLEDGMENTS

This work was supported by Grant No. NIDCD-001247 to CReSS LLC. The authors would also like to acknowledge the personal help of Dr. Sundberg and Dr. Fant, who supplied the x-ray tracings analyzed in this article.

Beaudoin, R. E., and McGowan, R. S. (**2000**). "Principal components analysis of x-ray microbeam data for articulatory recovery," in Proceedings of the Fifth Seminar on Speech Production: Models and Data, Kloster Seeon, Bavaria, Germany, pp. 225–228, Ludwig-Maximilians-Universität.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (**2003**). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. (Erlbaum, Mahwah, NJ), pp. 26, 69, 390, 391, and 428.

Fant, G. (**1965**). "Formants and cavities," E. Zwirner and W. Bethge, eds., in *Proceedings of the Fifth International Congress of Phonetic Sciences*, pp. 120–141.

Harshman, R., Ladefoged, P., and Goldstein, L. (**1977**). "Factor analysis of tongue shapes," J. Acoust. Soc. Am. **62**, 693–707.

Hays, W. L. (**1982**). *Statistics*, 3rd Ed. (CBS, New York), pp. 483–486.

Hill, T., and Lewicki, P. (**2006**). *Statistics: Methods and Applications* (StatSoft, Tulsa, OK), p. 346.

Hoole, P. (**1999**). "On the lingual organization of the German vowel system," J. Acoust. Soc. Am. **106**, 1020–1032.

Jackson, M. T. T. (**1988**). "Analysis of tongue positions: Language-specific and cross-linguistic models," J. Acoust. Soc. Am. **84**, 124–143.

Jacobsen, L. (**1978**). *DhoLuo Vowel Harmony: A Phonetic Investigation*, University of California, Los Angeles, Ph.D. dissertation.

Kaburagi, T., and Honda, M. (**1994**). "Determination of sagittal tongue shape from the positions of points on the tongue surface," J. Acoust. Soc. Am. **96**, 1356–1366.

Kiritani, S. (**1986**). "X-ray microbeam method for the measurement of articulatory dynamics: Techniques and results," Speech Commun. **5**, 119–140.

Lindau, M. (**1979**). "The feature expanded," J. Phonetics **7**, 163–176.

Lindau-Webb, M., and Ladefoged, P. (**1989**). "Methodological studies using an x-ray microbeam system," UCLA Working Papers in Phonetics **72**, 82–90.

Maeda, S. (**1978**). Une analyse statistique sur les positions de la langue: Etude préliminaire sur les voyelles françaises ("A statistical analysis of tongue positions: Preliminary study of French vowels"). *Actes des 9èmes Journées d'Études sur la Parole (Acts of the Ninth Speech Study Session)*, Lannion, France, 31 May — 2 June, 1978, pp. 191–200. Groupement d″Acousticiens de la Langue Française: Lannion.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in W. J. Hardcastle and A. Marchal, eds., *Speech Production and Speech Modeling* (Kluwer, Dordrecht), pp. 131–150.

McGowan, R. S., Nittrouer, S., and Manning, C. J. (**2004**). "Development of [ɹ] in young, Midwestern, American children," J. Acoust. Soc. Am. **115**, 871–884.

Nguyen, N., Hoole, P., and Marchal, A. (**1994**). "Regenerating the spectral shapes of [s] and [ʃ] from a limited set of articulatory parameters," J. Acoust. Soc. Am. **96**, 33–39.

Nix, D. A., Papçun, G., Hogden, J., and Zlokarnik, I. (**1996**). "Two cross-linguistic factors underlying tongue shape for vowels," J. Acoust. Soc. Am. **99**, 3707–3717.

Perkell, J. S., 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., Munhall, K. G., Gracco, V. L., and Ostry, D. J. (**1996**). "Functional data analyses of lip motion," J. Acoust. Soc. Am. **99**, 3718–3727.

Sekimoto, S., Imagawa, H., and Kiritani, S. (**1978**). "Dynamic characteristics of tongue movement in the production of connected vowels," Ann. Bull., Res. Inst. Logop. Phoniatr., Univ. Tokyo **12**, 11–20.

Shirai, K., and Honda, M. (**1978**). "Estimation of articulatory motion," M. Sawashima and F. S. Cooper, eds., in *Dynamic Aspects of Speech Production* (University of Tokyo Press, Tokyo).

Story, B. H. (**2007**). "Time-dependence of vocal tract modes during production of vowels and vowel sequences," J. Acoust. Soc. Am. **121**, 3770–3789.

Sundberg, J. (**1969**). "Articulatory differences between spoken and sung vowels in singers," Stockholm Royal Institute of Technology, Speech Transmission Laboratory/Quarterly Progress and Status Reports **1**, 33–46.

Tiede, M. K. (**1996**). "An MRI-based study of pharyngeal volume contrasts in Akan and English," J. Phonetics **24**, 399–421.

Westbury, J. R. (**1994**). "On coordinate systems and the representation of articulatory movements," J. Acoust. Soc. Am. **95**, 2271–2273.

Whalen, D. H., Kang, A. M., Magen, H. S., Fulbright, R. K., and Gore, J. C. (**1999**). "Predicting midsagittal pharynx shape from tongue position during vowel production," J. Speech Lang. Hear. Res. **42**, 592–603.

Zerling, J.-P. (**1979**). *Articulation et Coarticulation dans les Groupes Occlusive-Voyelle en Français (Articulation and Coarticulation in Occlusive-Vowel Groups in French)*. Thèse de Troisième Cycle, Université de Nancy, France.

Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography

Jan G. Švec[a)]
Groningen Voice Research Laboratory, Department of Biomedical Engineering, University Medical Center Groningen, University of Groningen, Antonius Deusinglaan 1, NL 9713 AV Groningen, the Netherlands and Department of Experimental Physics, Laboratory of Biophysics, Palacký University Olomouc, tř. Svobody 26, CZ 771 46 Olomouc, the Czech Republic

Johan Sundberg
Department of Speech, Music and Hearing, KTH, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

Stellan Hertegård
Department of Logopedics and Phoniatrics, Karolinska University Hospital Huddinge, SE-141 86 Stockholm, Sweden

(Received 17 April 2007; revised 14 September 2007; accepted 8 October 2007)

There has been a lack of objective data on the singing voice registers, particularly on the so called "whistle" register, occurring in the top part of the female pitch range, which is accessible only to some singers. This study offers unique strobolaryngoscopic and high-speed (7812.5 images/s) videokymographic data on the vocal fold behavior of an untrained female singer capable of producing three distinct voice qualities, i.e., the chest, head and whistle registers. The sound was documented spectrographically. The transition from chest to head register, accompanied by pitch jumps, occurred around tones B4–C#5 (500–550 Hz) and was found to be associated with a slight decrease in arytenoids adduction, resulting in decrease of the closed quotient. The register shifts from head to whistle, also accompanied by pitch jumps, occurred around tones E5-B5 (670–1000 Hz) without any noticeable changes in arytenoids adduction. Some evidence was found for the vocal tract influence on this transition. The mechanism of the vocal fold vibration in whistle register was found principally similar to that at lower registers: vibrations along the whole glottal length and vertical phase differences (indicated by sharp lateral peaks in videokymography) were seen on the vocal folds up to the highest tone G6 (1590 Hz). © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2804939]

PACS number(s): 43.70.Gr, 43.75.Rs [NHF] Pages: 347–353

I. INTRODUCTION

The pitch range of the human voice can be divided into vocal registers, denoting a series of tones of similar voice quality, distinct from a series of adjacent tones produced in other voice registers (Hollien, 1974). Vocal registers have, however, remained a confusing issue. Mörner *et al.* (1963) listed over 100 terms used for vocal registers.

It is generally agreed that there are at least three registers in adult nonsinger voices, (1) vocal fry/pulse register in the fundamental frequency (F_0) range from a few hertz up to about 80 Hz, (2) modal/chest register, typically used in speech, and (3) falsetto in the highest F_0 region (Hollien, 1974; Titze, 2000; Henrich, 2006). In singing voice, particularly in female classically trained singers, the registers used for the various F_0 ranges are less clear. While it is generally agreed that the chest register is used for the lowest F_0 range, up to about 300 Hz, it is unclear what register or registers are used for the higher F_0 range. Many voice experts claim that singers replace the nonsingers' falsetto register by a different register sometimes referred to as middle or head register (see e.g., Miller, 2000; Titze, 2000). In the top F_0 range, where F_0 exceeds about 700 Hz (or 1000 Hz or 1396 Hz, depending on the literature source), some authors argue that yet another register is used, referred to as the upper, high, whistle, flute or flageolet register (Walker, 1988; Miller and Schutte, 1993; Herzel and Reuter, 1997; Miller, 2000; Titze, 2000; Henrich,2006).

This confusion is largely due to lack of objective data. It is generally agreed that registers are associated with the glottal voice source, such that each register is produced with a specific set of vocal fold vibration characteristics. The most revealing techniques for documenting the vocal fold vibration characteristics in vivo are based on laryngoscopic imaging. However, it is often difficult or even impossible to document and describe such characteristics laryngoscopically in the extreme parts of the F_0 range since the laryngeal view of the glottis is often concealed. At high F_0, on the other hand, the commonly available imaging techniques were not fast enough to show vibration characteristics. Until recently the high-speed videolaryngoscopic cameras typically ran at a rate of 2000 frames/s, corresponding to two images per ms. For F_0=700 Hz this yielded no more than about three frames per period. As a result, the vocal fold vibration characteristics in the register used by classically trained female singers

[a)]author to whom correspondence should be addressed. Electronic mail: svecjan@vol.cz

in the top part of their F_0 range is largely unknown and the physiology of the whistle register is the most poorly understood of the vocal registers.

Two ambitious attempts have been made to describe the voice source in this register. One was carried out by Walker (1988), who analyzed seven female volunteer sopranos producing tones in the F_0 range 988 Hz (pitch B5) to 1568 Hz (pitch G6) both in what they considered whistle register and in nonwhistle register that they referred to as upper, head, loft, or falsetto register. These tones were arranged pairwise and a listening panel was asked to judge if the two tones were the same or different. Expert listeners were found to identify the registers. Spectrum analysis revealed that the head voice register tones had higher sound level, and a less dominating fundamental and in many singers the airflow rate was found to be higher in the head register.

The other investigation, carried out by Miller and Schutte (1993), analyzed the whistle register in two sopranos. Their aim was to "identify measurable characteristics that distinguish the production of the highest segment of the voice range from that of its lower neighbor... and then address the question of the appropriateness of designating them separate registers." They recorded electroglottographic (EGG) and audio signals, as well as sub- and supraglottal pressures picked up by small pressure transducers. They observed marked effects in both the EGG and the subglottal pressure signals. Also they estimated the formant frequencies and found that F_0 was higher than the first formant F_1 in the upper register and concluded that $F_0 > F_1$ produces a salient voice timbre effect which they reported sounding as a shift in register. They further assumed that the acoustic conditions emerging from the $F_0 > F_1$ situation were influencing the vocal fold vibrations and causing the effects observed in the EGG and in the subglottal pressure signals. Later, Titze (2004) showed theoretically that vocal tract impedance phenomena can have these effects on the vocal fold vibration.

Several other alternative physiological explanations of the whistle register have been proposed. According to Lullies (1953), in 1902, Schultz claimed that he had stroboscopic evidence that the glottis can produce "chink tones" (Spalttöne) in the F_0 range 775–2760 Hz, thus presumably an example of a whistle register (Schultz, 1902). He also reported that when such tones were produced the ventricular folds touched the vocal folds, which did not vibrate. Chink tone mechanism was claimed to be responsible for the whistle register also by van den Berg (1963) in his classic article on voice registers. Herzel and Reuter (1997) argued that it seemed unlikely that whistle register tones were chink tones, since tones produced by whistling are almost sinusoidal (Shadle, 1983) and tones sang in the whistle register tend to have stronger overtones. Therefore, they found it difficult to imagine that whistle register tones are produced by a whistling-like mechanism and hypothesized that vortex shedding in the glottis drives the folds to vibration (Berry *et al.*, 1996).

Videokymography (VKG), a novel method for laryngoscopic examination of vocal fold vibration, has substantially improved the possibilities to gain a clearer idea on registers in the very high F_0 range (Švec and Schutte, 1996; Švec, 2000). The method is based on a specially adapted charge coupled device black and white video camera and can work in two different modes—standard and high speed. In the standard CCIR/PAL mode, the camera works as a common commercial video camera with the image rate of 25 frames/s or 50 half-frames/s. In the high-speed mode it selects a single line from the whole image and monitors it at a rate of 7812.5 line images/s. The resolution is 768 pixels per image line (in PAL standard), which was, at the time of the experiment, considerably larger than the maximal image resolution of 256×256 pixels at the image rate of 2000 images/s in current high-speed digital imaging systems. The line images are arranged column-wise, composing a VKG image, which visualizes vibration of the selected part of the vocal folds. A mechanical foot switch allows instantaneous switching between the standard and high-speed modes. Both the standard and the high-speed VKG images can be visualized and recorded using a standard PAL/CCIR video equipment.

VKG can be used for revealing differences in vocal fold vibration patterns associated with vocal register changes. The aim of the present investigation was to explore the potentials of the VKG for finding out if the vocal fold vibratory pattern changes in a female voice, especially when F_0 approaches the top part of the range. In view of the great terminological confusion we decided to use the terms chest, head and whistle for the three registers analyzed in this investigation.

II. MATERIAL AND METHODS

In the data-gathering process for this study we have investigated five trained and one untrained female singers. All the trained singers found it impossible to produce the high pitches above 1000 Hz during laryngoscopy and exhibited no audible transition to whistle register. Consequently, they were found unsuitable for the purpose of the present study. The untrained female singer was therefore used as the subject for the present investigation. She reported no voice problems and was capable of singing over a very large pitch range in which the chest, head and whistle registers were separated by clearly audible, sudden transitions. The subject was examined laryngoscopically using videostroboscopy and VKG. She sang ascending scales of sustained tones and pitch glides starting from the lowest tones in chest register up to her highest tones possible. She was encouraged not to try to avoid pitch jumps. To allow laryngoscopic visualization of the vocal folds, the subject produced only the vowel /e/ during the strobovideolaryngoscopic and VKG investigations.

The VKG camera of Lambert Instruments was combined with a R. Wolf 300W Xenon Model 5131 continuous light source and with a rigid R. Wolf 90°, Model 4450.57 laryngoscope. The stroboscope was a R. Wolf, model 5052 and the video camera used for the stroboscopy was model 5512 of the same mark. The sound was recorded with an omnidirectional microphone (TCM110, AV-JEFE) which was placed on a head mount at the distance of approximately 10 cm from the mouth at the angle of about 45° to the side. In addition, an electroglottographic signal (Glottal Enterprises Twin EGG Model MC2-1) was synchronously recorded in the second audio channel but was not used for analysis. A

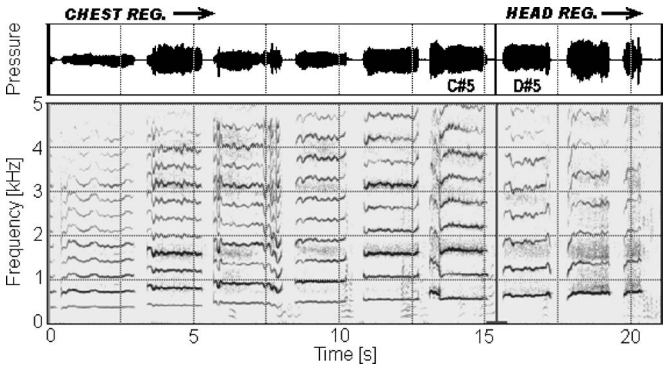


FIG. 1. Audio signal and spectrogram of an ascending scale recorded during the strobolaryngoscopic examination. The cursor marks the transition between the chest and head registers. The strobolaryngoscopic images shown in Figs. 2 and 3 were captured during the chest and head registers tones marked C5 and D5, respectively.

B&K 2234 was used as a sound level meter and a B&K 2669 as the audio microphone. The signals were digitized and recorded using the Soundswell (Hitech Development, Solna) signal workstation.

The stroboscopic and videokymographic recordings were saved as avi files, which combined the simultaneously recorded video and audio signals. For sound analysis, the audio tracks were extracted from the original avi files using the Virtual Dub software by Avery Lee, version 1.5.10, and saved as wav files. These files were analyzed using the MultiSpeech model 3700 voice analysis package by Kay Elemetrics. The audio signals were low-pass filtered (low-pass Blackman filter with cutoff frequency of 5000 Hz) and downsampled from the original 48 kHz to the sampling frequency of 12 kHz. Narrowband spectrograms (1024 samples analysis window, 17 Hz bandwidth) were then created and used for localizing the pitch jumps and spectral phenomena separating different registers. F_0 of the sung tones were determined manually from spectrograms by reading the frequency of the 10th harmonic (or lower, when the 10th was not visible) by means of a digital cursor and dividing this frequency by 10 (or the corresponding lower harmonic number). The phenomena displayed in the spectrograms were then identified in the avi video files of the videostroboscopic and VKG recordings. From the stroboscopic avi files images were exported showing the shape of the vibrating vocal folds at the phases of maximal glottal closure and maximal glottal opening in the different registers and saved as bitmaps using the Virtual Dub software. The VKG vibration patterns of the vocal folds in different registers were viewed using the VKIS software (created by A. Vetešník at the Center for Communication Disorders, Medical Healthcom, Ltd. in Prague, the Czech Republic). This program decomposed the interlaced video frames into separate video fields which were then exported and saved as bitmap files. A MATLAB-based program VKIT (created also by A. Vetešník) was used to obtain glottal contours from the videokymographic images and to automatically measure the glottal closed quotient. CorelPhotoPaint X3 software was used for the final processing and editing of the spectrographic, strobolaryngoscopic and videokymographic images; the image contrast was improved, image noise reduced and the individual images were combined into the final figures.

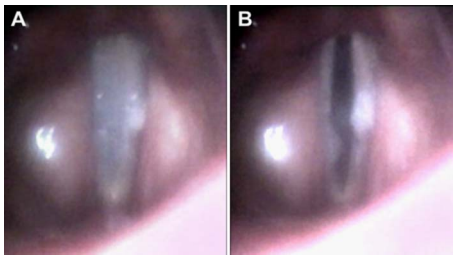


FIG. 2. (Color online). Stroboscopic view of the vocal folds during chest register phonation at the phase of (A) glottal closure and (B) maximum opening of the glottis. The whitish and bulging objects seen on the vocal folds were due to accumulated mucus. The vocal folds are closing and opening along their entire length.

III. RESULTS

A. Chest-head register transition

Figure 1 shows a spectrogram of an ascending scale sung by the subject. The transition between the chest and head register in the spectrogram between the 6th and 7th sustained tone is marked by the cursor. The transition occurred when F_0 changed from about 550 Hz (tone C#5) to about 620 Hz (tone D#5). In the chest register the third partial was dominant, boosted by the second formant (F_2) of the vowel /e/ around 1700 Hz. In head register, by contrast, the spectrum was dominated by the fundamental, which was amplified by the first formant (F_1) around 650 Hz. This change of dominance of partials was clearly perceivable and identified as a register transition. Note also the pitch jump preceding the 6th note.

The typical strobolaryngoscopic images of the vocal folds for the chest and head registers are shown in Fig. 2 and 3, respectively. Figure 2 pertains to the pitch C#5, the subject's highest tone in chest register, see Fig. 1. The left panel, showing the vocal folds during vibration at the phase of maximal glottal closure, reveals a complete glottal closure. The right panel, displaying the vocal folds at the phase of maximal glottal opening, reveals that the vocal folds are opening and vibrating along their whole length.

Figure 3 shows the stroboscopic images of the vocal folds phonating in the head register at the tone D#5 immediately after the transition shown in Fig. 1. A small posterior glottal gap can be seen at the phase of maximal glottal clo-

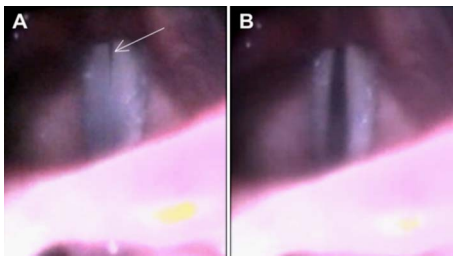


FIG. 3. (Color online). Stroboscopic view of the vocal folds phonating in the head register at the phases of (A) maximal glottal closure and (B) maximal glottal opening. The arrow points at the narrow gap in the posterior glottis that occurred after the transition to the head register. The anterior part of the vocal folds is partly hidden by epiglottis.

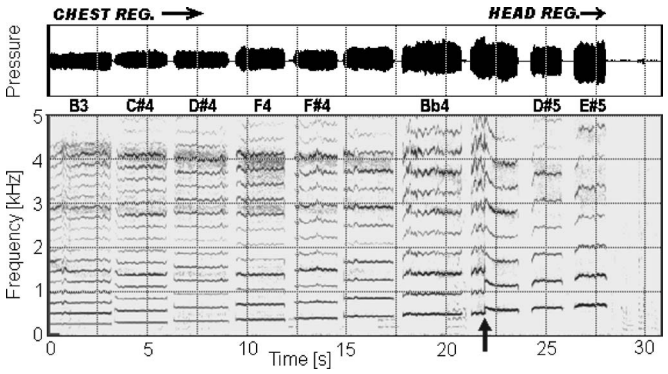


FIG. 4. Audio signal and spectrogram of an ascending scale recorded during the VKG examination. The chest-head transition occurred together with an ascending pitch jump (marked by arrow). The marks B3–E#5 designate the musical pitches of the tones for which the VKG images are shown in Fig. 5.

sure (panel A). This strongly suggests that for this tone the arytenoid adduction was less firm, or at least different, as compared with the preceding chest register tone illustrated in Fig. 2.

Figure 4 shows another example of a spectrogram of an ascending scale, this time recorded during the VKG examination. In this case the transition between the chest and head register occurred together with a discontinuity in the form of a pitch jump.

VKG images of the vibrating vocal folds at the individual tones are shown in Fig. 5. The transition from chest to

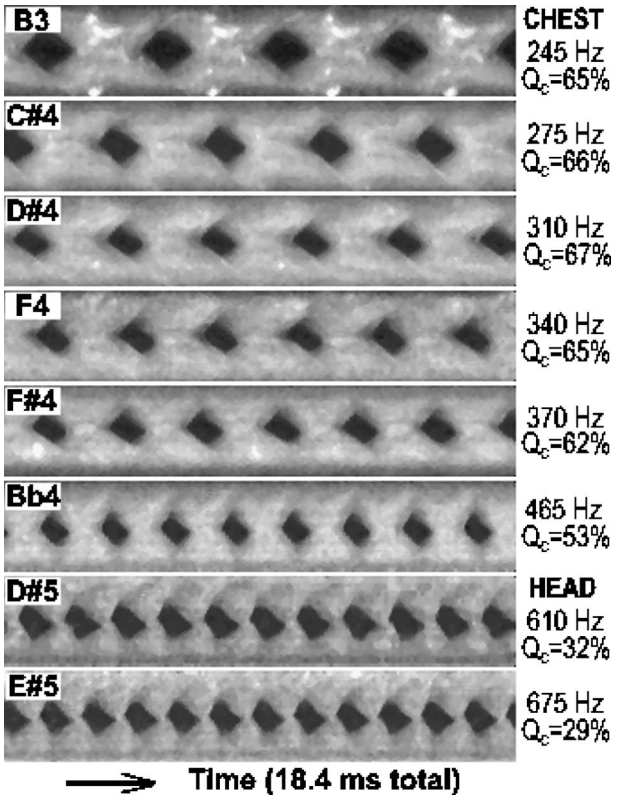


FIG. 5. VKG images showing the vibration pattern of the vocal folds in the middle of glottis when the subject raised the pitch and changed from chest to head register. The panels are marked by the corresponding pitches. The register, fundamental frequency (F_0) and closed quotient (Q_C) are indicated next to the images. Notice the lower closed quotient and the more rounded peaks in the head register (D#5, E#5) as compared to the chest register.

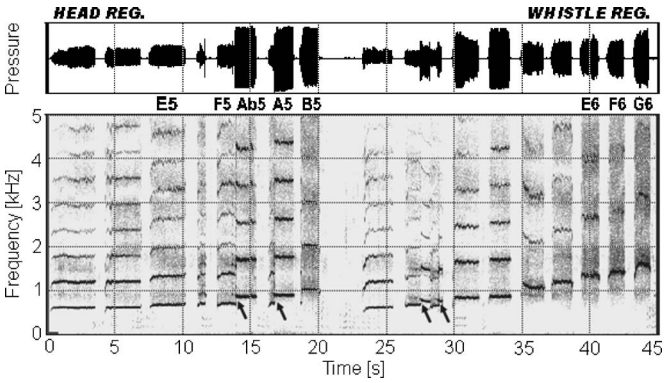


FIG. 6. Audio and spectrogram recordings of an ascending scale with abrupt transitions from head to whistle register (marked by arrows). In the upper panel, notice the large increase of the signal amplitude after the transitions and then the decrease of the amplitude when F_0 was increased above 1000 Hz.

head register between the tones Bb4 and D#5 was accompanied by a marked decrease of the closed quotient (Q_C, defined as the duration of the closed phase with respect to the duration of the vibration period) from more than 50% to about 30%. Furthermore, in the head register (Fig. 5, tones D#5 and E#5) the vibration pattern showed more rounded peaks of the maximal opening.

B. Head–whistle register transition

Figure 6 shows the audio signal and the spectrogram of ascending scales during which the voice changed from head to whistle register. Two qualitatively different transitions were observed here that occurred around 700 and 1000 Hz. The first transition was observed in the form of spontaneous jumps upwards from the frequencies around 670–700 Hz (tones E5–F5) to the frequencies of 750–840 Hz (tones Gb5–Ab5). With the upward jump, the amplitudes of the sound pressure signal largely increased, reaching values above the clipping level of the head-mounted microphone as can be seen in the upper panel of the figure. The second transition occurred when F_0 increased above 1000 Hz (time 35 s in Fig. 6) the sound pressure amplitudes decreased, the intensity of the overtones decreased such that the fundamental became dominant in the spectrum and the voice quality audibly changed to a perceptually "thinner" sound.

The VKG images corresponding to these spectrograms are presented in Fig. 7. The first two images show the vibration pattern of the head register at $F_0 \approx 660$ (tone E5), and $F_0 \approx 700$ Hz (tone F5), before the register transition. These images demonstrate patterns similar to those of the head register shown in Fig. 5 (tones D5 and E5). This is expected since they are produced in the same register. The pattern of the tone Ab5 shows the vocal folds vibrating at the frequency of 860 Hz, immediately after the upward pitch jump. The most marked difference from the head register pattern is that there is no closed phase ($Q_C=0$) and the vibration amplitudes of the vocal folds are smaller. This finding would seemingly suggest that a weaker sound is produced by these vibrations but the contrary was true—the audio signal amplitudes were much larger than those produced before the tran-

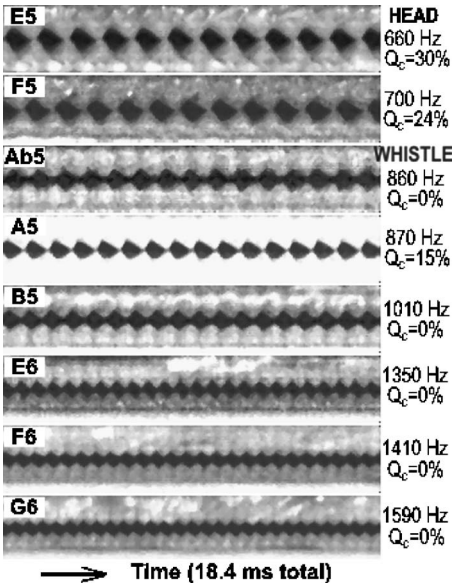


FIG. 7. VKG images showing the vibration pattern of the vocal folds in the middle of glottis during the ascending scale when changing from the head to whistle register. Notice the lack of vocal fold contact in all whistle register tones except A5. The register, fundamental frequency (F_0) and closed quotient (Q_C) are indicated next to the images.

sition in the head register. Such a paradox suggests that a vocal tract resonance plays an important role here.

The pattern in the second tone after the pitch jump (A5) with the F_0 of 870 Hz reveals that the contour wave form was skewed to the left, i.e., that the opening phase was remarkably shorter than the closing phase. Also, the vocal folds appeared to touch each other for a very brief moment, so the closed phase was nonzero. The pattern in the next image was derived from the tone B5 with $F_0 \approx 1010$ Hz. Here, the pattern is again remarkably different from the previous one—the contour wave form is much less skewed and there is no closed phase. However, the peaks at the maximal opening are rather sharp (Švec *et al.*, 2007), revealing that the vocal folds vibrate with vertical phase differences even at these high pitches in the whistle register. Besides decreasing vibration amplitudes of the vocal folds and shortening vibration periods, the whistle vibration pattern did not show any further noticeable qualitative changes with further pitch rises (tones E6, F6 and G6 in Fig. 7). Apart from the vibrating vocal folds, there could also be seen some irregular movements of mucus accumulated on the upper vocal fold surface which appeared to slightly perturb the regularity of the vocal fold vibrations. The highest F_0 achieved by the subject was 1590 Hz (tone G6).

Figure 8 shows stroboscopic images of the vocal folds vibrating at the highest pitch in the whistle register. The vocal folds appeared maximally elongated here and showed clear vibrations along their whole length, with maximal amplitudes located approximately in the middle of the membranous part of the vocal folds (panel B). When maximally closed, there remained a glottal gap along the whole vocal-fold length; the vocal folds did not touch each other during vibration (panel A). During the head-whistle pitch jumps we did not observe stroboscopically any gross change in the

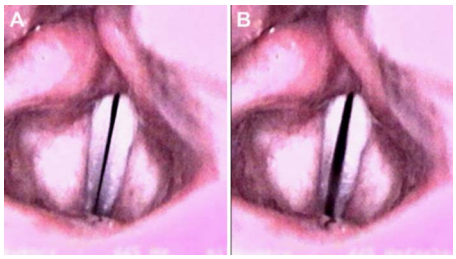


FIG. 8. (Color online). Stroboscopic views of the vocal folds vibrating at the highest pitch achieved in the whistle register, $F_0 \approx 1590$ Hz, tone G6. Panels A and B show images during maximal glottal closure and maximal glottal opening, respectively.

vocal fold adjustment comparable to those seen during the chest-head transition; only the vibration amplitude appeared to decrease suddenly and spontaneously.

IV. DISCUSSION AND CONCLUSION

In this study we were able to obtain unique strobolaryngoscopic and high-speed VKG data on the vibratory behavior of the vocal folds in an untrained female singer. The results revealed a number of new facts, particularly on the mechanism of the whistle register. The subject demonstrated a large frequency range, which was audibly divided into three separate registers, i.e., the chest, head and whistle. The lowest one, the chest register, was covering a range up to $F_0 \approx 550$ Hz. At the top of this range, spontaneous ascending frequency jumps occurred, separating the chest register from the head register. Stroboscopy showed that the membranous part of the glottis was fully closed in both these registers. However, a small gap in the posterior glottis was observed in the head register while no glottal gap was evident in the chest register. This suggests that muscles adducting the cartilaginous part of the glottis were slightly released when performing the transition from the chest to the head register. The assumption that adduction was less in the head register was supported also by the VKG, which revealed a lower Q_C than in the chest register.

In addition to these Q_C differences, the VKG images showed that the shape of the lateral peaks in the vibration patterns appeared to be more rounded in the head than in the chest register. This suggests slightly smaller vertical phase differences in the head register than in the chest register (Sundberg and Högset, 2001; Švec *et al.*, 2007). This can be interpreted as a sign of a lesser activity of the thyroarytenoid muscle causing the vocal folds to be slightly thinner vertically. These findings are in agreement with the generally accepted theory of the chest-head register transition being influenced by the adjustment of laryngeal muscles (e.g., Miller, 2000; Titze, 2000).

The transition from the head to whistle register tended to occur in combination with spontaneous ascending jumps from $F_0 \approx 670–700$ Hz to $F_0 \approx 750$ Hz or higher. No apparent changes in gross laryngeal adjustments were seen here laryngostroboscopically. After the jump, the sound pressure amplitudes largely increased. When further increasing F_0, above 1000 Hz, the sound pressure amplitudes decreased, the fundamental became the dominant frequency of the spec-

trum and the voice quality audibly changed. These observations suggest two transition effects which are not necessarily occurring simultaneously–the pitch jump around 700 Hz and the abrupt change in voice quality around 1000 Hz. Based on the theory of Titze (2004), we may hypothesize that the pitch jump occurs as a result of an interaction between the vocal tract resonance and the vocal fold vibrations when F_0 is entering a frequency region unfavored by the vocal tract. An audible change of voice quality, on the other hand, may be the result of F_0 passing F_1, as described by Miller and Schutte (1993). More studies are needed to justify these hypotheses. Whether the whistle register starts after the upward pitch jump from the head register or whether it is a voice quality resulting from $F_0 > F_1$ appears to be a matter of definition. Curiously, the two transitions observed around 700 and 1000 Hz here could possibly be related to the upper and whistle registers, which some authors recognize as two distinct registers of the female high voice (Miller, 2000). Under such scheme, the data of our subject could fit the four register division: chest, middle/head, upper/high and whistle, with the three transitions around 550, 700 and 1000 Hz. But cautiousness is needed here: since the high-pitched transitions appear to be influenced by the vocal tract formants (Titze, 2007), the upper/whistle register boundaries may be expected to be vowel dependent. Clearly more data are needed.

Interestingly, all the trained singers found it problematic to produce the high pitches above 1000 Hz during laryngoscopy and exhibited no audible transition to whistle register. The only singer who was able to produce fundamental frequencies above 1000 Hz and showed three distinct voice qualities was the untrained singer. It is known that classically trained singers learn to eliminate or minimize the timbral contrasts between registers. Hence, it was advantageous to use an untrained subject in this study. It is also known that classically trained singers make use of the resonances of the vocal tract through the "formant tuning" technique which is largely restricted during rigid laryngoscopy. We assume that this restriction interfered with the ability of the trained singers in reaching their uppermost voice range above 1000 Hz, whereas the untrained singer had the advantage of not being concerned about the correct technique.

Our exploratory investigation has offered information that is relevant to the description and understanding of the whistle register. First, there is no doubt that our subject's vocal folds were vibrating in the whistle register. Similar finding was reported by Keilmann and Michek (1993) but was not visually documented. Such a finding is fully documented here for the first time. This finding does not agree with observations of "chink" tones previously reported by Schultz (1902) and van den Berg (1963). The vibration amplitude decreased slightly during the transition from head to whistle register and the vibratory pattern was far from sinusoidal. The sharp peaks at the maximal opening brought evidence of non-negligible vertical phase differences, indicating the presence of a mucosal wave (Sundberg and Högset, 2001; Švec *et al.*, 2007). This suggests that the driving mechanism of the vocal fold vibration at F_0s above 1000 Hz is principally similar to that at lower F_0s. The closed phase was shorter in whistle register than in head register, and the

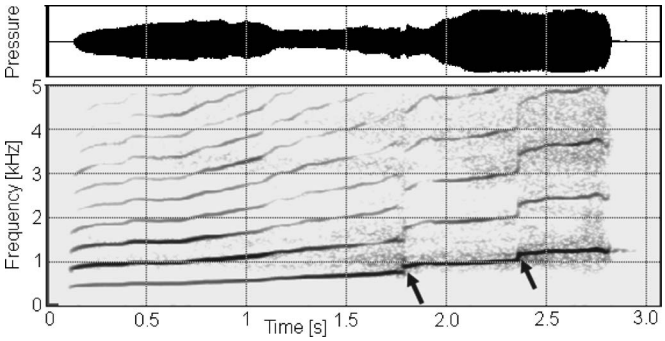


FIG. 9. Transition from the head to the whistle register with two consecutive frequency jumps from 740 to 880 Hz and from 1000 to 1150 Hz.

folds were vibrating along their whole length; no anterior-posterior vibration differences were apparent. Also, in the transition from head to whistle the vibration pattern showed a marked change from a considerably skewed to a rather symmetrical pattern.

The transition from head to whistle appeared to be variable and inconsistent. On one occasion two consecutive F_0 jumps were observed during the head to the whistle register transition (Fig. 9). Such spontaneous frequency jumps are manifestations of bifurcations—an effect inherent to nonlinear dynamic properties of the vibrating vocal folds when approaching boundaries of stable vibration regimes (Titze *et al.*, 1993; Herzel, 1996). Models have shown that abrupt changes of the vocal fold behavior can occur when smoothly changing such parameters as vocal fold tension (Švec *et al.*, 1999; Tokuda *et al.*, 2007), left-right asymmetry (Berry *et al.*, 1996; Herzel *et al.*, 1995; Steinecke and Herzel, 1995), or vocal tract and subglottal tract acoustic resonances (Mergell and Herzel, 1997; Neubauer *et al.*, 2004; Titze, 2004; Zhang *et al.*, 2006). Any of these factors could have had some influence on the F_0 jumps observed in our study. However, the paradox observed during the pitch jumps around 700 Hz—i.e., the increase in acoustic sound pressure occurring despite of the decrease in closed quotient and vibratory amplitude of the vocal folds—suggests that the vocal tract influence was particularly important during the head to whistle transition.

According to Miller and Schutte (1993), the whistle register results when F_0 passes F_1. Under such conditions, the vocal tract can be expected to have a strong influence on the vocal fold behavior, increasing the likelihood for pitch jumps to occur (Titze, 2004). This suggests that vocal fold vibration is rather unstable in this transition region. Our data did not allow localizing the exact position of F_1 with respect to F_0. However, the audible quality of the flageolet register of the sopranos studied by Miller and Schutte (1993) was found perceptually similar to the whistle register of our subject when above 1000 Hz (Miller, 2007, personal communication). It is not obvious that all vibratory differences observed between head and whistle register can be explained as the consequence of $F_0 > F_1$. Our whistle register may not be identical with the whistle register that is considered in some singing pedagogy literature to occur above the tone F6 (1397 Hz), as mentioned, e.g., in the overview of Titze (2000), without any available objective data. Such a high-

pitched transition was not observed in our study. In any event, however, a promising strategy for reducing the confusion in the area of vocal registers would be to collect more data on vocal fold vibration in different registers. The present study indicates that the VKG represents a promising tool in such an enterprise.

The sound and video examples analyzed in this study can be accessed at http://www.ncvs.org/ncvs/library/tech.

ACKNOWLEDGMENTS

Jan Švec's work at the Department of Speech, Music and Hearing, KTH, Stockholm was supported by an individual grant from the Wenner-Gren Foundation. In the Netherlands, his work was supported by the Technology Foundation STW (Stichting Technische Wetenschappen) Project No. GKG5973, Applied Science Division of NWO (Natuurwetenschappelijk Onderzoek), and the technology program of the Ministry of Economic Affairs, the Netherlands. The authors thank Hans Larsson at the Department of Logopedics and Phoniatrics, Karolinska Institute, Huddinge University Hospital, Stockholm, for his help in acquiring the laryngoscopic recordings. The authors also thank two anonymous reviewers for their comments which resulted in improvements of the manuscript.

Berry, D. A., Herzel, H., Titze, I. R., and Story, B. H. (**1996**). "Bifurcations in excised larynx experiments," J. Voice **10**(2), 129–138.

Henrich, N. (**2006**). "Mirroring the voice from Garcia to the present day: Some insights into singing voice registers," Logoped. Phoniatr. Vocol. **31**(1), 3–14.

Herzel, H. (**1996**). "Possible mechanisms of vocal instabilities," in *Vocal Fold Physiology: Controlling Complexity and Chaos*, edited by P. J. Davis and N. H. Fletcher (Singular, San Diego), pp. 63–75.

Herzel, H., Berry, D., Titze, I., and Steinecke, I. (**1995**). "Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling," Chaos **5**(1), 30–34.

Herzel, H., and Reuter, R. (**1997**). "Whistle register and biphonation in a child's voice," Folia Phoniatr. Logop. **49**(5), 216–224.

Hollien, H. (**1974**). "On vocal registers," J. Phonetics **2**, 125–143.

Keilmann, A., and Michek, F. (**1993**). "Physiologie und akustische Analysen der Pfeifstimme der Frau. [Physiology and acoustic analysis of whistle voice of the woman]," Folia Phoniatr. (Basel) **45**, 247–255.

Lullies, H. (**1953**). "Physiologie der Stimme und Sprache," in *Gehör - Stimme - Sprache*, edited by O. F. Ranke and H. Lullies (Springer-Verlag, Berlin), pp. 163–293.

Mergell, P., and Herzel, H. (**1997**). "Modeling biphonation—the role of the vocal tract," Speech Commun. **22**(2–3), 141–154.

Miller, D. G. (**2000**). "Registers in singing: Empirical and systematic studies in the theory of the singing voice. (Doctoral dissertation)," University of Groningen, Groningen, the Netherlands.

Miller, D. G., and Schutte, H. K. (**1993**). "Physical definition of the flageolet register," J. Voice **7**(3), 206–212.

Mörner, M., Fransson, F., and Fant, G. (**1963**). "Voice register terminology and standard pitch," STL-QPSR **4/1963**, 17–23.

Neubauer, J., Edgerton, M., and Herzel, H. (**2004**). "Nonlinear phenomena in contemporary vocal music," J. Voice **18**(1), 1–12.

Schultz, P. (**1902**). "Über einen Fall von willkürlichem laryngealen Pfeifen beim Menschen," Arch. f. Physiol. Suppl. 523.

Shadle, C. H. (**1983**). "Experiments on the acoustics of whistling," Phys. Teach. **21**(3)148–154.

Steinecke, I., and Herzel, H. (**1995**). "Bifurcations in an asymmetric vocal-fold model," J. Acoust. Soc. Am. **97**(3), 1874–1884.

Sundberg, J., and Högset, C. (**2001**). "Voice source differences between falsetto and modal registers in counter tenors, tenors and baritones," Logoped. Phoniatr. Vocol. **26**(1), 26–36.

Švec, J. G. (**2000**). "On vibration properties of human vocal folds: Voice registers, bifurcations, resonance characteristics, development and application of videokymography. (doctoral dissertation)," University of Groningen, Groningen, the Netherlands.

Švec, J. G., and Schutte, H. K. (**1996**). "Videokymography: High-speed line scanning of vocal fold vibration," J. Voice **10**(2), 201–205.

Švec, J. G., Schutte, H. K., and Miller, D. G. (**1999**). "On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges," J. Acoust. Soc. Am. **106** (3, Pt.1), 1523–1531.

Švec, J. G., Šram, F., and Schutte, H. K. (**2007**). "Videokymography in voice disorders: What to look for?," Ann. Otol. Rhinol. Laryngol. **116**(3), 172–180.

Titze, I. R. (**2000**). *Principles of Voice Production (second printing)*, National Center for Voice and Speech, Iowa City, IA.

Titze, I. R. (**2004**). "Theory of glottal airflow and source-filter interaction in speaking and singing," Acta. Acust. Acust. **90**(4), 641–648.

Titze, I. R. (**2007**). "Source-filter interaction in speaking and singing is nonlinear," Echoes Newsletter of the Acoustical Society of America **17**(3), 1–3.

Titze, I. R., Baken, R. J., and Herzel, H. (**1993**). "Evidence of chaos in vocal fold vibration," in *Vocal Fold Physiology: Frontiers in Basic Science*, edited by I. R. Titze (Singular, San Diego), pp. 143–188.

Tokuda, I., Horáček, J., Švec, J. G., and Herzel, H. (**2007**). "Comparison of biomechanical modeling of register transitions and voice instabilities with excised larynx experiments," J. Acoust. Soc. Am. **122**(1), 519–531.

van den Berg, Jw. (**1963**). "Vocal ligaments versus registers," NATS Bull. **20** (2/December 1963), 16–31.

Walker, J. S. (**1988**). "An investigation of the whistle register in the female voice," J. Voice **2**(2), 140–150.

Zhang, Z., Neubauer, J., and Berry, D. A. (**2006**). "The influence of subglottal acoustics on laboratory models of phonation," J. Acoust. Soc. Am. **120**(3), 1558–1569.

The interplay between the auditory and visual modality for end-of-utterance detection

Pashiera Barkhuysen, Emiel Krahmer, and Marc Swerts[a)]
Communication & Cognition, Faculty of Arts, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

(Received 2 June 2006; revised 22 October 2007; accepted 26 October 2007)

The existence of auditory cues such as intonation, rhythm, and pausing that facilitate end-of-utterance detection is by now well established. It has been argued repeatedly that speakers may also employ visual cues to indicate that they are at the end of their utterance. This raises at least two questions, which are addressed in the current paper. First, which modalities do speakers use for signalling finality and nonfinality, and second, how sensitive are observers to these signals. Our goal is to investigate the relative contribution of three different conditions to end-of-utterance detection: the two unimodal ones, vision only and audio only, and their bimodal combination. Speaker utterances were collected via a novel semicontrolled production experiment, in which participants provided lists of words in an interview setting. The data thus collected were used in two perception experiments, which systematically compared responses to unimodal (audio only and vision only) and bimodal (audio-visual) stimuli. Experiment I is a reaction time experiment, which revealed that humans are significantly quicker in end-of-utterance detection when confronted with bimodal or audio-only stimuli, than for vision-only stimuli. No significant differences in reaction times were found between the bimodal and audio-only condition, and therefore a second experiment was conducted. Experiment II is a classification experiment, and showed that participants perform significantly better in the bimodal condition than in the two unimodal ones. Both the first and the second experiment revealed interesting differences between speakers in the various conditions, which indicates that some speakers are more expressive in the visual and others in the auditory modality. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2816561]

PACS number(s): 43.70.Mn, 43.71.Sy, 43.71.An, 43.71.Bp [ARB] Pages: 354–365

I. INTRODUCTION

Speakers use nonlexical features to demarcate various kinds of speech units, varying from a simple phrase to a larger scale discourse segment or a turn in a natural conversation. Previous studies have largely focused on how prosodic variables, such as intonation, rhythm and pause, or more subtle modulations of voice quality, like creaky voice, can be exploited to signal the end of such units (e.g., de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts *et al.*, 1994a; Wightman *et al.*, 1992). In addition to features that are encoded in the speech signal itself, there is also an investigation into how particular visually observable variations from a speaker's face, like gaze patterns or bodily gestures, can be used as boundary cues (e.g., Argyle and Cook, 1976; Cassell *et al.*, 2001; Nakano *et al.*, 2003; Vertegaal *et al.*, 2000). However, little is known about the perception of these visual cues, and about the relative importance of the visual and the auditory modality for demarcation purposes. Therefore, the aim of this paper is to get more insight into which modalities speakers use for signaling finality or nonfinality, and how sensitive observers are to these respective signals. In particular, our goal is to investigate the relative contribution of three different conditions to end-of-utterance detection: two unimodal ones, vision only and audio only, and their bimodal combination.

It is by now well established that various auditory cues may serve as boundary markers of speech utterances (e.g., Koiso *et al.*, 1998; de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts *et al.*, 1994a; Ward and Tsukahara, 2000; Wightman *et al.*, 1992, among many others). One of the strongest prosodic indicators for the end of a speaker's utterance is a pause, either a silent interval or a filler such as "uh" and "uhm" (as shown by, among others, de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts, 1997, 1998; Wightman *et al.*, 1992). Many of these studies are based on analyses of monologues, where it was even found that pause length may covary with the strength of a boundary. When looking at natural interactions between multiple speakers, however, pauses tend to be rather short inbetween two consecutive speaker turns. Even though end-of-utterance pauses may be very short in interaction, turn switching proceeds remarkably smoothly, generally without overlap between speakers (Koiso *et al.*, 1998; Levinson, 1983; Ward and Tsukahara, 2000).

One of the reasons why the turn-taking mechanism may proceed so fluently, is that speakers "presignal" the end of their utterances (e.g., Couper-Kuhlen, 1993; Caspers, 1998; Swerts *et al.*, 1994a, Swerts *et al.*, 1994b). Listeners may pick up these cues and therefore may know in time when the current turn will be finished. Various researchers have looked

[a)]Author to whom correspondence should be addressed. Electronic mail: m.g.j.swerts@uvt.nl

in detail at the nature of these cues. It has been suggested, for instance, that the capacity of listeners to anticipate an upcoming boundary is based on what is called rhythmic expectancy (Couper-Kuhlen, 1993). Related to this, there is subtle durational variation, such as preboundary lengthening, which speakers can use to mark the final edge of a speech unit such as a turn (Wightman *et al.*, 1992; Price *et al.*, 1991). In addition to these timing-related phenomena, many researchers have focused on the potential use of melodic boundary markers as well. First, there are local boundary markers which occur at the extreme edge of a turn unit, right before an upcoming boundary, for which it has been shown that tones which reach a speaker's bottom range clearly function as finality cues (Swerts and Geluykens, 1994; Caspers, 1998; Koiso *et al.*, 1998). Moreover there appear to exist melodic structuring devices which are more global in nature in that they are spread over a whole speech unit. In particular, various studies have pointed out that speech melody gradually decreases in the course of an utterance, which may enable listeners to feel a boundary coming up (e.g., Leroy, 1984). However, this declination pattern may be typical of read-aloud speech which allows for a larger degree of look-ahead compared to spontaneous speech. Other finality cues are variations in pitch span, and more subtle differences in the alignment of pitch movements (Silverman and Pierrehumbert, 1990; Swerts, 1997). Finally, there is acoustic evidence which shows that marked deviations from normal phonation, in particular, creaky voice, typically occur at the end of an utterance (Carlson *et al.*, 2005).

The possible premonitoring cue value of prosodic cues has been explicitly tested in various perception studies. Grosjean (1983) and Leroy (1984) have already established that human subjects are surprisingly accurate in estimating the location of an upcoming boundary, using a variant of a gating paradigm, in which listeners are only presented with the initial part of an utterance. Along the same lines, Swerts *et al.* (1994a) and Swerts and Geluykens (1994) reported that people are able, on the basis of melodic cues, to judge the serial position of a phrase in a larger discourse unit. Carlson *et al.* (2005) found that native speakers of Swedish and of American English showed a remarkable similarity in judgments when they had to predict upcoming prosodic breaks in spontaneous Swedish speech, even when they had to base such estimations on stimuli that consisted of only a single word.

It thus seems safe to conclude that speakers and listeners take the auditory modality into account while marking the end of an utterance. But to what extent do they pay attention to the visual modality? Various researchers have argued that speakers may use visual cues for end-of-utterance signaling, where most studies have investigated how various bodily gestures may be used as markers of discourse boundaries. First, different studies focused on general changes in posture (Beattie *et al.*, 1982; Cassell *et al.*, 2001; Duncan, 1972). These studies suggest there is a general trend for people to change their pose when they start speaking, whereas they return to their initial posture at the end of a turn, for instance by raising their shoulders at the onset of a turn and lowering them again at the end. Second, one specific visual cue which has received much scholarly attention is related to movements of the eyes. Argyle and Cook (1976) describe in detail how the tuning of gaze behavior regulates many aspects of the interaction in a very subtle way. In general, it appears to be the case that speakers divert their gaze rather often while talking, whereas the listening conversation participant tends to look at the partner more frequently. When analyzing the gaze patterns in normal interactions more closely, it appears that a pattern emerges which is connected to the turn-taking mechanism, in that speakers tend to divert their gaze when they start talking, and return the gaze to their partner when they are finished (see also Goodwin, 1980; Kendon, 1967; Nakano *et al.*, 2003; Novick *et al.*, 1996; Vertegaal *et al.*, 2000). The cue value of gaze is likely to be due to the fact that human eyes have a unique morphology, with a large white sclera surrounding the dark iris. It has been argued that this contrast may have evolved to make it easier to detect the gaze direction of others (Kobayashi and Kohshima, 1997). While variation in posture shifts and gaze patterns have been directly linked to boundary marking, in particular in the turn-taking system, various researchers have argued that there may be further visual cues that may be important for demarcation purposes as well, such as head nods (e.g., Maynard, 1987), eyebrow movements (e.g., Ekman, 1979; Krahmer and Swerts, 2004), and eye blinks (e.g., Doughty, 2001).

The results from the various studies described above thus suggest that a speaker can display that he or she is going to stop speaking, by means of both auditory and visual features. However, there are still a large number of unsolved questions regarding the relative importance of the modalities and of their combined effects. While it has been shown that listeners are accurate in determining the end of an utterance based on the auditory modality, it is unknown whether they would be equally capable of doing so on the basis of visual information as well. And if so, it is still an empirical question as to how the visual modality relates to the auditory one, whether or not the two modalities may reinforce each other, and whether observers are helped or rather distracted when they have to focus on two rather than on a single modality in their finality judgments.

To this end, we have set up two experiments that are both based on perceptual judgments of stimuli in one of three conditions: a vision-only, audio-only, or audio-visual condition. The experiments make use of audio-visual recordings of semispontaneous utterances that were naturally elicited in a question-answering paradigm. The first experiment explores differences between modalities via a reaction time experiment in which participants are instructed to indicate as soon as possible when they think an utterance, presented in one of three conditions, ended. The second experiment makes use of basically the same stimuli as the ones from the first experiment, and looks in more detail at which factors influence participants' abilities to judge whether a speaker's turn is about to end or not; in this experiment, subjects are presented both with longer and shorter speech fragments, so we may get insight into the cue value of possible global versus local cues to finality. In addition, we look in more detail into the question of which auditory and visual cues are actually used by our speakers.

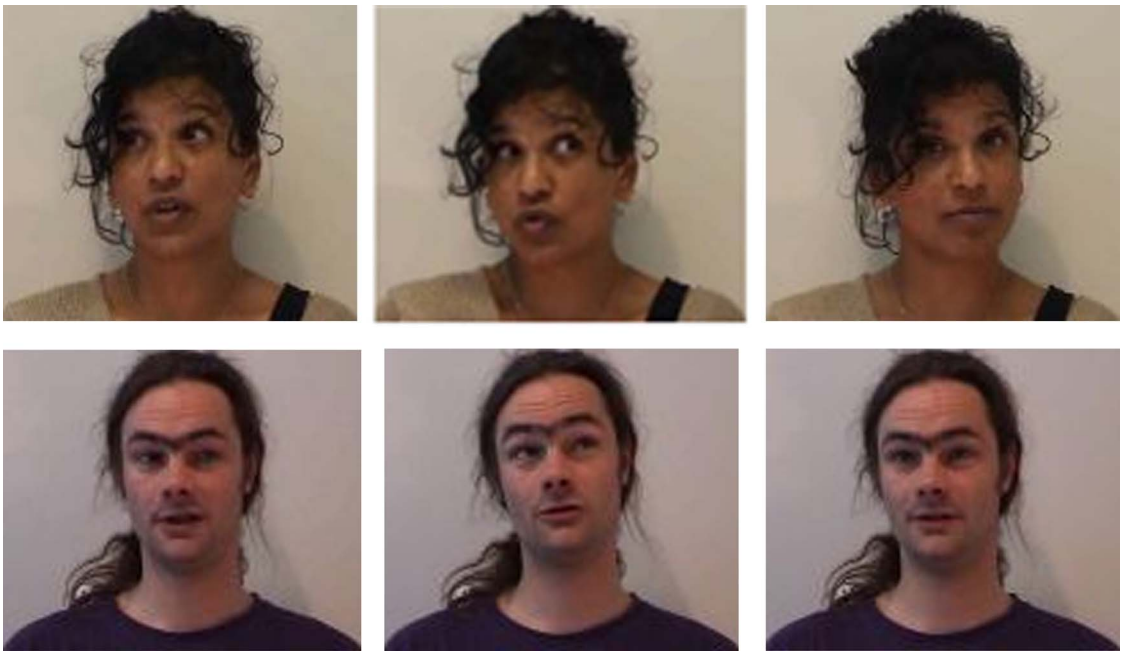


FIG. 1. (Color online) Representative stills of speakers SS (top) and BB (bottom) while uttering the first and middle word and just after uttering the final word of a three word answer, such as "red, white, blue."

II. AUDIO-VISUAL RECORDINGS

We gathered digital video recordings of speakers responding to questions in a natural, interview-style situation. Although recent research suggests that lexical and syntactic factors are relevant for end-of-utterance detection (de Ruiter *et al.*, 2006), for our current purposes, however, these factors should be eliminated as they would offer an unfair advantage to the auditory modality. Hence the questions were intended to elicit lists of words, where the lexical and syntactic structures of the answers offer no clues at all about where the end of the utterance is to be expected.

The questions were selected in such a way that they resulted in a variety of different answers, and such that potential answer words could occur in different positions in the list, depending on the question. Target answers varied in length, consisting of three or five words. Twelve questions were asked for predictable sets of numbers, in different orders, and with different number ranges. For instance: what are the multiples of five below 30?; what are the odd numbers below ten in reversed order?; and what are the multiples of five below 30 in reversed order?

Notice that the word "five" can occur both in a final and in a nonfinal position. The other questions addressed general knowledge or individual preferences of the interviewee, such as: what are the colors of the Dutch flag?; what are your three favorite colors?; and name five countries where you can go skiing.

Notice that for the second category the answers are never fully predictable. Even the colors of the Dutch flag are described by participants both as "red, white, blue" and "blue, white, red." Moreover, both "red" and "blue" can occur (and do in fact occur) as the second, middle word, in responses to the favorite color question. The interview consisted of 33 questions, of which 25 were experimental and eight were filler items. As filler items, questions were used for which the number of words in the answers could in principle not be predicted (e.g., Which languages do you speak?). These filler questions were added for the sake of variety and to make sure that speakers did not only produce three and five word lists.

A total of 22 speakers participated (13 male and nine female), between 21 and 51 years old. None of the speakers was involved with audio-visual research, and speakers did not know for what purpose the data were collected. The original recordings were made with a digital video camera [MiniDV; 25 frames/s, a resolution of 720×576 pixels, sampling of 4:2:0 (PAL), luma 8 bits chroma and 2 channel audio recording at 16 bits resolution and 48 kHz sampling rate]. The recordings were subsequently read into a computer and orthographically transcribed. See Fig. 1 for some representative stills.

III. EXPERIMENT I: REACTION TIMES

As a first exploration we performed a reaction time experiment with the intention to gain insight into the relative contribution of the auditory and visual modality, alone and in combination, for end-of-utterance detection.

A. Method

1. Stimuli

For this experiment four male and four female speakers were randomly selected from the corpus of 22 speakers described above. For each speaker, three instances of answers consisting of three words and three instances of five words were randomly selected on the basis of the transcriptions (8 speakers $\times$ 6 instances $=$ 48 stimuli in total). Notice that since this first selection was random, the set of selected an-

swers differed for each of the selected speakers. As a result, the lexical content of the selected answer lists was highly varied, and since words could occur in various (final and nonfinal) positions, observers could never rely on lexical information for their end-of-utterance detection. If the first selection contained answers with more than just list words (e.g., repetitions of the question, or fragments where speakers think aloud), these were replaced with another randomly selected answer. Moreover, lists where the prefinal and final word were separated by a conjunction (i.e., lists of the form "A, B, and C") were replaced as well. In addition, for each speaker two filler items were selected of different lengths. Fillers could include other spoken text (such as repetitions or corrections), and as a result the average length of filler items was 11 words. Each stimulus was cut from the interview session in such a way that it started immediately after the interviewer finished asking the current question until 1000 ms after the speaker finished answering (i.e., 1000 ms after the auditory speech signal of the answerer had stopped).

2. Participants

For the reaction time experiment, 30 right-handed native speakers of Dutch participated, seven male and 23 female, between 24 and 62 years old. None of the participants had participated as a speaker in the data collection phase, and none was involved in audio-visual speech research.

3. Procedure

Stimuli were presented to participants in three conditions: one bimodal one, containing audio-visual stimuli (AV), and two unimodal ones, one audio only (AO), and one vision only (VO). In the audio-visual condition, participants saw the stimuli as they were recorded. In the audio-only condition, participants heard the speakers while the visual channel only depicted a static black screen, and in the vision-only condition, participants only saw the speakers but could not hear them. All participants entered all three conditions (within design), but the order in which participants entered these conditions was systematically varied (using a 3×3 Latin square design). Moreover, within a condition, stimuli were always presented in a different random order. In this way, all potential learning effects could be compensated for.

Each condition consisted of two parts: a baseline measurement and the actual end-of-utterance detection. Each part was preceded by a short practice session so participants would be acquainted with the experimental setting and the kind of stimuli in the current condition. The practice session did not contain lexical material which reoccurred in the actual experiment.

The aim of the baseline measurement was to find out how long it took participants on average to respond to comparable stimuli in the three modalities of interest (AV, AO, VO) of varying durations but always completely devoid of finality cues. During the baseline measurement, the participants' task was to press a designated button as soon as the end of the stimulus was reached. Stimuli were constructed to make them comparable to the actual stimuli used in the nonbaseline conditions but without introducing potential finality cues. In the audio-visual modality, the baseline stimuli therefore consisted of a video still (a single frame of some speakers) accompanied by a stationary /m/ (a male voice for male speakers, and a female voice for female speakers), creating the impression of a speaker uttering a prolonged "mmm." In the vision-only baseline measurement, only the video still was displayed, and in the audio-only baseline measurement, only the stationary /m/ was heard. In all three conditions the baseline stimuli are therefore completely static: the face does not move, since it is a still image, and the sound does not change either, since it is stationary. When the end of a baseline stimulus is reached, the sound stops (in the AO condition) and a blank screen appears (in the VO condition); this happens simultaneously in the AV condition. Only then can participants know that the stimulus ended; there is no conceivable cue in the stimulus which could presignal this.

During the actual end-of-utterance detection part, participants were instructed to indicate, as soon as possible, when the speaker finished his or her utterance by pressing a dedicated button. In the experiment, it was crucial that participants pay attention to visual information on the screen. Therefore, they were given an additional monitoring task, where participants had to press another button as soon as they saw a small red dot appearing on the screen. These red dots were added to a limited number of dummy stimuli. Even though the audio-only condition did not include any potentially relevant visual information (only a black screen), participants also had to spot the red dots in this condition to make sure all conditions were alike in this respect. The duration of the red dot appearance was $1/25$ s (a single frame); it appeared at varying locations on the screen. The dummy stimuli were only used to control the visual attention of participants and were not used in the reaction time analyses. This use of dots to make sure participants process visual information is a common procedure in audiovisual speech research (e.g., Bertelson *et al.*, 2003).

The experiment was individually performed. Participants were invited into a quiet room, and asked to take a seat behind a computer on which the stimuli would be displayed. There were loudspeakers to the left and right of the screen through which the sound was played. Participants received instructions before each of the three conditions and before they started with the relevant practice session. If everything was clear, the actual experiment started and the experimenter moved out of the visual field of the participant. There was no further interaction between participant and experimenter during the experiment.

4. Data processing

Reaction times (RTs) were always measured in milliseconds from the actual end of utterance (i.e., the moment where the speech signal ended). An RT of 0 thus means that a participant pressed exactly at the end of the utterance (when the auditory speech signal stopped). Notice that in the baseline measurement, the end of the dummy utterance /mmm/ also marked the end of the stimulus. In the actual experiment, stimuli continued for 1000 ms after the speaker

TABLE I. Reaction times in milliseconds for the different conditions: audio-visual (AV); vision-only (VO); audio-only (AO) in both the baseline measurement and the actual experiment, with standard errors and with 95% confidence intervals.

Measurement	Condition	RT	Std. error	95% CI
Baseline	AV	391.7	7.6	(376.1,407.3)
	VO	330.8	5.9	(318.9,342.9)
	AO	380.3	5.5	(368.9,391.7)
Experiment	AV	508.8	38.6	(429.7,587.8)
	VO	668.5	33.3	(600.4,736.7)
	AO	524.6	40.2	(442.4,606.9)

finished speaking (i.e., after the spoken audio signal ended), and the end of utterance thus does not coincide with the end of the stimulus.

Inspection of the measurements revealed that occasionally a negative RT was recorded. This happened 13 times during the baseline measurement (i.e., 1.8% of the baseline data points), and 302 times during the actual experiment (nearly 7% of the experimental data points). In both cases, the negative RTs were evenly distributed over the modality conditions. In the case of the baseline measurement we can be certain that these are errors, since participants had to respond to the "ending" of the baseline stimuli and, as explained above, there were no cues that could possibly presignal the end. Hence these errors were replaced by the mean RT value for that stimulus. It is important to note that this did not significantly alter the results, so the inclusion of the negative RTs in the baseline condition would have led to basically the same results as reported below (given the very small number of negative instances).

In the actual end-of-utterance experiment a negative RT is not necessarily an error, because here, as noted in Sec. I, presignals may occur, and hence the participant may believe the end of the utterance is near even though the speaker has not actually stopped speaking yet. Since there is no other criterion for their exclusion, we decided not to remove these negative RTs. Finally, there was a total of 23 nonresponses (0.5%), which were treated as missing values in the statistical analysis. We did not manipulate the raw data in any other way.

5. Statistical analyses

All tests for significance were performed with a repeated measures analysis of variance (ANOVA). Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse–Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. *Post hoc* analyses were performed with the Bonferroni method.

B. Results

A general overview of the RT results for the different conditions can be found in Table I. First consider the baseline measurement. Here the VO condition evoked the fastest reaction times followed by the AO and the AV conditions. An ANOVA was performed with condition and stimulus duration as within participants variables and reaction time as the dependent variable was performed. It indeed revealed a main effect of condition $[F(2,58)=11.215, p<0.001, \eta_p^2=0.279]$. *Post hoc* analyses showed that there was a significant difference between the audio-visual and vision-only condition ($p<0.001$), and between the vision-only and the audio-only condition ($p<0.001$). The audio-only and the audio-visual condition did not differ significantly ($p=0.368$). The stimuli used for the baseline measurement differed in duration, but this did not have a significant influence on the reaction times [$F(7,203)=2.891$, *n.s.*], nor was the interaction between condition and stimulus duration significant [$F(14,406)=2.021$, *n.s.*].

TABLE II. Reaction times in milliseconds for the different conditions: audio-visual (AV); vision-only (VO); audio-only (AO) in the actual experiment as a function of length (three words or five words), with standard errors between brackets.

	Length	
Condition	Three words	Five words
AV	585.0 (36.6)	432.5 (42.7)
VO	803.9 (33.0)	533.0 (44.3)
AO	627.6 (48.9)	421.7 (42.6)

Next consider the results of the actual experiment. Here the AV condition yielded the quickest responses, followed by the AO condition, while the VO condition leads to the slowest reaction times. An ANOVA with condition, length (measured by the number of words: three or five), and speaker as within participants variables and reaction time as the dependent variable was carried out. A significant main effect of condition was found $[F(2,58)=17.052, p<0.001, \eta_p^2=0.370]$. *Post hoc* analyses showed that there was a significant difference between the audio-visual and vision-only condition ($p<0.001$), and between the vision-only and the audio-only condition ($p<0.001$). The audio-only and the audio-visual condition did not differ significantly ($p=0.396$). In addition, a main effect of stimulus length was found $[F(1,29)=90.086, p<0.001, \eta_p^2=0.756]$. Inspection of Table II reveals that three word utterances led to longer reaction times than five word utterances. Finally, there was also a main effect of speaker $[F(7,203)=23.500, p<0.001, \eta_p^2=0.448]$ which indicates that some speakers gave overall better or more cues that they were nearing the end of the utterance than other speakers did.

When looking at the interaction effects, a significant interaction between condition and stimulus length $[F(2,58)=26.480, p<0.001, \eta_p^2=0.477]$ was found. As can be seen in Table II, the RT for three word utterances and for five word utterances differs substantially across the different conditions: it is relatively small for the audio-visual condition and relatively large for the vision-only condition, suggesting that the presence of extra cues in longer fragments is particularly useful for the vision-only condition. The RT patterns for the eight speakers are similar over the three modality conditions, as can be seen in Fig. 2. However, some speakers score par-

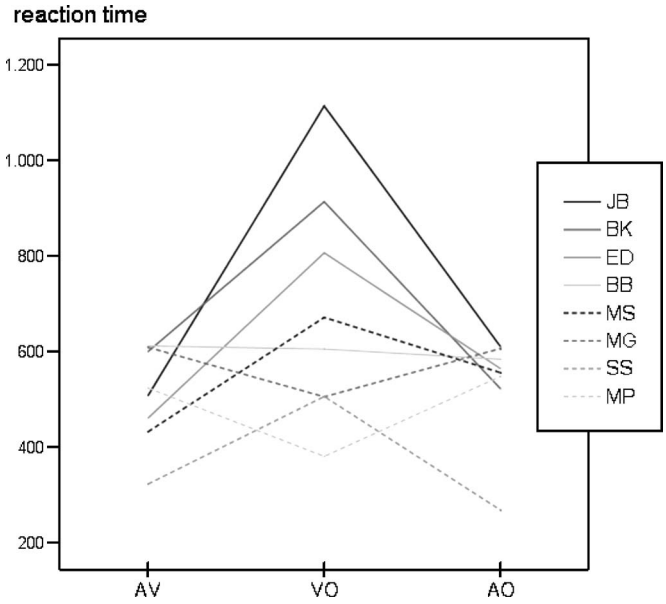


FIG. 2. The mean reaction time (in ms) for the different speakers in the three modalities.

ticularly well in one of the conditions, for instance, because they better cue the end of their utterances using facial cues rather than auditory ones.

It is interesting to see that the reaction time patterns for the baseline measurement are rather different from those of the actual experiment. The aim of the baseline measurement was to find out how long it takes to respond to a stimulus without any finality cues presented in a certain modality, and to compare these scores with the reaction times in the actual experiment in order to eliminate the influence of the presentation modality itself. The picture that emerges is visualized in Fig. 3, which shows that the reaction times for the baseline and nonbaseline versions are more similar in the audio-visual condition, and more divergent in the vision-only condition,

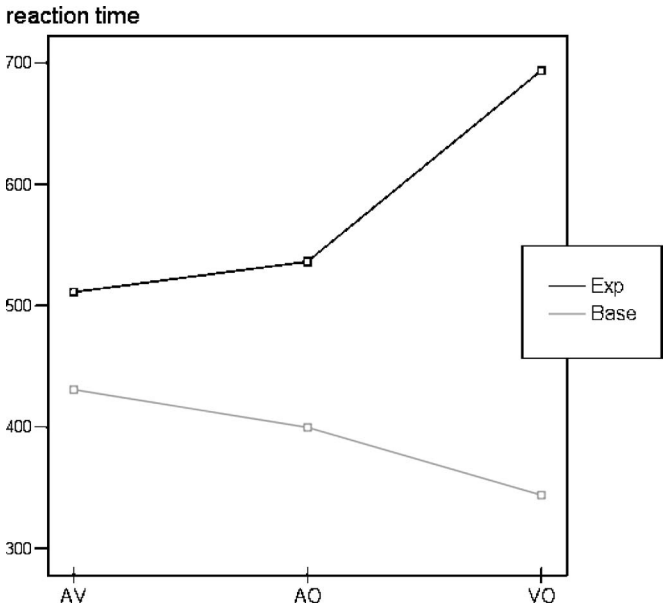


FIG. 3. The mean reaction time (ms) in the three conditions for the baseline and the actual experiment.

while the results for the audio-only condition are inbetween these two extremes. That is, where the visual modality leads to the fastest RT results in the baseline measurement, they are the slowest in the actual experiment. The reverse is true for the data in the audio-visual modality, whereas the data for the auditory modality are in the middle in both sessions.

To test these differences, we computed a difference score for each participant and stimulus, by subtracting the audio-visual baseline RT scores from that participant from his or her nonbaseline RT scores for the audio-visual stimuli, and similarly for the other two modalities. The resulting average difference score was 80.3 ms for the audio-visual condition, 136.8 ms for the audio-only condition, and 349.9 ms for the vision-only condition. We then performed a univariate ANOVA with average difference score for each participant as the dependent variable, and condition (AV, AO, VO) as the independent variable, which indeed revealed a significant effect of condition on difference score $[F(2,87)=13.704, p<0.001, \eta_p^2=0.40]$. A Bonferroni *post hoc* analysis revealed that all pairwise comparisons were significant at the $p<0.001$ level, except for the one between the audio-visual and the audio-only condition ($p=0.906$).

C. Summary

In the first experiment, we measured reaction times for end-of-utterance detection in three different conditions: audio only, vision only, and audio-visual. If prediction of the end of a turn was impossible, the reaction times for the different modalities in the actual experiment would have been the same, or at least have the same pattern as in the baseline measurement, where no cues were present. However, this is clearly not what was found. Rather, the audio-visual stimuli in the actual experiment led to the quickest responses, the audio-only stimuli led to slightly longer reaction times (although the difference with the audio-visual stimuli was not statistically significant), and the vision-only stimuli led to the slowest responses. While this result suggests that combining modalities is useful for end-of-utterance detection, it also leaves open the possibility that participants essentially rely on auditory information only for end-of-utterance detection. This issue is investigated more closely in a second experiment, where participants have to classify brief fragments as nonfinal or final (end of utterance) ones.

IV. EXPERIMENT II: CLASSIFICATION

The design of the classification task experiment resembles the design used in gating tasks. In a gating task a spoken language stimulus is presented in segments of increasing duration, usually starting at the beginning of the stimulus. Participants must try to recognize the entire spoken stimulus on the basis of the fragment (Grosjean, 1996).

In one possible presentation format, the *duration-blocked format*, participants are presented with all the stimuli at a particular segment size, then all the stimuli again in a different segment size (Grosjean, 1996; Walley *et al.*, 1995). In the current experiment we used two sizes, a long and a short one, both of which did not cover the entire original utterance. Participants had to make a binary decision about

the setting from which the fragment originated (i.e., final or not final).

A. Method

1. Stimuli

The stimuli for Experiment II were selected from the utterances of the same eight speakers which were used in Experiment I. For each of these speakers we randomly extracted answers from their original set of answers (see Sec. II), and constructed two types of fragments from these: short ones, consisting of one word, and long ones, consisting of two words. Orthogonal to this, half of the fragments were from a final (end-of-utterance) and half from a nonfinal position. In the same way as for Experiment I, we made sure that participants could not pick up on lexical cues for their final/nonfinal classifications.

For each of the eight speakers, we created four short pairs (final/nonfinal) and four long pairs of fragments, where the short fragments always consisted of the last word of the corresponding long (two word) fragment. Naturally, the final pairs were always selected from the tail of the list, while the nonfinal pairs were selected from varying positions in the list. The length of the original context surrounding a fragment was more or less balanced, with a small majority of fragments extracted from answers consisting of five words.

To guarantee the understandability of the fragments and to make sure they were comparable across conditions, the fragments were selected such that they included a naturally occurring pause after the last word of the fragment (when it was a nonfinal fragment), or a pause after the end of the original answer (when it consisted of the final part of an answer). The fragments were always cut in such a way that the pauses in the corresponding one word and two word stimuli lasted equally long, to make sure that the length of the pause (which, as noted in Sec. I, is an important signal for end of utterance) could not be used as a cue for classification.

As for Experiment I, all fragments were stored in three ways: AO, VO, or AV. Therefore, in total 128 stimuli were created for each modality: 8 speakers $\times 2$ lengths (short-long) $\times 2$ types (nonfinal and final) $\times 4$ fragments.

2. Participants

The participants consisted of a group of 60 native speakers of Dutch; 25 male and 35 female, between 20 and 56 years old. None of them participated as a speaker in the data collection phase nor as a participant in Experiment I, and none was involved in audio-visual speech research.

3. Procedure

Participants were given a simple classification task: they were told to determine for each fragment whether it marked the end of a speaker's utterance or not. Again, stimuli were presented in three conditions: an AV, an AO, and a VO, which were presented to participants in the same format as in Experiment I, but this time in a between-participants design.

Each condition consisted of two parts: one part for the short (one word) fragments and one part for the long (two word) fragments. The order in which participants passed the two different parts was systematically varied. For each part, two lists were created with a different random order. Participants were exposed to either the A versions or the B versions of a list. Therefore, each participant passed the items in a different random order in each part, and since the order in which participants underwent the short and long fragments part was also systematically varied, potential learning effects could be compensated for.

Each condition was preceded by a short practice session, consisting of two stimuli (different from the experimental stimuli), so that participants could get used to the type of tasks and stimuli. The general procedure was the same as for Experiment I.

TABLE III. For each factor, the levels of the factor, the percentage of correct judged utterances with standard errors, and 95% confidence intervals are given.

Factor	Level	Percent correct (%)	Std. error	95% CI
Fragment type	NF	80.8	0.11	(78.6,83.0)
	F	75.2	0.12	(72.9,77.7)
Stimulus length	Short	75.1	0.09	(73.3,77.0)
	Long	81.0	0.07	(79.5,82.3)
Modality	AV	84.7	0.11	(82.5,86.9)
	VO	75.7	0.11	(73.6,77.9)
	AO	73.6	0.11	(71.5,75.8)

4. Statistical analyses

Tests for significance were performed with a repeated measures ANOVA with speaker (eight levels), stimulus length (short: one word, long: two words), and fragment type (not final, final) as within-subjects factors and modality (VO, AO, AV) as a between-subjects factor (mixed design) and with the percentage of correct classifications over the four fragments as the dependent variable (recall that for each speaker four short and long pairs of final and nonfinal stimuli were selected). Mauchly's test for sphericity was used to test for homogeneity of variance, and when this test was significant or could not be computed, we applied the Greenhouse–Geisser correction on the degrees of freedom. For the purpose of readability, we report the normal degrees of freedom in these cases. The Bonferroni correction was applied for multiple pairwise *post hoc* comparisons, and contrasts were computed in several cases.

B. Results

Table III gives the overall results for three factors of interest, i.e., fragment type, stimulus length, and modality. According to the ANOVA all three factors had a significant influence on the classification. First, consider the main effect of fragment type $[F(1,57)=7.855, p<0.01, \eta_p^2=0.121]$. It appears that judging nonfinality is somewhat easier than judging finality (80.8% versus 75.2%), but overall it is clear that the vast majority of fragments are classified correctly.

Stimulus length also had a significant influence $[F(1,57)=28.800, p<0.001, \eta_p^2=0.336]$. Inspection of Table

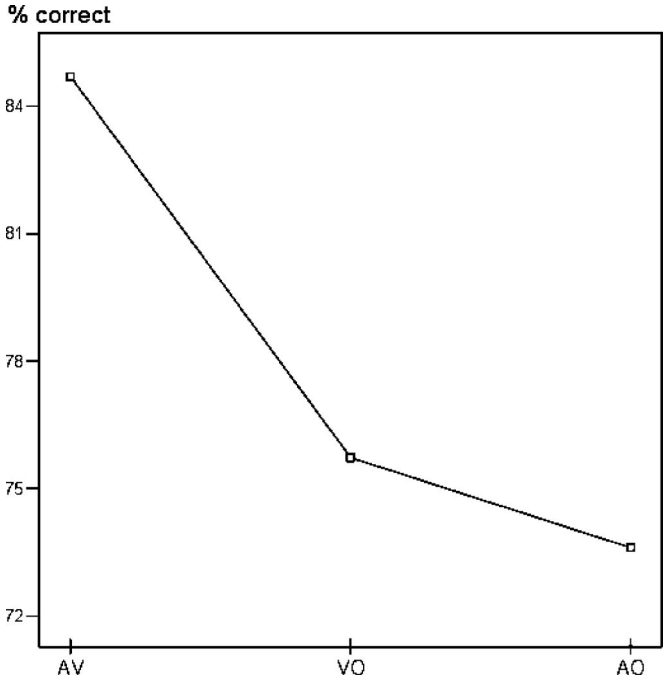


FIG. 4. Percentage of correct answers in the audio-visual (AV), vision-only (VO), and audio-only (AO) conditions.

III reveals that short (one word) fragments are somewhat more difficult than longer (two word) fragments.

The most interesting main effect is that of modality, which was significant as well [$F(2,57)=29.475, p<0.001, \eta_p^2=0.508$]. It is interesting to note that both unimodal conditions yield around 75% correct classifications (75.7 for the vision-only condition and 73.6 for the audio-only condition), and that both are clearly outperformed by the bimodal, audio-visual condition (with 84.7% correct). *Post hoc* analyses showed that there was a significant difference between the audio-visual and the vision-only condition ($p<0.001$), and between the audio-visual and the audio-only condition ($p<0.001$). The vision-only and the audio-only condition did not, however, differ significantly ($p=0.54$). This pattern of results is visualized in Fig. 4.

Besides the main effects for the three factors listed in Table III, the factor speaker also had a significant main effect [$F(7,399)=52.375, p<0.001, \eta_p^2=0.48$]. As can be seen in Table IV, the total number of correct classifications differs per speaker, ranging from 63% correct for speaker JB to 87.8% for speaker SS. *Post hoc* analyses showed that this difference was significant ($p<0.001$). Various other pairwise comparisons of speakers were significant as well, and this shows that there are overall substantial differences between speakers in end-of-utterance signaling. It is rather interesting to observe that the scores per speaker may differ across conditions. Indeed, a significant two-way interaction was found between speaker and modality [$F(7,399)=14.764, p<0.001, \eta_p^2=0.341$]; in Table IV it can be seen that, for instance, speaker BB apparently offers clearer visual than auditory cues, as the percentage of correctly classified stimuli for this speaker drops considerably in the AO condition. This is different for speaker MG, for instance, who seems to send more useful auditory cues (in her case the classification scores drop in the VO condition). Simple contrasts showed that this difference was significant [$F(2,57)=78.839, p<0.001, \eta_p^2=0.734$].

TABLE IV. For each speaker, the total percentage of correctly judged utterances, and the percentage of correctly judged utterances as a function of the three modalities.

Speaker	AV	VO	AO	Total
BB	86.5	86.5	56.8	76.7
BK	74.1	74.4	59.3	69.3
ED	90.6	73.3	77.7	80.5
JB	64.7	57.5	66.9	63.0
MG	86.6	68.1	86.0	80.2
MP	85.9	76.7	76.2	79.6
MS	93.1	87.2	81.0	87.1
SS	96.2	82.0	85.0	87.8

TABLE V. For each modality, the percentage of correctly judged utterances, as a function of stimulus length (one or two words) and fragment type (nonfinal and final).

Length	Finality	AV	VO	AO	Total
1	NF	81.8	76.2	69.7	75.9
1	F	83.1	73.6	66.0	74.3
Subtotal		82.5	74.9	67.9	
2	NF	89.4	82.6	85.2	85.7
2	F	84.5	70.6	73.6	76.2
Subtotal		86.9	76.6	79.4	
Total		84.7	75.7	73.6	

In addition, a significant two-way interaction was found between fragment type and stimulus length [$F(1,57)=11.317, p<0.01, \eta_p^2=0.166$]. This interaction can also be explained by looking at Table V, where it can be seen that for the nonfinal fragments, the longer stimuli evoked more correct answers (85.7%) than the short stimuli (75.9%), while for the final fragments the stimulus length makes almost no difference (74.3% versus 76.2%, respectively).

Table V also illustrates a second significant two-way interaction between stimulus length and modality [$F(2,57)=6.889, p<0.01, \eta_p^2=0.195$]. As expected, for both stimulus lengths, the audio-visual modality is the easiest one. For the short fragments, the audio-visual modality (82.5% correct answers) is followed by the visual modality (74.9%), and subsequently the auditory modality (67.9%). A *post hoc* test within the short word fragments revealed that all pairwise comparisons are statistically significant (AV-VO, $p<0.01$, AV-AO $p<0.001$, and VO-AO, $p<0.05$). However, for the long fragments, the audio-visual modality (86.9% correct answers) is followed by the auditory modality (79.4%), and subsequently the visual modality (76.6%). A *post hoc* within the long fragments revealed that all pairwise comparisons differ at the $p<0.001$ level, with the exception of the difference between VO and AO which is not significant. No other significant interactions were found.

C. Summary

The classification experiment reveals that speakers can make the best end-of-utterance classifications for bimodal,

audio-visual stimuli. It is interesting to observe that the numerically lowest scores are obtained for the audio-only condition, which has received the most attention in the literature. The vision-only results are somewhat better, which shows that visual cues to end of utterance are indeed useful for participants. Besides the modality effects, some other interesting results were obtained. A small response bias was found for nonfinal fragments, so that nonfinal fragments are slightly more often classified correctly. For the nonfinal fragments, the longer stimuli evoked more correct answers than the short stimuli, while for the final fragments the stimulus length makes almost no difference. Finally, the classification scores were found to vary per speaker, both overall and as a function of modality.

V. GENERAL DISCUSSION AND CONCLUSION

The fact that speakers use auditory cues (intonation, pausing, rhythm, etc.) which indicate that they are approaching the end of their utterance is well established (e.g., de Pijper and Sanderman, 1994; Price *et al.*, 1991; Swerts *et al.*, 1994a; 1994b; Wightman *et al.*, 1992). Various researchers have pointed out that speakers may also employ visual cues (such as posture, head movements, or gaze) for this purpose (e.g., Argyle and Cook, 1976; Cassell *et al.*, 2001; Nakano *et al.*, 2003; Vertegaal *et al.*, 2000). While the auditory cues have been studied from a perceptual perspective as well, comparable studies addressing the perception of visual cues (or the audio-visual combination) for end-of-utterance detection are thin on the ground. This naturally raises the question which modalities people actually employ to determine whether a speaker is at the end of an utterance and what the effect is of combining information from different modalities. In order to answer these questions, we first collected utterances in a semispontaneous way using a new experimental paradigm eliciting target list answers of three or five words long, making sure that target words could occur at the beginning, middle, or end of the list. On the basis of these utterances, two perception experiments were carried out.

As a first exploration, we performed a reaction time experiment in which participants were confronted with utterances, taken out of their original interview context to make sure that participants could not rely on lexical cues, and presented in three formats: VO, AO, or AV. The task for participants was to indicate as soon as possible when the speaker reached the end of his or her current utterance. It was found that participants could do this most quickly in the bimodal, audio-visual condition, followed (with a relative small, nonsignificant margin) by the audio-only condition, and with the slowest responses in the vision-only condition.

To find out how participants respond to stimuli in the respective conditions without any cues that participants might relate to (non)finality, we also performed a baseline reaction time measurement using artificially created static stimuli. Even though these artificial stimuli are of necessity not fully comparable with the real, experimental stimuli, comparing the experimental scores with those obtained in the baseline reveals some suggestive differences. It is interesting to observe that in the baseline condition, the audio-visual stimuli led to the slowest responses. That RTs for the AV condition are slower in the baseline than in the actual experiment may be explained by the thesis that when two different modalities (which contain no cues when their presentation will end) are offered at the same time, they will produce a cognitive overload because two sources of information have to be processed instead of one (Doherty-Sneddon *et al.*, 2001). However, when two modalities are presented in a situation where the information does contain predictive cues, as in the nonbaseline condition, the different modalities might serve as sources providing complementary information, and thus can help each other in resolving ambiguous slots in the stream of speech (compare Kim *et al.*, 2004; Schwartz *et al.*, 2004).

In general, the responses to the baseline stimuli were substantially faster than the responses in the nonbaseline conditions. This is in line with various reaction time studies concluding that a complex stimulus leads to slower reaction times (e.g., Brebner and Welford, 1980; Luce, 1986; Teichner and Krebs, 1974). Since the baseline stimuli are essentially static, without any variations that might be informative for end-of-utterance detections, there is much less information to process than in the experimental stimuli.

It was also interesting to see that the five word stimuli lead to quicker responses than the three word ones, which is in line with the studies of Carlson *et al.* (2005) and Swerts and Geluykens (1994). Again, this result is also consistent with findings from the literature on reaction time studies. Froeberg (1907), for instance, already found that longer visual stimuli elicit faster reaction times than stimuli of a shorter duration, and Wells (1913) found the same for auditory stimuli. In general, it is known that stimulus duration has a clear impact on reaction times (e.g., Ulrich *et al.*, 1998). Moreover, in this particular setup, the five word stimuli may also simply contain more potential finality cues than the three word stimuli, which would be an additional explanation for the fact that five word stimuli result in quicker responses than three word ones.

The results from the first experiment cannot be used to rule out the possibility that auditory information is sufficient for end-of-utterance detection, since it did not result in a significant difference between the audio-visual and the audio-only condition. Therefore a second experiment was conducted, to get more insight into how participants respond to stimuli in the different modalities. In this experiment participants were offered short (one word) and long (two word) fragments which either did or did not mark the end of an utterance, and participants had to classify these as final or nonfinal. In this experiment the bimodal presentation format gave significantly better results than the unimodal ones: when participants have access to both auditory and visual cues they make more adequate classifications than in situations where they only have information from one modality at their disposal. It was interesting to observe that overall most mistakes were made in the audio-only condition, i.e., the situation which has received the most attention in the literature so far, although the difference between the respective unimodal conditions was not statistically significant. Two possible explanations can be given for the superiority of the

audio-visual stimuli in this particular experiment. First, a combined audio-visual presentation format clearly offers more cues than a presentation in a single modality. But we have also seen that speakers differ in which signals they give, with some speakers showing more visual cues and others more auditory ones. Clearly, this also speaks in favor of a bimodal presentation.

In addition a slight response bias was found for nonfinal fragments, with nonfinal fragments more often classified correctly than the final ones. And for the nonfinal fragments, it was found that the longer stimuli were more often classified correctly than the shorter ones, while stimulus length did not have an effect on the final fragments. This suggests that when finality cues are available, it makes no difference whether the fragment is short or long, but when finality cues are not available, participants need longer fragments to make a decision. This could be caused by the fact that finality is displayed in local cues, thus in the last part of a fragment, just before it stops. In contrast, when no local finality cues are displayed, people need to base their decision more on global cues. In general, it is a well-known finding in cognitive psychology that it is easier to determine whether a cue is present than to decide that something is not there (e.g., Hearst, 1991).

It is also noteworthy that the longer fragments are better classified than the short fragments in the audio-only condition, which suggests that the finality cues in speech seem to be more global in nature, and hence that participants can make better judgments for longer fragments when more of these global cues are available. For the vision-only condition, length does not appear to have an influence, which suggests that the visual cues may be more local. Notice that this would also offer an explanation for the fact that the audio-only condition outperforms the vision-only condition in Experiment I, but not in Experiment II. Since the stimuli in the second experiment were overall shorter (consisting of one or two words) than those in the first experiment (which consisted of entire utterances of three or more words), the participants in the second experiment could not use the spoken global cues to the full effect.

The focus in this paper has been on a perceptual comparison of the cue value of different modalities for signaling end of utterance. However, it would be interesting to see which auditory and visual behaviors might have served as cues in both experiments. To gain some insight into this, we annotated for both the final and the nonfinal stimuli the 50% that received the best classification scores in Experiment II. In particular, we concentrated on those cues that are known from the literature (see Sec. I), and that could clearly and consistently be determined on the basis of visual or auditory inspection of our stimuli. The following auditory cues were labeled:

(1) **Boundary tone:** whether a fragment ends in a low (*L*), medium (*M*), or high boundary tone (*H*); and

TABLE VI. Representative stills illustrating the annotated visual features. Notice that various stills contain multiple features, since cues may cooccur. For example, the female speaker with her mouth open also moves her head and eyes away.

Label	Example	
Brows [up]		
Eyes [away]		
Mouth [open]		
Head [away]		
Posture [away]		

(2) **Creaky voice:** whether a stimulus contains some creaky fragments.

In both cases, the annotation was determined by perceptual judgments, and performed by professional intonologists. The distinction between high, mid, and low boundary tones was determined by comparing the tonal pattern in the final syllables of the fragment to the pitch range of the preceding part. If the final stretch of speech was clearly below or above the preceding pitch range, it would be categorized as either low or high, whereas a tone inbetween those two extremes would get a mid label.

In the visual domain, the following features were labeled (Table VI contains representative stills for each of the visual features):

TABLE VII. The annotation as a function of fragment type (nonfinal and final).

Modality	Feature	Setting	NF	*F*	Total
Auditory	Boundary tone	*H*	0	6	6
		M	13	2	15
		L	3	8	11
	Creaky voice		5	5	10
Visual	Brows	Up	11	8	19
		Down	3	4	7
	Eyes	Blinking	7	12	19
		Away	23	8	31
		Back	3	13	16
	Mouth	Open	6	2	8
		Closed	0	4	4
	Head	Nodding	12	21	33
		Away	10	4	14
		Back	1	4	5
	Posture	Away	7	6	13
		Back	0	2	2

(1) **Brows:** whether the eyebrows are raised (up) or lowered (down);
(2) **Eyes:** whether the eyes of the speaker are turned away from the camera (away), or whether the speaker returns his/her gaze towards the camera (back); we also labeled cases where a speaker was blinking;
(3) **Mouth:** whether the mouth at the end of the stimulus is closed or open;
(4) **Head:** whether the speaker turns his/her head away from the camera during the answer, or moves the head back to the camera; moreover, we also labeled cases where the speaker makes a nodding movement during the fragment; and
(5) **Posture:** whether the speaker changes his/her posture away from the camera, or rather moves his/her body back towards the camera.

The cues were always labeled blind to condition, in order to avoid circularity in their annotation. Table VII gives the overall results for the factors of interest, split by the two possible modalities, i.e., auditory (boundary tones, creaky voice) and visual (brows, eyes, mouth, head, posture) as a function of fragment type (nonfinal of final).

In the auditory domain, it can be observed that the midending tones are more typical for the nonfinal fragments, while both high and low boundary tones occur more often at the end of final fragments. This result is in line with many previous studies which show that a clearly low or high tone (such as in question intonation) may signal the end of an utterance, whereas a midtone serves to cue continuity (e.g., Caspers 1998; Silverman and Pierrehumbert, 1990). At first sight, the presence of a creaky voice (which in our stimuli rarely happens in the first place) does not appear to be related to finality or nonfinality, but a closer inspection of the stimuli reveals that all the noncreaky fragments occur in cases where speakers used a midtone, while the creaky fragments only occur when speakers produce a high or low tone, so that creakiness may serve as an extra cue to reinforce the finality/nonfinality marking of boundary tones. With respect to the visual features, Table VII suggests that there is a clear tendency for speakers to divert their eyes and head in nonfinal fragments, while they return eyes, head, and also posture in the final fragments. Additionally, there is a trend for the mouth to be still open when a fragment has not yet been finished (even though the speaker is not speaking), whereas a mouth is more often closed at the end of a final fragment. Also, final fragments display relatively more cases of blinking and nodding, while the brows tend to be up or down at the end of nonfinal versus final fragments, respectively.

There are also many individual differences between speakers. In the annotated utterances, speakers produce almost 23 cues on average, but there are clear differences. Speaker JB for instance, produces only 14 visual cues to signal finality, which is consistent with the fact that speaker JB was most difficult to classify in Experiment II. On the other hand, speaker JB tends to use low boundary tones more often than other speakers. This may account for the observation, for Experiment I, that participants took relatively long to respond to JB's stimuli in the vision-only modality, and were rather quick for this speaker in the audio-only and audio-visual conditions. Speaker SS, to give a second example, is visually the most expressive (33 visual cues) and indeed her stimuli lead to the overall quickest responses in Experiment I, and to the most correct classifications in Experiment II. Apart from the fact that some speakers display more cues than others, some speakers also tend to display different cues than other speakers. For example, on the visual level, while most speakers return their gaze in a final position, some speakers (e.g., ED) do not return their gaze but instead nod more often in the final position.

This small scale annotation reveals that many of the cues mentioned in Sec. I indeed occur in the stimuli, and it seems likely that participants made their classification on the basis of these various cues. In future research, it would be interesting to find out how the different audio-visual features discussed above are distributed over the whole utterance. It has been argued (Argyle and Cook, 1976) that an utterance consists of different phases, i.e., a starting phase, a middle phase, and a closing phase, which are connected to patterns in eye gaze (see also Cassell *et al.*, 2001 for similar kinds of observations in other bodily gestures.) It remains to be seen whether such patterns are also true for other visual features, and how these relate to more global auditory cues, such as declination or rhythmic patterns. It would also be interesting to test the relative importance of the various auditory and visual cues in followup experiments.

In sum: our study, using a reaction-time experiment and a classification task, has revealed that subjects are sensitive both to auditory and visual signals when they need to estimate whether or not a speaker utterance has ended. While both modalities separately contain cues that enable subjects to make reliable finality judgments, it turns out that a bimodal, audio-visual condition leads to the most accurate results. The relative cue value of the two unimodal conditions depends on the experiment, where auditory cues were more important in the RT experiment, and visual cues in the classification task. In addition, its relative importance also differs

between stimuli from different speakers, due to the fact that some speakers display more auditory cues, and others more visual ones.

ACKNOWLEDGMENTS

This research was conducted as part of the VIDI project "Functions of Audiovisual Prosody" (FOAP), sponsored by the Netherlands Organization for Scientific Research (NWO). We thank Lennard van de Laar for various kinds of technical assistance, Carel van Wijk for statistical advice, and Jean Vroomen for allowing us to make use of the Pamar software. We greatly benefitted from the comments of three anonymous reviewers on a previous version of this paper.

Argyle, M., and Cook, M. (**1976**). *Gaze and Mutual Gaze* (Cambridge University Press, Cambridge, UK).

Beattie, G. W., Cutler, A., and Pearson, M. (**1982**). "Why is Mrs. Thatcher interrupted so often?" Nature (London) **300**, 744–747.

Bertelson, P., Vroomen, J., and de Gelder, B. (**2003**). "Visual recalibration of auditory speech identification: A McGurk aftereffect." Psychol. Sci. **14**, 592–597.

Brebner, J., and Welford, A. (**1980**). "Introduction: An historical background sketch," in *Reaction Times* edited by A. Welford (Academic, New York), pp. 1–23.

Carlson, R., Hirschberg, J., and Swerts, M. (**2005**). "Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates." Speech Commun. **46**, 326–333.

Caspers, J. (**1998**). "Who's next? The melodic marking of questions vs. continuation in Dutch." Lang Speech **41**, 375–398.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., and Rich, C. (**2001**). "Non-verbal cues for discourse structure." *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*, Toulouse, France, July 9–11, pp. 114–123.

Couper-Kuhlen, E. (**1993**). *English Speech Rhythm* (Benjamins, Philadelphia).

de Pijper, J. R., and Sanderman, A. A. (**1994**). "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues." J. Acoust. Soc. Am. **96**, 2037–2047.

de Ruiter, J. P., Mitterer, H., and Enfield, N. (**2006**). "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation." Language **82**, 515–535.

Doherty-Sneddon, G., Bonner, L., and Bruce, V. (**2001**). "Cognitive demands of face monitoring: Evidence for visuospatial overload." Mem. Cognit. **29**, 909–917.

Doughty, M. J. (**2001**). "Consideration of three types of spontaneous eyeblink activity in normal humans: During reading and video display terminal use, in primary gaze, and while in conversation." Optom. Vision Sci. **78**, 712–725.

Duncan, S. (**1972**). "Some signals and rules for taking speaking turns in conversations." J. Pers Soc. Psychol. **23**, 283–292.

Ekman, P. (**1979**). "About brows: Emotional and conversational signals," in *Human Ethology: Claims and Limits of a New Discipline*, edited by M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (Cambridge University Press, Cambridge, UK), pp. 169–202.

Froeberg, S. (**1907**). "The relation between the magnitude of stimulus and the time of reaction." Arch. Psychol. (Frankf) **8**.

Goodwin, C. (**1980**). "Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning." Sociological Inquiry **50**, 272–302.

Grosjean, F. (**1983**). "How long is the sentence? Prediction and prosody in the on-line processing of language." Linguistics **21**, 501–529.

Grosjean, F. (**1996**). "Gating." Lang. Cognit. Processes **11**, 597–604.

Hearst, E. (**1991**). "Psychology and nothing." Am. Psychol. **79**, 432–443.

Kendon, A. (**1967**). "Some functions of gaze-direction in social interaction." Acta Psychol. **26**, 22–63.

Kim, J., Davis, C., and Krins, P. (**2004**). "Amodal processing of visual speech as revealed by priming." Cognition **93**, B39–B47.

Kobayashi, H., and Kohshima, S. (**1997**). "Unique morphology of the human eye." Nature (London) **387**, 767–768.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (**1998**). "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs." Lang Speech **41**, 295–321.

Krahmer, E., and Swerts, M. (**2004**). "More about brows," in *From Brows to Trust: Evaluating Embodied Conversational Agents*, edited by C. Pelachaud and Zs. Ruttkay (Kluwer, Dordrecht), pp. 191–216.

Leroy, L. (**1984**). "The psychological reality of fundamental frequency declination." *Antwerp Papers in Linguistics* (Antwerp University Press, Antwerp, Belgium), Vol. **40**.

Levinson, S. (**1983**). *Pragmatics* (Cambridge University Press, Cambridge, UK).

Luce, R. (**1986**). *Response Times: Their Role in Inferring Elementary Mental Organization* (Oxford University Press, New York).

Maynard, S. K. (**1987**). "Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation." J. Pragmat. **11**, 589–606.

Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (**2003**). "Towards a model of face-to-face grounding." *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 7–12, pp. 553–561.

Novick, D. G., Hansen, B., and Ward, K. (**1996**). "Coordinating turn-taking with gaze." *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA, October 3–6, pp. 1888–1891.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, S. (**1991**). "The use of prosody in syntactic disambiguation." J. Acoust. Soc. Am. **90**, 2956–2970.

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (**2004**). "Seeing to hear better: Evidence for early audio-visual interactions in speech identification." Cognition **93**, B69–B78.

Silverman, S., and Pierrehumbert, J. (**1990**). "The timing of prenuclear high accents in English," *Laboratory Phonology: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman (Cambridge University Press, Cambridge, UK), Vol. **I**, pp. 71–106.

Swerts, M. (**1997**). "Prosodic features at discourse boundaries of different strength." J. Acoust. Soc. Am. **101**, 514–521.

Swerts, M. (**1998**). "Filled pauses as markers of discourse structure." J. Pragmat. **30**, 485–496.

Swerts, M., Bouwhuis, D., and Collier, R. (**1994a**). "Melodic cues to the perceived finality of utterances." J. Acoust. Soc. Am. **96**, 2064–2075.

Swerts, M., Collier, R., and Terken, J. (**1994b**). "Prosodic predictors of discourse finality in spontaneous monologues." Speech Commun. **15**, 79–90.

Swerts, M., and Geluykens, R. (**1994**). "Prosody as a marker of information flow in spoken discourse." Lang Speech **37**, 21–43.

Teichner, W., and Krebs, M. (**1974**). "Laws of visual choice reaction time." Psychol. Rev. **81**, 75–98.

Ulrich, R., Rinkenauer, G., and Miller, J. (**1998**). "Effects of stimulus duration and intensity on simple reaction time and response force." J. Exp. Psychol. Hum. Percept. Perform. **24**, 915–928.

Vertegaal, R., Slagter, R., van de Veer, G., and Nijholt, A. (**2000**). "Why conversational agents should catch the eye." *Proceedings of the Internation Computer-Human Interaction Conference (CHI)*, The Hague, The Netherlands, April 1–6, pp. 257–258.

Walley, A. C., Michela, V., and Wood, D. (**1995**). "The gating paradigm: Effects of presentation format on spoken word recognition by children and adults." Percept. Psychophys. **57**, 343–351.

Ward, N., and Tsukahara, W. (**2000**). "Prosodic features which cue back-channel responses in English and Japanese." J. Pragmat. **23**, 1177–1207.

Wells, G. (**1913**). "The influence of stimulus duration on RT." Psychol. Monogr. **15**.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (**1992**). "Segmental durations in the vicinity of prosodic phrase boundaries." J. Acoust. Soc. Am. **91**, 1707–1717.

Absorption of reliable spectral characteristics in auditory perception

Michael Kiefte[a)]
School of Human Communication Disorders, Dalhousie University, Halifax, Nova Scotia B3H 1R2, Canada

Keith R. Kluender
Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706-1696, USA

(Received 14 April 2006; revised 5 October 2007; accepted 12 October 2007)

Several experiments are described in which synthetic monophthongs from series varying between /i/ and /u/ are presented following filtered precursors. In addition to F_2, target stimuli vary in spectral tilt by applying a filter that either raises or lowers the amplitudes of higher formants. Previous studies have shown that both of these spectral properties contribute to identification of these stimuli in isolation. However, in the present experiments we show that when a precursor sentence is processed by the same filter used to adjust spectral tilt in the target stimulus, listeners identify synthetic vowels on the basis of F_2 alone. Conversely, when the precursor sentence is processed by a single-pole filter with center frequency and bandwidth identical to that of the F_2 peak of the following vowel, listeners identify synthetic vowels on the basis of spectral tilt alone. These results show that listeners ignore spectral details that are unchanged in the acoustic context. Instead of identifying vowels on the basis of incorrect acoustic information, however (e.g., all vowels are heard as /i/ when second formant is perceptually ignored), listeners discriminate the vowel stimuli on the basis of the more informative spectral property. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2804951]

PACS number(s): 43.71.An, 43.66.Ba [MSS] Pages: 366–376

I. INTRODUCTION

Identification of vowels may rely upon both spectrally narrow and broad properties (see Rosner and Pickering, 1994, for a review). Kiefte and Kluender (2005) recently found that spectral tilt, or the relative balance between low- and high-frequency energy, and spectral peaks, corresponding to formants, play a strong role in the perception of isolated synthetic monophthongs. This result is consistent with either whole-spectrum (e.g., Bladon and Lindblom, 1981) or center-of-gravity (e.g., Beddor and Hawkins, 1990) hypotheses of vowel perception. This result was also predicted by results obtained by Ito *et al.* (2001), who found that synthetic vowels could be accurately identified even when selected spectral peaks were excised leaving spectral tilt as the only remaining spectral property to discriminate back from front vowels. Kiefte and Kluender's result was nevertheless surprising. Previous research had indicated that spectral-peak amplitude, of which spectral tilt is a contributor, was not important in speech perception (Joos, 1948; Carlson *et al.*, 1970; Chistovich, 1971; Assmann, 1991) and that the frequency location of the lowest spectral peaks was the primary discriminant of vowel identity (e.g., Assmann and Summerfield, 1989). It also had been suggested that spectral tilt only plays a role in indicating emotional state, speaker identity, or transmission channel characteristics such as room acoustics (Fant, 1973; Klatt, 1982, 1986).

Kiefte and Kluender (2005) concluded that results from synthetic monophthongs were not generalizable because such sounds have limited resemblance to real-world utterances. Because vowels are rarely monophthongal in English (Hillenbrand *et al.*, 1995), and because results from more natural-sounding diphthongs did not show an effect for spectral tilt, Kiefte and Kluender claimed that, while spectral tilt could be used to discriminate vowels under unusual circumstances (i.e., synthetic monophthongs), spectral peaks were a much more reliable property. Because spectral tilt can fluctuate with external or paralinguistic factors, it was considered unreliable for the purposes of vowel identification.

In their discussion, Kiefte and Kluender (2005) offered possible physiological mechanisms that might favor spectral peaks over tilt as a correlate to vowel identity. One proposal was that spectral properties lose perceptual salience if they remain unchanged over relatively short periods of time as is the case with synthetic monophthongs. Because spectral peaks are rapidly changing in naturally produced speech, they should not be vulnerable to this perceptual degradation. However, spectral tilt, if largely due to paralinguistic or external factors, is expected to be relatively sustained and thus lose contrastiveness across time. Ultimately, Kiefte and Kluender concluded that listeners normally respond predominantly to changing acoustic properties, and responses to synthetic monophthongs reflected perception of impoverished acoustic cues.

It is both true and efficient that, in general, sensorineural systems respond to change and to little else. Perceptual systems do not record absolute level, be it loudness, pitch, brightness, or color. By definition, information for perception is specified by change, as there is no new information in properties that are static or redundant (Kluender and Alex-

a)Electronic mail: mkiefte@dal.ca

ander, 2007; Kluender and Kiefte, 2006; Shannon and Weaver, 1949). This sensitivity to change requires that systems register and compensate for predictable characteristics in order to be most responsive to unpredictable new information. This necessitates a strong role for context, with perception always relative to a listening context. Properties that are shared can be relatively neglected while properties that contrast with context are emphasized in perception.

That context plays a strong role in perception of speech has been demonstrated numerous times (e.g., Ladefoged and Broadbent, 1957; Broadbent and Ladefoged, 1960; Darwin *et al.*, 1989; Watkins, 1991;Watkins and Makin, 1994; Summerfield *et al.*, 1984; Watkins and Makin, 1996a,b). In their classic study on context effects in vowel perception, Ladefoged and Broadbent showed that identification of a target vowel from a synthetic /bɪt/–/bɛt/ series could be altered by adjusting the average of first formant (F_1) frequency within a preceding carrier sentence. When average F_1 frequency in the carrier sentence was raised, /bɪt/ was heard more often by listeners without any change in the acoustic properties of the target vowel. This result is consistent with perception of a lower frequency F_1 in the target vowel when following higher F_1 in the carrier. This result was also demonstrated with naturally produced speech (Ladefoged, 1989) and was replicated by Watkins and Makin (1994).

Many researchers have viewed context-dependent vowel identification as a consequence of *extrinsic vowel specification* which partly solves the problem of between-speaker vowel category overlap in $F_1 \times F_2$ frequency space (e.g., Peterson and Barney, 1952). Because formant frequencies are produced over a range that is roughly determined by vocal tract size, any information regarding the range or average frequency of a particular formant for a particular speaker constrains the formant-frequency vowel space thereby aiding vowel identification (see Nearey, 1989 for a review).

However, Watkins and Makin (1994) demonstrated that these results could also be explained with reference to an "inverse-filtering heuristic" in which effects of long-term spectral characteristics are filtered out of the target phoneme before relevant acoustic properties are extracted. Similar results to those observed by Ladefoged and Broadbent (1957) could be obtained when a context sentence is filtered so that the long-term average spectrum matched that of either the low- or high-frequency average F_1 carrier sentences without directly altering formant center frequencies themselves. This work built upon experiments which demonstrated that the perception compensated for the long-term spectral characteristics of the acoustic context (Watkins, 1988). For example, if the context sentence was processed by a filter with spectral response of /ɪ/ minus the spectral response of /ɛ/, there was an increase in the number of /ɛtʃ/ responses in a /ɪtʃ/–/ɛtʃ/ series. Similar to results reported by Ladefoged and Broadbent, this observation is consistent with greater perceived contrast between target and context. However, contrast in this case is defined by long-term spectral properties instead of the speaker-dependent property of average F_1 frequency. Watkins and Makin (1994) argued that it was not the range of F_1 frequency that shifted responses, but rather the long-term spectrum of the context sentence. Therefore, when F_1 frequency was raised in the precursor sentence, the long-term spectrum was more similar to /ɛ/ resulting in more /ɪ/ responses.

Contrast effects in vowel perception were previously described by Summerfield *et al.* (1984). In their experiments, it was shown that listeners could identify a vowel from an equal-amplitude-harmonic stimulus if it was preceded by a stimulus with troughs substituted for peaks. The vowel perceived from the flat-spectrum stimulus corresponded to a peripheral auditory "negative afterimage" of the precursor (Summerfield *et al.*, 1987). Identification was due entirely to contrast. This result was further extended by Coady *et al.* (2003). While it is known that perception of stop-consonant place-of-articulation is dependent on the preceding vowel, Coady *et al.* showed that vowel complements (i.e., complex periodic stimuli with troughs similar to those used by Summerfield *et al.*) have an effect on consonant perception that is opposite that following vowel spectra with peaks.

However, Watkins (1991) demonstrated that there are potentially two sources of auditory compensation in perception. When speech-shaped noise carriers are used, negative-auditory-afterimage effects are either eliminated or greatly attenuated when carrier and target phoneme are presented to different ears or if a silent gap is introduced between carrier and target. However, contrast effects, although attenuated, remain significant if a speech-like precursor is used in these conditions instead. This suggests that a more central mechanism also plays a strong role in auditory compensation. Because auditory compensation effects also persist when carrier and target stimulus are presented with different interaural delays, Watkins concludes that this central mechanism operates in addition to a more peripheral mechanism responsible for the negative auditory afterimage.

If increasing spectral change between context and target shifts responses toward the contrasting vowels, then decreasing contrast should have the opposite effect. Darwin *et al.* (1989) showed that, if a precursor sentence was processed with the same filter that was used to alter relative amplitudes of first and higher formants, then the perceptual effect of this spectral distortion was attenuated. Any spectral properties that were unchanged (redundant) between context and target were effectively ignored.

Relating these results to perceptual effects of spectral tilt, it is expected that in naturally produced speech, the long-term relative balance between low- and high-frequency energy will be perceptually ignored in a manner similar to that demonstrated in the experiments described earlier. Because this particular spectral property has been attributed to either paralinguistic or external factors such as room acoustics or channel characteristics (Fant, 1973; Klatt, 1982, 1986), these properties should not fluctuate widely over short periods of time.

In contrast, spectral peaks change relatively rapidly and therefore provide more salient correlates to vowel identity. Because of this, spectral peaks may also be more resistant to auditory compensation. For example, while the visual system is remarkable in the ability to maintain color constancy over widely varying illumination spectra by compensating for

contextual illumination effects, it is not as successful when illumination spectra include local spectral prominences or peaks. For example, fluorescent lights (which have multiple spectral peaks) and narrow-band illumination such as that from mercury or sodium vapor lights (Boynton and Purl, 1989; Fieandt *et al.*, 1964) compromise the ability of viewers to maintain color constancy. One might suggest then that the auditory system is also less adept at compensating for local spectral perturbations. Isolated synthetic monophthongs, such as those used in experiments by Kiefte and Kluender (2005) do not have a natural acoustic context against which to contrast spectral properties. In these studies, listeners responded to a large number of stimuli over a relatively short period of time. The acoustic context in these experiments consisted of a rapid succession of acoustically stationary stimuli that varied in *both* spectral peaks *and* tilt. The contrast between successive presentations may have resulted in an exaggerated and unnatural effect for spectral tilt. However, when Kiefte and Kluender presented stimuli in which spectral peaks were allowed to vary in a manner more akin to naturally produced vowels, the perceptual effect of spectral tilt was largely attenuated. Therefore, it was suggested that spectral tilt is ignored in speech stimuli that more closely approximated naturally produced speech. If, however, spectral tilt is contrastive, as is the case for isolated synthetic monophthongs that vary in spectral tilt from trial to trial, spectral tilt may have an exaggerated influence on perception. For example, van Dijkhuizen *et al.* (1987) found that continuously modulating spectral tilt reduced speech reception threshold, perhaps introducing contrast in spectral tilt that encouraged perceptual errors.

Similar to Summerfield *et al.* (1984), Darwin *et al.* (1989), Watkins (1991), and Watkins and Makin (1994, 1996a,b), vowel sounds following different acoustic contexts were used in the present study to investigate how auditory systems may absorb predictable (i.e., redundant) spectral characteristics of acoustic context in order to be more sensitive to information-bearing characteristics of sounds. Although isolated synthetic monophthongs show significant effects for spectral tilt, it may be possible to cancel perceptual effects of gross spectral properties with an appropriately filtered context, thereby simulating the long-term spectral effects of para- or nonlinguistic factors such as vocal effort or emotional state. In contrast to work by Darwin *et al.*, speech-like precursors more akin to those used by Watkins and Watkins and Makin will be used. Although previous work with speech-like precursors also showed compensatory effects for filtered precursors, these studies did not distinguish between the perceptual effects of what Watkins and Makin (1996a) refer to as potentially "invariant" spectral properties, or formant peaks, versus gross spectral shape properties such as spectral tilt.[1] However, where previous studies used precursors processed with filters designed from the entire spectrum of one vowel minus that of another to maximize contrast effects within a vowel series, we will be using filters designed to target specific spectral properties. Filters used here correspond specifically to either the spectral tilt of the target vowel or to the F_2 peak. This is intended to evaluate the relative contributions of these acoustic properties when they form part of the long-term spectral characteristics of the sentence.

Because spectral tilt does not typically vary rapidly in naturally produced speech, it might be expected that this property is perceptually absorbed when there is no contrast between carrier and target vowel. In experiment 1, a matrix of synthetic, vowel-like stimuli (Kiefte and Kluender, 2005) that varied acoustically in center frequency of F_2 and gross spectral tilt, and varied perceptually from /u/ to /i/, was created. These stimuli were presented following a precursor sentence that was passed through the same filter that was used to alter the relative spectral tilt in the target vowel. Results show that the effect for spectral tilt for isolated vowel sounds observed by Kiefte and Kluender (2005) was greatly attenuated. In experiment 2, the precursor was instead processed by a single-pole filter with center frequency and bandwidth identical to that of the F_2 peak of the target vowel. Because formant frequencies do vary rapidly in naturally produced speech and are important acoustic properties for vowel identification, one might expect that this property is not as easily canceled. However, results from this experiment were almost orthogonal to those of experiment 1: There was no effect for F_2 frequency. Because the same basic precursor was used for all stimuli in the initial two experiments, and because it may be possible for listeners to infer effects of spectral distortion from differences between successive presentations of the same precursor, experiments 3 and 4 employ randomized, time-reversed precursors on each trial. Results from experiments 1 and 2 were replicated in these experiments.

II. EXPERIMENT 1

Ito *et al.* (2001) and Kiefte and Kluender (2005) showed that perception of synthetic monophthongs in isolation can be strongly influenced by spectral tilt—the relative balance between high- and low-frequency energy. However, because this gross spectral property is heavily influenced by nonphonetic factors, it is expected that perceptual effects of spectral-tilt may be attenuated when a target vowel is presented following an appropriately filtered context sentence.

A. Method

Listeners identified a matrix of 49 vowel sounds (7 formant frequencies by 7 spectral tilts) preceded by an acoustic context with reliable spectral properties. This precursor context was a synthesized rendition of the sentence "You will now hear the vowel…." Relative balance between low- and high-frequency energy in context sentences was then altered to match the gross spectral tilt of the following vowel. Listeners were asked to identify whether they heard the vowels as "ee" or "oo."

1. Subjects

Twelve native speakers of Midwestern American English were recruited from the Department of Psychology at

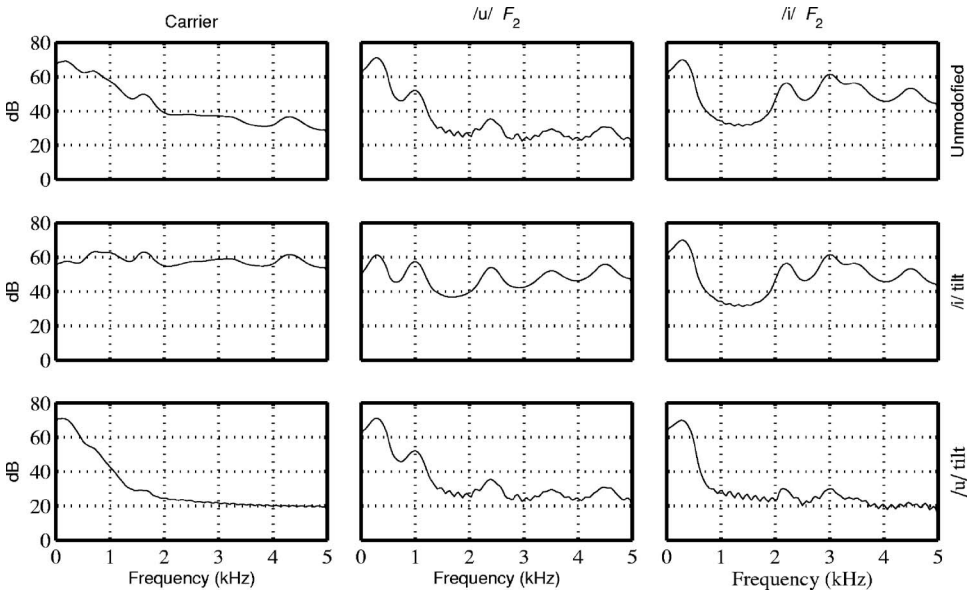


FIG. 1. Mean-square spectra of carrier and target stimuli used in experiment 1. The first row gives the spectra for the original, unfiltered carrier as well as the end points of the unmodified /i/-/u/ continuum. The /u/ in the second row has been filtered such that the spectral tilt, defined as dB/octave between the peaks of F_1 and F_3 matched that of /i/ while the carrier in the second row has been processed by the same filter. The /i/ in the second row with unmodified spectral tilt has been added for comparison. The /i/ in the third row has been filtered such that the spectral tilt matches that of /u/.

the University of Wisconsin–Madison. None reported any hearing impairment. Listeners received course credit for their participation.

2. Stimuli

Stimuli were identical to those used in experiment 2 of Kiefte and Kluender (2005). A seven-step series of 245 ms vowel stimuli ranging perceptually from /i/ to /u/ was generated using the cascade branch of a flexible implementation of Klatt's (1980) speech synthesizer (Kiefte *et al.*, 2002). Center frequency of F_2 ranged from 1000 to 2200 Hz in 200 Hz steps, while center frequency of F_3 varied collinearly with F_2 between 2400 and 3000 Hz in 100 Hz steps to encourage subjective naturalness. Center frequency of F_1 was fixed at 300 Hz, while F_4 and F_5 were held at 3500 and 4500 Hz, respectively. Formant values for series end points are similar to those for adult male speakers of American English as reported by Hillenbrand *et al.* (1995), and each end point was judged to be a reasonable approximation of either /i/ or /u/ by the authors. Fundamental frequency (f_o) was set at 100 Hz so formant center frequencies corresponded to harmonics. Voicing amplitude (AV) was set to 60 dB and reduced to 0 dB at the 5 ms frame starting at 240 ms. All remaining synthesis parameters were maintained at default values.

Spectral tilt was fully crossed with F_2 center frequencies. Spectral tilt for each vowel sound was adjusted to match that of every other vowel in the series for a total of 49 (7×7) stimuli. For the purposes of this study, spectral tilt was defined as the difference between F_1 and F_3 peak amplitudes in dB/octave. Amplitudes were determined analytically based on the source and cascade synthesis parameters for peak harmonics of F_1 and F_3. The dB/octave difference was imposed via a filter similar to that used by the KLSYN88 speech synthesizer to adjust source tilt (Klatt and Klatt, 1990). A single-zero filter was used instead of a single pole in those cases where spectral tilt needed to be increased instead of lowered (e.g., /u/ formants with /i/ spectral tilt). Bandwidth (BW) of the spectral tilt filter was determined by a simple recursive algorithm based on the desired attenuation of amplification of F_3 relative to F_1. The frequency of the pole or zero was constrained by $F=0.375\times\mathrm{BW}$ in order to achieve critical damping (Klatt and Klatt, 1990). All stimuli were normalized in rms amplitude.

The context sentence, "You will now hear the vowel…" was recorded from a male talker (MK) and was analyzed via a 16-coefficient linear predictive coding (LPC) which approximately captures the changing spectral peaks and valleys of the original utterance over time. The sentence was resynthesized by passing a harmonic source ($f_0=100$ Hz) through filters designed using the LPC coefficients resulting in a synthetic utterance that approximated the spectral and temporal characteristics of the original utterance, but with monotone pitch.

This context sentence was filtered (using either single pole or zero to match spectral tilt) in the same fashion as described earlier for the target vowel that followed, so the long-term spectral tilt of the precursor matched that of the target vowel (i.e., the same filter that altered the spectral tilt of the target vowel was used to process the precursor). Mean-square spectra of the carrier and target stimuli are given in Fig. 1.

3. Procedure

All contexts and vowel sounds were rms matched and presented at 72 dB SPL over headphones (Beyer-Dynamic DT-100). Stimulus presentation and response collection was under the control of an 80486-25 microcomputer. Following D/A conversion (Ariel DSP-16), stimuli were low-pass filtered (4.8 kHz cutoff frequency, Frequency Devices, No. 677), amplified (Stewart HDA4), and presented to subjects. Experiments were conducted with one to three listeners participating concurrently in single-subject sound-proof booths. Listeners identified vowel sounds by pressing buttons labeled "ee" and "oo" on a response box. Each of the 49 sentences was presented four times in each of two 15 min sessions for a total of 392 responses from each subject.

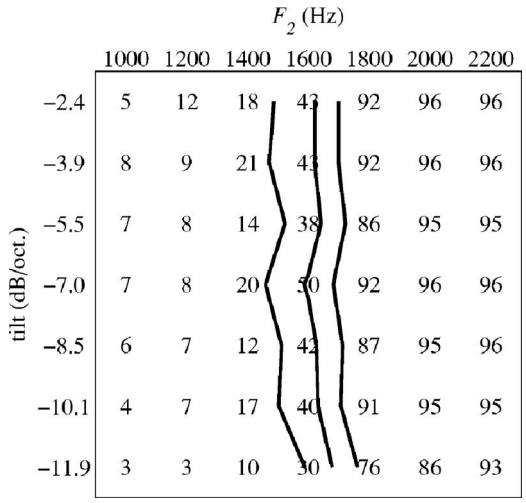


FIG. 2. /i/ responses pooled across speakers (out of a maximum of 96) from experiment 1. Contour lines indicate 30%, 50%, and 70% cutoffs. Steps along the ordinate are not evenly spaced as they correspond to stimulus steps which increase linearly in F_2 and F_3, but not linearly in spectral tilt.

B. Results

Pooled responses to the 49 vowel sounds following a precursor with matched spectral tilt are shown in Fig. 2 as a function of vowel F_2 frequency and spectral tilt in dB/octave.

Responses were modeled for each subject via logistic regression (McCullagh and Nelder, 1989) with both spectral tilt and F_2 frequency as covariates with the probability of /i/ response. In contrast with results reported by Kiefte and Kluender (2005) for isolated synthetic monophthongs, spectral tilt appears to play no role in vowel identification for this stimulus matrix following tilt-matched precursors. Fitted regression boundaries for each subject are given in Fig. 3.

Subject regression coefficients for both spectral tilt and F_2 frequency were analyzed via single-sample t-test to determine if either or both of these coefficients were significantly different from zero, thereby indicating a statistically significant effect for the parameter (Davis, 2002; Gumpertz and Pantula, 1989). As expected, the effect for F_2 frequency was highly significant ($t_{11}=4.69$; $p<0.001$). However, despite the apparently attenuated effect for spectral tilt in these data, tilt also was found to be significant ($t_{11}=2.47$; $p=0.02$). The effect size, or Cohen's d, for tilt was 0.71. Although this is considered a medium effect size (Cohen, 1988), it is much smaller than the effect size for F_2 frequency ($d=1.35$). For purposes of comparison, the effect size for spectral tilt observed in the original study with isolated synthetic monophthongs was $d=1.94$.

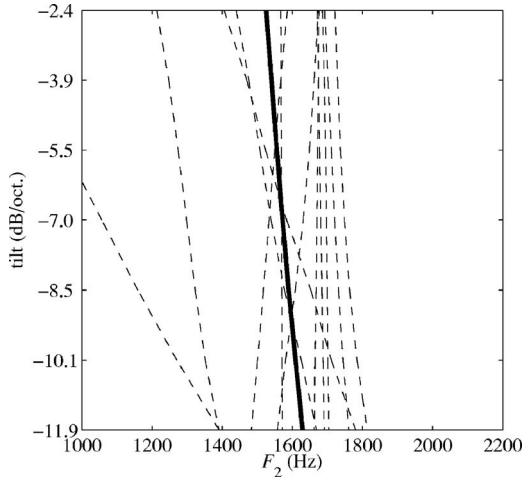


FIG. 3. Fitted regression boundaries for individual subjects in experiment 1. Dashed lines indicate fitted 50% cutoffs for each subject in the experiment. The solid line indicates fitted boundary for aggregated data.

The significant main effect for spectral tilt also suggests that, if the range of this stimulus parameter was expanded beyond that used in this experiment, then the effect would have been more dramatic. The range of /i/ responses, keeping spectral tilt constant at −7 dB/octave and changing F_2 frequency from 1400 to 1800 Hz, was between 20 and 92 of a possible 96 responses (21%–96%). Extrapolating the regression model fit with these data, if F_2 frequency were instead held constant at 1600 Hz, spectral tilt would have to vary between −50.0 and +7.8 dB/octave for a comparable shift in /i/ responses to occur. Although the effect for spectral tilt was statistically significant, it is ecologically relatively insignificant.

In terms of model predictions, with both spectral tilt and F_2 frequency included as parameters, the percent modal agreement (PMA—i.e., the percentage of predicted responses from the model that concur with the modal response given by listeners) was 92%. When spectral tilt is omitted from the model, PMA is reduced to 88%. However, if F_2 frequency is omitted from the model, PMA is reduced to 49% or approximately chance.[2]

C. Discussion

When long-term spectral tilt was the same for both the carrier sentence and the target vowel, the contribution of tilt to vowel perception was strongly attenuated; spectral tilt of the target vowel had a tiny contribution to perception when tilt of the precursor and target were the same. In contrast with results obtained by Kiefte and Kluender (2005) for the same monophthongs, performance was predicted almost entirely on the basis of formant frequencies alone. The second experiment was designed to evaluate whether the effect of formant frequency could be attenuated in a similar manner with an appropriately filtered precursor.

III. EXPERIMENT 2

The next experiment was designed to investigate whether perceptual cancellation of predictable acoustic characteristics is restricted to gross properties such as tilt, or whether perception also compensates for local properties such as spectral peaks when these properties are reliable across time. Previous studies have also examined the effects of local spectral properties, such as formants, on the effects of perceptual compensation of spectral context. For example, Ladefoged and Broadbent (1957) varied the total range of F_1 in their precursor sentences, while Watkins (1991), Watkins and Makin (1994, 1996a, b) processed precursors with filters designed by subtracting the spectrum of one vowel from another. In both types of experiments, however, both local and

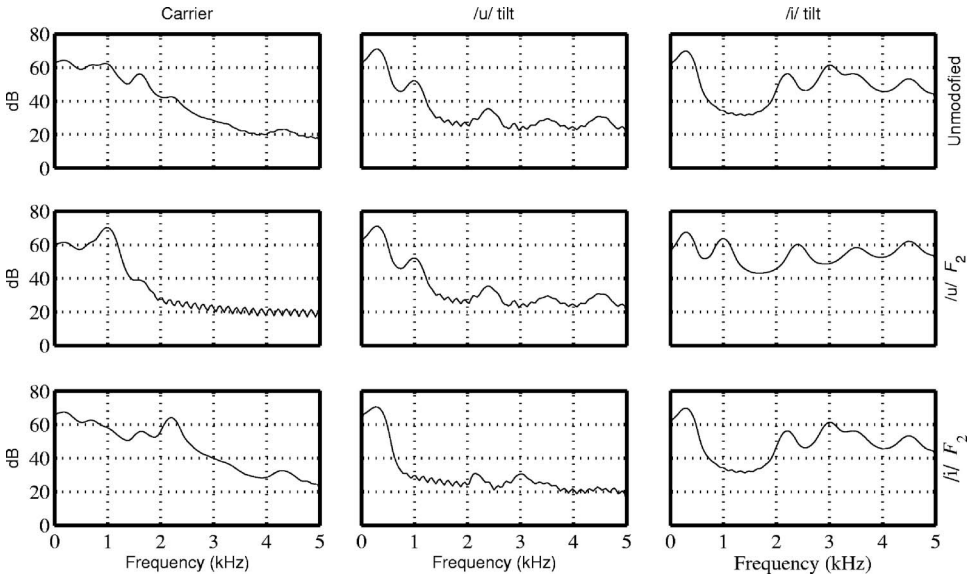


FIG. 4. Mean-square spectra of carrier and target stimuli used in experiment 2. The first row gives the spectra for the original, unfiltered carrier as well as the end points of the unmodified /i/-/u/ continuum. The carrier in the second row has been processed by a filter with the same frequency and bandwidth as the F_2 of the following vowel (in this case /u/), while the carrier in the third row was processed by a filter with the same frequency and bandwidth as the F_2 of /i/.

broad spectral properties were manipulated simultaneously. Given that it has been shown that spectral tilt plays a role in the perception of synthetic monophthongs,[3] the shifts in identification reported by these authors could be due to any spectral property that differentiates these vowels. Although Watkins and Makin (1996b) addresses the role of formant peaks directly by manipulating level differences between spectral peaks and valleys, the manipulations performed by these authors also influence spectral tilt (see footnote 1).

Summerfield *et al.* (1984) and Darwin (1990) did use narrowly defined spectral properties in their experiments. However, unlike previously, the purpose here is not to test whether perception can be shifted when target stimuli follow a specially filtered precursor. Analogous to experiment 1, we wish to examine perceptual cancellation of a target spectral peak. While the previous experiment showed that listeners will ignore spectral tilt when it is matched with that of the preceding acoustic context, will similar effects be observed for a spectral peak which typically does not persist in the way spectral does? Will listeners identify vowels on the basis of spectral tilt alone or will listeners instead identify vowels following a perceptual reassignment of the F_3 peak to F_2 resulting in a majority of "ee" responses?

Beginning with the same sentence context used in experiment 1, sentences were processed with a single-pole filter corresponding exactly to the frequency and bandwidth of F_2 of the target vowel (seven different formant frequencies). This yielded intact sentences with an additional constant-frequency spectral peak added throughout the entire duration. It should be noted that this filtering merely amplifies existing energy within the bandwidth of the filter. Intensity within the band still waxes and wanes in a fashion characteristic of speech (i.e., approximately 3 to 4 Hz). This modulation reduces the possibility that simple low-level processes such as peripheral adaptation are in force.

Although results from the previous experiment showed that perceptual effects of spectral tilt could be strongly attenuated if this property is left unchanged throughout the duration of a precursor sentence, there may be reasons to believe that perceptual effects of formant peaks cannot be mitigated in the same way. To the extent that spectral tilt is commonly associated with nonphonetic information such as listening environment or emotional state of the speaker, it is less likely to fluctuate rapidly over short periods of time. Therefore, it is generally adaptive for listeners to rely upon more informative signal properties. The same cannot be said for formant peaks, however.

A. Method

The method is largely the same as in the previous experiment: Subjects were asked to identify isolated, synthetic, steady-state vowels identical to those used in experiment 1. However, instead of manipulating spectral tilt of the precursor sentence to match that of the target vowel, an additional resonance, which matched the vowel F_2 peak in both frequency and bandwidth, was added to the precursor.

1. Subjects

Fifteen subjects were recruited in the same manner as described in experiment 1 and also met the same criteria. No subject participated in both experiments.

2. Stimuli

The same target vowels that were used in experiment 1 were used in this experiment, as well as the same precursor sentence "You will now hear the vowel…." However, for each of the 49 stimuli, the precursor sentence was filtered to include an additional resonance which matched the frequency and bandwidth of the target vowel F_2. Remaining aspects of the procedure are identical to those of experiment 1. Example mean-square spectra of the precursor sentence target stimuli are given in Fig. 4.

B. Results

Pooled results for 15 listeners shown in Fig. 5 illustrate an effect orthogonal to that for the condition in which spectral tilt was matched between context and vowel. When the center frequency of the additional spectral peak in the con-

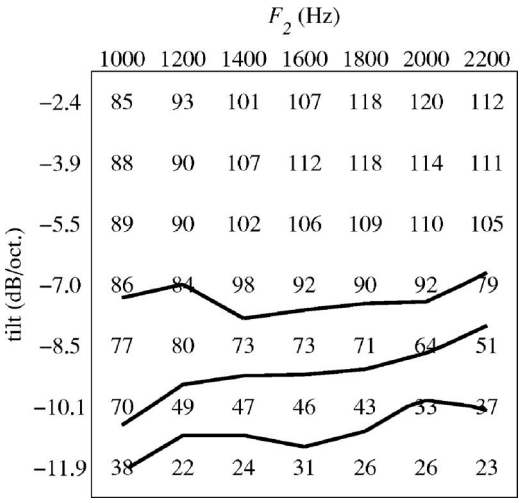


FIG. 5. /i/ responses pooled across speakers (out of a maximum of 120) from experiment 2. Contour lines are the same as those in Fig. 2.

text sentence is matched to F_2 of the following target vowel, listeners appear to rely exclusively upon global spectral characteristics (tilt) for identification of the target vowel. Responses by individual subjects were modeled via logistic regression with both spectral tilt and F_2 frequency as covariates, and subject regression coefficients were analyzed via single-sample t-test. As expected, effects for spectral tilt were highly significant ($t_{14}=5.84$; $p<0.001$). However, effects for F_2 frequency were not significant ($t_{14}=0.68$; $p=0.255$). Individual fitted regression boundaries are given in Fig. 6. Although there appears to be much more individual variability in this condition, the effect size for spectral tilt was still very large ($d=1.51$).

The model with both spectral tilt and F_2 frequency as covariates obtained a PMA of approximately 96%, while the model with spectral tilt alone also obtained a PMA of approximately 96%—i.e., spectral tilt alone predicts as many modal responses as spectral tilt and F_2 frequency together. Second formant alone predicts 71% of the modal responses. However, because /i/ responses accounted for 65% of the total responses, this is not significantly better than a model that simply predicts all /i/ responses ($p=0.131$).

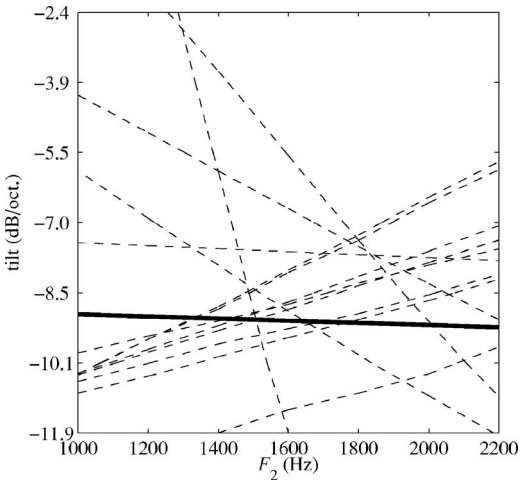


FIG. 6. Fitted regression boundaries for individual subjects in experiment 2.

C. Discussion

Similar to results obtained previously (e.g., Watkins, 1991), this pattern of performance indicates that, when energy within a limited frequency range is a predictable spectral property of acoustic context, it becomes effectively absorbed by perceptual processes. However, these results also show that effects of preceding context are clearly not restricted to gross spectral properties such as spectral tilt. Perceptual cancellation of predictable spectral characteristics also occurs for local, relatively narrow-band spectral characteristics of the acoustic context.

Two aspects of the context sentence could potentially restrict extrapolation from these first experiments. First, the context was explicitly interpretable speech, and one may question whether these effects depend upon listeners' knowledge of how English speech should sound without spectral distortions such as those imposed upon precursors in this experiment. If this were true, such *a priori* knowledge could account for some of these perceptual effects. Although Watkins and Makin (1994) used time-reversed speech carriers, these contexts were derived from a single phrase which was prepended to each target sound after filtering. With the exception of the filter, carriers were therefore perfectly predictable trial to trial and, in this way, may not be representative of variable acoustic contexts more broadly. If the effects were largely peripheral (e.g., Summerfield *et al.*, 1987; Darwin, 1990), then the predictability of the context sentence should not have affected the results. However, it has been suggested that perceptual adaptation to acoustic context may also involve processes more central in nature (Watkins, 1991), which may confound these results. For example, it is known that in vision, subjects are able to identify the exact shapes of arbitrary, unfamiliar, camouflaged objects presented across variations in surface pattern and shading despite the fact that, for a given stimulus presentation, the target object was indistinguishable from its background (Brady and Kersten, 2003). Similarly, if the same precursor is differently filtered across multiple trials, it may be possible for listeners to compensate for filter distortion as perceived through differences in the carrier *across* multiple presentations.

To address these concerns, new precursors were constructed to share spectral and temporal characteristics of sentences while being neither intelligible nor predictable trial to trial. These new contexts were constructed from multiple sentences that were time reversed and reconstructed via spectral analysis and resynthesis. These time-reversed contexts had durations of 1320–2220 ms and were completely unintelligible. Both conditions, tilt-matched and F_2-matched between context and vowel sound, were tested.

IV. EXPERIMENT 3

This experiment attempts to replicate results from experiment 1 in which perceptual effects of spectral tilt on identification of the target vowel were highly attenuated due to spectral properties of preceding context. However, with the exception of long-term spectral tilt of the precursor itself, the context sentence in experiment 1 was completely predict-

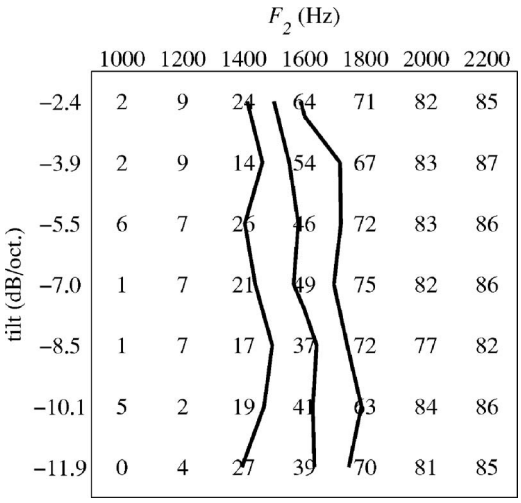


FIG. 7. /i/ responses pooled across speakers (out of a maximum 88) from experiment 3. Contour lines are calculated like those in Figs. 2 and 5.

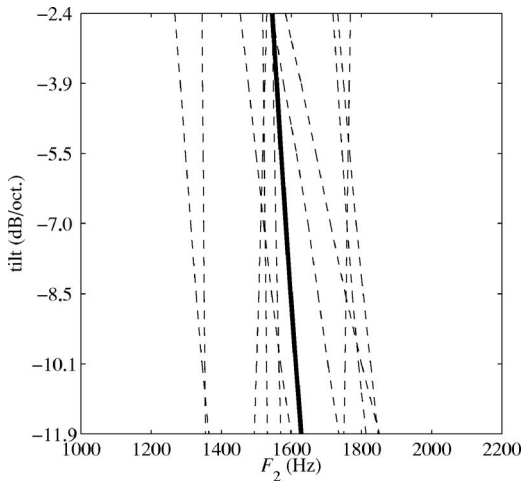


FIG. 8. Individual fitted regression lines for experiment 3.

able from trial to trial. In this experiment, completely unintelligible precursors of variable duration were randomized from trial to trial.

A. Methods

Methods for this experiment are exactly the same as for experiment 1 in Sec. II, with the exception of the precursors used. Instead of the phrase "You will now hear the vowel…," precursors were generated by first reversing sentences from the Hearing in Noise Test (Nilsson *et al.*, 1994), analyzing them with the same LPC model as before, and resynthesizing them with a 100 Hz fundamental frequency. This procedure is the same as that used in producing the precursor sentence in experiments 1 and 2 with the exception that the LPC analysis and resynthesis was performed on randomly selected sentences that are reversed in time.

Eleven subjects, recruited in the same manner as described in experiment 1, also met the same criteria. No subject participated in either of the previous experiments.

V. RESULTS

Pooled responses are presented in Fig. 7 while individual fitted boundaries are given in Fig. 8. Effects of preceding context found for the sentence "You will now hear the vowel…" were closely replicated with these unintelligible, time-reversed precursors of variable duration. As expected, effects for F_2 frequency were highly significant ($t_{10}=11.47$; $p<0.001$). In addition, the magnitude of the effect size was very large ($d=3.46$). While the effects for spectral tilt were also significant ($t_{10}=2.94$; $p<0.01$), the magnitude of the effect was much smaller ($d=0.89$). Figure 7 shows that, holding spectral tilt at −7 dB/octave and varying F_2 frequency from 1400 to 1800 Hz, the range of /i/ responses was 21–74 out of a possible 88 (24%–84%). To obtain a similar shift in responses varying only spectral tilt while holding F_2 frequency at 1600 Hz would require a change in spectral tilt from 13.9 to −43.0 dB/octave.

The model with both spectral tilt and F_2 frequency as covariates obtained a PMA of approximately 98%, while the model with F_2 frequency obtained a PMA of 94%. The model with spectral tilt alone as covariate obtained a PMA of only 55%; however, this is only slightly more than the total proportion of /i/ responses and is not significant ($p=0.33$).

VI. EXPERIMENT 4

This experiment used the same basic stimuli as in experiment 3. However, time-reversed, resynthesized precursor sentences were processed by a filter with the same bandwidth and center frequency of the target vowel F_2 peak as in experiment 2. Ten subjects were recruited for this experiment.

Results Pooled responses are presented in Fig. 9 while individual fitted regression boundaries are given in Fig. 10. There appears visually to be more of an effect for formant frequency than was observed in experiment 2. However, while effects for spectral tilt were highly significant ($t_9=5.34$; $p<0.001$) and the magnitude of this effect was very large ($d=1.69$), the effect for F_2 was not significant ($t_9=1.32$; $p=0.11$).

Percent modal agreement for the model including both spectral tilt and formant frequency as covariates was 88%.

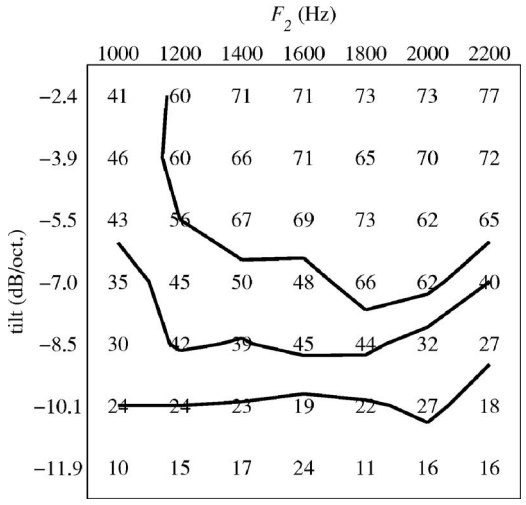


FIG. 9. /i/ responses pooled across speakers (out of a maximum of 80) from experiment 4. Lines are calculated like those in previous figures.

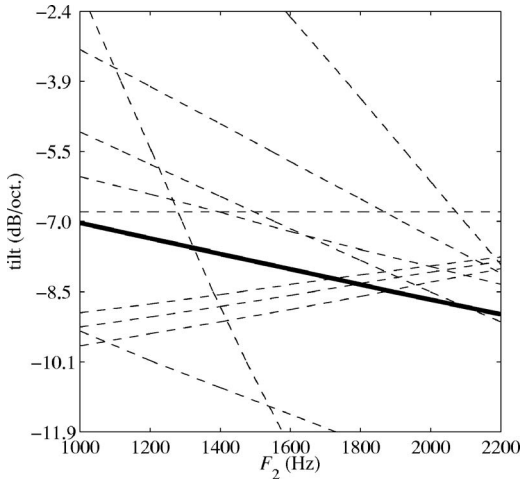


FIG. 10. Individual fitted regression lines for experiment 4.

However, the model including only spectral tilt achieved an even higher PMA (92%).[4] The model with only formant frequency achieved a PMA of 63%, which when compared to the overall proportion of /i/ responses is not significant ($p = 0.22$).

Perceptual cancellation of predictable acoustic context does not depend upon preceding context being intelligible speech, or does it depend upon the context being identical trial to trial.

VII. GENERAL DISCUSSION

Across all of the experiments reported here, and consistent with previous findings of Summerfield *et al.* (1984); Darwin *et al.* (1989); Watkins (1988, 1991); Watkins and Makin (1994, 1996a, b), listener performance provides evidence that the auditory system is adept at factoring out predictable characteristics of a listening context, and is consequently more sensitive to informative changes in spectral composition across time. This perceptual tuning to reliable characteristics of the listening context is demonstrated by perceptual compensation for redundant spectral features, spanning both broad and narrow spectral extents. By preceding target stimuli with a specially filtered precursor, we were able to selectively attenuate perceptual effects of specific spectral properties.

That effectiveness of spectral tilt (or relative formant amplitude) is strongly attenuated when it is unchanged throughout the acoustic context is not surprising (e.g., Fant, 1973; Klatt, 1982, 1986; Assmann, 1991). However, in experiment 2 we showed that listeners will discriminate vowels solely on the basis of spectral tilt if the second formant is matched in the long-term spectrum of the preceding carrier phrase. In these stimuli, because F_2 is continuous throughout the full stimulus, the peak is effectively absorbed perceptually. By perceptually compensating for the additional peak in the carrier phrase, F_2 of the target stimulus has been virtually excised. The results are therefore consistent with those of Ito *et al.* (2001), who have shown that listeners will identify vowels on the basis of spectral tilt if important spectral peaks are missing, and if the spectral tilt of the target vowel contrasts with that of the preceding carrier sentence.

One may ask whether the results we report here can be accounted for by the simple heuristic of inverse filtering which is frequently implemented in automatic speech recognition algorithms (e.g., Hermansky and Morgan, 1994). In this scheme, the long-term spectrum of the acoustic context is subtracted from the target phoneme before relevant acoustic properties are extracted from the stimulus (Watkins, 1991). The inverse filtering heuristic predicts results in which perception is shifted through changes in long-term spectral properties of an acoustic context (e.g., Watkins, 1991; Darwin *et al.*, 1989) or where a vowel is perceived from a flat-spectrum stimulus (Summerfield *et al.*, 1984). The inverse filtering heuristic also explains results reported here. While we have shown that auditory compensation is sensitive to narrowly defined acoustic characteristics of the acoustic context in addition to more broadly tuned properties such as gross spectral tilt, we have also supported observations that listeners will identify vowels in the absence of narrowly defined acoustic characteristics such as spectral peaks corresponding to formants. For example, in the present experiments, listeners discriminate vowels solely on the basis of spectral tilt when the additional resonance corresponding to vowel F_2 does not change between precursor and target stimulus. Therefore, these results are also consistent with those of Ito *et al.* (2001), who show that, in the absence of spectral information associated with the second formant peak, listeners can reliably identify vowels on the basis of broadly defined spectral properties. This can be compared with data from listeners with hearing impairment, who suffer from signal degradation including spectral smearing, showing that, relative to listeners with normal hearing, spectral tilt plays a greater role in perception of consonants (Lindholm *et al.*, 1988; Alexander and Kluender, 2007).

How listeners exploit global spectral properties in the absence of purportedly more reliable detailed spectral information is still unclear, however. It is unknown whether tilt provides independent information for phonetic perception, or whether tilt comes to serve perception as a consequence of experienced covariance with more primary acoustic attributes such as formant peaks. Because formant frequency is correlated with spectral tilt, they are effectively redundant except in those cases where the signal is distorted by, for example, channel characteristics. Across studies presented here, the data illustrate how the auditory system registers reliable properties of a listening context in ways that enhance sensitivity to change, adapting to redundant properties of the context. These findings are consistent with two related general principles concerning the way perceptual systems work. First, perceptual systems respond predominantly to change. They do not record absolute levels whether loudness, pitch, brightness, or color, and this has been demonstrated perceptually in every sensory domain (e.g., Kluender *et al.*, 2003). This sensitivity to change not only increases the effective dynamic range of biological systems, but it also increases the amount of information conveyed between organism and environment (Kluender and Alexander, 2007; Kluender and Kiefte, 2006).

ACKNOWLEDGMENTS

This work was supported by a grant from the Social Sciences and Humanities Research Council to the first author and a grant from the National Institutes of Deafness and Communication Disorders to the second author. We are grateful for comments by Anthony Watkins and an anonymous reviewer on a previous version of this manuscript.

[1]Watkins and Makin (1996a) do attempt to uncover exactly how peaks in the frequency response of the carrier filter result in category boundary shifts by either increasing or decreasing the local spectral contrast of the peaks themselves. This was accomplished by scaling the amplitude response of the carrier filters thereby either increasing or decreasing the spectral prominence of peaks. Although Watkins and Makin attempted to compensate for distortion of spectral tilt in these experiments by smoothing the first 0.4 ms of the filter impulse response, scaling of the resultant amplitude responses also changes the relative amplitude of spectral peaks themselves, thereby introducing a potential confound.

[2]Responses for /i/ accounted for 52% of the total number of responses. A model which simply predicted all /i/ responses would therefore achieve 52% PMA—greater than that predicted by the spectral-tilt-only model in experiment 1.

[3]Although stimuli designed by Watkins and Watkins and Makin are based on naturally produced recordings, the procedure they use to generate stimulus series results in spectrally static vowels.

[4]This is made possible by the fact that the models are not fit to maximize percent modal agreement, but rather are fit to minimize weighted least-squares error. Responses that are not reflected in modal responses may be weighted more heavily than modal responses themselves.

Alexander, J., and Kluender, K. (**2007**). "Contributions of gross spectral properties and duration of spectral change to perception of stop consonants," J. Acoust. Soc. Am. **118**, 1933.

Assmann, P. F. (**1991**). "The perception of back vowels: Centre of gravity hypothesis," Q. J. Exp. Psychol. A **43A**, 423–428.

Assmann, P. F., and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–338.

Beddor, P. S., and Hawkins, S. (**1990**). "The influence of spectral prominence on perceived vowel quality," J. Acoust. Soc. Am. **87**, 2684–2704.

Bladon, R. A. W., and Lindblom, B. (**1981**). "Modeling the judgement of vowel quality differences," J. Acoust. Soc. Am. **69**, 1414–1422.

Boynton, R. M., and Purl, K. F. (**1989**). "Categorical colour perception under low-pressure sodium lighting with small amounts of added incandescent illumination," Light. Res. Technol. **21**, 23–27.

Brady, M. J., and Kersten, D. (**2003**). "Bootstrapped learning of novel objects," J. Vision **3**, 413–422.

Broadbent, D. E., and Ladefoged, P. (**1960**). "Vowel judgements and adaption level," Proc. R. Soc. London, Ser. B **151**, 384–399.

Carlson, R., Granström, B., and Fant, G. (**1970**). "Some studies concerning the perception of isolated vowels," Speech Transmission Laboratory Q. Prog. Status Rep. (3–4), **11**, 73–83.

Chistovich, L. (**1971**). "Auditory processing of speech-evidences from psychoacoustics and neurophysiology," in *Proceedings of the Seventh International Congress on Acoustics*, Budapest, Vol. **1**, pp. 27–42.

Coady, J. A., Kluender, K. R., and Rhode, W. S. (**2003**). "Effects of contrast between onsets of speech and other complex spectra," J. Acoust. Soc. Am. **114**, 2225–2235.

Cohen, J. (**1988**). *Statistical Power Analysis for the Behavioral Sciences* (Erlbaum, Hillsdale, NJ).

Darwin, C. J. (**1990**). "Environmental influences on speech perception," in *Advances in Speech, Hearing and Language Processing*, edited by W. A. Ainsworth (JAI Press, London), Vol. **1**, pp. 219–241.

Darwin, C. J., McKeown, J. D., and Kirby, D. (**1989**). "Perceptual compensation for transmission channel and speaker effects on vowel quality," Speech Commun. **8**, 221–234.

Davis, C. S. (**2002**). *Statistical Methods for the Analysis of Repeated Measurements* (Springer, New York).

Fant, G. (**1973**). *Speech Sounds and Features* (MIT, Cambridge, MA).

Fieandt, K. V., Ahonen, L., Jarvinen, J., and Lian, A. (**1964**). "Color experiments with modern sources of illumination," Annales Academiae Scientiarum Fennicae. Ser. B **134**, 3–89.

Gumpertz, M., and Pantula, S. G. (**1989**). "A simple approach to inference in random coefficient models," Am. Stat. **43**, 203–210.

Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Trans. Speech Audio Process. **2**, 587–589.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Ito, M., Tsuchida, J., and Yano, M. (**2001**). "On the effectiveness of whole spectral shape vowel perception," J. Acoust. Soc. Am. **110**, 1141–1149.

Joos, M. (**1948**). "Acoustic phonetics," Language **24**, 1–136.

Kiefte, M., and Kluender, K. R. (**2005**). "The relative importance of spectral tilt in monophthongs and diphthongs," J. Acoust. Soc. Am. **117**, 1395–1404.

Kiefte, M., Kluender, K. R., and Rhode, W. S. (**2002**). "Synthetic speech stimuli spectrally normalized for nonhuman cochlear dimensions," ARLO **3**, 41–46.

Klatt, D. H. (**1980**). "Software for a cascade-parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D. H. (**1982**). "Speech processing strategies based on auditory models," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström, (Elsevier, Amsterdam), pp. 181–196.

Klatt, D. H. (**1986**). "Problem of variability in speech recognition and in models of speech perception," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 300–324.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Kluender, K. R., and Alexander, J. M. (**2007**). "Perception of speech sounds," in *Handbook of the Senses: Audition*, edited by P. Dallos and D. Oertel (Elsevier, London).

Kluender, K. R., Coady, J. A., and Kiefte, M. (**2003**). "Perceptual sensitivity to change in perception of speech," Speech Commun. **41**, 59–69.

Kluender, K. R., and Kiefte, M. (**2006**). "Speech perception within a biologically realistic information-theoretic framework," in *Handbook of Psycholinguistics*, edited by M. A. Gernsbacher and M. Traxler, 2nd ed. (Elsevier, London), pp. 153–199.

Ladefoged, P. (**1989**). "A note on 'information conveyed by vowels'," J. Acoust. Soc. Am. **85**, 2223–2224.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowel's," J. Acoust. Soc. Am. **29**, 98–104.

Lindholm, J. M, Dorman, M., Taylor, B. E., and Hannley, M. T. (**1988**). "Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing-impaired adults," J. Acoust. Soc. Am. **83**, 1608–1614.

McCullagh, P., and Nelder, J. A. (**1989**). *Generalized Linear Models*, 2nd ed. (Chapman & Hall, London).

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Rosner, B. S., and Pickering, J. B. (**1994**). *Vowel Perception and Production* (Oxford University Press, New York).

Shannon, C. E., and Weaver, W. (**1949**). *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, IL).

Summerfield, Q., Haggard, M., Foster, J., and Gray, S. (**1984**). "Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect," Percept. Psychophys. **35**, 203–213.

Summerfield, Q., Sidwell, A., and Nelson, T. (**1987**). "Auditory enhancement of changes in spectral amplitude," J. Acoust. Soc. Am. **81**, 700–708.

van Dijkhuizen, J. N., Anema, P. C., and Plomp, R. (**1987**). "The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences," J. Acoust. Soc. Am. **81**, 465–469.

Watkins, A. J. (**1988**). "Spectral transitions and perceptual compensation for effects of transmission channels," in *Proceedings of the Seventh Symposium of the Federation of Acoustical Societies of Europe: Speech '88*, edited by W. Ainsworth and J. Holmes, (Institute of Acoustics, Edinburgh).

Watkins, A. J. (**1991**). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **90**, 2942–2955.
Watkins, A. J., and Makin, S. J. (**1994**). "Perceptual compensation for speaker differences and for spectral-envelope distortion," J. Acoust. Soc. Am. **96**, 1263–1284.
Watkins, A. J., and Makin, S. J. (**1996a**). "Effects of spectral envelope contrast on perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **99**, 3749–3757.
Watkins, A. J., and Makin, S. J. (**1996b**). "Some effects of filtered contexts on the perception of vowels and fricatives," J. Acoust. Soc. Am. **99**, 588–594.

Spectral structure across the syllable specifies final-stop voicing for adults and children alike

Susan Nittrouer[a] and Joanna H. Lowenstein
Department of Speech and Hearing Science, Ohio State University, Columbia, Ohio 43210

(Received 28 March 2007; revised 14 August 2007; accepted 12 October 2007)

Traditional accounts of speech perception generally hold that listeners use isolable acoustic "cues" to label phonemes. For syllable-final stops, duration of the preceding vocalic portion and formant transitions at syllable's end have been considered the primary cues to voicing decisions. The current experiment tried to extend traditional accounts by asking two questions concerning voicing decisions by adults and children: (1) What weight is given to vocalic duration versus spectral structure, both at syllable's end and across the syllable? (2) Does the naturalness of stimuli affect labeling? Adults and children (4, 6, and 8 years old) labeled synthetic stimuli that varied in vocalic duration and spectral structure, either at syllable's end or earlier in the syllable. Results showed that all listeners weighted dynamic spectral structure, both at syllable's end and earlier in the syllable, more than vocalic duration, and listeners performed with these synthetic stimuli as listeners had performed previously with natural stimuli. The conclusion for accounts of human speech perception is that rather than simply gathering acoustic cues and summing them to derive strings of phonemic segments, listeners are able to attend to global spectral structure, and use it to help recover explicitly phonetic structure. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804950]

PACS number(s): 43.71.An, 43.71.Ft, 43.71.Es [MSS] Pages: 377–385

I. INTRODUCTION

Perhaps the trouble all started in 1944 when Frank Cooper and Al Liberman decided to build a reading machine for the blind. At that time they adopted what Liberman would later call the "horizontal view" in his book, *Speech: A Special Code* (1996). According to this view separate segments are aligned in the speech signal in a linear fashion, strictly auditory perceptual processes recover the acoustic character of each segment, and cognitive processes then translate those acoustic descriptors into phonemic units, void of physical attributes. Assuming this much about the acoustic speech signal, Cooper and Liberman turned their attention to what they saw as the truly difficult problem: optically isolating the letters on the page that would need to be converted into acoustic segments. But their own experiments soon revealed the intractable problem that listeners are unable to recognize separate acoustic elements presented at a rate replicating typical speech production. The declassification after World War II of the technology needed to build a sound spectrograph provided a possible clue to the source of the problem: separate segments are not represented in the acoustic speech stream. What ensued were decades of searching for acoustic properties that were at once both invariant correlates of specific phonetic categories as well as robust predictors of listeners' phonetic judgments. Such properties came to be known as acoustic "cues," and were generally defined as portions of the signal that can be isolated visually on the spectrogram, can be manipulated independently in speech synthesis, and can be shown to influence phonetic decisions (Repp, 1982).

Experiments exploring possible acoustic cues were all conducted in essentially the same way: by manipulating one acoustic property along a continuum, most typically in steps of equal (linear) size in the construction of synthetic stimuli, playing those stimuli for listeners in a labeling task, and plotting the probability of a specific phonetic decision as a function of the acoustic setting of the manipulated property. It was not uncommon for experimenters to manipulate one other selected property in a dichotomous manner such that it was set to be appropriate for one or the other phoneme. Under these circumstances two parallel labeling functions are generally derived, with the separation serving as an index of the amount of influence the dichotomously manipulated property has on the phonetic decision.

The voicing of syllable-final stops is one consonantal feature that has been extensively studied in this way. An acoustic difference in many languages between syllables that end in voiceless stops, such as "buck," and those that end in voiced stops, such as "bug," that is clearly apparent on a spectrogram is the duration of the vocalic syllable portion preceding closure, as Fig. 1 shows. The vocalic syllable portion is shorter when the final stop is voiceless than when it is voiced. In a carefully controlled series of experiments done in the 1970s, Raphael and his colleagues demonstrated that adult English speakers show a strong influence of this temporal property on their phonetic decisions (Raphael, 1972; Raphael *et al.* 1975; Raphael *et al.* 1980). These experiments were conducted in the traditional method, using synthetic stimuli in which all acoustic properties were held constant across the set of stimuli, except for the duration of those stimuli and first-formant offsets. Vocalic duration varied in a linear fashion from rather short to rather long. The offset frequency of the first formant (F1), and so the rate and extent of the transition, was manipulated in a dichotomous manner

[a]Author to whom correspondence should be addressed. Electronic mail: nittrouer.1@osu.edu

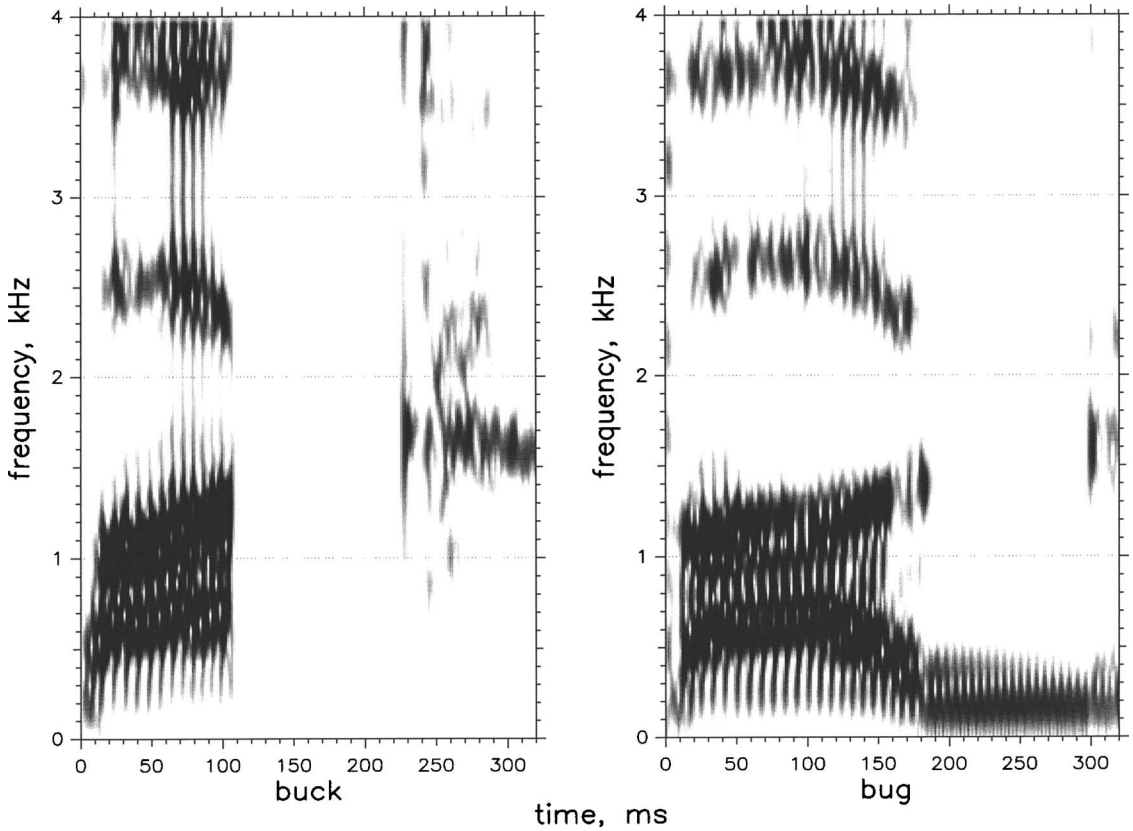


FIG. 1. Spectrogram of *buck* and *bug* spoken by a male, adult speaker.

in these experiments to signal either a voiced or voiceless final stop: higher F1 offsets more strongly support voiceless decisions, and lower F1 offsets support voiced decisions. Results consistently showed that both vocalic duration and F1 offset frequency contributed to voicing decisions.

Investigators interested in language acquisition became interested in the question of whether children use acoustic cues in the same way as adults to reach decisions about phonetic identity, and voicing decisions for syllable-final stops served as one focus of this work. Several investigators constructed synthetic stimuli similar to those used in the experiments of Raphael and colleagues, with vocalic duration varying in a continuous manner and F1 set dichotomously to be appropriate for either a final voiced or voiceless stop. The common conclusions of these experiments were that children did not base their voicing decisions as strongly on vocalic duration as adults did, but that the frequency of F1 at stimulus offset exerted a stronger influence on children's responding than on that of adults (e.g., Greenlee, 1980; Krause, 1982; Wardrip-Fruin and Peach, 1984). Those findings fit the larger picture that was emerging concerning differences between adults' and children's speech perception. Although not without its detractors, experiments with several kinds of stimuli were generally revealing that children's phonetic decisions appeared more strongly dependent on formant transitions, and less dependent on other sorts of acoustic cues (e.g., Morrongiello *et al.* 1984; Nittrouer and Studdert-Kennedy, 1987; Parnell and Amerman, 1978). The collective findings of these experiments led to the proposal that children modify the amount of perceptual weight they assign to individual acoustic cues as they get older and gain experience with a first language (e.g., Nittrouer, 1996). Of course, this model fit the general theoretical perspective of the day, which again was that separate acoustic cues are recovered during speech perception and translated by a cognitive processor to derive phonemic labels.

In 2004, a series of related experiments were conducted involving adults' and children's labeling of words ending in voiced or voiceless final stops (Nittrouer, 2004). In the first of these experiments, synthetic stimuli were constructed in the manner of earlier experiments so that vocalic duration varied across a continuum and F1 at stimulus offset was set to each of two frequencies appropriate for either a voiced or voiceless stop. These experiments replicated the findings of those earlier developmental studies: children's responses were less dependent on duration of the vocalic signal portion and more dependent on the frequency of F1 at stimulus offset than were responses of adult listeners. In the second kind of stimulus preparation, natural productions of words ending with voiced or voiceless final stops were edited so that vocalic duration varied from short to long. In those experiments it was observed that responses of children and adults alike were strongly related to whether the stimulus had been derived from a word ending with a voiceless or a voiced stop. For children this meant that labeling functions resembled those obtained for the children who had heard the synthetic stimuli, leading to the easily supportable conclusion that children were basing their responses strongly on formant transitions near stimulus offsets, as they had with the synthetic stimuli. For adults, however, this pattern of responding

meant there was a substantial change from the pattern of responding seen in numerous earlier experiments with synthetic stimuli. This difference in response patterns led to the conclusion that perhaps experiments done with synthetic stimuli had constrained our ability to determine which cues listeners use in phonetic decisions with natural speech. It may be that even adults rely primarily on formant transitions near a syllable's end to make decisions about the voicing of final stops.

The problem with that conclusion, however, is that there are numerous cues that vary in natural stimuli as a function of phonetic structure, and so it is hard to determine which one accounts for most of the variability in phonetic decisions. It is this very fact that has always made natural stimuli so undesirable for empirical study: With all those uncontrolled acoustic properties it is difficult to know how much each one explains about perceptual responding. The finding that adults and children showed similar labeling functions for edited natural stimuli ending in voiced and voiceless final stops might not mean that adults and children were basing decisions on the same acoustic property. Perhaps adults and children actually rely on different acoustic attributes, both of which happen to vary across voicing conditions in natural stimuli. The current experiment was originally undertaken to address this possibility.

One difference between synthetic stimuli generally used in experiments exploring the perception of syllable-final stop voicing and natural tokens was that synthetic stimuli preserved voicing-related differences in F1 offset transitions only. In fact, all formant transitions at the syllable's end are affected by the voicing category of the final stop. Because the vocal folds are abducted before the vocal tract achieves complete closure in the production of a syllable ending in a voiceless final stop, formants have not attained their final frequencies at voicing offset. In the production of words with voiced final stops, on the other hand, speakers continue voicing through closure, and so those final frequency destinations are reached. The difference in F1 depending on voicing is always that F1 is higher at voicing offset for words with voiceless, rather than voiced, final stops because F1 frequency is tightly linked to the degree of vocal tract opening. For higher formants, however, the relation between the formant's frequency at voicing offset and the voicing feature of the final stop depends on the place of closure for the stop. But regardless of the exact nature of that relation, dynamic spectral information at the ends of syllables was clearly richer in the natural stimuli than in the earlier synthetic stimuli, and Nittrouer (2004) suggested that findings for edited natural tokens showed that all listeners were basing their voicing decisions on those final formant transitions.

There are differences in acoustic structure earlier in the syllable as a function of whether the final stop is voiced or voiceless, as well. In particular, F1 rises more rapidly at voicing onset and achieves a higher frequency when the final stop is voiceless than when it is voiced (Summers, 1987). Summers (1988) showed that this acoustic difference alone influences voicing decisions for adult listeners. Consequently, the possibility needed to be considered that perhaps it was this acoustic characteristic that was influencing decisions of some listeners for natural tokens in the Nittrouer (2004) study, and adults were seen as more likely to use this perceptual strategy. That is, perhaps adults are able to take advantage of dynamic spectral patterns across the lengths of syllables in making phonetic decisions whereas children might be restricted to using only that information in the signal region generally affiliated with the segment in question. To examine this possibility we presented adults and children in the current experiment with synthetic stimuli in which formants replicated the patterns of natural syllables over their entirety, and also with stimuli in which formant transitions were deliberately held constant at syllable offsets, but allowed to vary in a natural manner earlier in the syllables.

In summary, this experiment was conducted in order to extend the work of Nittrouer (2004) in two ways. First, by presenting synthetic stimuli that replicated natural tokens of words ending in voiced and voiceless final stops we would be able to examine whether the difference in adults' response patterns observed for synthetic and natural stimuli had something to do with the naturalness of the stimuli. Would adults and children show the same labeling functions for stimuli replicating the acoustic structure of the natural stimuli in the 2004 study, but possessing a synthetic nature? Generally concern exists that children may show decrements in performance for synthetic compared to natural speech (e.g., Mirenda and Beukelman, 1987; Reynolds and Jefferson, 1999), but in this case the question may be asked of adults' responding, as well. In the second stimulus manipulation we asked what it was in the natural stimuli that supported adults' and children's voicing decisions. In particular the hypothesis was explored that perhaps children based their voicing decisions on formant transitions going into vocal tract closure, but adults based their decisions on formant characteristics earlier in the stimulus. This hypothesis would be supported if children showed greater differences in their response patterns between the two stimulus types used in this experiment than adults showed.

II. METHOD

A. Listeners

Adults between the ages of 18 and 39 years participated, as well as 8 year olds, 6 year olds, and 4 year olds. Children were between −1 and +5 months of their birthdays. All listeners had to meet certain criteria to participate. All participants were native English speakers with no histories of speech, language, or hearing problems. All were required to pass hearing screenings of the frequencies 0.5, 1, 2, 4, and 6 kHz presented at 25 dB hearing level to each ear separately. Children could have had no more than five episodes of otitis media before their second birthdays. Children needed to perform at or better than the 30th percentile on the Goldman-Fristoe Test of Articulation 2, Sounds-in-Words subtest (Goldman and Fristoe, 2000), and adults needed to read at or better than an 11th grade reading level on the Wide Range Achievement Test – Revised (Jastak and Wilkinson, 1984). Meeting these criteria were 24 adults, 24 8 year olds, 32 6 year olds, and 30 4 year olds.

B. Equipment and materials

Perceptual testing took place in a soundproof booth, with the computer that controlled the experiment in an adjacent room. The hearing screening was done with a Welch Allen TM 262 audiometer and TDH-39 earphones. Stimuli were stored on a computer and presented through a Creative Labs Soundblaster card, a Samson headphone amplifier, and AKG-K141 headphones. The experimenter recorded responses with a keyboard connected to the computer. Two pictures on 8 in.×8 in. cards were used to represent the response labels of *buck* (a male deer) and *bug* (a ladybug). Game boards with ten steps were also used with children: they moved a marker to the next number on the board after each block of test stimuli. Cartoon pictures were used as reinforcement and were presented on a color monitor after completion of each block of stimuli. A bell sounded while the pictures were being shown and served as additional reinforcement.

C. Stimuli

Two sets of stimuli were created for this experiment: synthetic *buck/bug* that copied the patterns of frequency change across the first three formants of natural *buck* and *bug* tokens ("natural-formant" stimuli), and synthetic *buck/bug* that had a relatively high or low F1 frequency at syllable center ("high/low F1" stimuli), but ambiguous formant transitions near the syllable's end. Both sets of stimuli were created using the Sensyn Laboratory Speech Synthesizer. In both sets of stimuli, f0 started at 130 Hz and fell linearly throughout the stimulus to an offset frequency of 100 Hz. All stimuli were completely voiced, with no stimulus portion replicating voicing during closure for the final stop or a release burst. Each stimulus that was created was subsequently manipulated so that individual tokens varied along a continuum from 100 to 260 ms, in nine 20-ms steps. So, vocalic duration served as a nondynamic (i.e., static) property that has been shown to have a large influence on voicing decisions of English-speaking adults in experiments with synthetic stimuli that vary only vocalic duration and F1-offset transitions. All stimuli were presented as isolated words in a labeling task.

1. Natural-formant

The natural-formant stimuli were based on natural tokens of an adult, male speaker producing the words *buck* and *bug* in isolation. Nittrouer *et al.* (2005) provide complete results of acoustic analyses on these and similar words, and measures from that study served as a guide for creating these stimuli. In both kinds of stimuli, F3 was held constant at 2700 Hz until 50 ms before offset. For the more *buck*-like stimulus, F3 fell to an ending frequency of 2500 Hz, while in the more *bug*-like stimulus it fell to 2400 Hz. For the more *buck*-like stimulus, F2 started at 1000 Hz and rose linearly throughout the stimulus to 1200 Hz at stimulus offset. For the more *bug*-like stimulus, F2 was constant at 1000 Hz until 50 ms before offset, at which time it rose linearly to 1400 Hz. Regarding F1, it started at 400 Hz, rose linearly to 800 Hz over the first 50 ms, and remained at 800 Hz until stimulus offset for the more *buck*-like stimulus. For the more *bug*-like stimulus, F1 started at 400 Hz, rose linearly to 625 Hz over the first 50 ms, and remained at 625 Hz until 50 ms before offset, at which time it fell linearly to 250 Hz. Thus, all formants varied across the voicing conditions as they do in natural syllables: F2 and F3 more closely approximated a "velar pinch" and F1 was lower in frequency at stimulus offset in the voiced condition. Furthermore, F1 differed across the length of the stimuli as natural tokens of *buck* and *bug* do. There were 18 natural-formant stimuli: two formant patterns x nine vocalic durations.

2. High/low F1

For both of the high/low F1 stimuli, F3 was constant at 2700 Hz until 50 ms before offset, at which time it fell to its ending frequency of 2450 Hz (midpoint of settings for the voiced and voiceless conditions in the natural-formant stimuli). F2 was constant at 1000 Hz until 50 ms before offset, at which time it rose linearly to 1300 Hz (again, midpoint of the settings for the voiced and voiceless conditions in the natural-formant stimuli). Thus, both F2 and F3 were set to ambiguously signal voicing for the final stop. For the most *buck*-like of these high/low F1 stimuli, F1 started at 400 Hz, rose linearly to 800 Hz over the first 50 ms, and stayed at 800 Hz until 50 ms before offset. At that time it fell by 175 Hz to its ending frequency of 625 Hz. For the most *bug*-like of the stimuli, F1 started at 400 Hz, rose linearly to 625 Hz over the first 50 ms, and stayed at 625 Hz until 50 ms before offset. At that time it fell by 175 Hz to its ending frequency of 450 Hz. So F1 differed at syllable center across stimuli, but fell by the same amount at offset. There were 18 of these high/low F1 stimuli: two F1 patterns X nine vocalic durations.

D. Procedures

Adults attended one test session and children attended two. Screening procedures were completed first. Next, two sets of stimuli were used in pretesting. A set of completely natural tokens of *buck* and *bug* was presented, with whatever voicing during closure was in the signal and release bursts. These tokens were taken from three different adult, male speakers. Two tokens each of *buck* and *bug* were used from each speaker, making a total of 12 stimuli for the pretest. Listeners had to respond correctly to 11 of them to proceed to the next pretest. The next pretest consisted of the same 12 stimuli, only with the release bursts and voicing during closure removed. Listeners had to respond correctly to 11 of these stimuli in order to proceed to testing.

In testing, all listeners were presented with both sets of stimuli. The order of presentation of the natural-formant and the high/low F1 stimuli was randomized across listeners. The same procedures were followed for each set of stimuli. Practice items were presented before the testing began. Practice items consisted of *buck* and *bug* "best exemplars," which were the stimuli that should most strongly evoke the correct words. For example, for the natural-formant stimuli, the stimulus with the *buck*-like spectral settings that was 100 ms long and the stimulus with the *bug*-like spectral settings that

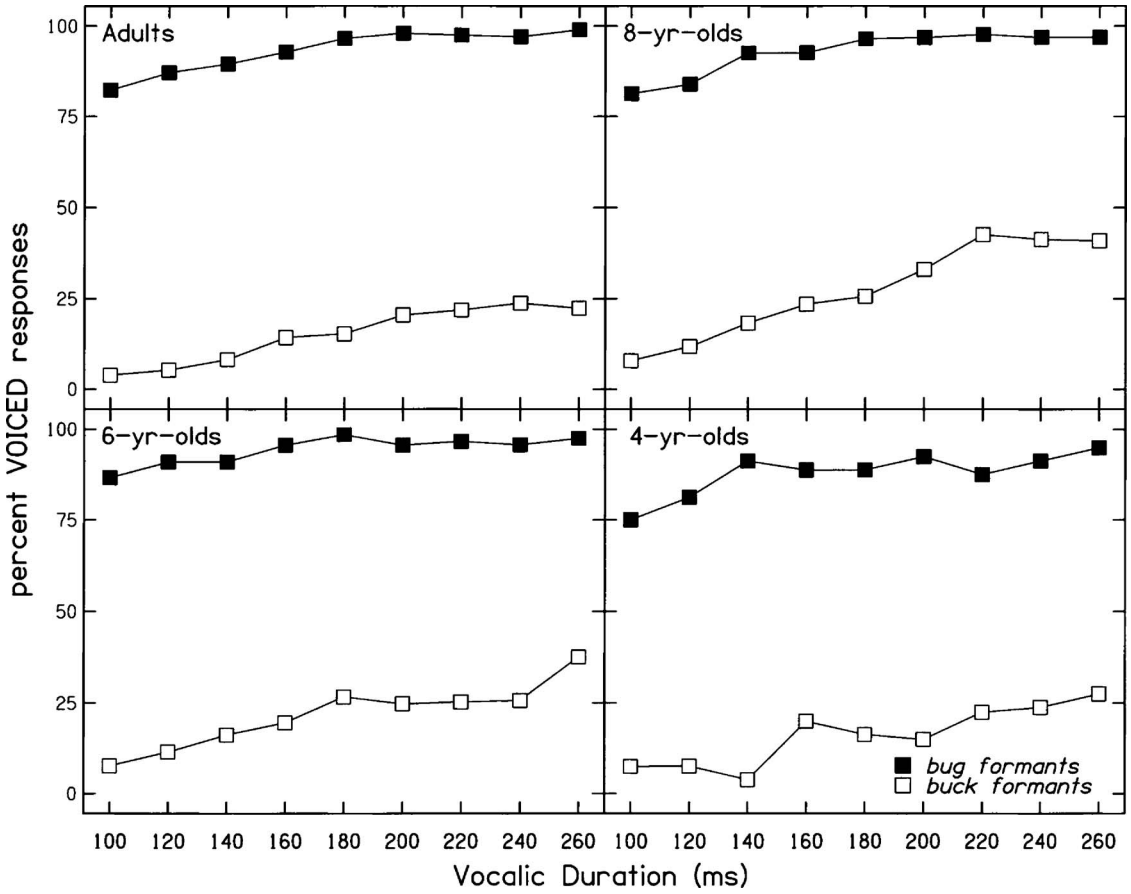


FIG. 2. Labeling functions for the natural-formant stimuli.

was 260 ms long were played, six times each (12 stimuli). The listener had to respond correctly to at least 11 of these stimuli to proceed to testing. During testing, ten blocks of stimuli were presented for both sets. Listeners responded by saying the label and pointing to the picture that represented their selection. To have their data included in the final analysis, participants needed to respond with at least 80% correct responses to these best exemplars during testing. This requirement served as a check that data were analyzed only from participants who maintained attention to the task: because all listeners were required to respond correctly to 90% of these stimuli during practice, they should have been able to do so during testing, if they maintained attention.

For children, cartoon pictures were displayed on the monitor and a bell sounded at the end of each block. They moved a marker to the next space on a gameboard after each block as a way of keeping track of how much more time they had left in the test.

The percentages of *bug* responses to each stimulus were tabulated and subsequently correlated with settings for vocalic duration and spectral structure (appropriate for *buck* or *bug*). The obtained regression coefficients indexed the weights that were assigned to each of these acoustic properties in perception. These correlation coefficients from individual listeners were used for statistical analyses.

III. RESULTS

Of the 30 4 year olds who met the criteria to participate, three did not meet the labeling criterion with unedited, natural *buck/bug* stimuli; that is, those that preserved voicing during closure and burst releases. Another ten did not meet criterion with natural *buck/bug* stimuli that had the voicing during closure and burst releases edited out. So, a total of 17 4 year olds participated in the testing.[1]

A. Natural-formant buck/bug

Nine 4 year olds, 11 6 year olds, and one 8 year old were unable to reach the criteria for having their data included because they failed either to label 90% of the best exemplars correctly during the pretest or to label 80% correctly during actual testing. Consequently, data were included for eight 4 year olds, 21 6 year olds, 23 8 year olds, and 21 adults.

Figure 2 shows mean labeling functions for each age group for the natural-formant *buck/bug* stimuli. Labeling functions were similar across age groups, suggesting that all listeners were responding similarly. Furthermore, it is clear that when the patterns of spectral change in formants across the utterance were appropriate for *bug*, all listeners responded with close to 100% *bug* responses; when these formant patterns were appropriate for *buck*, listeners responded with close to 0% *bug* responses. These functions are all extremely flat, as well, suggesting that children and adults did not weight vocalic duration very strongly at all in these stimuli. These patterns of responding replicate results for all contrasts created by editing natural stimuli in Nittrouer (2004).

TABLE I. Mean partial correlation coefficients for each age group, for each kind of acoustic property (vocalic duration and spectral) in each condition.

	Natural Formant		High/Low F1	
	Vocalic Duration	Spectral	Vocalic Duration	Spectral
4 year olds	0.21 (0.16)	0.90 (0.12)	0.46 (0.24)	0.65 (0.25)
6 year olds	0.19 (0.24)	0.87 (0.25)	0.39 (0.28)	0.72 (0.28)
8 year olds	0.27 (0.22)	0.84 (0.25)	0.48 (0.24)	0.67 (0.31)
Adults	0.17 (0.25)	0.88 (0.28)	0.45 (0.30)	0.73 (0.28)
Mean	0.21 (0.23)	0.87 (0.24)	0.44 (0.26)	0.70 (0.28)

In viewing labeling functions such as those in Fig. 2 it is difficult to get a sense of the variability in how listeners within a group labeled the stimuli, and so a sense of the variability in labeling across groups. For this reason we computed the means of the within-group standard deviations (of percent VOICED responses) across each of the 18 stimuli in this experiment. These means were 17 percent for adults, 17.5 percent for 8 year olds, 16 percent for 6 year olds, and 13 percent for 4 year olds. From these numbers it is concluded that variability in labeling was similar across groups.

The left side of Table I shows mean partial correlation coefficients (i.e., Pearson r) for each age group for the natural-formant stimuli. It can be seen that listeners in all age groups generally weighted spectral structure more than vocalic duration. A two-way analysis of varience (ANOVA) done on the coefficients, with property (vocalic duration vs spectral structure) and age as the main effects, showed a significant effect of property only, $F(1,76)=163.93$, $p<.001$. The Property X Age interaction was not significant. This result reveals that more of the variance in response patterns was explained by spectral structure than by vocalic duration, and this pattern did not vary with listener age.

B. High/low F1 stimuli

Seven 4 year olds, 13 6 year olds, one 8 year old, and one adult were unable to reach the criteria for having their data included during either the pretest or test. Consequently, data were included for ten 4 year olds, 19 6 year olds, 23 8 year olds, and 20 adults.

Figure 3 shows mean labeling functions for each age group for the high/low F1 stimuli. Again, children and adults appear to be weighting transitions heavily in their voicing decisions, and vocalic duration less so. This pattern appears similar across groups. In order to compare variability in labeling across groups we again computed means of the within-group standard deviations across each of the 18 stimuli in this experiment. Here we found means of 20 percent for adults, 17.5 percent for 8 year olds, 21 percent for 6 year olds, and 22.5 percent for 4 year olds. Again, these numbers were taken as an indication of similar variability across groups.

The right side of Table I displays mean correlation coefficients for each age group for these stimuli. A two-way ANOVA done on the coefficients with property and age as the main effects showed a significant effect of property only, $F(1,76)=17.10$, $p<0.001$, again revealing that correlation coefficients were greater for spectral structure than for vocalic duration. Again, the Property x Age interaction was not significant.

C. Comparison across stimulus types

Simple effects analysis was performed on correlation coefficients for each kind of acoustic property (vocalic duration and spectral structure) across the two stimulus types (natural-formant and high/low F1) separately to see if listeners weighted the acoustic properties differently depending on how much information was available in the overall spectral structure: Spectral structure more strongly signaled voicing in the natural-formant stimuli than in the high/low F1 stimuli. Both kinds of acoustic properties showed significant effects of stimulus type: vocalic duration, $F(1,76)=109.70$, $p<0.001$; spectral structure, $F(1,76)=59.97$, $p<0.001$. These results show that listeners weighted spectral structure less and vocalic duration more for the high/low stimuli than for the natural-formant stimuli. This difference in weighting strategy across the stimuli sets indicates that listeners pay more attention to spectral structure when that structure is more informative, as it was for the natural-formant stimuli compared to the high/low F1 stimuli in this experiment. When spectral structure is less informative, then listeners apparently turn their attention to other aspects of the acoustic structure that reliably signal phonetic identity, which in this case was vocalic duration. Neither the main effect of age nor the Stimulus Type x Age interaction were found to be significant in this analysis, indicating that adults and children showed similar shifts in perceptual attention across stimulus types. So adults and children based their voicing decisions on the same aspects of acoustic structure in both sets of stimuli, and the structure that they primarily used was the dynamic spectral structure of the formants across the syllables.

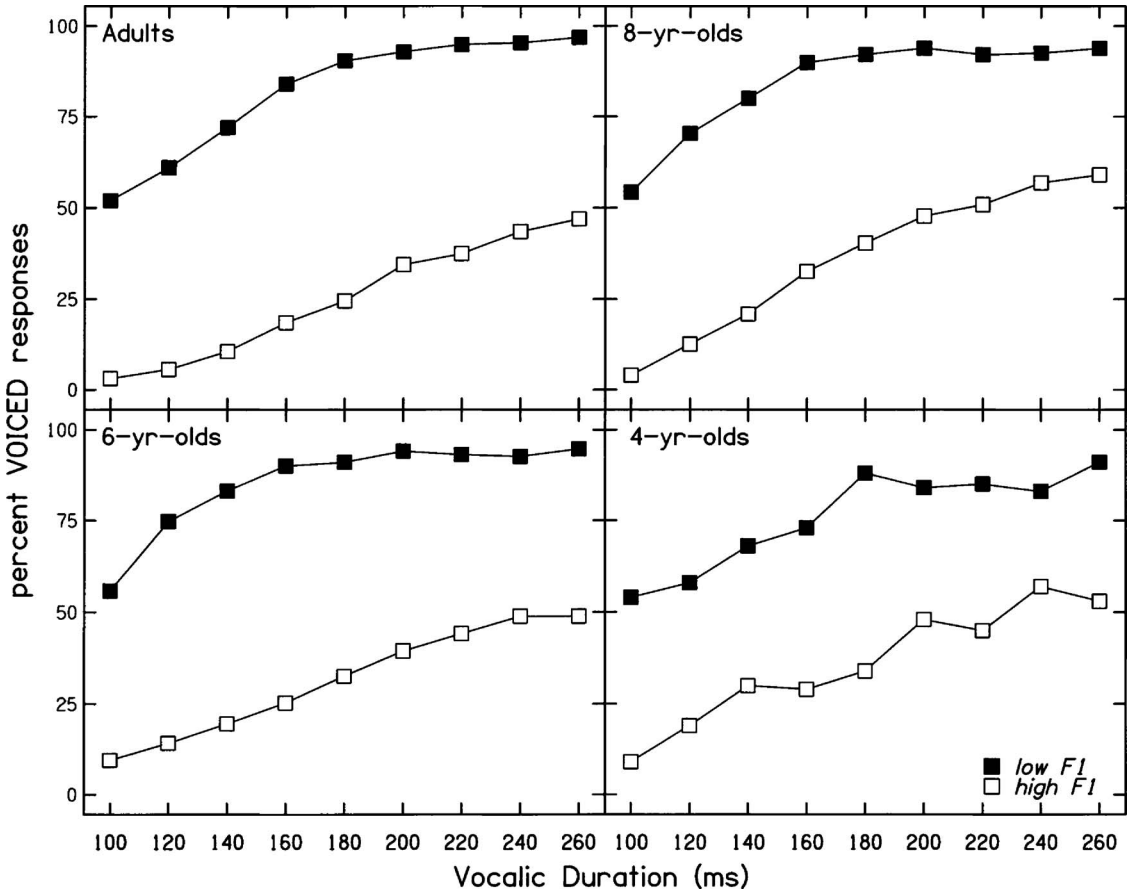


FIG. 3. Labeling functions for the high/low F1 stimuli.

IV. DISCUSSION

In this experiment, listeners were presented with synthetic stimuli that varied in two ways: vocalic duration varied along a continuum from short to long, and stimulus-internal spectral structure differed across all or most of the stimulus depending on the voicing of the final stop. All listeners were strongly influenced in their phonetic judgments by the spectral structure of those stimuli. Vocalic duration did not heavily influence responses for any age group. These findings contrast sharply with previous results from adults and children for synthetic stimuli that carefully controlled the spectral structure across most of the syllable, except for the final portion, and there varied it for only one formant, F1 (e.g., Greenlee, 1980; Krause, 1982; Nittrouer, 2004; Wardrip-Fruin and Peach, 1984). For those stimuli it was found that children, but not adults, used that extremely circumscribed spectral structure in their voicing decisions. Adults instead turned their perceptual attention to vocalic duration for making voicing decisions in those experiments, as they have been found to do in similar experiments dating back 50 years (e.g., Denes, 1955; Raphael, 1972; Wardrip-Fruin, 1982). Thus it may be concluded from results across those earlier experiments and this current experiment that younger listeners seek out dynamic spectral structure, and use whatever information of that nature they can find in the signal. Adults, on the other hand, do not use dynamic information that is impoverished. Perhaps it is the case that adults are more facile at shifting their perceptual attention as the situation demands: In the case of synthetic stimuli in earlier final-voicing experiments, adults may have shifted their attention completely to the temporal information when the spectral information was impoverished. Be that as it may, when available, adults weight dynamic spectral information more than the temporal information for final voicing decisions, and much as children do.

Of course, it is always tempting to attribute any age-related differences in speech perception to possible age-related differences in auditory sensitivities for the properties being manipulated. So, for example, the earlier results with synthetic speech stimuli for syllable-final stop voicing might indicate that adults are more sensitive than children to temporal properties. By the same reasoning, an auditory hypothesis would have to predict that children are more sensitive than adults to spectral glides because children have been found to weight that information more than adults. Nittrouer and Lowenstein (2007) examined the auditory hypothesis by manipulating temporal and spectral properties in nonspeech signals to replicate the structure of stimuli in earlier speech perception experiments with syllable-final voicing. Discrimination thresholds were obtained for nonspeech signals that varied in either duration or spectral structure. Similar sensitivity was found for adults and children for stimulus duration and spectral glides across three sine waves corresponding to the first three formants. When only the glide of the lowest frequency sine wave was manipulated (the condition corresponding most closely to those earlier synthetic speech

stimuli), adults were actually slightly more sensitive than children to those glides. So, no evidence was found from experiments with nonspeech stimuli to support the contention that perceptual weighting strategies for speech signals are based on auditory sensitivity.

The findings reported here replicate the finding of Summers (1988) showing that adults make use of differences in F1 frequency across the syllable. The current study extends that work by demonstrating that children show similar effects. Of course, because the high/low F1 stimuli manipulated the onset and middle portions of the stimuli together, this study was unable to distinguish whether listeners were specifically weighting the onset transition or the steady-state information in their decisions. Whatever the case, it is clear that listeners of all ages use spectral structure that is not temporally constrained to the acoustic region traditionally associated with the segment they are being asked to label in those experiments: In the high/low F1 condition, the spectral structure at the ends of the syllables was ambiguous between the voiced and voiceless replicas.

A broader implication of the results reported here is that perhaps the traditional view of human speech perception must be revised. Perhaps our theory building regarding this important process has been ill served by the view that listeners extract discrete acoustic cues and then perform a cognitive translation of these properties into phonetic units. The results of the current study are consistent with a model of speech perception in which listeners attend to overall spectral structure, the kind of structure that arises from the relatively slow articulatory movements within the vocal tract. When the fine spectral structure within the syllable portion generally considered to be associated with the specific phonetic segment listeners were being asked to label was set to be ambiguous with regard to stop voicing, the weight that listeners assigned to vocalic duration increased slightly, but still decisions were largely based on the dynamic spectral pattern associated with the production of the whole syllable. This trend was as apparent for children as for adults. So, perhaps listeners attend primarily to overall patterns of spectral change during speech perception. Perhaps it is primarily in our psychophysical experiments where we restrict the sensory information available for decision making that we find that some listeners are able to turn their attention to other properties. And it should not surprise us that the listeners who are most capable of doing so are mature listeners who are native speakers of the language being manipulated, and so who have had the most experience hearing how other properties covary with overall spectral structure. For words ending in voiced and voiceless stops, this suggestion derives from the numerous studies showing that English-speaking adults are better able to use vocalic duration (when spectral structure is constrained) than English-speaking children (Greenlee, 1980; Krause, 1982; Nittrouer, 2004; Wardrip-Fruin and Peach, 1984) or adults who are native speakers of a language without a vowel-length distinction associated with final voicing (Crowther and Mann, 1994; Flege and Wang, 1989).

In summary, perhaps the early emphasis on the horizontal view, as exemplified by the work of Cooper and Liberman (Liberman, 1996) took us down a blind alley. To be sure, these investigators eventually turned their attention to alternative models of speech perception, as Liberman explains in his 1996 book. Nonetheless the view of the speech signal as a collection of discrete cues persisted in theory and experimental procedure. Have we spent decades dissecting speech stimuli and manipulating individual "cues," all the while ignoring the importance of integrated spectral structure to human speech perception? Only further investigation using paradigms that do not rely exclusively on manipulation of separate acoustic cues can answer this question.

ACKNOWLEDGMENT

This work was supported by Grant No. R01 DC000633 from the National Institute on Deafness and Other Communication Disorders, the National Institutes of Health.

[1]These numbers reveal that 37 percent of our 4 year olds were unable to label practice syllables if there was no voicing during closure or release burst. We wished to examine the hypothesis that this high failure rate may be at least partly explained by young children's inattention to the ends of syllables when there is not a salient marker to those syllable endings. To examine that possibility, we appended 12 ms of a natural /k/ burst taken from a *buck* sample to the ends of all natural-formant stimuli, 100 ms after stimulus offset. When we played these stimuli to a group of 15 4 year olds not included in the main study, using the same practice and test procedures, we found that only three children were unable to label the practice stimuli (20 percent). Testing was completed with the remaining children, and mean partial correlation coefficients computed on the data from their labeling results were 0.07 for vocalic duration and 0.97 for spectral structure. Although these values suggested a pattern of responding in which these 4 year olds weighted vocalic duration less and offset transitions more than any listener group in the main study, they were within a standard deviation of the mean correlation coefficients found for 4 year olds in the study (see Table I). Moreover, running the reported statistical analyses with these children included in the sample did not change any outcomes.

Crowther, C. S., and Mann, V. (**1994**). "Use of vocalic cues to consonant voicing and native language background: The influence of experimental design," Percept. Psychophys. **55**, 513–525.
Denes, P. (**1955**). "Effect of duration on the perception of voicing," J. Acoust. Soc. Am. **27**, 761–764.
Flege, J. E., and Wang, C. (**1989**). "Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast," J. Phonetics **17**, 299–315.
Goldman, R., and Fristoe, M. (**2000**). *Goldman Fristoe 2: Test of Articulation* (American Guidance Service, Inc., Circle Pines, MN).
Greenlee, M. (**1980**). "Learning the phonetic cues to the voiced-voiceless distinction: A comparison of child and adult speech perception," J. Child Lang **7**, 459–468.
Jastak, S., and Wilkinson, G. S. (**1984**). *The Wide Range Achievement Test-Revised* (Jastak Associates, Wilmington, DE).
Krause, S. E. (**1982**). "Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults," J. Acoust. Soc. Am. **71**, 990–995.
Liberman, A. M. (**1996**). *Speech: A special code* (MIT Press, Cambridge, MA).
Mirenda, P., and Beukelman, D. R. (**1987**). "A comparison of speech synthesis intelligibility with listeners from three age groups," Augmentative and Alternative Communication **3**, 120–128.
Morrongiello, B. A., Robson, R. C., Best, C. T., and Clifton, R. K. (**1984**). "Trading relations in the perception of speech by 5-year-old children," J. Exp. Child Psychol. **37**, 231–250.
Nittrouer, S. (**1996**). "The discriminability and perceptual weighting of some acoustic cues to speech perception by three year olds," J. Speech Hear. Res. **39**, 278–297.
Nittrouer, S. (**2004**). "The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults," J. Acoust. Soc. Am. **115**, 1777–1790.

Nittrouer, S., Estee, S., Lowenstein, J. H., and Smith, J. (**2005**). "The emergence of mature gestural patterns in the production of voiceless and voiced word-final stops," J. Acoust. Soc. Am. **117**, 351–364.
Nittrouer, S., and Lowenstein, J. H. (**2007**). "Children's weighting strategies for word-final stop voicing are not explained by auditory sensitivities," J. Speech Lang. Hear. Res. **50**, 58–73.
Nittrouer, S., and Studdert-Kennedy, M. (**1987**). "The role of coarticulatory effects in the perception of fricatives by children and adults," J. Speech Hear. Res. **30**, 319–329.
Parnell, M. M., and Amerman, J. D. (**1978**). "Maturational influences on perception of coarticulatory effects," J. Speech Hear. Res. **21**, 682–701.
Raphael, L. J. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**, 1296–1303.
Raphael, L. J., Dorman, M. F., Freeman, F., and Tobin, C. (**1975**). "Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies," J. Speech Hear. Res. **18**, 389–400.
Raphael, L. J., Dorman, M. F., and Liberman, A. M. (**1980**). "On defining the vowel duration that cues voicing in final position," Lang Speech **23**, 297–307.
Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bull. **92**, 81–110.
Reynolds, M. E., and Jefferson, L. (**1999**). "Natural and synthetic speech comprehension: Comparison of children from two age groups," Augmentative and Alternative Communication **15**, 174–182.
Summers, W. V. (**1987**). "Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses," J. Acoust. Soc. Am. **82**, 847–863.
Summers, W. V. (**1988**). "F1 structure provides information for final-consonant voicing," J. Acoust. Soc. Am. **84**, 485–492.
Wardrip-Fruin, C. (**1982**). "On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants," J. Acoust. Soc. Am. **71**, 187–195.
Wardrip-Fruin, C., and Peach, S. (**1984**). "Developmental aspects of the perception of acoustic cues in determining the voicing feature of final stop consonants," Lang Speech **27**, 367–379.

Spectral tilt change in stop consonant perception

Joshua M. Alexander[a] and Keith R. Kluender
Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706

(Received 22 January 2007; revised 26 September 2007; accepted 5 November 2007)

There exists no clear understanding of the importance of spectral tilt for perception of stop consonants. It is hypothesized that spectral tilt may be particularly salient when formant patterns are ambiguous or degraded. Here, it is demonstrated that relative change in spectral tilt over time, not absolute tilt, significantly influences perception of /b/ vs /d/. Experiments consisted of burstless synthesized stimuli that varied in spectral tilt and onset frequency of the second formant. In Experiment 1, tilt of the consonant at voice onset was varied. In Experiment 2, tilt of the vowel steady state was varied. Results of these experiments were complementary and revealed a significant contribution of relative spectral tilt change only when formant information was ambiguous. Experiments 3 and 4 replicated Experiments 1 and 2 in an /aba/-/ada/ context. The additional tilt contrast provided by the initial vowel modestly enhanced effects. In Experiment 5, there was no effect for absolute tilt when consonant and vowel tilts were identical. Consistent with earlier studies demonstrating contrast between successive local spectral features, perceptual effects of gross spectral characteristics are likewise relative. These findings have implications for perception in nonlaboratory environments and for listeners with hearing impairment. [DOI: 10.1121/1.2817617]

PACS number(s): 43.71.Es [JES] Pages: 386–396

I. INTRODUCTION

Experience with multiple acoustic differences between speech sounds gives rise to trading relations between attributes or cues for identification (e.g., Repp, 1982), thereby making speech perception robust (Kluender and Alexander, in press). When one attribute becomes less informative or ambiguous, perceptual constancy can be maintained by information provided by other attributes. For example, voice onset time (VOT) and fundamental frequency (*f*0) often covary in the production of stop consonants, with shorter VOTs and lower *f*0's corresponding to voiced consonants. Whether because of learned covariation (Holt *et al.*, 2001) or general auditory processes (Kingston and Diehl, 1994), increases in VOTs can be offset to some extent by decreases in *f*0 in order to maintain the perception of voicing (Whalen *et al.*, 1993). In other words, perception of voicing is not an absolute function of VOT or *f*0, but is dependent on these and other attributes of the signal. The literature is replete with examples of these trading relations (e.g., Repp 1982), which are by-products of the perceptual system's natural ability to extract relationships among multiple sources of information.

Spectrally global sources of information, such as gross spectral tilt (i.e., the balance of low- and high-frequency energy), may contribute to perception of place of articulation in stop consonants. However, the majority of research concerning gross spectral properties of stop consonants varying in place of articulation has been conducted to support or to critique claims of acoustic invariance (e.g., Stevens and Blumstein, 1978, 1981; Blumstein and Stevens, 1979, 1980; Blumstein *et al.*, 1982; Walley and Carrell, 1983; Kewley-Port, 1983; Lahiri *et al.*, 1984; Kewley-Port and Luce, 1984; Dorman and Loizou, 1996). It is now generally accepted that multiple sources of information influence perception and that no one source is absolute or invariant across all acoustic contexts for all listeners. It is likely that the importance of a particular acoustic attribute of speech is at least related to (1) its reliability (mean differences and variance) in distinguishing phonemes across speech contexts (2) listeners' sensitivity (ability to discriminate differences) to the attribute across phonemes, and (3) the presence of other acoustic attributes. Much work has focused on the first of these (see the following), but much less work has been done on the last two. For example, little is understood about how changes in the relative sensitivity of different acoustic attributes accompanying sensorineural hearing loss influence their importance in speech perception. This is the focus of other work in this lab. The focus of the current work on normal-hearing listeners is to examine how a spectrally global attribute, gross spectral tilt, conspires with spectrally local information, formant peaks, to inform perception of place of articulation in stop consonants.

Much of the research on gross spectral tilt in speech emanated from Stevens and Blumstein's work (e.g., 1979, 1980; Stevens and Blumstein, 1978, 1981) which advanced a strong argument that an invariant acoustic marker for place of articulation in stop consonants can be found by integrating the spectral energy of the first 25 ms or so following the onset of the release burst. For voiced consonants with absent or short-duration bursts, this integration window includes not only burst energy but also the first one or two pitch pulses associated with the onset of voicing. Stevens and Blumstein argued that perception of place of articulation is primarily determined by the shape of the stimulus onset spectrum, which is governed by the size and shape of resonator cavities created by different constriction points in the oral tract. In a

[a]Current address: Boys Town National Research Hospital, Omaha, NE 68131. Electronic mail: alexanderj@boystown.org

preemphasized signal (+6 dB/oct.), a labial place of articulation is characterized by a "diffuse-falling spectrum" (i.e., negative spectral tilt), an alveolar place of articulation by a "diffuse-rising spectrum" (i.e., positive spectral tilt), and a velar place of articulation by a "prominent midfrequency spectral peak." Kewley-Port (1983), who advocated for classification templates based on kinematic spectral features through the first 40 ms following stimulus onset, also adopted this classification convention for place of articulation, in addition to the timing of voice onset and the presence of midfrequency peaks. In contrast to early work (e.g., Liberman *et al.*, 1952; Halle *et al.*, 1957) that was event based and focused on spectral shape information exclusively in the burst, the spectral features in both the Stevens and Blumstein framework and the Kewley-Port framework extend to the vocalic portion of the syllable because they are time based and not limited to the burst.

Refinement of the above-mentioned approach as a template for invariance was motivated in part by the failure to accurately classify stops in Malayalam and French (Lahiri *et al.*, 1984) and by conflicting-cue experiments that seemed to indicate that onset frequencies of formant peaks influenced perception of stop consonants much more than absolute spectral tilt (Blumstein *et al.*, 1982; Walley and Carrell, 1983). Both Blumstein (Lahiri *et al.*, 1984) and Kewley-Port (Kewley-Port and Luce, 1984) altered their original templates based on static spectral shapes to include a kinematic component based on the relative change in tilt. That is, a positive change in tilt was used as characteristic marker for a labial place of articulation and a negative change in tilt for an alveolar place of articulation.

Lahiri *et al.* (1984) devised a perceptual experiment to support their classification data. Synthesizing "prototypical" consonant-vowel (CV) syllables as produced by a French speaker, they altered either the tilt of the burst or the tilt of the voicing onset of [b] and [d] in five different vowel contexts ([i], [e], [a], [o], [u]). For the [b] exemplars, relative tilt was altered to more closely resemble the pattern for [d]. That is, the tilt of the burst was increased to a diffuse-rising spectrum or the tilt of the voicing onset was decreased while the tilt of the burst was held constant so the change in tilt was negative. The opposite was done for the [d] exemplars. Lahiri *et al.* found that most of the stimuli were classified as the phoneme cued by relative tilt, and tilt manipulation of the burst was slightly more effective at altering perception than tilt manipulation of voicing onsets.

Several limitations of the Lahiri *et al.* (1984) perceptual experiment motivate the need for a clearer demonstration of the use of relative spectral tilt in speech perception. First, relative spectral tilt was considered only as a categorical variable using an arbitrary metric based on the ratio of the energy difference in the spectral envelope between burst and voicing onset at 3500 Hz and the difference at 1500 Hz. At the time, one was limited to this crude metric of relative tilt given the difficulty in generating the different tilts via formant amplitude manipulation in the parallel branch of the Klatt synthesizer (Klatt, 1980).[1] As noted by the authors, "because there was overlap in the skirts of the filters, a change in the amplitude of one formant peak often resulted in a change of spectral shape for other formant peaks" (p. 401). Finally, as noted by Dorman and Loizou (1996), the exemplar stimuli of Lahiri *et al.* were perceptually impoverished; 12 of the 30 subjects could not reliably identify the exemplar stimuli above 70% in three of the five vowel contexts.

Dorman and Loizou (1996) took a slightly different approach to the perceptual experiment of Lahiri *et al.* (1984). First, they used naturally produced speech rather than synthesized speech, which resulted in near perfect identification of the exemplar stimuli. Next, using the fast-Fourier transform (FFT) magnitude spectra of the stimuli, they altered the tilt of the burst or of the voicing onset to match the metric defined by Lahiri *et al.* (1984). Unlike Lahiri *et al.*, Dorman and Loizou found that altering relative tilt change did not substantially influence perception of /b/ and /d/ in any of the vowel contexts, except for the syllable /bi/.

While Dorman and Loizou (1996) demonstrated that relative spectral tilt is not an invariant cue for place of articulation, this does not imply that tilt plays no role in perception. As noted earlier, the importance of a particular acoustic attribute (e.g., spectral tilt) depends on a number of factors, including the quality of information provided by other acoustic attributes (e.g., formant frequency). From this perspective, the importance of Dorman and Loizou's findings with clearly spoken natural speech is a demonstration that the effects of relative tilt are overwhelmed when most other potential acoustic information is available to listeners. However, there are multiple circumstances when formant information is compromised. Listeners with hearing loss have diminished frequency resolution that obscures spectral peaks, and the presence of background noise likely has a similar effect for normal-hearing listeners. Outside the laboratory, fluent speech is often hypoarticulated such that ambiguous information specifying place might be the rule more than the exception. From this perspective, the greater importance of spectral tilt in Lahiri *et al.* (1984) can be explained by the fact that their stimuli had compromised information for perception of place.

In the following series of experiments, we evaluate the interaction between spectral tilt and formant frequency in perception of voiced stop consonants. While controlling for phonetic context and all other sources of information, each attribute independently varied along a series from [ba] to [da], including several intermediate or ambiguous levels.

II. STIMULI

A. Rationale

The goals in creating our stimuli were to control for all extraneous variables except those under study and to make the acoustic manipulations systematically. For this reason, we first synthesized a burstless /ba/ to /da/ series that varied acoustically in the onset frequency of the second formant (F2) transition. We then used digital filters to manipulate the tilt trajectory of the formant transitions. We chose a labial to alveolar series because the spectral shape manipulation is easily defined from negative to positive. In our case, because spectral shape is described without preemphasis, this series is

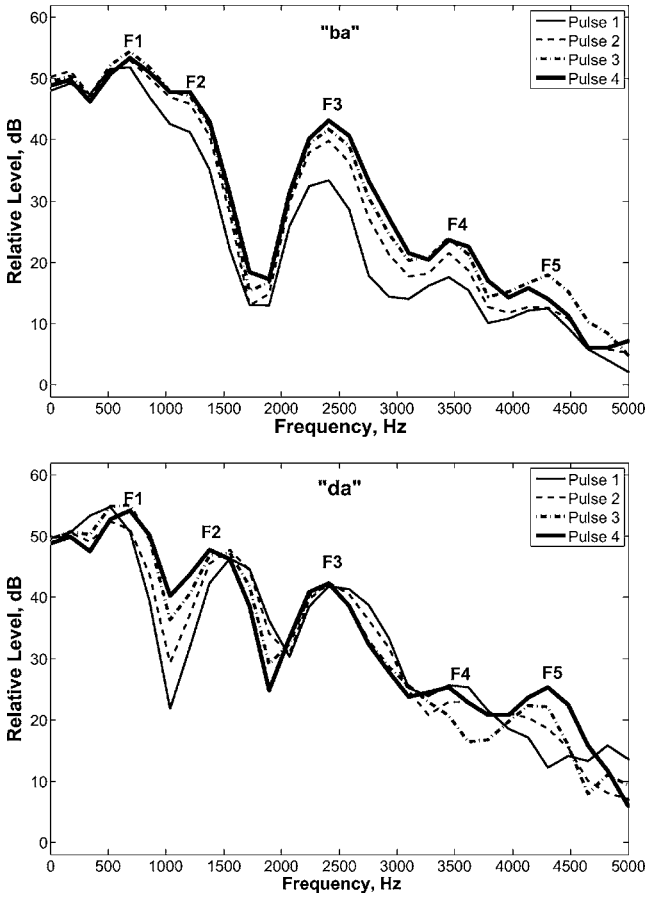


FIG. 1. The short-term spectra for the first four pitch pulses following the onset of voicing from a male talker with fundamental frequency of about 105 Hz are displayed. The top and bottom panels show the short-term spectral history of a production of the syllable /ba/ and the syllable /da/, respectively. Each pitch pulse (different lines) was about 9.5 ms in duration ($1/f0$), and was analyzed without preemphasis using a 256-point FFT with a 50% Hamming window overlap.

described as going from steep negative tilt (labial) to shallow tilt (alveolar). The low, back vowel /a/ was chosen because it has a relatively neutral spectral tilt and because phonetically trained and untrained listeners identified it as /a/ despite wide manipulations of spectral tilt. This contrasts with perception of high vowels, for example, which Kiefte and Kluender (2005) demonstrated to be quite sensitive to changes in spectral tilt. There is precedence for synthesizing burstless stops (e.g., Fruchter and Sussman, 1997) as it simplifies conclusions concerning the parameter under study. Furthermore, because stop bursts in vocalic positions other than the utterance initial position are often either low in energy or absent, synthesizing stops without the bursts extends the generality of our findings to typical running speech for which bursts are a less frequent event particularly for voiced consonants.

As noted earlier, the use of tilt as a cue for stop consonant perception is not limited to bursts, but also extends to the following vocalic segment. Changes in spectral tilt related to resonance properties (versus source properties) are to be expected as vocal tract shape changes. Figure 1 shows the short-term spectra for the first four pitch pulses following the onset of voicing from an adult male production (fundamental frequency of about 105 Hz) of the syllables [ba] (top panel) and [da] (bottom panel) extracted from a sample of running nonsense speech recorded with a sampling rate of 44 100 Hz. Here and elsewhere in this report, each pitch pulse was analyzed without preemphasis using a 256-point FFT with a 50% Hamming window overlap. The top panel shows that for this production of [ba], the short-term spectra of the initial pitch pulse ($t=0-9.5$ ms) following the onset of voicing (thin solid line) is steeply negative and quickly transitions to the shallower vowel tilt by the second or third pitch pulse (about 19–28.5 ms). In contrast, the short-term spectra of the initial pitch pulse for the production of [da] in the bottom panel is relatively shallow and changes very little over the time course of the formant transitions.[2]

B. Synthesis of F2 series

A burstless CV series varying perceptually from /ba/ to /da/ and acoustically in eight steps of F2-onset frequency (1000–1700 Hz, respectively) was synthesized at a sampling rate of 22 050 Hz with 16 bits of resolution and with a 5 ms update rate using the parallel branch of the Klatt synthesizer (Klatt and Klatt, 1990). The CVs were 250 ms total duration, including 30 ms formant transitions with a linear amplitude rise of 6 dB. Fundamental frequency was 100 Hz and decreased to 90 Hz during the final 50 ms. Synthesis parameters are provided in Table I. Because F4 was the highest

TABLE I. Klatt synthesis parameters for the eight-step series varying in the onset frequency of the second formant (F2). The amplitudes of F1 and F2 (A1V and A2V, respectively) were varied as a function of formant frequency, in order to maintain a relatively constant spectral tilt of −3 dB/oct. All other parameters were kept constant throughout the duration of the 250 ms stimuli, except $f0$ which started at 100 Hz and decreased to 90 Hz during the final 50 ms.

F1	A1V	BW1	F2	A2V	BW2	F3	A3V	BW3	F4	A4V	BW4
$t=0$ (constant onset)											
300	50	80	1000	70	90	2400	72	150	3600	80	350
			1100	69							
			1200	68							
			1300	67							
			1400	66							
			1500	65							
			1600	64							
			1700	63							
$t=30$ ms (vowel steady state)											
800	55	80	1200	67	90	2400	72	150	3600	80	350

formant frequency synthesized, its nominal amplitude (A4V) was set to maximum (80 dB) and nominal values in the synthesizer for the remaining formant amplitudes (A1V–A3V) were manipulated so that each of the CVs had a reasonably constant spectral tilt of −3 dB/oct. throughout its duration. Spectral tilt was referenced to the energy in F4 because it did not change frequency during the duration of the stimulus. This is expressed in the following:

$$(\mathrm{Amp}_{\mathrm{F2}})_{\mathrm{dB}} = -3 \times \log_2(\mathrm{F2/F4}) + (\mathrm{Amp}_{\mathrm{F4}})_{\mathrm{dB}}, \quad (1)$$

where $(\mathrm{Amp}_{\mathrm{F2}})_{\mathrm{dB}}$ is the desired amplitude of F2 in dB, −3 is the designated tilt, $\log_2(\mathrm{F2/F4})$ is the number of octaves between the fourth and second formant, and $(\mathrm{Amp}_{\mathrm{F4}})_{\mathrm{dB}}$ is the measured amplitude of F4 in dB. F1 and F3 were substituted for F2 in Eq. (1) to derive the desired amplitudes of the first and third formants. Formant frequency and amplitude values were linearly interpolated by the synthesizer between $t=0$ ms (consonant onset) and $t=30$ ms (beginning of vowel steady state). It is important to note that the amplitudes of the formants were empirically measured from spectra of the individual pitch pulses (about every 10 ms) and were used to set the nominal amplitude values in the synthesizer so that spectral tilt for all four formants was reasonably constant [cf. Eq. (1)].

III. EXPERIMENT 1: EFFECT OF CONSONANT ONSET TILT ON CV PERCEPTION

A. Rationale

We predict that spectral tilt will have the greatest influence for stimuli in which F2-onset frequency is ambiguous between [ba] and [da]. When formant cues to place of articulation are ambiguous, steeper tilts at onset (more negative) should encourage perception of /ba/ and shallower tilts at onset (more positive) should encourage perception of /da/. To test this hypothesis, stimuli varying in F2-onset frequency (see earlier text) were filtered to generate a series of five spectral tilts that varied at consonant onset from −12 to 0 dB/oct. These end point tilts are equivalent to Blumstein and Stevens's templates for labial and alveolar stop consonants, respectively (e.g., Blumstein and Stevens, 1979). For all stimuli, spectral tilt transitioned along with F1 and F2 frequency from an initial onset value at $t=0$ ms to a common value in the vowel portion of the syllables at $t=30$ ms.

B. Methods

1. Listeners

Listeners for all experiments were undergraduate students from the University of Wisconsin–Madison and participated as part of course credit. No listener participated in more than one experiment. All reported that they were native speakers of American English and had normal hearing. One to three listeners ran in the experiment concurrently. Each individual was seated in an isolated single-walled sound chamber and had a unique presentation order of the stimuli. Listeners were recruited for each experiment until data were collected on at least 20 participants. Twenty-three listeners (3 male, 20 female) participated in Experiment 1.

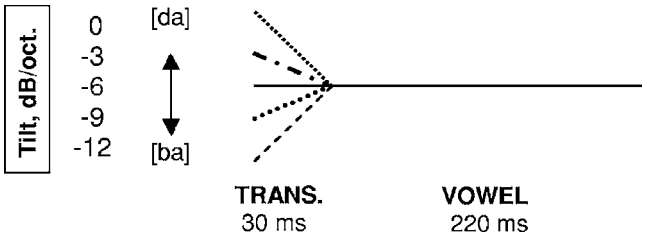


FIG. 2. Schematic representing the change in tilt for the stimuli in Experiment 1 in which different consonant onset tilts converged to a −6 dB/oct. vowel tilt. Consonant onset tilts steeper than the vowel tilt are expected to result in more labial responses and consonant onset tilts shallower than the vowel tilt are expected to result in more alveolar responses.

2. Stimuli

Using the eight-step series varying in F2-onset frequency, parametric manipulations of spectral tilt were made on time slices cut at the zero crossings corresponding to the midpoint and the end of each pitch pulse (about every 5 ms). Stimuli were filtered between 212 and 4800 Hz, using 90-order finite impulse response (FIR) filters created in MATLAB, to have one of five different spectral tilts at consonant onset ranging from −12 to 0 dB/oct. (Because stimuli already had a constant −3 dB/oct. tilt, the actual slopes of the filters varied from −9 to +3 dB/oct.) During formant transitions ($t \leqslant 30$ ms), spectral tilt converged linearly to the vowel steady state, which was filtered as a whole segment to a tilt of −6 dB/oct. (see Fig. 2). Because initial portions of wave forms are not filtered accurately when the length of the impulse response approaches that of the wave form, each input wave form was concatenated several times and then convolved with the FIR filters. From the medial portion of this filtered wave form, the output wave form was extracted at zero crossings corresponding to the original input wave form and then scaled to the RMS amplitude of the original time slice. The CVs were unsampled to 48 828 Hz with 24 bits of resolution and low-pass filtered with an 86-order FIR filter with a passband at 4800 Hz and a stopband of −90 dB of at 6400 Hz. The CVs were then scaled to a constant rms amplitude.

Figure 3 displays the short-term spectra of the first four pitch pulses from sample stimuli with F2-onset frequencies of 1400 Hz. Each pitch pulse is approximately 10 ms in duration. In each panel, the thin solid line is the short-term spectra of the first pitch pulse ($t \approx 0-10$ ms) and represents the consonant onset and the thick solid line is the short-term spectra of the fourth pitch pulse ($t \approx 30-40$ ms) and represents the beginning of the steady state portion of the following vowel. The short-term spectra of the second and third pulses show a continuous change in tilt and formant frequency between the first and fourth pitch pulse. Following Blumstein and Stevens's templates (e.g., Blumstein and Stevens, 1979), stimuli with steeply negative consonant onset tilts like the one represented in the top panel are expected to lead to more /ba/ responses and stimuli with relatively shallow onset tilts as in the bottom panel are expected to lead to more /da/ responses.

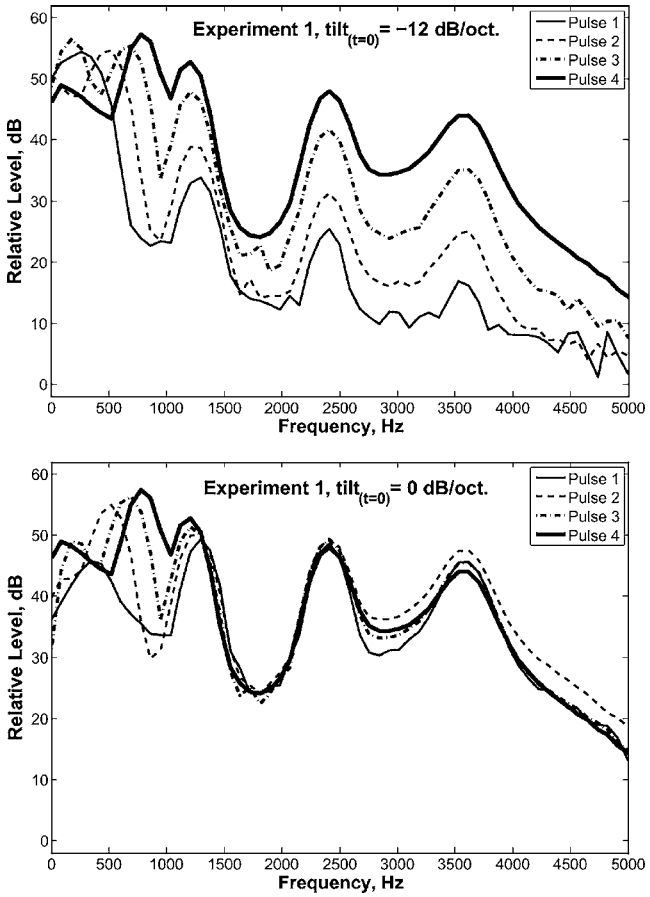


FIG. 3. Short-term spectra for the first four pitch pulses (about 10 ms each) for the tilt end point stimuli in Experiment 1. This example represents the stimuli with an F2-onset frequency of 1400 Hz. It is hypothesized that the stimuli in the top panel will lead to more labial responses because the tilt is steeply negative for the initial pitch pulse (thin solid line) and becomes shallower over the duration of the formant transition until it reaches a steady state at the vowel onset (thick solid line). The stimuli represented in the bottom panel are hypothesized to lead to more alveolar responses because the tilt is flat for the initial pitch pulse and becomes steeper over the duration of the formant transition until it reaches a steady state at the vowel onset.

3. Procedure

Participants listened to each of the 40 CV tokens (eight F2-onset frequencies by five spectral tilt trajectories) once per trial block in randomized order. Following two warm-up blocks (80 trials), data were collected on eight subsequent blocks (320 trials). Stimuli were presented diotically to participants through Beyerdynamic DT150 headphones at an average level of 73 dBA. In a two-alternative, forced choice task participants indicated their responses by pressing one of two buttons labeled "BA" and "DA."

C. Results

For every listener, at each tilt manipulation, the probability of responding /da/ as a function of F2-onset frequency was fit to a logistic function using the psignifit toolbox for MATLAB[3] which implements the maximum-likelihood method described by Wichmann and Hill (2001). The upper and lower asymptotes were free to vary in order to achieve the best fit. Frequency of F2 corresponding to the $p=0.5$ point on the ordinate, where /ba/ and /da/ responses are equally likely, hereafter the boundary, was obtained from the

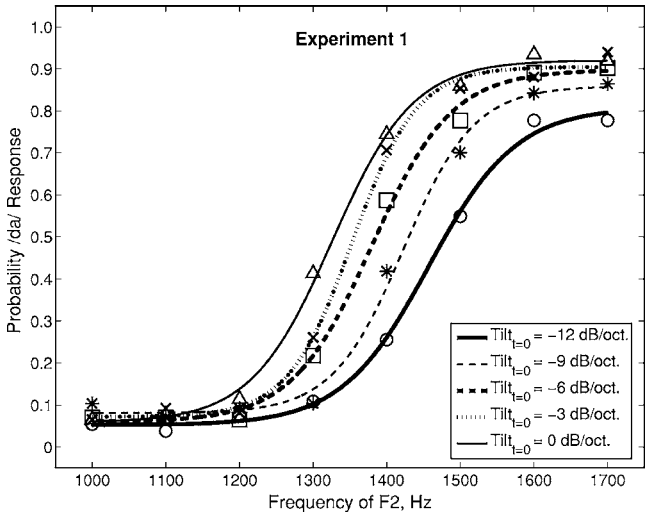


FIG. 4. Mean data for Experiment 1 in which the probability of responding /da/ as a function of F2-onset frequency is plotted separately for each consonant onset tilt. Circles, asterisks, squares, crosses, and triangles represent the mean data for consonant onset tilts of −12, −9, −6, −3, and 0 dB/oct., respectively. Maximum-likelihood fits of the identification functions are displayed for the mean data at each consonant onset tilt as different lines (see the legend).

fits. The higher the boundary frequency, the greater the likelihood of a /ba/ response over the range of F2-onset frequencies. Figure 4 displays identification functions of the mean data for each consonant onset tilt; circles, asterisks, squares, crosses, and triangles represent the mean data for consonant onset tilts of −12, −9, −6, −3, and 0 dB/oct., respectively, and the different lines represent fitted functions for the mean data (see the legend).

Listeners' identifications were primarily influenced by onset frequencies of F2 as indicated by the floor/ceiling identifications of the formant frequencies near the series end points. However, tilt of consonant onset also influenced listeners' identifications, especially at intermediate values of F2 (e.g., 1400 Hz) where formant information for place of articulation was ambiguous. Listeners were more likely to perceive /da/ for consonant onsets with shallower tilt compared to consonant onsets with steeper tilt. The influence of consonant onset tilt can be quantified by shifts in the identification functions along the F2 axis—the boundary frequencies. Figure 5 displays the mean frequency and 95% confidence intervals (CIs) of boundaries at each consonant onset tilt. As can be seen, there was a systematic decrease in boundary frequency (increase in bias for /da/ responses) for progressively shallower consonant onset tilts (e.g., toward 0 dB/oct.). A within-subjects analysis of variance (ANOVA) confirmed that the effect of consonant onset tilt was statistically significant $[F(4,88)=16.1,\ p<0.0001]$. Tukey HSD post hoc tests revealed that the mean boundary frequency for the −12 dB/oct. consonant onset tilt was significantly greater than the mean boundary frequencies of the other consonant onset tilts except the −3 dB/oct. consonant onset tilt. Also, the mean boundary frequency for the 0 dB/oct. consonant onset tilt was significantly less than the mean boundary frequencies of the other consonant onset tilts except the −3 dB/oct. consonant onset tilt.[4] In addition, mean boundary frequencies for the −9 and −3 dB/oct. consonant onset tilts

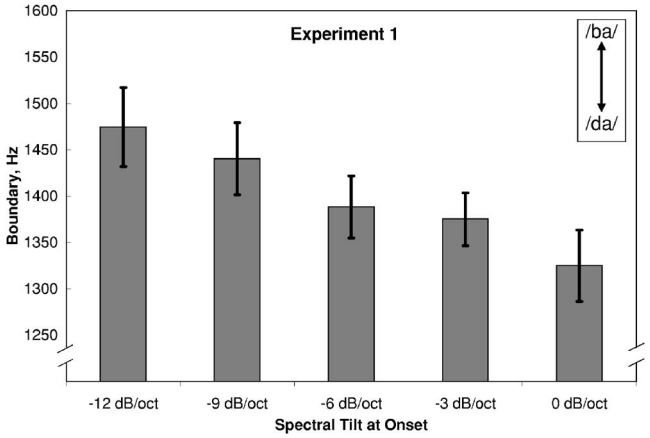


FIG. 5. Mean boundary frequencies of the identification functions at each consonant onset tilt from Experiment 1. Lower-frequency boundaries are associated with a bias toward /da/ responses. Error bars represent the 95% confidence intervals.

were significantly different from one another. This pattern of results indicates that shallower consonant onset tilts progressively increase the bias for alveolar responses and vice versa for the steeper consonant onset tilts creating a bias for labial responses.

IV. EXPERIMENT 2: EFFECT OF FOLLOWING VOWEL TILT ON CV PERCEPTION

A. Rationale

The results of Experiment 1 clearly indicate that despite an absence of bursts, the gross shape of the spectrum during the voiced portions of consonant onset influences perception of /ba/ and /da/. Following the templates of Blumstein and Stevens (e.g., Blumstein and Stevens, 1979), which were based on preemphasized spectra (+6 dB/oct.), negative spectral tilts at onset in our experiment (less than −6 dB/oct.) were associated with more labial (/ba/) responses and positive spectral tilts at onset (greater than −6 dB/oct.) were associated with more alveolar (/da/) responses. However, as noted by others (Lahiri *et al.* 1984; Kewley-Port and Luce, 1984), the important feature of the onset spectrum is its shape relative to the following vowel. Spectral tilt that becomes shallower (more positive) from consonant onset to vowel steady state (e.g., relative negative consonant onset tilt) should encourage perception of /ba/ and spectral tilt that becomes steeper (more negative) from consonant onset to vowel steady state (e.g., relative positive consonant onset tilt) should encourage perception of /da/. If relative tilt change is the perceptual cue used by listeners, then the pattern of results in Experiment 1 should be maintained if consonant onset tilt is held constant and vowel tilt varied. Experiment 2 tests this hypothesis.

B. Methods

Twenty-one college students who reported normal hearing (6 male, 15 female) identified as /ba/ or /da/ a series of 40 CVs that varied along both F2-onset frequency in eight steps and along spectral tilt of the following vowel in five steps ranging from −12 to 0 dB/oct. (see Fig. 6). Stimuli were filtered using the same techniques as described for Experiment 1. Participants listened to each of the CV tokens once per trial block in randomized order. Following two warm-up blocks (80 trials), data were collected on eight subsequent blocks (320 trials). Stimuli were presented diotically at an average level of 72 dBA.

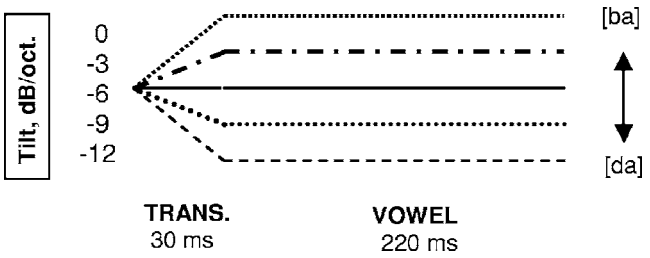


FIG. 6. Schematic representing the change in tilt for the stimuli in Experiment 2 in which a −6 dB/oct. consonant onset tilt diverged to different vowel tilts. Vowel tilts shallower (more positive) than the consonant onset tilt are expected to result in more labial responses and vowel tilts steeper (more negative) than the consonant onset tilt are expected to result in more alveolar responses.

Figure 7 displays the short-term spectra of the first four pitch pulses from sample stimuli in Experiment 2 with F2-

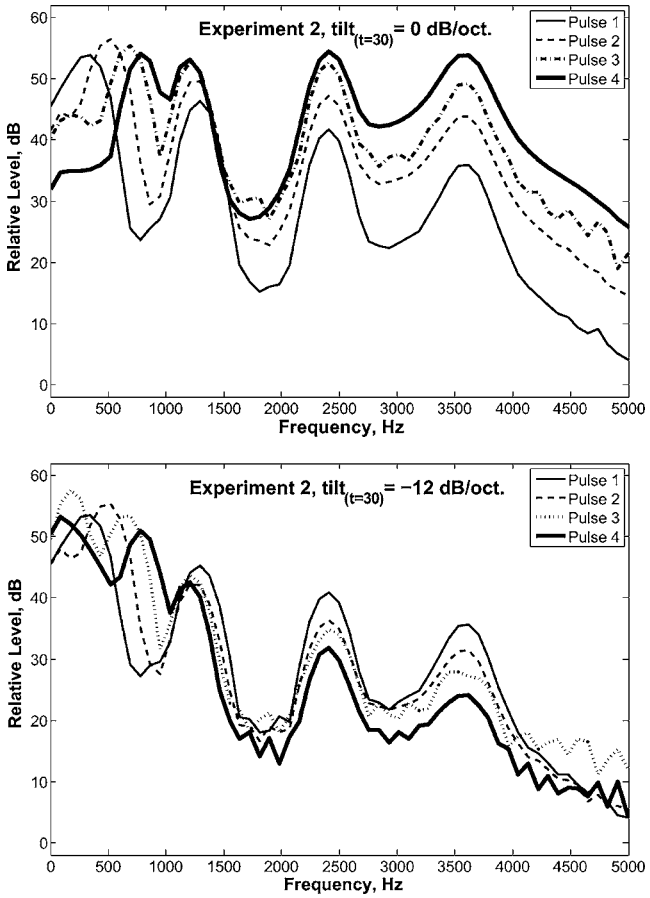


FIG. 7. Short-term spectra for the first four pitch pulses for the tilt endpoint stimuli in Experiment 2. As with Fig. 3, this example represents the stimuli with an F2-onset frequency of 1400 Hz. It is hypothesized that the stimuli in the top panel will lead to more labial responses and that the stimuli in the bottom panel will lead to more alveolar responses despite having identical stimulus onset spectra because the change in tilt is different. For the top panel, tilt becomes shallower until it reaches a flat spectrum that is sustained during the duration of the vowel, whereas, for the bottom panel, tilt becomes steeper until it reaches a steeply negative spectrum for the vowel.

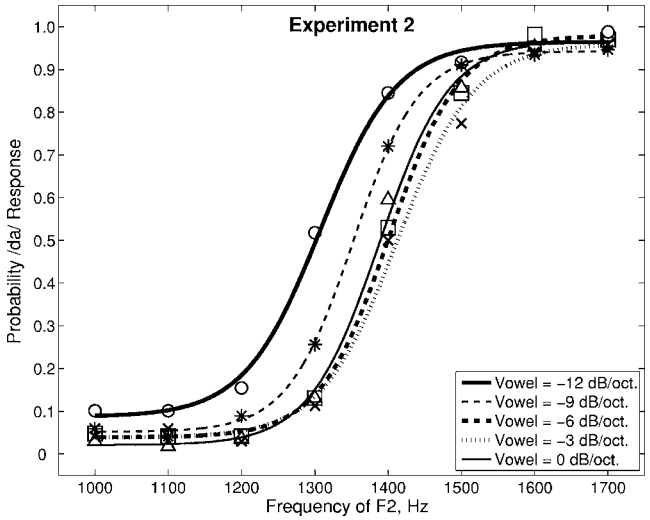


FIG. 8. Mean data for Experiment 2 in which the probability of responding /da/ as a function of F2-onset frequency is plotted separately for each vowel tilt. Circles, asterisks, squares, crosses, and triangles represent the mean data for vowel tilts of −12, −9, −6, −3, and 0 dB/oct., respectively. Maximum-likelihood fits of the identification functions are displayed for the mean data at each vowel tilt as different lines (see the legend).

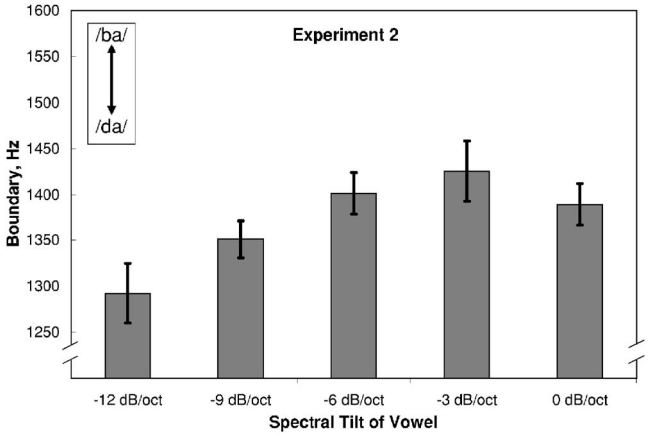


FIG. 9. Mean boundary frequencies of the identification functions at each vowel tilt from Experiment 2. Error bars represent the 95% confidence intervals.

onset frequencies of 1400 Hz (cf. Fig. 3). Notice that the initial pitch pulses (thin solid lines) in both the top and bottom panels have the same tilt but diverge to two different steady state vowel tilts by the fourth pulse (thick solid line). We hypothesize that manipulation of vowel tilt should lead to a complementary pattern of results as Experiment 1. The top panel shows a relative flattening of spectral tilt (−6 to 0 dB/oct.) from consonant onset (t=0 ms) to vowel steady state (t=30 ms). This pattern of change is predicted to increase the perception of a labial stop consonant. In contrast, the bottom panel is predicted to increase the perception of an alveolar stop consonant because spectral tilt becomes steeper over the course of the consonant transition.

C. Results

Figure 8 displays the mean data and identification functions for each vowel tilt. Circles, asterisks, squares, crosses, and triangles represent the mean data for vowel tilts of −12, −9, −6, −3, and 0 dB/oct., respectively, and the different lines represent the fitted functions for the mean data (see the legend). As with Experiment 1, listeners' perceptions were primarily influenced by onset frequency of F2. However, they were also influenced by tilt of the following vowel, especially at intermediate values of F2. Listeners were more likely to perceive /da/ for steeper vowel tilts. Figure 9 displays the mean frequency and 95% CIs of boundaries at each vowel tilt. With the exception of the boundary for the 0 dB/oct. vowel tilt, there was a systematic increase in the mean boundary frequency (increase in bias for /ba/ responses) with shallower vowel tilts. A within-subjects ANOVA confirmed that the effect of vowel tilt was statistically significant [$F(4,80)=21.7, p<0.0001$]. Tukey HSD post hoc tests revealed that the mean boundary frequency for the −12 dB/oct. vowel tilt was significantly less than the mean boundary frequencies of the other vowel tilts. In addition, the mean boundary frequency for the −9 dB/oct. vowel tilt was significantly less than the mean boundary frequencies of the −6 and −3 dB/oct. vowel tilts. None of the differences in mean boundary frequencies for the vowel tilts between −6 and 0 dB/oct. were statistically significant.

D. Discussion of Experiments 1 and 2

The results of Experiments 1 and 2 demonstrate that for burstless, voiced stops, the spectral tilt of the consonant formant transitions is perceived relative to the tilt of the following vowel. Perception of consonant onset tilt is contrastive to the tilt of the following vowel. Relatively *shallow* vowel tilts encourage perception of /ba/ (*steeper* consonant onset tilt) and relatively *steep* vowel tilts encourage perception of /da/ (*shallower* consonant onset tilt). Experiments 1 and 2 also indicate that the influence of relative tilt depends on the ambiguity of formant cues. When F2-onset frequency is near the series end points and most appropriate for /ba/ (1000 Hz) or most appropriate for /da/ (1700 Hz), relative tilt has little or no influence on perception. This result is consistent with results from Dorman and Loizou (1996), who used naturally produced stimuli with very high identification rates. However, when F2-onset frequency is intermediate between [ba] and [da], relative tilt has a substantial effect on stop consonant identification. This result is consistent with Lahiri *et al.* (1984), who used perceptually impoverished stimuli with regard to place of articulation. In summary, perception of the place of articulation in CVs (specifically, /ba/ and /da/) when formant information is ambiguous can be largely influenced by the relative change in spectral tilt.

V. EXPERIMENTS 3 AND 4: EFFECT OF RELATIVE TILT CHANGE ON VCV PERCEPTION

A. Rationale

Experiments 1 and 2 demonstrate the relative and contrastive influence of spectral tilt change as a perceptual cue to stop consonant identification in a CV context without bursts. In contrast to an utterance initial position, stop consonants in medial position are much more common in natural speech and much less likely to have a significant burst release. The next set of experiments were designed to test the influence

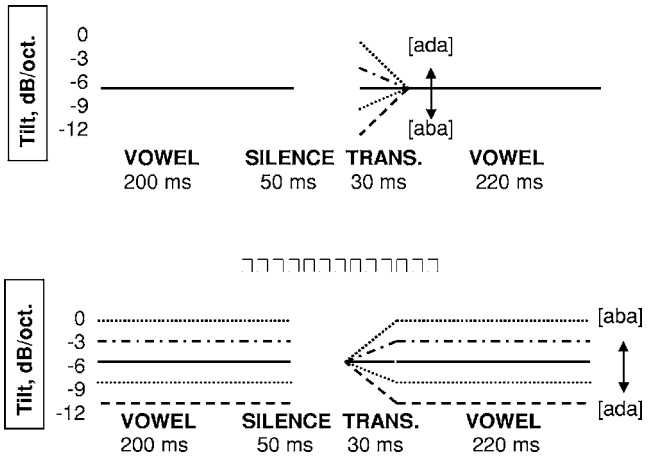


FIG. 10. Schematic for the stimuli in Experiment 3 (top panel) and Experiment 4 (bottom panel). The addition of a preceding vowel should enhance the perceptual cue associated with the spectral tilt change during the consonant transition.

spectral tilt change on the perception of burstless stop consonants in medial position. Specifically, Experiments 3 and 4 tested if effects of tilt are enhanced in a VCV context in which preceding vowels share the same critical acoustic features as the following vowels.

B. Methods

Experiments 3 and 4 replicated Experiments 1 and 2, respectively, using 25 (10 male, 15 female) and 22 (6 males, 16 females) college students who reported normal hearing, respectively. Stimuli were created by appending a 200 ms [a] followed by 50 ms of silence to the stimuli in Experiments 1 and 2. Formant frequencies and spectral tilt of the preceding [a] were matched to the following [a] to enhance the perception of tilt and formant frequency change associated with the consonant onset. As shown in Fig. 10, spectral tilt of the preceding [a] in Experiment 3 was always −6 dB/oct. (top panel) but varied in Experiment 4 from trial to trial in the same way the tilt of the following vowel varied (bottom panel). In a diotic presentation, listeners identified the series of 40 VCVs as /aba/ or /ada/ eight times in separate blocks (320 trials) following two warm-up blocks (80 trials).

C. Results and discussion

Figure 11 displays the mean data and fitted identification functions for Experiment 3, and Fig. 12 displays the mean boundary frequencies and 95% CIs. Compared to the data points in Fig. 4 (CVs), the data points in Fig. 11 (VCVs) show a greater divergence as a function of consonant onset tilt at intermediate F2 frequencies, especially at 1300 Hz. The effect of consonant onset tilt was highly significant $[F(4,96)=64.9,\ p<0.0001]$ in a within-subjects ANOVA. Tukey HSD post hoc tests revealed that each paired comparison of the mean boundary frequencies for the different consonant onset tilts was statistically significant except for the mean boundary frequencies of the −3 and 0 dB/oct. consonant onset tilts. Analyses of the mean boundary frequencies across Experiments 1 and 3 reveal that providing an additional comparison for consonant onset tilt in the form of a

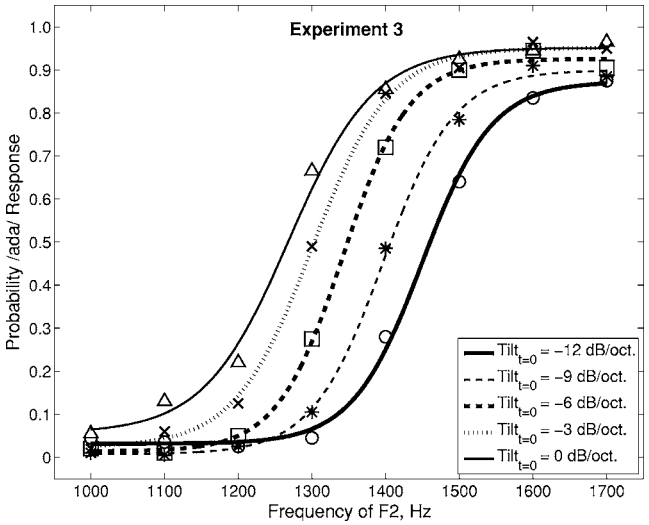


FIG. 11. Mean data and maximum-likelihood fits for Experiment 3 plotted in the same way as Fig. 4 in which the probability of responding /ada/ as a function of F2-onset frequency is plotted separately for each consonant onset tilt.

preceding vowel resulted in significantly negative shifts in the mean boundary frequencies (increase in bias for alveolar responses) for the stimuli with consonant onset tilts of −3 dB/oct. $[t(46)=4.14,\ p=0.0001]$ and 0 dB/oct. $[t(46)=2.16,\ p<0.05]$.

Figure 13 displays mean data and fitted identification functions for Experiment 4, and Fig. 14 displays the mean boundary frequencies and 95% CIs. Compared to data for Experiment 2, addition of the preceding vowel contrast in Experiment 4 resulted in a significant positive shift in mean boundary frequency (increase in bias for labial responses) for the stimuli with 0 dB/oct. vowel tilts $[t(41)=-2.17,\ p<0.05]$. This is evidenced in Fig. 14, which now shows an orderly increase in labial responses with progressively shallower vowel tilts. The overall effect of vowel tilt in Experiment 4 was significant $[F(4,84)=14.4,\ p<0.0001]$ according to a within-subjects ANOVA. Tukey HSD post hoc tests revealed that the mean boundary frequency for the −12 dB/oct. vowel tilt was significantly less than the mean

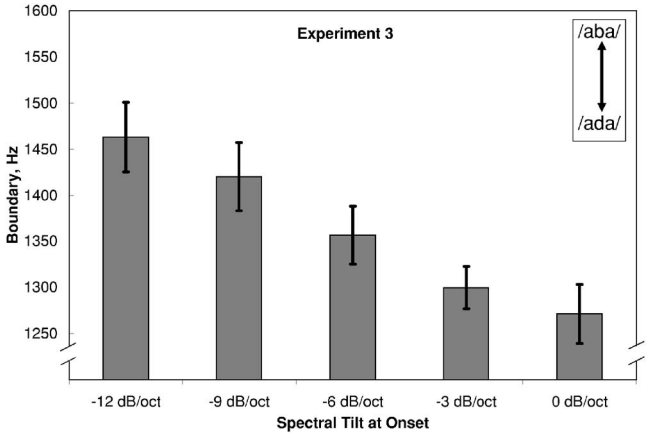


FIG. 12. Mean boundary frequencies of the identification functions at each consonant onset tilt in Experiment 3. Lower-frequency boundaries are associated with a bias toward /ada/ responses. Error bars represent the 95% confidence intervals.

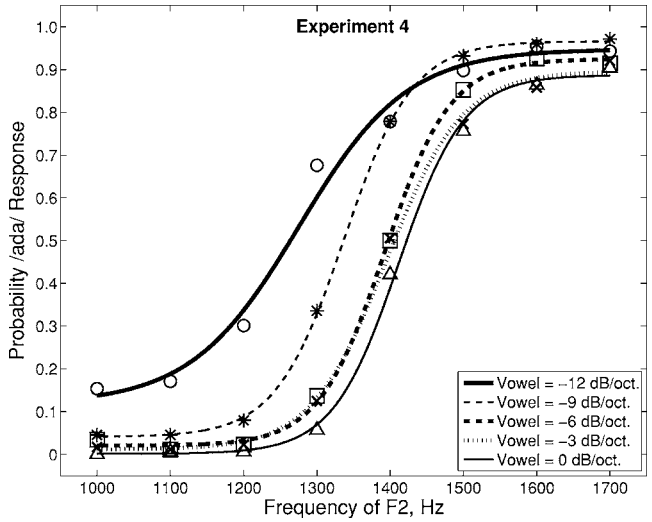


FIG. 13. Mean data and maximum-likelihood fits for Experiment 4 plotted in the same way as Fig. 8 in which the probability of responding /ada/ as a function of F2-onset frequency is plotted separately for each vowel tilt.

boundary frequencies of the stimuli with vowel tilts between −6 and 0 dB/oct. Additionally, the mean boundary frequency for the −9 dB/oct. vowel tilt was significantly less than the mean boundary frequencies of the −3 and 0 dB/oct. vowel tilts.

Results of Experiments 3 and 4 demonstrate the effects of relative spectral tilt change on the perception of medial position stop consonants. As is the case with following vowel tilt, the influence of preceding vowel tilt on the perception of labial versus alveolar voiced stops is contrastive.

VI. EXPERIMENT 5: EFFECTS OF ABSOLUTE TILT

A. Rationale

The above-mentioned experiments establish that relative spectral tilt change from consonant to vowel influences voiced stop consonant perception. The purpose of this final experiment is to test whether the influence of spectral tilt in stop consonant perception depends on a change in tilt or whether static differences in consonant onset tilt are enough to influence stop consonant perception (cf. Blumstein and Stevens, 1979).

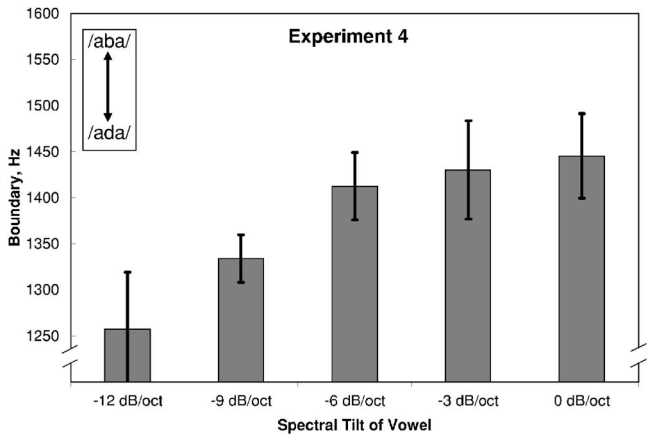


FIG. 14. Mean boundary frequencies of the identification functions at each vowel tilt in Experiment 4. Error bars represent the 95% confidence intervals.

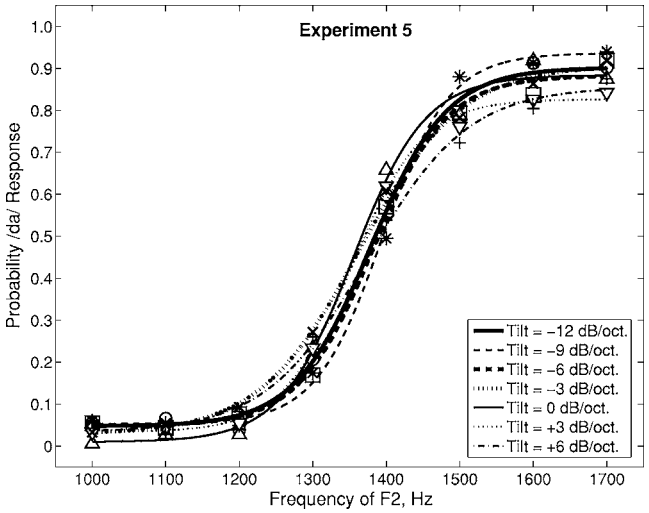


FIG. 15. Mean data for Experiment 5 in which the probability of responding /da/ as a function of F2-onset frequency is plotted separately for each absolute spectral tilt. Circles, asterisks, squares, crosses, triangles, inverted triangles, and plus signs represent the mean data for CVs with absolute tilts of −12, −9, −6, −3, 0, +3, and +6 dB/oct. respectively. Maximum-likelihood fits of the identification functions are displayed for the mean data at each tilt as different lines (see the legend).

B. Methods

Twenty-three college students who reported normal hearing (9 male, 14 female) identified a series of 56 CVs that varied from [ba] to [da] along both F2-onset frequency (eight steps) and absolute tilt (constant tilt throughout duration) in seven steps (−12 to +6 dB/oct.). Participants listened to every CV token once per trial block in randomized order. Following two warm-up blocks (112 trials), data were collected on eight subsequent blocks (448 trials). Stimuli were presented diotically at an average level of 73 dBA.

C. Results

Figure 15 displays the mean data and fitted identification

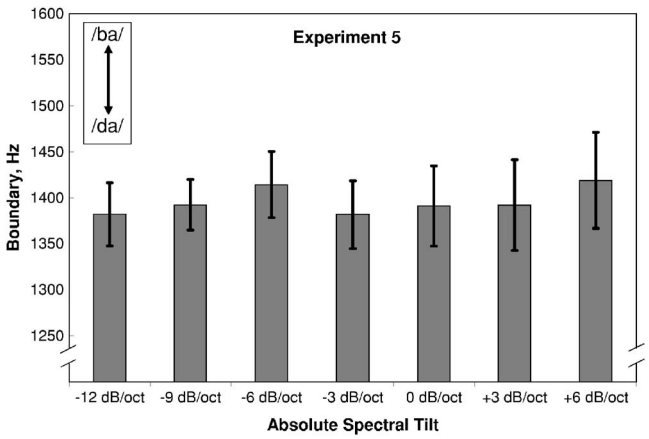


FIG. 16. Mean boundary frequencies of the identification functions at each absolute tilt from Experiment 5. Error bars represent the 95% confidence intervals.

functions for Experiment 5 and Fig. 16 displays the mean boundary frequencies and 95% CIs. From the data, it is clear that absolute tilt, over a wide range, had virtually no effect on listeners' perception [$F(5,132)=1.0$].[5] From this, we can conclude that consonant onset tilt alone, absent relative change, is not used by listeners to classify stop consonants. Matching the tilt of the following vowel to the tilt of the formant transitions, effectively nullifies the effect of consonant onset tilt on perception.

VII. GENERAL DISCUSSION

Results of the experiments in this report establish that the influence of spectral shape on stop consonant perception is dependent on the availability of other cues to place of articulation, especially formant peaks. When formant information was at the series' end points and most appropriate for [ba] or [da], spectral tilt had little to no influence on identification (cf. Dorman and Loizou, 1996). When formant information was ambiguous and intermediate between [ba] and [da], spectral tilt had a substantial influence on identification (cf. Lahiri *et al.*, 1984). A similar pattern of results was observed by Abramson and Lisker (1985) for the influence of VOT and f_0 on the perception of voicing, with the influence of f_0 maintaining only at ambiguous VOT values. Moreover, the critical feature of spectral tilt is the tilt of the consonant onset relative to the tilt of the following vowel. A negative (steeper) consonant onset tilt relative to vowel tilt encourages perception of a labial place of articulation and a positive (shallower) consonant onset tilt relative to vowel tilt encourages perception of an alveolar place of articulation. Manipulations of consonant onset tilt (Experiments 1 and 3) and vowel tilt (Experiments 2 and 4) are complementary, and either is sufficient to influence perception.

There are several reasons to believe that our results are conservative with respect to natural speech and to the wide inventory of speech sounds and acoustic contexts. First, consider fluent conversational speech in the presence of competing sounds as encountered by a normal-hearing listener. We have already shown that adding preceding context can further enhance the influence of tilt in burstless, voiced stop consonants. Furthermore, articulation is less precise (i.e., formant frequencies for different phonemes are less extreme) in connected speech. Addition of ambient sound sources further undermines resolution of spectral peaks. Second, consider listeners with sensorineural hearing loss (SNHL), for whom spectral detail (e.g., formant peaks) is often compromised. Because only a gross characterization of the spectrum is necessary to encode spectral tilt, it likely takes on greater importance in speech perception by hearing-impaired listeners. In ongoing research, the current experiments have been replicated by listeners with SNHL. For these listeners, spectral tilt can and often does dominate perception of place of articulation even for frequencies near the F2 end points. Finally, our depiction of tilt effects for normal-hearing listeners may be conservative inasmuch as the appearance of mitigated effects near F2-onset end points could be the result of ceiling and floor effects.

There is evidence that our choice of stimuli may also have worked against stronger effects of spectral tilt. First, one can expect that had we extended our manipulations to include bursts, our findings would be an exaggerated version of the present findings because of the additional spectral tilt information provided by the burst. Furthermore, in a series of deleted-cue and conflicting-cue experiments, Smits *et al.* (1996) found that the relative effectiveness of bursts (spectrally gross information, including shape, level, and duration) and formant transitions (spectrally detailed information) in the perception of place of articulation in stop consonants depended on consonant voicing and vowel context. Specifically, the relative influence of spectrally global information in the bursts was more dominant for voiceless stops compared to voiced stops and more dominant for front vowel contexts (e.g., /i/) compared to nonfront vowel contexts (e.g., /a/). Interestingly, the only effect for Dorman and Loizou's (1996) spectral tilt manipulation occurred in the /i/ context, and the effects of Lahiri *et al.* (1984) were strongest for the /i/ context and weakest for the /a/ context. Smits (1996) argued that these findings have less to do with the identity of the vowel and more to do with the reliability (mean differences and variance) of the acoustic information. That is, an acoustic analysis of voiceless stop consonants across the three places of articulation revealed that in the /a/ context, the formant frequencies were more different from one another while bursts were more similar to one another. The situation was opposite for the /i/ context in which the formant frequencies were more similar to one another and the bursts were more different from one another.

Results of these experiments also establish that change in spectral tilt, not absolute tilt, is perceptually effective. When tilt varied over a wide range from trial to trial, but did not change within a stimulus, tilt had no effect on perception (Experiment 5). Further, perception of relative spectral tilt was enhanced when a preceding vowel tilt also contrasted with the consonant onset tilt (Experiments 3 and 4). The idea that acoustic features are encoded relative to one another across time is not new. For example, locus equations, which are linear regression equations that compare the F2 frequency at voicing onset to the F2 frequency of the following vowel steady state, were used by Lindblom (1963) to describe the contextual relationship of F2. Sussman and colleagues (e.g., Sussman *et al.*, 1991) have further developed the concept of locus equations in efforts to define an invariant cue for place of articulation. Our findings with respect to spectrally global information (i.e., tilt) are similar to Sussman's locus equations for spectrally local information in that they both are relational in nature, however, we cannot make any claims of invariance given the restricted set of stimuli employed here. We demonstrated that listeners can and do use the change in spectral tilt as information to classify labial and alveolar voiced stop consonants in an /a/ context, especially when other information specifying place of articulation is absent or ambiguous. However, it could be that, if our general relations were expanded to resemble specific locus equations, then separate equations would need to be calculated for different speakers, vowel contexts, and for voicing/unvoicing just as has been done for locus equations. Such determinations for any model of consonant identification are

likely to be influenced by the relative reliability of the acoustic information across contexts and the presence of other acoustic attributes.

Finally, the importance of relative tilt, or tilt contrast, is consistent with multiple demonstrations that perception of coarticulated speech, signaled by formant changes, is facilitated by spectral contrast between local spectral prominences (formants) (e.g., Coady *et al.*, 2003; Holt and Kluender, 2000; Holt *et al.*, 2000; Lotto and Kluender, 1998; Lotto *et al.*, 1997; Kluender *et al.*, 2003). The fact that the perceptual efficacy of spectral tilt is dependent on change is expected given that sensorineural systems, in general, respond predominantly to change relative to what is predictable or does not change (Kluender *et al.*, 2003; Kluender and Alexander, in press; Kluender and Kiefte, 2006).

ACKNOWLEDGMENTS

The authors would like to thank two anonymous reviewers whose comments on an earlier draft proved extremely helpful. The authors also thank Amanda Baum, Rebecca Edds, Tricia Nechodom, and Rebecca ("Hallie") Strauss for their efforts during the data collection process. This research was supported by a grant to J.M.A. from The National Organization for Hearing Research Foundation ("*The 2006 Graymer Foundation Grant in Auditory Science*") and a grant to K.R.K. from the National Institutes of Health, NIDCD (R01DC04072). Portions of this manuscript were written while receiving support from NIDCD T32 DC000013 (BTNRH).

[1]In Klatt's synthesizers (1980, Klatt and Klatt, 1990), spectral tilt is increased or decreased by adjusting a single-pole filter such that the broad high-frequency skirt of the filter alters the overall shape of the spectrum.

[2]The point of this speech production example is not to make a claim of acoustic invariance (cf. Stevens and Blumstein, 1978; 1981; Blumstein and Stevens, 1979; 1980). The important point is that when spectra tilt information is present it can influence perception.

[3]Software version 2.5.41. See http://bootstrap-software.org/psignifit.

[4]Unless otherwise stated, statistical significance is assumed to be $p < 0.05$ on a two-tailed test.

[5]One listener who did show an effect for absolute tilt responded /aba/ for all stimuli with a +6 dB/oct. tilt, hence, no boundary frequency could be computed. The boundary frequency for +3 dB/oct., 1717 Hz, was substituted for the missing data point. Unexplainably, increases in absolute tilt resulted in more /aba/ responses for this listener.

Abramson, A. S., and Lisker, L. (**1985**). "Relative power of cues: F_0 shift versus voice timing," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. Fromkin (Academic, Orlando, FL), pp. 25–33.

Blumstein, S. E., Isaacs, E., and Mertus, J. (**1982**). "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **72**, 43–50.

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.

Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Coady, J. A., Kluender, K. R., and Rhode, W. S. (**2003**). "Effects of contrast between onsets of speech and other complex spectra," J. Acoust. Soc. Am. **114**, 2225–2235.

Dorman, M. F., and Loizou, P. C. (**1996**). "Relative spectral change and formant transitions as cues to labial and alveolar place of articulation," J. Acoust. Soc. Am. **100**, 3825–3830.

Fruchter, D., and Sussman, H. M. (**1997**). "The perceptual relevance of locus equations," J. Acoust. Soc. Am. **102**, 2997–3008.

Halle, M., Hughes, G. W., and Radley, J.-P. A. (**1957**). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107–116.

Holt, L. L., and Kluender, K. R. (**2000**). "General auditory processes contribute to perceptual accommodation of coarticulation," Phonetica **57**, 170–180.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2000**). "Neighboring spectral content influences vowel identification," J. Acoust. Soc. Am. **108**, 710–722.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2001**). "Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariance or auditory enhancement?," J. Acoust. Soc. Am. **109**, 764–774.

Kewley-Port, D. (**1983**). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.

Kewley-Port, D., and Luce, P. A. (**1984**). "Time-varying features of initial stop consonants in auditory running spectra: A first report," Percept. Psychophys. **35**, 353–360.

Kiefte, M., and Kluender, K. R. (**2005**). "The relative importance of spectral tilt in monophthongs and diphthongs," J. Acoust. Soc. Am. **117**, 1395–1404.

Kingston, J., and Diehl, R. L. (**1994**). "Phonetic knowledge," Language **70**, 419–454.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Kluender, K. R., and Alexander, J. M., (in press). "Perception of speech sounds," in *Handbook of the Senses: Audition*, edited by P. Dallos and D. Oertel (Elsevier, London).

Kluender, K. R., Coady, J. A., and Kiefte, M. (**2003**). "Sensitivity to change in perception of speech," Speech Commun. **41**, 59–69.

Kluender, K. R., and Kiefte, M. (**2006**). "Speech perception within a biologically-realistic information-theoretic framework," in *Handbook of Psycholinguistics*, edited by M. A. Gernsbacher and M. Traxler (Elsevier, London), pp. 153–199.

Lahiri, A., Gewirth, L., and Blumstein, S. E. (**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.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1952**). "The role of selected stimulus variables in the perception of unvoiced stop consonants," Am. J. Psychol. **65**, 497–516.

Lindblom, B. (**1963**). "On vowel reduction," Rep. No. 29, The Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, Sweden.

Lotto, A. J., and Kluender, K. R. (**1998**). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," Percept. Psychophys. **60**, 602–619.

Lotto, A. J., Kluender, K. R., and Holt, L. L. (**1997**). "Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica)," J. Acoust. Soc. Am. **102**, 1134–1140.

Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode," Psychol. Bull. **92**, 81–110.

Smits, R. (**1996**). "Context-dependent relevance of burst and transitions for perceived place in stops: It's in production, not perception," Proceedings of the Fourth International Conference on Spoken Language, Philadelphia, PA, Vol. **4** (IEEE, Piscataway, NJ), pp. 2470–2473.

Smits, R., ten Bosch, L., and Collier, R. (**1996**). "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment," J. Acoust. Soc. Am. **100**, 3852–3864.

Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

Stevens, K. N., and Blumstein, S. E. (**1981**). "The search for invariant acoustic correlates of phonetic features," *Perspectives in the Study of Speech*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ), pp. 1–38.

Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (**1991**). "An investigation of locus equations as a source of relational invariance for stop place categorization," J. Acoust. Soc. Am. **90**, 1309–1325.

Walley, A. C., and Carrell, T. D. (**1983**). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "*F*0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**, 2152–2159.

Wichmann, F. A., and Hill, N. J. (**2001**). "The psychometric function. I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.

Training English listeners to perceive phonemic length contrasts in Japanese[a)]

Keiichi Tajima[b)]
Department of Psychology, Hosei University, 2-17-1 Fujimi, Chiyoda-ku, Tokyo 102-8160, Japan

Hiroaki Kato
ATR Cognitive Information Science Laboratories/National Institute of Information and Communications Technology, 2-2-2 Hikaridai, Seika-cho, Kyoto 619-0288, Japan

Amanda Rothwell
School of Kinesiology, The University of Western Ontario, London, Ontario N6A 3K7, Canada

Reiko Akahane-Yamada
ATR Cognitive Information Science Laboratories, 2-2-2 Hikaridai, Seika-cho, Kyoto 619-0288, Japan and Graduate School of Cultural Studies and Human Science, Kobe University, 1-2-1 Tsurukabuto, Nada-ku, Kobe 657-8501, Japan

Kevin G. Munhall
Department of Psychology and Department of Otolaryngology, Queen's University, Humphrey Hall, 62 Arch Street, Kingston, Ontario K7L 3N6, Canada

(Received 17 July 2007; revised 10 October 2007; accepted 11 October 2007)

The present study investigated the extent to which native English listeners' perception of Japanese length contrasts can be modified with perceptual training, and how their performance is affected by factors that influence segment duration, which is a primary correlate of Japanese length contrasts. Listeners were trained in a minimal-pair identification paradigm with feedback, using isolated words contrasting in vowel length, produced at a normal speaking rate. Experiment 1 tested listeners using stimuli varying in speaking rate, presentation context (in isolation versus embedded in carrier sentences), and type of length contrast. Experiment 2 examined whether performance varied by the position of the contrast within the word, and by whether the test talkers were professionally trained or not. Results did not show that trained listeners improved overall performance to a greater extent than untrained control participants. Training improved perception of trained contrast types, generalized to nonprofessional talkers' productions, and improved performance in difficult within-word positions. However, training did not enable listeners to cope with speaking rate variation, and did not generalize to untrained contrast types. These results suggest that perceptual training improves non-native listeners' perception of Japanese length contrasts only to a limited extent. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804942]

PACS number(s): 43.71.Hw, 43.71.Es [PEI] Pages: 397–413

I. INTRODUCTION

In Japanese, vowel and consonant length can be used phonemically to distinguish words. The primary acoustic correlate and perceptual cue to such length contrasts are said to be the duration of the vowel or consonant (Fujisaki *et al.*, 1975; Uchida, 1998). Segment duration, however, is affected by numerous factors, including inherent duration, neighboring segments, position within a word or phrase, length of the word or phrase, emphasis or semantic novelty, and speaking rate (e.g., Klatt, 1976; Sagisaka and Tohkura, 1984; Takeda *et al.*, 1989). In fact, the duration of phonemically short and long segments produced across various speaking rates overlap considerably (Hirata, 2004a; Hirata and Whiton, 2005), implying that perceptual judgment of phonemic length cannot be made simply based on absolute segment duration, but must be made in relation to the context in which the segment appears. Such complex behavior of segment duration potentially makes phonemic length extremely difficult for second-language (L2) learners to learn, particularly for native speakers of a language that does not use segment duration phonemically, such as English. Native English speakers have in fact been shown to have difficulty learning to perceive Japanese length contrasts (Yamada *et al.*, 1994; Oguma, 2000; Toda, 2003), but the nature of their difficulty in relation to the various contextual factors has not been thoroughly investigated. Thus, the first purpose of the present study is to investigate how non-native listeners' perception of Japanese length contrasts is affected by contextual factors that affect segment duration, including speaking rate, presentation con-

[a)]Portions of this work were presented in "Perception of phonemic length contrasts in Japanese by native and non-native listeners," Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, August, 2003.

[b)]Electronic mail: tajima@hosei.ac.jp

text (target words uttered in isolation versus embedded in a carrier sentence), and position within the word.

Meanwhile, numerous studies on L2 speech learning have demonstrated that, contrary to the traditional belief that adults lose neurological plasticity required for acquiring the sound system of a foreign language (Lenneberg, 1967), their production and perception abilities can be modified with experience. Both exposure to L2 (e.g., Yamada, 1995) and L2 speech training (e.g., Logan *et al.*, 1991; Lively *et al.*, 1993, 1994; Bradlow *et al.*, 1997) lead to substantial improvement in non-native listeners' ability to perceive and produce difficult L2 phonetic contrasts. In particular, Logan *et al.* (1991) and Lively *et al.* (1993, 1994) have demonstrated that "high variability perceptual training," which exposes trainees to instances of L2 phonetic categories produced in many phonetic environments and by many talkers, significantly improves listeners' ability to identify difficult L2 phonemes, such as the perception of the English /r/-/l/ contrast by Japanese listeners. Recently, this training method has been demonstrated to be effective for training prosodic properties of L2 speech as well, such as the perception of Mandarin lexical tones by English listeners (Wang *et al.*, 1999), perception of English syllables by Japanese listeners (Tajima and Erickson, 2001), and perception of Japanese length contrasts (e.g., Hirata *et al.*, 2007). However, the extent to which this training method improves perception of Japanese length contrasts has not been thoroughly investigated, especially in relation to the above-mentioned contextual factors that might affect performance. As such, the second purpose of the present study is to address the degree to which perceptual training improves English listeners' perception of Japanese length contrasts, and to assess the degree to which training generalizes to novel stimulus conditions that affect the temporal context.

A. Japanese length contrasts

Japanese can be said to have several distinct types of length contrasts that are signaled primarily by duration, depending on the type of segment involved. Phonologically, the length contrast involves the addition of an extra mora. In the present study, the following four types of length-based minimal pairs are considered (a hyphen "-" indicates mora boundaries): (1) *vowel pairs*, which contrast in the length of a vowel, e.g., /ka-do/ (corner) versus /ka-a-do/ (card); (2) *obstruent pairs*, which contrast in the length of an obstruent consonant, e.g., /ha-ke-n/ (dispatch) versus /ha-k-ke-n/ (discovery); (3) *nasal pairs*, which contrast in the length of a nasal consonant, e.g., /ta-ni-n/ (stranger) versus /ta-n-ni-n/ (person in charge); and (4) *palatal pairs*, which differ in the length of a palatal /i/-like segment, e.g., /kja-ku/ (visitor) versus /ki-ja-ku/ (statute). The first three types of contrast involve the presence or absence of a moraic vowel, moraic obstruent, or moraic nasal, respectively, signaled orthographically by distinct kana syllabary symbols. The moraic nasal is not difficult to acquire for non-native learners when they precede non-nasal consonants, e.g., /ho-n-da/ (Honda), but when it is immediately followed by another nasal consonant, its presence is signaled by the duration of the nasal segment, therefore causing potential problems for non-native learners (Uchida, 1998). The fourth contrast type, palatal pairs, is signaled orthographically by using either a subscript or full-size kana symbol that indicates the palatal sound, corresponding to the short and long members of the pair, respectively.

Partly because different orthographic conventions are used to transcribe the four contrast types in Japanese, they have often not been examined together under a common framework. However, all four contrast types can be construed as being signaled mainly by durational cues. For example, several studies have shown that the primary acoustic and perceptual cues for distinguishing short versus long vowels, obstruents, and nasals are the duration of the vowel, obstruent, or nasal segment, respectively (Fujisaka *et al.*, 1975; Uchida, 1998). Furthermore, these contrasts have been shown to be perceived in a categorical manner by native listeners (Uchida, 1998). In fact, these studies have claimed that essentially the same perceptual mechanism is employed for perceiving these contrast types. Cues for distinguishing palatal pairs have not been extensively investigated, but the primary perceptual cue for palatal pairs such as /kjaku/-/kijaku/ is likely the duration of a /i/-like vocalic interval with a high second formant (which immediately precedes another vocalic segment with a relatively low second formant, such as /a/, /o/ or /u/).

If the same perceptual mechanism is involved in perceiving the four contrast types, then improvement in the ability to perceive one type of contrast might generalize to other contrast types. On the other hand, it has been reported that some contrasts are more difficult to learn than others; for example, Toda (2003) has reported that obstruent length contrasts are more difficult to acquire than vowel length contrasts. If so, then learning to perceive one type of contrast may not be sufficient to guarantee improved ability to perceive other contrast types. The present study investigated this question by training listeners with only vowel pairs, and testing whether training generalizes to the other three contrast types, or whether the effect of training is limited to the contrast type that listeners were trained with.

B. Sources of contextual variability

The present study investigated three contextual factors that are expected to affect the perception of length contrasts by non-native listeners: speaking rate, presentation context, and within-word position. Simultaneous investigation of these factors in a single study was necessary because some of the factors were expected to interact with each other.

First, speaking rate exerts large effects on segment duration. For example, Hirata (2004a) found that the duration of a short vowel produced at a slow speaking rate is sometimes longer than that of a long vowel produced at a fast rate, and that the durational difference between short and long segments is relatively large at slow speaking rates but smaller at faster rates. These results suggest that listeners may have difficulty perceiving length contrasts if they rely on a fixed durational threshold between short and long segments, and that their perceptual performance may decline as speaking rate increases. Native Japanese listeners have been

shown to shift their perceptual boundary between phonemically short and long segments according to changes in speaking rate (e.g., Fujisaki *et al.*, 1975). However, studies conducted with non-native listeners have shown different results. For example, Toda (2003) asked listeners to identify words that belonged to continua such as /kate/-/kate:/ and /kate/-/katte/, and manipulated speaking rate by either shortening or lengthening the duration of the first vowel /a/. She found that while native listeners' perceptual boundary generally shifted according to the duration change in the preceding segment, English listeners did not show such a systematic shift. Toda's study, however, artificially manipulated the duration of segments, and did not use materials that naturally varied in speaking rate.

A recent study by Hirata *et al.* (2007) reported that two-rate training, in which native English listeners are trained using sentences produced at two speaking rates (slow and fast), is superior over one-rate training, in which listeners are trained using sentences produced at one rate (slow or fast). Listeners trained with two rates performed better than listeners trained with one rate when tested with sentences produced at various speaking rates. The present study pursued a similar question, by training listeners using words produced at a normal rate, and testing whether performance improves for words and sentences produced at slow, normal, and fast speaking rates.

Second, the presentation context, i.e., whether the target word is produced in isolation or is embedded in a carrier sentence, may also affect perception of Japanese length contrasts. For native Japanese listeners, Fujisaki *et al.* (1975) found that the perceptual boundary between phonemically short and long segments shifted as a function of speaking rate for both words uttered in isolation and words embedded in a short carrier sentence, but this adaptation was slightly stronger in sentence context than in word context, suggesting that carrier sentences provide contextual cues that facilitate judgment of speaking rate.

Whether non-native listeners would also benefit from carrier sentences, however, is unclear. Both inhibitory and facilitatory effects are conceivable. On the one hand, carrier sentences increase the amount of information that listeners need to process, and require listeners to spot the word in the sentence, while no such segmentation would be necessary if the word were presented by itself. Furthermore, words produced in sentences are typically spoken at a faster rate than the same words produced in isolation, since an increase in the number of syllables or words in a breath group typically leads to an increase in speaking rate. These factors might together make the sentence condition more difficult than the word condition. Studies on L2 phoneme perception have in fact demonstrated that non-native listeners' identification performance is poorer when the target word is in a semantically neutral carrier sentence than when it is presented in isolation (Ikuma and Akahane-Yamada, 2004). Hirata (2004b) also reported that English listeners' perception of Japanese length contrasts was worse in sentence context than word context.

On the other hand, a carrier sentence potentially provides contextual cues about overall tempo of the utterance, which could be useful in judging the phonemic length of segments. In a study examining the role of sentence context for native Japanese listeners' perception of vowel length, Hirata and Lambacher (2004) found that native Japanese listeners' perception of vowel length was poorer when the target word was excised from the carrier sentence and presented in isolation than when the target word was presented in the original carrier sentence. This suggests that the carrier sentence contained important information for accurately perceiving phonemic length. While it is not clear whether a similar disadvantage would be observed for a target word that is originally produced in isolation (as opposed to being excised from a carrier sentence), this opens the possibility that non-native listeners may also benefit from contextual cues surrounding the target word.

Finally, the position of the length contrast within the word may also affect performance. Statistical analyses of segment duration in Japanese speech databases have indicated that vowels exhibit final lengthening at the end of an isolated word or a (sentence-nonfinal) phrase, but they exhibit final *shortening* at the end of a sentence (Takeda *et al.*, 1989; Kaiki and Sagisaka, 1992). Whether within-word position affects the durational contrast between short and long segments has not been extensively investigated, but there is some limited data from acoustic measurements of vowels in isolated-word utterances (Kubozono, 2002) which indicate that the difference between short and long vowels is smaller in word-final position than nonfinal position. This suggests that perception of length contrasts might be less accurate in word-final position than other positions. Such a position effect has not been closely examined for native listeners, but studies with non-native listeners have found that errors in identifying short and long vowels were most frequent in word-final position, and significantly less frequent when the vowel appeared in a word-initial or word-medial syllable (Oguma, 2000; Minagawa-Kawai *et al.*, 2002).

To explain this position effect, it has been suggested that the effect can be attributed to the presence/absence of phonetic materials following the target segment. That is, when the target segment is word-internal, it is followed by other speech sounds that potentially provide additional timing cues that facilitate judgment of phonemic length, while no such cues are available when the segment is word-final [Minagawa-Kawai *et al.* (2002); see also Kubozono (2002)]. If this explanation holds, then one would predict that the position effect would not be observed if the word-final segment were followed by other phonetic materials, e.g., those belonging to the carrier sentence. The present study investigated this question by examining the effect of position for both words produced in isolation and words embedded in carrier sentences.

C. Perceptual training

Effects of laboratory training on English listeners' perception of Japanese length contrasts have been examined in several previous studies. For example, Yamada *et al.* (1994) trained American English listeners to identify nonwords of the form $C_1V_1C_2V_2$ where V_1 or C_2 varied in segment iden-

tity as well as phonemic length, and found that training significantly improved performance, and generalized to untrained nonwords and talkers. A series of recent studies by Hirata and her colleagues have also examined perceptual learning of Japanese length contrasts by English listeners. Hirata (2004b) investigated the effect of training using isolated words versus words embedded in carrier sentences, and found that listeners trained with words in isolation improved performance for words embedded in sentences, and vice versa. Hirata *et al.* (2007) examined the effect of training using sentences produced at two rates versus/sentences produced at only one rate, and found that both one- and two-rate training improved performance but that two-rate training yielded more robust generalization to untrained rates. Many of these studies, however, typically tested listeners with stimuli in which the length contrast occurred in fixed positions in the target word. It is therefore unclear to what extent listeners' performance varies across different positions. It is also unclear whether speaking rate and presentation context combine in a simple additive manner or whether they interact in complex ways.

Tajima *et al.* (2003b) trained native English listeners residing in Japan to identify Japanese words contrasting in phonemic length using a minimal-pair identification task. Listeners were trained with vowel pairs spoken in isolation at a normal rate, but were tested with words of various contrast types produced at three speaking rates and in two presentation contexts (in isolation or embedded in a carrier sentence). Training improved performance from 90.6% to 94.1%, but listeners' performance was very high even before training, making it difficult to assess the effectiveness of training and to evaluate the degree to which training generalizes to various conditions. One possible reason for the high accuracies in the study of Tajima *et al.* (2003b) is that the identification task had relatively low stimulus uncertainty (cf. Watson *et al.*, 1976), making the identification task fairly easy even for non-native listeners. That is, in each block of trials, speaking rate and contrast type were fixed; thus, listeners could easily predict which stimulus properties to pay attention to, and they could potentially set up a fixed perceptual criterion for judging the phonemic length of the target segment. Furthermore, stimuli were produced by professionally trained talkers, who were expected to be better able than nonprofessional talkers to produce a clear distinction between short and long segments even at multiple speaking rates. Thus, performance may have been poorer had the stimuli been produced by nonprofessional talkers.

In the present study, two experiments were carried out in order to investigate the effect of perceptual training under conditions of high contextual variability, e.g., conditions in which speaking rate, presentation context, and contrast type vary. Experiment 1 tested listeners in conditions of relatively high trial-to-trial stimulus uncertainty, and focused on how differences in contrast type and speaking rate affect performance. Rather than presenting just a single speaking rate and a single contrast type within each block of trials, as was done in the study of Tajima *et al.* (2003b), stimuli from the three speaking rates and the four contrast types were mixed and presented in a random order within the same block of trials. Such a test not only was expected to increase task difficulty, but it was also expected to be a more sensitive test of how speaking rate affected performance, and how training generalized to various untrained stimulus conditions. Experiment 2 focused on testing whether perceptual training using words produced by professionally trained talkers also improves performance on ordinary, nonprofessional talkers, who may not produce as clear a length distinction as professional talkers. Experiment 2 also focused on the effect of within-word position, and examined whether length contrasts are more difficult to identify in word-final position than other positions, and whether this position effect would be reduced if words are embedded in carrier sentences (so that word-final target segments would be followed by other phonetic materials). The experiment also examined how these positional effects are modified with perceptual training. In both experiments 1 and 2, a pretest–posttest design was employed in which a group of non-native listeners took the same test twice (dubbed test1 and test2), once before and once after training. As a control group, a different group of non-native listeners took only test1 and test2 separated by about the same number of days as the training group.

II. EXPERIMENT 1

A. Participants

Three listener groups participated. (1) Group ET (English Training): ten native Canadian English listeners who took five days of training between test1 and test2 (one male, nine females, aged 19–25, mean age=21.3). (2) Group EC (English Control): ten native Canadian English listeners who took only test1 and test2, but no training (five male, five females, aged 18–21, mean age=19.2). Listeners in groups ET and EC had no prior experience with Japanese. (3) Group JC (Japanese Control): ten native Japanese speakers (seven males, three females, aged 19–22, mean age=20.6). The English listeners participated in the experiment at Queen's University, and the Japanese listeners at ATR Laboratories. None of the listeners had any history of speech or hearing disorders.

B. Stimuli and procedure

The experiment consisted of three phases: test1, five days of training, and test2. Group ET participated in all three phases, group EC participated in test1 and test2, and group JC took the test once. Prior to the experiment, the non-native listeners were given a brief description of Japanese length contrasts, along with audio samples and English transcriptions that illustrated the four contrast types. Long vowels were transcribed as "i: e: a: o: u:" rather than "ii ee aa oo uu" because double letters such as "ee" and "oo" were likely to be misinterpreted as /i:/ and /u:/ by English listeners (a color ":" was used to mimic the IPA symbol for extra length). Long consonants were transcribed using double letters, i.e., "pp tt kk ss zz mm nn jj," or as "ssh" and "tch" for long counterparts of "sh" and "ch," respectively. The sample words were not used in the test or training.

1. Test

The test stimuli consisted of 76 real word pairs and three nonword triplets. Words in each real word pair minimally contrasted in one of the four contrast types, and were matched in word accent pattern. The word pairs were selected based on a search through a Japanese lexical database (Amano and Kondo, 2000), which contains, for over 80 000 Japanese words, subjective ratings for word familiarity, appropriateness of word accent pattern, etc. Most word pairs used in the present study had familiarity ratings of 5.0 or higher (on a scale from 1.0 to 7.0) and accent appropriateness ratings of 4.7 and higher (on a scale from 1.0 to 5.0), although nasal and palatal pairs contained greater proportions of less familiar words (due to the paucity of available minimal pairs). The word pairs were also selected so that the set as a whole was reasonably phonetically balanced. Target segments appeared in several possible positions within the word depending on the stimulus. For example, vowel length contrasts appeared in either the word-initial syllable, e.g., /kado/ versus /ka:do/ or word-final syllable, e.g., /kaze/ (wind) versus /kaze:/ (taxation), or they appeared in monosyllabic words, e.g., /ki/ (tree) versus /ki:/ (key). Nasal length contrasts appeared either in the word-initial syllable /tanin/ versus /tannin/ or in a phrase-medial position, e.g., /koi no e/ (picture of a carp) versus /koin no e/ (picture of a coin). For nasal length contrasts in word-medial position, short phrases were devised as stimuli because there were no appropriate word pairs that contrasted in nasal length in word-medial position. The three nonword triplets were of the form /ereCe/-/ere:Ce/-/ereC:e/, where C was one of the following consonants, /t s n/. The 76 real word pairs consisted of 20 vowel, obstruent, and nasal pairs each, and 16 palatal pairs. They were equally divided into four lists, each of which was to be read by a different talker. The three nonword triplets were to be read by all talkers.

Each word or nonword was read in two contexts, (1) in isolation (word context), and (2) in a sentence context in which each item was randomly embedded in one of ten carrier sentences, e.g., /ima kara __ to iimasu/ (I will say __ now). All carrier sentences had four moras preceding the target word, and either five or six moras following the target word depending on the carrier sentence. A different carrier sentence was assigned to each item across lists to be read by different talkers. The words and sentences were compiled into separate lists.

The talkers were four professionally trained native Japanese talkers (two females aged 38 and 51 and two males aged 37 and 53), who had been trained as voice actors/actresses and spoke standard Tokyo Japanese comfortably. The lists were read first at a self-selected normal rate, then at a fast rate, and finally at a slow rate. To obtain speech samples produced at sufficiently distinct speaking rates, the talkers were encouraged to utter the "fast" items at about twice the speed as the "slow" items. The recording took place in an anechoic chamber at ATR Laboratories, and was later saved as audio files at 22.05 Hz sampling frequency and 16 bit resolution.[1]

Listeners took the test in a sound-treated booth. The task was a single-stimulus, two-or three-alternative forced-choice identification task. On each trial, English transcriptions of two Japanese words comprising a minimal pair, or three nonwords comprising a triplet, appeared as clickable buttons in the computer program window. In the sentence condition, an English transcription of the carrier sentence was also presented, with the target word replaced by an underline. Simultaneously, listeners heard one of the words, or a sentence containing one of the target words, presented through headphones at a comfortable listening level. Their task was to select the word they heard by clicking the appropriate button. Listeners were able to listen to the stimulus again by clicking the "replay" button, but they were discouraged from doing so frequently. The trials were self-paced. The test consisted of 1128 trials, divided into 16 blocks of either 114 real word trials or 27 nonword trials. In each block, stimuli from each combination of the following three factors were presented: presentation context (word versus sentence), talker, and word type (word versus nonword). The order of the two presentation contexts and the four talkers was counterbalanced across listeners, but the order of the word types was fixed, such that the word trials always immediately preceded the corresponding nonword trials. Within each block of word trials, words from the four contrast types uttered at the three speaking rates were presented in a random order (38 words $\times$ 3 rates = 114 trials). Within each block of nonword trials, nonwords spoken at the three rates were presented in a random order (9 nonwords $\times$ 3 rates = 27 trials). Listeners were allowed to take short breaks between blocks of trials and halfway within long blocks of trials. The four test talkers were different from the talkers who appeared during training. (An unfortunate error in the experimental design resulted in some of the test words in experiment 1 appearing in both the tests and the training, contrary to the original intention which was to have no overlap between the test and training stimuli, as a genuine test of generalization of training to untrained items. Specifically, 9 of the 20 vowel pairs appeared in both the tests and the training. All other word pairs and nonwords were different from those that appeared during training. As discussed in Sec. II C, there were no consistent differences in listeners' performance between trained and untrained stimuli.)

Groups ET and EC took the same test twice, with an average of 9.1 days (7–12 days) between test1 and test2 for group ET and 9.8 days (7–14 days) for group EC. For both groups, test1 took approximately 60–90 min, while test2 took approximately 50–70 min. The JC listeners took the test only once.

2. Training

Listeners in group ET underwent five days of perceptual identification training between test1 and test2. The training stimuli consisted of 60 vowel pairs, most of which were different from the test pairs (see previous text). There were roughly an equal number of vowel pairs that contained the vowel length contrast in the initial syllable (31 pairs) and those that contained the contrast in the final syllable (27 pairs); the remaining two pairs were monosyllabic words. There were no vowel pairs that contained the contrast in word-medial position. The training words were produced in isolation at a normal speaking rate only, by five profession-

ally trained native Japanese talkers of various ages (two males and three females, aged 35–65). These talkers were different from the four talkers who appeared during the tests. These talkers were actually instructed to read a larger set of Japanese words and sentences at multiple speaking rates, similar to the test talkers, but only some of the isolated words produced at a normal rate were used for training. The recording took place in a recording studio in Tokyo, and was later saved as audio files in the same format as the test stimuli.

The training was conducted in the same laboratory environment as the tests. Three training sessions were conducted on each training day, with no more than three free days separating consecutive training days. There were 240 trials in each session, in which the 60 vowel pairs as spoken by one talker were presented two times each in a random order. Over the five training days, listeners cycled three times, in a fixed order, through the five talkers, yielding a total of 15 training sessions or 3600 training trials. Training trials were identical to the test trials except that immediate feedback was provided concerning listeners responses and that correction trials were performed, such that listeners repeated a given trial until the correct response was selected. Each training day lasted for roughly 35–60 min, with a mild tendency for sessions to become shorter as listeners accumulated training.

C. Results and discussion

In subsequent sections, listeners' performance is reported as percent-correct identification accuracies. All statistical tests in experiments 1 and 2 were conducted on arcsine-transformed values of the identification accuracies. Repeated-measures analyses of variance (ANOVA) were conducted with correction for sphericity, based on Greenhouse and Geisser's (1959) method.

An error in the experimental design resulted in 9 of the 20 vowel pairs being included in both the tests and the training, while all other items in the tests were different from those used during training, as originally intended. To examine whether group ET's performance differed between trained versus untrained test items, mean identification accuracies were computed separately for the 9 trained vowel pairs and 11 untrained vowel pairs, for each test. Results indicated that identification scores were not significantly different from each other for either test1 or test2. Thus, all subsequent analyses pooled together data from trained and untrained test items.

1. Overall performance

Figure 1 shows boxplots of the identification accuracies in test1 and test2 for groups EC and ET and accuracies in test1 for group JC. Accuracies are based on trials in all conditions of talker, rate, presentation context, and contrast type, including words and nonwords.

Figure 1 shows considerable individual variation in performance among the non-native listeners. For group ET, mean identification accuracy was 69.1% (s.d.=7.3) in test1, but rose to 76.6% (s.d.=10.3) in test2. For group EC, accuracies in test1 and test2 were 75.2% (s.d.=7.0) and 79.3% (s.d.=6.7), respectively. Mean accuracy in test1 turned out to be higher for group EC than for group ET, even though both groups were recruited from the same subject pool at the same university in Canada. Listeners in group ET improved from test1 to test2 by 7.5 percentage points on average, but group EC's mean accuracy also improved, although to a smaller degree on average (4.1 points). Group JC's mean accuracy was 98.0% (s.d.=0.8).

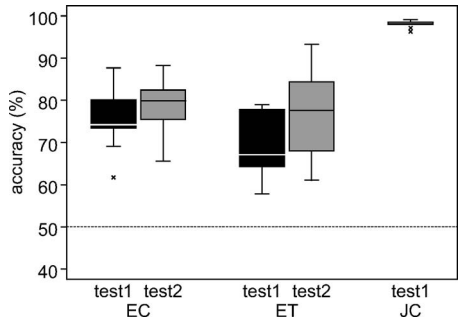


FIG. 1. Identification accuracies in experiment 1 as a function of listener group (EC, ET, JC) and test (test1, test2). The JC listeners took the test only once. Accuracies are based on trials in all conditions of talker, rate, presentation context, and contrast type, including words and nonwords. The horizontal dashed line indicates chance level performance. In this and subsequent boxplots, the horizontal line in each box indicates the median, the vertical length of the box indicates the interquartile range (the range between the lower and upper quartiles), each whisker indicates the furthest data point from the edge of the box that is not further than 1.5 times the length of the box, and individual data points indicate outliers that are further out than the extent of the whiskers.

The non-native listeners' accuracies were submitted to a two-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable and test (test1, test2) as a within-subjects variable. Results revealed a significant main effect of test [$F(1,18)=19.64$, $p<0.001$], but no significant (n.s.) main effect of group [$F(1,18)=1.49$, n.s.], or significant interaction between test and group [$F(1,18)=1.75$, n.s.]. Further pairwise comparisons of accuracies (with Bonferroni correction when family-wise error rate was set at $\alpha=0.05$) revealed that the improvement in accuracy from test1 to test2 was significant for both group ET [$t(9)=3.14$, $p<0.05$] and group EC [$t(9)=3.86$, $p<0.05$]. However, there was no significant difference in accuracy between groups ET and EC for either test1 [$t(18)=1.90$, n.s.] or test2 [$t(18)=0.57$, n.s.]. Another set of comparisons between the non-native listeners' test2 scores and native Japanese listeners' scores revealed that both group ET's test2 scores [$t(18)=6.55$, $p<0.05$] and group EC's test2 scores [$t(18)=11.93$, $p<0.05$] were significantly lower than group JC's scores.

To briefly examine how listeners' performance changed during the training sessions, mean accuracies among the ten ET listeners were computed for each of the 15 training sessions. Listeners started out at 83.7% accuracy in session 1 and ended at 91.1% accuracy in session 15, with the highest accuracy (93.5%) obtained in session 12. Generally speaking, accuracy increased across the 15 sessions, but the amount of increase was greater during the first half of training than the second half.

In summary, these results suggest that non-native listeners have considerable difficulty identifying phonemic length contrasts when they appear in Japanese words and sentences spoken at various speaking rates. Accuracies in test1 and test2 in the present study were lower than those obtained in the study of Tajima *et al.* (2003b)—87%. The primary difference between the study of Tajima *et al.* (2003b) and the present study was that speaking rate and contrast type varied from trial to trial in the latter study, while they were fixed within blocks of trials in the former. This suggests that non-native listeners have perceptual difficulties especially under conditions of considerable trial-to-trial stimulus uncertainty.

Performance during training, in which only vowel pairs produced in isolation at a normal rate were used, turned out to be relatively high, even at the beginning of training. This suggests that there was little room left for further improvement to take place during training, although accuracy did improve to some extent as training proceeded.

Perhaps because of the relative small improvement during training, group ET's improvement in accuracy from test1 to test2 was not significantly greater than the improvement observed for group EC, as indicated by the lack of a significant interaction between group and test. Thus, the results in Fig. 1 do not provide strong evidence that training *per se* led to significantly greater improvement in performance than factors such as repeated exposure to the test materials and increased familiarity with the task.

2. Contrast type

Even though group ET did not show a significantly greater increase in overall accuracy than group EC, it is possible that repeated training with vowel pairs may lead to more specific improvement in certain contrast types for group ET. To examine this, Fig. 2(a) shows the performance of groups ET and JC as a function of the four contrast types and the nonwords. Figure 2(b) shows a similar graph for group EC. Accuracies in both Figs. 2(a) and 2(b) are based on stimuli produced at all rates and presentation contexts by all talkers. Figure 2 suggest that non-native listeners' accuracies were generally lower for the nonwords than for the real words. This may be related to the fact that chance level was 33% for the nonwords but 50% for the real words (shown as dotted lines in Fig. 2). Performance appears to improve to some extent from test1 to test2 for both groups ET and EC. The most salient improvement seems to be witnessed for group ET's vowel pairs.

A three-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable and test (test1, test2) and contrast (vowel, obstruent, nasal, palatal) as within-subjects variables was conducted for the real words. If perceptual training improves performance equally on all contrast types, then one might expect to see a significant interaction between group and test, with no interactions involving contrast. On the other hand, if training improves performance only on vowel pairs or on a subset of the contrast types, then one would expect a significant three-way interaction among group, test, and contrast. Results indicated significant main effects of test [$F(1,18)=25.23$, $p<0.001$] and contrast [$F(2,31)=37.98$, $p<0.001$], and a significant

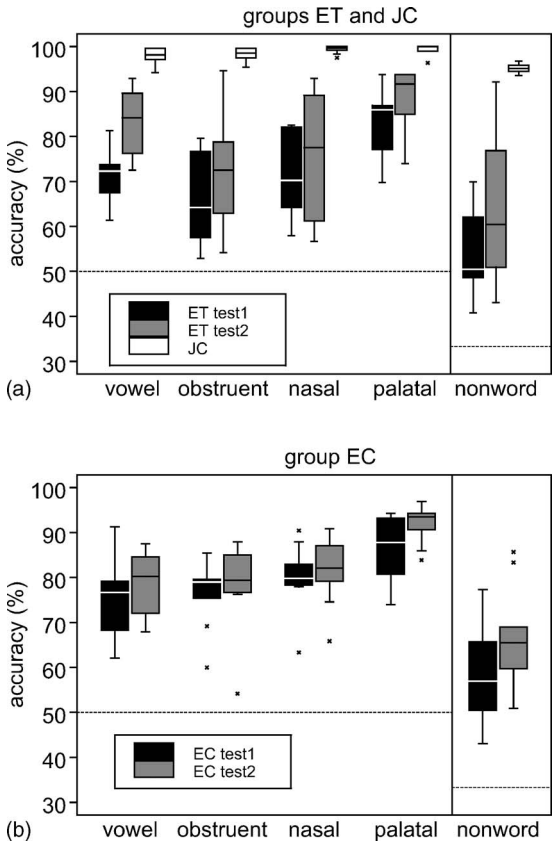


FIG. 2. Identification accuracies in experiment 1 for (a) group ET's test1 and test2 and group JC, and (b) group EC, as a function of the four contrast types and the nonwords. Accuracies are based on stimuli produced at all rates and presentation contexts by all talkers. The horizontal dashed line indicates chance level performance (50% for real words and 33% for nonwords).

test-by-contrast interaction [$F(2,38)=3.22$, $p<0.05$]. Other main effects and interactions were not significant. Further analysis of the test-by-contrast interaction with simple effect tests and multiple comparisons using Tukey's HSD revealed that for both test1 and test2, accuracy was significantly higher for palatal pairs than for the other three contrast types ($p<0.05$). For test1, accuracy for obstruent pairs was significantly lower than that for nasal pairs, but for test2, accuracy for obstruent pairs was significantly lower than that for vowel pairs ($p<0.05$). Looking at each individual contrast type, the increase in accuracy from test1 to test2 was significant for all four contrast types, although the magnitude of the increase varied across contrasts, with the greatest increase for vowel pairs (73.4% to 81.2%; $p<0.001$), followed by palatal pairs (84.6% to 90.4%; $p<0.001$), obstruent pairs (70.8% to 75.1%; $p<0.01$), and nasal pairs (76.0% to 79.2%; $p<0.05$).

In short, these results suggest that non-native listeners' ability to identify Japanese length contrasts vary depending on the contrast type involved. As for the effect of training, since there were no significant interactions involving group and test, the present results do not provide evidence that training *per se* improved performance on all contrast types, or that training improved performance on specific contrast types.

TABLE I. Mean mora duration in milliseconds (and standard deviations) for sample vowel pairs for all test talkers in experiment 1 shown for each combination of speaking rate and presentation context. Mean mora duration was computed by dividing the word duration by the number of moras in the word. Data in each cell are based on four sample vowel pairs (eight tokens). The bottom row shows the range of mean mora durations observed for tokens in each condition.

	Word			Sentence		
Talker	Slow	Normal	Fast	Slow	Normal	Fast
PF4	313.2 (20.7)	207.9 (19.8)	159.2 (13.8)	174.6 (8.4)	127.5 (7.9)	96.5 (7.0)
PF5	293.2 (34.0)	200.5 (25.1)	132.2 (21.4)	183.7 (26.0)	134.9 (22.0)	91.3 (15.4)
PM4	262.6 (23.2)	202.7 (24.7)	146.2 (15.3)	192.8 (18.7)	142.2 (8.8)	98.5 (10.7)
PM6	278.2 (29.5)	207.2 (19.3)	130.4 (9.4)	190.4 (19.1)	137.0 (10.8)	104.5 (10.9)
Mean	286.8 (32.2)	204.6 (21.5)	142.0 (19.0)	185.3 (19.5)	135.4 (14.0)	97.7 (11.9)
Range	225.4–346.1	164.7–247.7	95.2–177.5	148.5–218.3	105.2–168.7	72.7–122.9

3. Speaking rate and presentation context

To assess how well the "slow," "normal," and "fast" speaking rates were implemented by the talkers, and how speaking rate varied across talkers, word duration was measured for four sample vowel pairs for each of the four test talkers and each of the five training talkers, and mean mora duration was computed by dividing the word duration by the number of moras in the word. Table I shows results for each test talker, separately for each speaking rate and presentation context. Table II shows results for each training talker. The bottom row of each column shows the range of mean mora durations observed for the tokens measured in each condition ($N=32$ for test stimuli and $N=40$ for training stimuli).

Looking across different speaking rates and presentation contexts, Table I shows that the test talkers as a group produced three distinct rates in both word and sentence contexts. The mean mora duration across all talkers for the slow, normal, and fast rates were 286.8, 204.6, and 142.0 ms, respectively, for the word context, and 185.3, 135.4, and 97.7 ms, respectively, for the sentence context. The mean duration ratios between the fast and normal rates and between the slow and normal rates, when the normal rate is normalized to 1.00, were approximately 0.69–0.72:1.00 and 1.37–1.40:1.00, respectively. The ranges at the bottom of Table I suggest that there was considerable token-to-token variation in mean mora duration within each rate, but it appears that the professional talkers produced sufficiently distinct speaking rates. Mean mora duration was shorter overall in sentence context than in word context, reflecting a tendency for sentences to be produced at a faster speaking rate than isolated words.

TABLE II. Mean mora duration in milliseconds (and standard deviations) for sample vowel pairs for all training talkers in experiment 1. Data in each cell are based on four sample vowel pairs (eight tokens). The bottom row shows the range of mean mora durations observed for the measured tokens.

Talker	Duration (ms)	
PF1	194.8	(19.6)
PF2	161.9	(4.6)
PF3	158.0	(5.8)
PM1	189.6	(14.2)
PM3	151.1	(17.0)
Mean	171.1	(22.1)
Range	121.3–221.7	

Looking across talkers, Table I suggests that there was considerable overlap in speaking rate across the four test talkers in each condition. Differences across speaking rates and presentation contexts were much greater than differences across talkers. Table II suggests that differences in speaking rate among the five training talkers were somewhat greater than those among the test talkers. Mean mora duration ranged from 151.1 to 194.8 ms across the training talkers. It appears that three talkers (PF2, PF3, PM3) produced somewhat faster speaking rates than the other two talkers, suggesting that the training words contained some amount of speaking rate variation across talkers, even though the words were produced at a "normal rate."

Turning to the effect of speaking rate and presentation context on listeners' performance, Fig. 3(a) shows groups ET and JC's accuracies as a function of rate and context. Accuracies are based on responses to word pairs of all four contrast types. Figure 3(b) shows a similar plot for group EC. Figure 3 reveals that accuracy varied considerably depending on speaking rate. Performance was poorer at a fast speaking rate than at other rates, in both presentation contexts. Between the two presentation contexts, overall performance did not seem to be better for one context than the other. Accuracy seems to be higher overall in test2 than in test1 for both groups ET and EC. There also seems to be considerable individual variation in performance, especially for group ET.

The non-native listeners' accuracies were submitted to a four-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable, and test (test1, test2), context (word, sentence), and rate (slow, normal, fast) as within-subjects factors. If training improves perception of length contrasts in words produced at various speaking rates and presentation contexts, then one might expect to see a significant group-by-test interaction. However, if training improves performance for specific presentation contexts or speaking rates, then one would expect to see a significant interaction among group, test, and either context or rate (or both). Results indicated significant main effects of test [$F(1,18)=20.06$, $p<0.001$] and rate [$F(2,27)=95.51$, $p<0.001$], and a significant context-by-rate interaction [$F(2,32)=13.31$, $p<0.001$]. No other main effects or interactions were signifi-

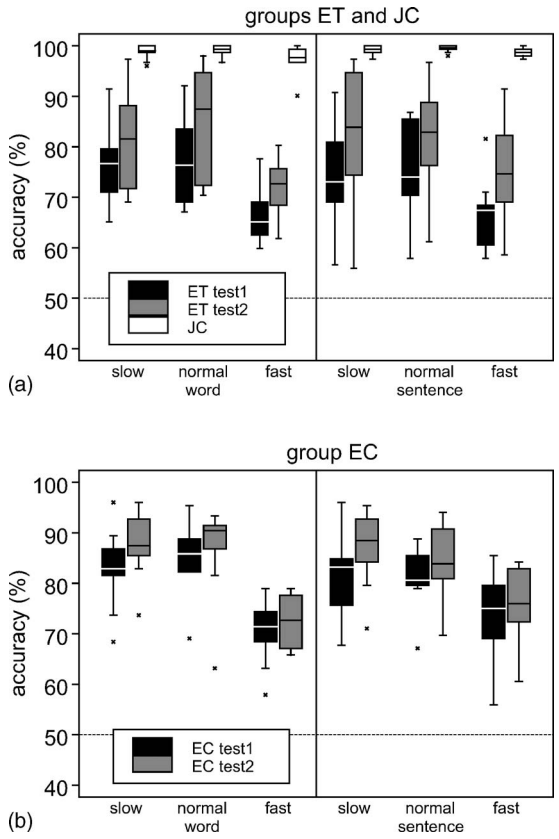


FIG. 3. Identification accuracies in experiment 1 for (a) group ET's test1 and test2 and group JC's test1, and (b) group EC, as a function of presentation context (word, sentence) and speaking rate (fast, normal, slow). Accuracies are based on responses to word pairs of all four contrast types.

cant. Further analysis of the context-by-rate interaction indicated that at the normal rate, mean accuracy was significantly higher in the word context (83.4%) than in the sentence context (80.4%; $p<0.01$), but at the fast rate, the direction reversed, such that accuracy was significantly lower in the word context (70.3%) than in the sentence context (72.8%; $p<0.05$). At the slow rate, accuracy in the word (82.1%) and sentence (81.3%) contexts did not significantly differ from each other. Looking at the effect of speaking rate for each presentation context, accuracy at the fast rate was significantly lower than accuracies at the normal and slow rates for both word and sentence contexts ($p<0.05$). Accuracies at the normal and slow rates did not significantly differ from each other.

For the the native Japanese listeners' data in Fig. 3(a), a separate two-way ANOVA with context and rate as within-subjects factors indicated that the main effect of rate was significant [$F(2,18)=10.98$, $p<0.001$]. Post-hoc tests indicated that mean accuracies at the fast rate were significantly lower than those at other rates ($p<0.05$). With the exception of one native listener who scored around 90% in the fast-rate isolated-word condition, all native listeners' scores were above 96%.

Altogether, Fig. 3 demonstrates that speaking rate has strong influences on non-native listeners' performance. Accuracies were substantially lower for stimuli produced at the fast rate than for those produced at normal or slow rates. Between the two presentation contexts, accuracy did not seem to be higher for one context than the other. Instead, presentation context was found to interact with speaking rate, such that performance for words produced in isolation was higher in the normal-rate condition but lower in the fast-rate condition than performance for words embedded in carrier sentences. Although the effect was relatively small, this suggests that presentation context and speaking rate are not independent of each other.

Between test1 and test2, accuracy was found to increase, but this was found to be the case for both groups ET and EC. Statistical tests did not indicate that the increase in accuracy was greater for group ET than for group EC, even though the increase in accuracy for group ET depicted in Fig. 3(a) seems to be somewhat greater in magnitude than those for group EC shown in Fig. 3(b). Furthermore, statistical tests failed to reveal any significant interactions involving group, test, context, and rate, suggesting that the increase in accuracy from test1 to test2 was not significantly different between word and sentence contexts, or among the three speaking rates. Thus, the present results do not provide positive evidence that training improves non-native listeners' perception of Japanese length contrasts produced at various rates and contexts.

III. EXPERIMENT 2

Results from experiment 1 suggest that overcoming variation in speaking rate and presentation context are not trivial for non-native listeners. Experiment 2 focused on a different set of factors, and investigated whether perceptual training using words produced by professionally trained talkers would also generalize to non-professional talkers' productions, in which the durational distinction between phonemically short and long segments may be less clear than productions by professional talkers. Experiment 2 also examined the extent to which performance would vary depending on the position of the length contrast within the word, and the extent to which such positional effects are modified with training.

A. Participants

All participants in experiment 2 were different from those who participated in experiment 1. Three groups of listeners participated. (1) Group ET: ten native Canadian English speakers who underwent 5 days of perceptual training between test1 and test2 (three males, seven females, aged 19–23, mean age=20.3). (2) Group EC: nine native Canadian English listeners who took only test1 and test2, but no training (five male, four females, aged 18–21, mean age =19.4). Listeners in groups ET and EC had no prior experience with Japanese. (3) Group JC: a control group of ten native Japanese speakers (six males, four females, aged 19–22, mean age=21.0). As in experiment 1, the English listeners participated in the experiment at Queen's University, and the Japanese listeners at ATR Laboratories. None of the listeners had any history of speech or hearing disorders.

B. Stimuli and procedure

The general procedure was essentially the same as experiment 1 (see Sec. II B). The stimuli were as described in the following.

The test stimuli consisted of 102 real word pairs and three nonword triplets. The real word pairs consisted of 30 vowel, obstruent, and nasal pairs each, and 12 palatal pairs, some of which were used in experiment 1. These word pairs were equally divided into six lists. Each word list was to be read by one professional talker and one nonprofessional talker, who were matched in gender and age as closely as possible. The nonword triplets were the same as those in experiment 1, and were to be read by all talkers. There were six professional talkers (three males and three females, aged 35–66); five of the six talkers appeared in experiment 1 as training talkers. There were also six nonprofessional talkers (three males and three females, aged 35–65), who spoke standard Tokyo Japanese comfortably but had never received special training as voice actors/actresses. Each talker read the words and nonwords in two contexts (in isolation and in one of ten carrier sentences as in experiment 1) at a self-selected normal speaking rate. The professional talkers also read the lists at slow and fast rates, but only normal-rate stimuli were used as test stimuli. Recording of the professional talkers took place in a recording studio in Tokyo. For nonprofessional talkers, recording was made only at a normal rate because they were likely to have difficulty reliably producing length contrasts at different rates without some practice. The recording of the nonprofessional talkers took place in an anechoic chamber at ATR Laboratories.

The test consisted of 1032 trials, divided into 24 blocks of either 34 real word trials or nine nonword trials. In each block, stimuli from each combination of the following factors were presented: presentation context (word, sentence), talker, and word type (word, nonword). The order of the two presentation contexts was counterbalanced across listeners, but the word trials always immediately preceded the corresponding nonword trials. The 12 talkers were pseudorandomly ordered in such a way that no more than two talkers from the same group (professional or nonprofessional) appeared consecutively. Within each block of word trials, 34 words from the four contrast types were presented in a random order. Within each block of nonword trials, the nine nonwords were presented in a random order. None of the 12 test talkers appeared during training as training talkers.

The ET and EC listeners took the same test twice, with an average of 8.8 days (8–11 days) between test1 and test2 for group ET and 9.1 days (7–15 days) for group EC. The JC listeners took the test once.

The training stimuli were the same 60 vowel pairs as those used in experiment 1. As mentioned before, 31 of the 60 pairs contained the vowel length contrast in the initial syllable, 27 pairs contained the contrast in the final syllable, and 2 pairs were monosyllabic words. The 60 word pairs were all different from the test words. The words were produced in isolation at a normal rate, by a different set of professionally trained talkers (three females and two males, aged 27–53) from the training talkers in experiment 1. Some of the training talkers in experiment 2 appeared as test talkers in experiment 1. None of the training talkers appeared as test talkers in the present experiment.

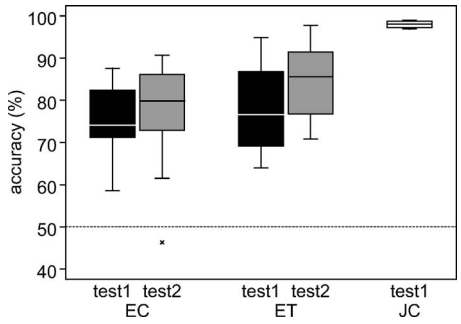


FIG. 4. Identification accuracies in experiment 2 as a function of listener group (EC, ET, JC) and test (test1, test2). The JC listeners took the test only once. Accuracies are based on trials in all conditions of talker, presentation context, and contrast type, including words and nonwords.

C. Results and discussion

1. Overall performance

Figure 4 shows the identification accuracies in test1 and test2 for groups EC and ET and accuracies in test1 for group JC. Accuracies are based on trials in all conditions of talker, presentation context, and contrast type, including words and nonwords.

For group ET, mean identification accuracy was 78.5% (s.d.=10.4) in test1, but rose to 84.4% (s.d.=9.0) in test2. For group EC, accuracies in test1 and test2 were 74.2% (s.d.=10.1) and 75.9% (s.d.=14.3), respectively. Group JC's mean accuracy was 98.0% (s.d.=0.7).

A close look at Fig. 4 reveals that one listener in group EC (EC01) scored substantially lower in test2 (46.3%) than other listeners in the same condition. This score was more than 2 s.d. lower than the mean score in test2. This listener scored 58.6% in test1, resulting in a 12.3-point decrease in accuracy from test1 to test2. Since inclusion of EC01's data effectively lowers group EC's mean scores in test2, excluding his data results in a more conservative test of the experimental hypotheses. If this listener's data are excluded, then group EC's accuracies in test1 and test2 were 76.1% (s.d. =8.8) and 79.6% (s.d.=9.7), respectively.

The non-native listeners' data in Fig. 1 were submitted to two kinds of tests, an ANOVA with EC01's data included, and an ANOVA without EC01's data. Both tests were two-way repeated-measures ANOVAs with group (ET, EC) as a between-subjects variable and test (test1, test2) as a within-subjects variable. When EC01's data were included in the analysis, results revealed a significant main effect of test $[F(1,17)=17.75, p<0.001]$, and a significant interaction between group and test $[F(1,17)=4.69, p<0.05]$. However, when EC01's data were excluded, results revealed a significant main effect of test $[F(1,16)=43.25, p<0.001]$ but no significant interaction between group and test $[F(1,17)=3.54$, n.s.]. Inclusion versus exclusion of EC01's data did not lead to different results for the analysis of talker type, contrast type, and position (Secs. III C 2, III C 3, and III C 4).

TABLE III. Mean identification accuracies in experiment 2 for each professional (pro) and nonprofessional (nonpro) talker (F=female, M=male), for group ET's test1 and test2 and for group JC. Accuracies are based on trials in all combinations of contrast type and presentation context, including words and nonwords.

Talker type	Talker	ET test1	ET test2	JC
Pro	PF1	76.6	84.2	97.9
Pro	PF2	80.0	83.5	98.5
Pro	PF3	78.4	85.3	98.4
Pro	PM1	82.9	84.1	99.1
Pro	PM2	80.5	85.9	98.5
Pro	PM3	79.8	86.6	98.1
Mean		79.7	84.9	98.4
Nonpro	NF1	80.6	86.2	98.8
Nonpro	NF2	77.8	84.1	97.8
Nonpro	NF3	79.3	85.5	97.8
Nonpro	NM1	78.0	85.3	97.8
Nonpro	NM2	77.1	82.1	93.4
Nonpro	NM3	69.7	74.8	91.3
Mean		77.1	83.0	96.1

When group ET's performance during training was briefly examined, it was found that listeners started out at 90.6% accuracy in session 1 and ended at 97.5% accuracy in session 15, with the highest accuracy (98.3%) obtained in session 13. As in experiment 1, accuracy more or less increased across the 15 sessions, but the amount of increase was greater during the first half of training than the second half.

In short, these data provide mixed results concerning whether perceptual training significantly improves non-native listeners' overall identification performance. The results varied depending on whether certain data that appear as outliers are included in the analysis or not. Even if there were an effect of training, the magnitude of the effect appears to be small (an increase of 5.9 percentage points). This may be related to the fact that listeners' performance during training was relatively high even at the beginning, leaving little room for further improvement to take place. Thus, it appears that the perceptual training method in the present study did not lead to a substantial improvement in overall performance. The following sections examine whether training modifies non-native listeners' perceptual tendencies in more subtle ways.

2. Talker type

To examine how accuracy varied across talkers and talker types, Table III shows identification accuracies separately for each talker, for groups ET and JC. Accuracies are based on trials in all combinations of contrast type and presentation context, including words and nonwords. Comparison of the two talker types indicates that mean accuracies were slightly lower for the nonprofessional talkers than for the professional talkers. A closer look at the individual talkers' accuracies reveals that among the six nonprofessional talkers, talker NM3 showed the lowest accuracies in all three tests (ET test1, ET test2, and JC). Talker NM2 also showed somewhat lower accuracies than other nonprofessional talkers in group ET's test2 and the Japanese listeners' test. The remaining nonprofessional talkers (NF1–NF3 and NM1) showed accuracies that were comparable to those of the professional talkers. A two-way ANOVA was carried out on the individual talkers' data in Table III, with talker type (professional, nonprofessional) as a between-subjects factor and test (ET test1, ET test2, JC) as a within-subjects factor. While there was a significant main effect of test [$F(2,20)=732.45$, $p<0.001$], there was no significant main effect of talker type [$F(1,10)=2.21$, n.s.] or a significant interaction [$F(2,20)=0.21$, n.s.], suggesting that there were no significant differences in listeners' performance between the professional and nonprofessional talkers. A similar analysis for group EC (not shown) also showed no significant differences in accuracy between the two talker types.

Recall that training stimuli in the present study were all produced by professionally trained talkers. The absence of a significant difference in overall accuracy between the two talker types or a significant interaction between talker type and test suggests that perceptual training using only productions by professional talkers does not have unequal effects on non-native listeners' ability to perceive productions by professional versus ordinary, nonprofessional talkers.

3. Contrast type

To test again for whether or not training has differential effects on non-native listeners' perception of trained versus untrained contrast types, Fig. 5(a) shows group ET's accuracies in test1 and test2, as well as group JC's accuracies, as a function of the four contrast types and the nonwords. Figure 5(b) shows a similar graph for group EC. Accuracies are based on stimuli produced by all talkers and in both presentation contexts. Listener EC01's data are omitted from Fig. 5(b) and the statistical analysis.

Figure 5 suggests that non-native listeners' accuracies were lowest for nonwords, and higher for the palatal contrasts than for other real word pairs. Also, the greatest increase in accuracy from test1 to test2 seems to be observed for group ET's vowel pairs. These patterns are similar to Fig. 2 in experiment 1.

A three-way repeated-measures ANOVA with group (ET, EC) as a between-subjects variable and test (test1, test2) and contrast (vowel, obstruent, nasal, palatal) as within-subjects variables was conducted for the real words. Listener EC01's data were excluded.[2] Results revealed significant main effects of test [$F(1.17)=36.07$, $p<0.001$] and contrast [$F(2,26)=12.01$, $p<0.001$], and a significant three-way interaction among group, test, and contrast [$F(2,33)=4.61$, $p<0.05$]. Further analysis of the three-way interaction was carried out by examining the group-by-test interaction for each contrast type. For vowel pairs, the test-by-group interaction was highly significant ($p<0.001$). Further examination of this interaction revealed that the increase in accuracy from test1 to test2 was significant for group ET (78.6% to 89.2%; $p<0.001$) but not for group EC (75.1% to 77.3%). For obstruent, nasal, and palatal pairs, the group-by-test interactions were not significant, suggesting that increases in accuracy from test1 to test2 did not significantly differ between group ET (obstruent: 80.5% to 83.3%; nasal: 81.8% to

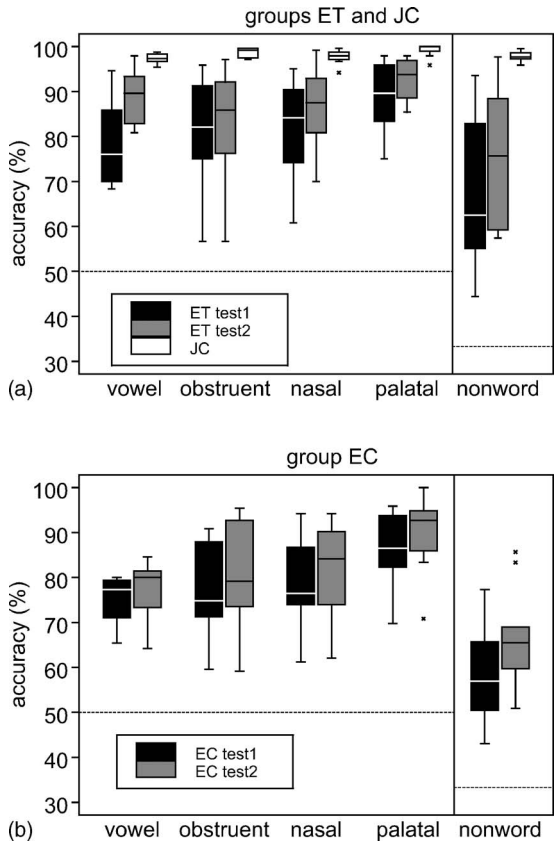


FIG. 5. Identification accuracies in experiment 2 for (a) group ET's test1 and test2 and group JC, and (b) group EC, as a function of the four contrast types and the nonwords. Accuracies are based on stimuli produced by all talkers and in both presentation contexts. The horizontal dashed line indicates chance level performance (50% for real words and 33% for nonwords).

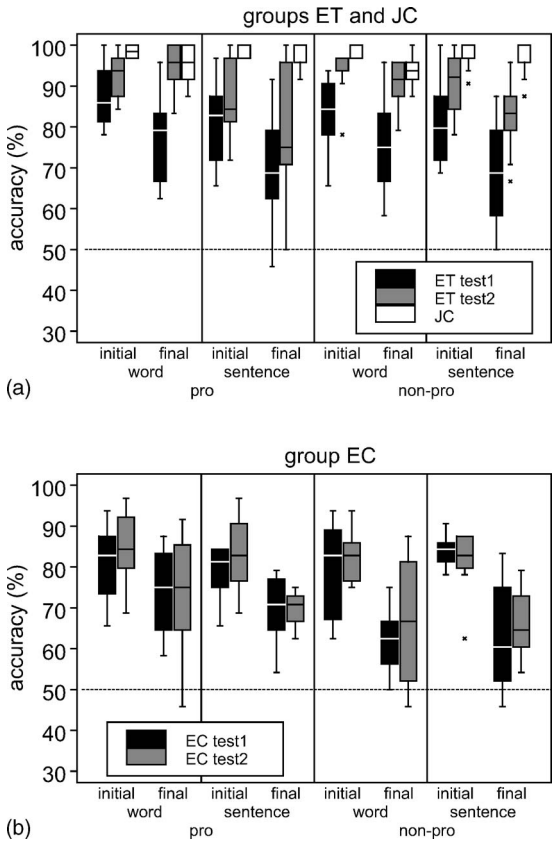


FIG. 6. Identification accuracies in experiment 2 for (a) group ET's test1 and test2 and group JC, and (b) group EC, as a function of within-word position (initial, final) separately for professional and nonprofessional talkers and for word and sentence contexts. Accuracies are based on polysyllabic vowel pairs produced by all test talkers.

86.0%; palatal: 88.5% to 92.8%) and EC (obstruent: 77.3% to 80.7%; nasal: 78.7% to 81.6%; palatal: 86.3% to 89.7%).

Put together, these results suggest, first, that non-native listeners show varying degrees of difficulty with different types of length contrasts, consistent with the results in experiment 1. Second, a significant group-by-test interaction for vowel pairs suggests that training significantly improved non-native listeners' perception of contrast types that listeners were specifically trained with, even though the specific word pairs used in the tests were different from those presented during training. Finally, the lack of significant group-by-test interactions for obstruent, nasal, and palatal pairs suggests that training did not reliably generalize to contrast types that listeners were not trained with.

4. Within-word position

To examine how position of the length contrast within the target word affected listeners' performance, Fig. 6(a) shows identification accuracies for groups ET and JC as a function of the position of the length contrast within the target word, separately for professional and nonprofessional talkers and for word and sentence contexts. Accuracies are based on polysyllabic vowel pairs produced by all test talkers. Figure 6(b) shows a similar plot for group EC. Listener EC01's data are omitted from Fig. 6 and the statistical analyses. Figure 6 shows a general tendency for non-native listeners' accuracies to be higher for length contrasts appearing in word-initial syllables than those in word-final syllables. This seems to be evident for group EC's test1 as well as test2. For group ET, this tendency appears to be somewhat weaker in test2 than in test1.

The non-native listeners' data were submitted to a five-factor repeated-measures ANOVA with listener group (ET, EC) as a between-subjects variable, and with test (test1, test2), talker type (professional, nonprofessional), context (word, sentence), and position (initial, final) as within-subjects factors. If perceptual training improves performance for specific talker types, contexts, or positions, then listener group and test should show significant interactions with these factors.

Results revealed that the main effects for all five factors were significant: listener group [ET: 83.2%; EC: 75.0%; $F(1,16)=6.66$, $p<0.05$], test [test1: 76.1%; test2: 83.0%; $F(1,16)=61.56$, $p<0.001$], talker type [pro: 80.5%; nonpro: 78.6%; $F(1,16)=10.00$, $p<0.01$], context [word: 81.7%; sentence: 77.4%; $F(1,16)=16.37$, $p<0.001$], and position [initial: 84.7%; final: 74.4%; $F(1,16)=51.05$, $p<0.001$]. These results suggest that non-native listeners' performance was strongly affected by the position of the length contrast within the word; performance was poorer by roughly 10 percentage points on average when the length contrast appeared in the word-final syllable than when it appeared in the word-initial syllable. Talker type also significantly affected perfor-

mance, but the size of the effect (a 1.9 percentage point difference) was very small, and Sec. III C 2 clearly shows that this was primarily due to low accuracies associated with one or two of the nonprofessional talkers. Context was also found to affect performance; accuracies were slightly higher when the target word appeared in isolation than when it was embedded in a carrier sentence. This perhaps suggests that carrier sentences have inhibitory, rather than facilitatory, effects on non-native listeners' performance.

In addition to the main effects, the following interactions were significant: listener group by context, talker type by context, listener group by test, listener group by position, talker type by position, and listener group by test by talker type by position. Analysis of the listener-group-by-context interaction indicated that group ET's accuracies were significantly higher in the word context (86.7%) than in the sentence context (79.8%) ($p<0.001$), while group EC's accuracies in the word (75.4%) and sentence (74.5%) contexts did not significantly differ from each other. Next, analysis of the talker-type-by-context interaction indicated that for words produced in isolation, accuracies were significantly higher for professional talkers' productions (83.7%) than for nonprofessional talkers' productions (79.7%; $p<0.001$), but for words embedded in carrier sentences, accuracies did not significantly differ between professional (77.4%) and nonprofessional (77.4%) talkers. This result suggests that professional talkers may have produced isolated words with clearer perceptual cues for length contrasts than nonprofessional talkers.

Finally, the remaining four interactions all involved some or all of the four factors: listener group, test, talker type, and position. Thus, analysis of the highest-order, four-way interaction is reported here. This was done by examining the interaction among listeners group, test, and position for each level of talker type. For professional talkers, group ET showed significant increases in accuracy from test1 to test2 in both the initial position (84.2% to 89.5%; $p<0.05$) and the final position (73.8% to 86.5%; $p<0.001$), with a much larger increase in the final than initial position. Group EC, on the other hand, did not show significant increases in accuracy in either the initial (79.9% to 84.0%) or the final (71.9% to 71.6%) position. For nonprofessional talkers, group ET showed about the same level of significant increase in accuracy in the initial and final positions (76.7% to 89.2%; $p<0.001$), but group EC again did not show significant increase in accuracy (72.1% to 74.1%). These results suggest that training significantly improved non-native listeners' performance, for length contrasts appearing in both word-initial and word-final syllables. Training significantly improved accuracy not just for professional talkers' productions but also for nonprofessional talkers' productions. The largest improvement was observed for word-final length contrasts produced by professional talkers. In fact, even though accuracies were consistently lower for word-final contrasts than word-initial contrasts in most conditions even after training, group ET's test2 accuracies showed no significant differences between the initial (89.5%) and final (86.5%) positions for professional talkers' productions.

Group JC's data were submitted to a three-way ANOVA with talker type, context, and position as within-subjects factors. Results revealed a significant main effect of position [$F(1,9)=15.58$, $p<0.01$] and a significant context-by-position interaction [$F(1,9)=5.25$, $p<0.05$]. Further analysis of the context-by-position interaction using simple effects test revealed that accuracy was significantly lower in final position than initial position for isolated-word stimuli [$F(1,9)=18.72$, $p<0.001$].

In short, non-native listeners' performance was generally much poorer when the length contrast appeared in word-final syllables than when they appeared in word-initial syllables. Perceptual training, however, significantly improved performance for both word-initial and word-final length contrasts. Training using professional talkers' productions also significantly generalized to nonprofessional talkers' productions.

IV. GENERAL DISCUSSION

The present study assessed the extent to which adult non-native listeners' perception of Japanese length contrasts can be modified with identification training, and the extent to which factors such as contrast type, speaking rate, presentation context, within-word position, and talker type affected performance before and after training.

A. Overall effect of perceptual training

In both experiments 1 and 2, both the trained listeners as well as the untrained control listeners improved performance from test1 to test2. Even though group means showed greater improvement in accuracy for trained listeners than untrained listeners, statistical tests failed to show significant differences in the amount of improvement between trained and untrained listeners. Thus, results from the present study do not provide strong evidence that perceptual identification training improves non-native listeners' overall ability to identify Japanese length contrasts. These results are not in line with past studies that have used the same training paradigm to improve L2 learners' perception of L2 segmental contrasts (e.g., Logan *et al.*, 1991; Bradlow *et al.*, 1997), L2 tones (Wang *et al.*, 1999), and L2 syllables (Tajima and Erickson, 2001). However, these results seem to echo the finding of Hirata *et al.* (2007) that training improved English listeners' perception of Japanese vowel length contrasts only to a small extent (9.1 percentage points).

Several explanations are possible for this outcome. First, the tests in the present study may have been excessively long; each test in experiments 1 and 2 consisted of 1128 and 1032 trials, respectively. These are much greater than the number of test trials employed in other training studies, e.g., 32 trials (16 pairs) in the study of Logan *et al.* (1991), 100 trials in the study of Wang *et al.* (1999), and 180 trials in the study of Hirata *et al.* (2007). A large number of trials was necessary in the present study because listeners were to be tested with various combinations of stimulus properties. However, repeated exposure to stimuli in various conditions and increased familiarity with the task may have led to some

improvement during the tests, even in the absence of explicit feedback. This may partly account for group EC's improvement in accuracy between test1 and test2.

Second, the perceptual training employed in the present study may not have been set at an appropriate level of difficulty. Listeners' identification performance during the training sessions started out at a relatively high level in both experiment 1 (83.7%) and experiment 2 (90.6%), leaving little room for performance to improve during training.

Finally, in connection with the previous point, the training stimuli may not have contained sufficient stimulus variability to lead to robust perceptual improvement. The training stimuli in the present study were all vowel pairs produced in isolation at a normal rate. Only stimuli in this limited set of conditions were used during training so as to test for generalization to untrained conditions, and to assess the amount of variability necessary for achieving robust training effects. Past studies have shown that high stimulus variability facilitates the formation of new L2 phonetic categories (e.g., Logan *et al.*, 1991; Lively *et al.*, 1993, 1994; Bradlow *et al.*, 1997). It appears that normal-rate, isolated-word training using vowel pairs only is not sufficient to yield robust training effects.

Even though perceptual training did not lead to significant improvement in overall performance, training did seem to have subtle effects on performance, improving listeners' accuracies in some conditions but not others, as discussed in the following.

B. Effect of contrast types

Results from both experiments 1 and 2 indicated that, among the four contrast types examined (vowel, obstruent, nasal, palatal), palatal pairs had the highest accuracies, while the other three contrast types did not show consistent relative rankings. One potential reason for the high accuracies among palatal pairs is that some of the pairs used in the present study, although construed in Japanese phonology as phonemic length contrasts, can be regarded as involving the presence versus absence of the palatal /i/-like segment characteristic of this contrast, rather than involving a durational contrast; for example, the pair "shaku" (serving saké) and "shiyaku" (reagent) can be seen to differ by whether the palatal portion (underlined) is absent or present, while the pair "kyaku" and "kiyaku" differs by the relative duration of the palatal portion. Non-native listeners may therefore have been able to identify some of these pairs based on the presence versus absence of palatal segments rather than their duration.

As for the effect of training, both Fig. 2(a) from experiment 1 and Fig. 5(a) from experiment 2 suggest that the increase in accuracy from test1 to test2 was greater for group ET's vowel pairs than for other word pairs in group ET or for group EC. However, results of statistical tests from the two experiments diverged, with a significant group-by-test-by-contrast interaction found in experiment 2 but not in experiment 1. It is not clear why divergent results were obtained between the two experiments. One possible reason may be that the variation in speaking rate in experiment 1 may have reduced the effect of training, even for vowel length contrasts which listeners were trained with during training. Since all the stimuli in experiment 2 were produced at the normal rate, listeners may have been able to benefit more from training, especially for contrast types that they were trained with.

Despite the divergent statistical results, training does seem to improve accuracy for length contrasts that listeners were trained with. That is, for group ET, accuracy for vowel pairs increased from 71.4% to 83.4% in experiment 1 and from 78.6% to 89.2% in experiment 2, while for group EC, accuracy only rose from 75.3% to 79.0% in experiment 1 and from 73.3% to 74.1% in experiment 2. Given that most of the test words were different from the training words, this suggests that training generalized to untrained words of the same contrast type.

As for whether training generalized to untrained contrasts, the present data do not provide strong evidence that it does. Data from both experiments suggest that the increases in accuracy from test1 to test2 observed for group ET are not greater than those observed for group EC. These results therefore suggest that the effect of training may be restricted to the specific contrast type that listeners are trained with.

The lack of significant generalization of training to untrained contrast types seems to have important theoretical and practical implications. From a theoretical standpoint, several studies have claimed that essentially the same perceptual mechanisms are used in perceiving various types of length contrasts, in the sense that segment duration serves as the primary perceptual cue for phonemic length (Fujisaki *et al.*, 1975; Uchida, 1998). Under this view, training would be expected to lead to similar levels of improvement in performance for trained as well as untrained contrast types. However, the lack of generalization may suggest that there may be fundamental differences among the contrast types. For example, it has been reported that perceptual sensitivity to durational modifications in segment duration varies depending on the type of speech sound involved, such that sensitivity is higher for vowels than consonants (Huggins, 1972; Kato *et al.*, 2002). If so, then training using vowel pairs, which may be relatively easy to perceive, may not yield improvement in other contrast types, which may be relatively difficult. Furthermore, the four contrast types differ in the phonetic environment in which they appear. For example, vowel length contrasts are typically preceded or followed by consonants, and they form the nucleus of syllables, while obstruent and nasal length contrasts are preceded and followed by vowels, and do not form the nucleus of syllables. Such structural differences may account for why vowel length training does not straightforwardly transfer to consonant length contrasts.

From a practical standpoint, lack of generalization suggests that training listeners with just one type of length contrast does not guarantee improved perception of other contrast types. Training involving multiple contrast types might be necessary to improve non-native listeners' perception of various contrast types. Further research is necessary to determine the extent to which perception of the various contrast types are independent of one another.

C. Effect of speaking rate and presentation context

Non-native listeners' identification accuracies were found to be affected by the speaking rate and presentation context of the test stimuli. In both the word and sentence contexts, performance was lowest for the fast rate (70.3% on average), and higher for the slow (78.6%) and normal (79.6%) rates. The low accuracy at the fast rate likely stems from the fact that the durational difference between phonemically short and long segments tends to be smallest at a fast rate (Hirata, 2004a; Hirata and Whitonm, 2005) making this condition more difficult for non-native listeners than other rate conditions.

Following this line of reasoning, one would expect that non-native listeners' accuracy would be higher for the slow rate than for the normal rate, since the phonemic length distinction tends to be most salient at a slow rate. No systematic differences in accuracy, however, were found between the slow and normal rates in experiment 1. One possible reason for this is that listeners may tend to respond to the stimuli based on an "average" speaking rate among all the stimuli presented in the test. Because speaking rate varied from trial to trial in experiment 1, listeners may not have been able to fully compensate for the rate variation, and may to some extent have performed the task based on a perceptual criterion that applies to tokens produced at an average rate that lies somewhere in the middle of the range of speaking rates encountered. To the extent that this strategy is adopted, this would yield better performance for normal-rate stimuli (which are close to the average rate) and would tend to reduce performance for slow-rate and fast-rate stimuli (which both diverge from the average rate). There is some evidence from previous work that non-native listeners' accuracies were sometimes higher for normal-rate stimuli than slow-rate or fast-rate stimuli (Tajima *et al.*, 2003a). If this interpretation is correct, then accuracies for the slow-rate condition in the present study may not have been as high as expected because the salience of perceptual cues may have been be canceled out by this average rate effect.

Embedding the target word in carrier sentences did not lead to consistently higher or lower accuracies than presenting the word in isolation, but presentation context was found to interact with speaking rate, such that accuracies in the word context were higher than in the sentence context at the normal rate, but *lower* at the fast rate. Such an interaction points to the importance of examining the two factors at the same time. The source of this interaction is not entirely clear. One possibility, although speculative, is that carrier sentences may have facilitatory effects under conditions in which the target word itself contains relatively weak perceptual cues for phonemic length, while they may have inhibitory effects in other conditions. That is, when the target word itself contains relatively weak phonetic cues for phonemic length, as in the case for fast-rate stimuli, non-native listeners may benefit from contextual cues provided by the carrier sentence. On the other hand, when the target word contains sufficiently salient perceptual cues, as might be the case for normal- and slow-rate stimuli, non-native listeners may not need to rely on contextual cues provided by carrier sentences. Instead, carrier sentences may impose additional processing demands on non-native listeners, thus hindering performance rather helping it (e.g., Ikuma and Akahane-Yamada, 2004).

It is worth noting that native Japanese listeners' accuracies were very high at all speaking rates and presentation contexts. This was so even under conditions in which target words, whose speaking rate varied from trial to trial, were presented in isolation with no other contextual information. This suggests that the stimuli contained sufficient perceptual cues for identifying the length contrasts, and that native Japanese listeners, but not native English listeners, were able to utilize those cues to identify phonemic length.

As for the effect of training, statistical tests in experiment 1 indicated that group and test did not significantly interact with context or rate. This suggests that the effects of context and rate mentioned earlier applied equally to both listener groups and to both test1 and test2, with no systematic differences between groups or tests. Thus, the present findings do not provide evidence that training improves listeners' ability to cope with variation in speaking rate and presentation context. Even though speaking rate varied to some extent across the five training talkers in experiment 1 (mean mora duration of 151–194 ms according to Table II), the variability was much greater for the test stimuli (mean mora duration of 98–287 ms according to Table I), which were produced at three speaking rates and presented in a random order across trials. The relatively small variability in the training stimuli may not have been sufficient to modify non-native listeners' perceptual strategies. One way to improve the effectiveness of training may be to increase the variability in speaking rate during training. In fact, Hirata *et al.* (2007) have recently reported that training non-native listeners with sentences produced at two rates (e.g., slow and fast) leads to a more robust training effect than does training with sentences produced at only a single rate (e.g., slow only or fast only). One question that remains open for future research is whether multiple-rate training is equally effective with isolated words as it is with words embedded in sentences.

D. Effect of talker type and within-word position

Professionally trained talkers were recruited in experiment 1 since they were expected to be better able than nonprofessional talkers to produce speech at distinct speaking rates while maintaining clear distinctions between phonemically short and long segments. Experiment 2 tested whether there were in fact differences in identification accuracy between professional and nonprofessional talkers' productions, and whether perceptual training using professional talkers' productions would generalize to nonprofessional talkers' utterances. Results from experiment 2 indicated that there were no significant overall differences in accuracy between the two talker types. However, some subtle differences were observed. For example, for words produced in isolation, mean accuracies were significantly higher for professional talkers' productions (83.7%) than for nonprofessional talkers' productions (79.7%; see Sec. III C 4), suggesting that professional talkers produced clearer perceptual cues for phonemic

length in isolated-word productions than nonprofessional talkers. This gives some support for the original motivation to use professional talkers' productions during training. When the effect of training was examined, group ET showed significantly greater improvement than group EC for both professional and nonprofessional talkers' productions. Thus, training significantly generalized to utterances produced by ordinary, nonprofessional talkers.

Results from experiment 2 suggest that the position of the length contrast within the target word had a very strong effect on non-native listeners' performance. Even native Japanese listeners' performance was somewhat poorer for word-final contrasts in the word context compared to other conditions. The effect of position found in the present study is in agreement with past studies that also found poorer performance in word-final position than initial position (Oguma, 2000; Minagawa-Kawai *et al.*, 2002).

One explanation offered by previous studies (Oguma, 2000; Minagawa-Kawai *et al.*, 2002) for the lower accuracy in word-final position than word-initial position was the absence of phonetic materials in final position. This was predicted to make judgment of segment duration relatively difficult in word-final position (cf. Kubozono, 2002) However, contrary to the prediction, accuracy was relatively low in final position regardless of presentation context; no significant interaction between context and position was found in experiment 2. Since the target word in the present study was always followed by the voiceless stop /t/ of the particle /to/ when produced in carrier sentences, the target word-final segment was likely to be immediately followed by acoustically and perceptually salient speech events. Furthermore, since the particle /to/ usually forms a prosodic word with the preceding word, durational variability between the target word and the particle should be no greater than that observed within the target word (cf. Warner and Arai, 2001). If so, then the availability of temporal cues for phonemic length could be considered to be comparable between final and nonfinal positions. The difference in performance between word-initial and final positions therefore cannot be simply attributed to the presence versus absence of following phonetic materials.

An alternative explanation for the difference in accuracy between word-initial and word-final positions is the presence of pitch-related cues in addition to durational cues for length contrasts in this position. Words in Tokyo Japanese typically have either a low-high or high-low tone pattern in the first two moras, resulting in a rising or falling fundamental frequency contour. For word pairs that contain a length contrast in initial position, the contrast occurs entirely or partially within the first two moras of the word, thus causing the length contrast to be associated with different tone patterns, e.g., */kǵ-dò/* versus */kǵ-ǹ-do/* (target segments are underlined, mora boundaries are indicated with a hypen "-," and high and low tones in the first two moras are marked with /◌́/ and /◌̀/ diacritics, respectively). This means that short and long vowels that occur in word-initial position are often associated with systematically different fundamental frequency contours in addition to differences in duration, while length contrasts appearing in other positions are typically not associated with such tone differences. The availability of such secondary cues may have made length contrasts easier to identify in initial position than other positions. Pitch-related cues have been shown to serve as secondary cues to phonemic length for native Japanese listeners (Omuro-Hayashida, 1999; Kinoshita *et al.*, 2002). Non-native listeners have been reported to rely more on durational cues than pitch cues (Tabuchi *et al.*, 1997; Omuro-Hayashida, 1999). However, the degree to which the English listeners' high accuracies for word-initial length contrasts can be attributed to pitch-related cues remains unclear.

As for the effect of training, group ET in experiment 2, which underwent training, significantly improved performance for both word-initial and word-final length contrasts, for both professional and nonprofessional talkers' productions. In contrast, group EC, which did not undergo training, did not significantly improve performance in any of the conditions. This demonstrates that perceptual training was effective at improving non-native listeners' perception of Japanese vowel length contrasts.

A closer look at the pattern of improvement for group ET indicated that the level of improvement was comparable between initial and final positions (from 76.7% to 89.2% across the two positions) for nonprofessional talkers' productions, but was smaller for word-initial contrasts (from 84.2% to 89.5%) than word-final contrasts (from 73.6% to 86.5%) for professional talkers' productions. Since accuracy for word-initial contrasts for professional talkers' productions was already relatively high in test1 (84.2%), accuracy in this condition may not have increased as much as in other conditions due to ceiling effects. Aside from this difference, levels of improvement from test1 to test2 were comparable between word-initial and word-final contrasts, and between professional and nonprofessional talkers' productions. This again supports the notion that training significantly generalized to nonprofessional talkers' utterances.

Furthermore, results did not reveal significant interactions involving listener group, test, and context. This provides some suggestion that even though listeners were trained using words in isolation, the improvement in performance for words embedded in carrier sentences (from 75.1% to 84.4% on average) was not significantly smaller than that for words produced in isolation (from 80.6% to 92.8%), although the former was slightly smaller than the latter. Concerning generalization of training to untrained contexts, Hirata (2004b) has found that listeners who were trained using words in isolation did not improve performance on words embedded in sentences as well as on words in isolation. Additional research is needed to determine the extent to which training generalizes to untrained contexts.

In conclusion, non-native listeners who were trained to identify Japanese vowel length contrasts did not show greater overall improvement in performance compared to control listeners who did not receive training. However, training seems to affect performance in more subtle ways, modifying performance in some conditions but not in others. Specifically, training improves perception of contrast types that listeners are trained with, generalizes to productions by nonprofessional talkers, and improves perception of length

contrasts that occurs in positions in the word that are originally difficult. However, training does not generalize to contrast types that listeners are not trained with, nor does it significantly improve perception of words and sentences produced at various speaking rates. Further research is needed to clarify ways to refine the training methods so as to yield more robust training effects.

ACKNOWLEDGMENTS

We are grateful to Paul Iverson, Yukari Hirata, and two anonymous reviewers for their helpful comments on earlier versions of this manuscript. We also thank Bryan Burt for running the control participants in Kingston, Canada. This research was funded by the Japan Society for the Promotion of Science and the Ministry of Education, Culture, Sports, Science and Technology.

[1]Attempts were made to eliminate allophonic differences such as vowel devoicing within each minimal pair by asking listeners to read the minimal pairs together in sets rather than separately.

[2]As mentioned previously, inclusion versus exclusion of EC01's data did not alter the main results.

Amano, S., and Kondo, T. (**2000**). *Nihongo-no Goi-Tokusei (Lexical Properties of Japanese)* (Sanseido, Tokyo).

Bradlow, A. R., Pisoni, D. B., 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.

Fujisaki, H., Nakamura, K., and Imoto, T. (**1975**). "Auditory perception of duration of speech and non-speech stimuli," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 197–219.

Greenhouse, S. W., and Geisser, S. (**1959**). "On methods in the analysis of profile data," Psychometrika **24**, 94–112.

Hirata, Y. (**2004a**). "Effects of speaking rate on the vowel length distinction in Japanese," J. Phonetics **32**, 565–589.

Hirata, Y. (**2004b**). "Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts," J. Acoust. Soc. Am. **116**, 2384–2394.

Hirata, Y., and Lambacher, S. G. (**2004**). "Role of word-external contexts in native speakers' identification of vowel length in Japanese," Phonetica **61**, 177–200.

Hirata, Y., Whitehurst, E., and Cullings, E. (**2007**). "Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates," J. Acoust. Soc. Am. **121**, 3837–3845.

Hirata, Y., and Whiton, J. (**2005**). "Effects of speaking rate on the single/geminate stop distinction in Japanese," J. Acoust. Soc. Am. **118**, 1647–1660.

Huggins, A. W. F. (**1972**). "Just noticeable differences for segment duration in natural speech," J. Acoust. Soc. Am. **51**, 1270–1278.

Ikuma, Y., and Akahane-Yamada, R. (**2004**). "An empirical study on the effects of acoustic and semantic contexts on perceptual learning of L2 phonemes," Annual Review of English Language Education in Japan **15**, 101–108.

Kaiki, N., and Sagisaka, Y. (**1992**). "The control of segmental duration in speech synthesis using statistical methods," in *Speech Perception, Production, and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, (Ohmsha, Tokyo), pp. 391–402.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (**2002**). "Effects of phoneme class and duration on the acceptability of temporal modifications in speech," J. Acoust. Soc. Am. **111**, 387–400.

Kinoshita, K., Behne, D., and Arai, T. (**2002**). "Duration and F0 as perceptual cues to Japanese vowel quantity," in Proceedings of the 2002 International Conference on Spoken Language Processing, Denver, CO, pp. 757–760.

Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.

Kubozono, H. (**2002**). "Temporal neutralization in Japanese," in *Papers in Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner (Mouton, Berlin).

Lenneberg E. (**1967**). *Biological Foundations of Language* (Wiley, New York).

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). "Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new phonetic categories," J. Acoust. Soc. Am. **94**, 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., and Yamada, T. (**1994**). "Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories," J. Acoust. Soc. Am. **96**, 2076–2087.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**, 874–886.

Minagawa-Kawai, Y., Maekawa, K., and Kiritani, S. (**2002**). "Effects of pitch accent and syllable position in identifying Japanese long and short vowels: Comparison of English and Korean speakers," Journal Phonetic Soc. of Japan **6**, 88–97 (in Japanese).

Oguma, R. (**2000**). "Perception of Japanese long vowels and short vowels by English-speaking learners," Japanese-Language Education around the Globe **10**, 43–54 (in Japanese).

Omuro-Hayashida, K. (**1999**). "Pitch or duration? The perception of morae in long vowels in Japanese: A comparison between Japanese and English native speakers," in *Transactions of the Technical Committee on Psychological and Physiological Acoustics* [Acoustical Society of Japan (in Japanese), Kumamoto, Japan], Vol. **29**, pp. 1–8.

Sagisaka, Y., and Tohkura, Y. (**1984**). "Phoneme duration control for speech synthesis by rule," Trans. Inst. Electron., Inf. Commun. Eng. A **J67-A**, 629–636.

Tabuchi, S., Tokiyoshi, S., Yamakawa, K., Kai, T., Baba, R., Ueno, K., Usagawa, T., and Ebata, M. (**1997**). "Japanese long vowels: About the role of the pitch in morae perception," in *Transactions of the Technical Committee on Psychological and Physiological Acoustics* [Acoustical Society of Japan (in Japanese), Kumamoto, Japan], Vol. **27**, pp. 1–8.

Tajima, K., and Erickson, D. (**2001**). "Syllable structure and the perception of second language speech," in *Bunpo to Onsei 3 (Speech and Grammar 3)*, edited by Spoken Language Working Group (Kuroshio, Tokyo), pp. 221–239.

Tajima, K., Kato, H., Rothwell, A., and Munhall, K. G. (**2003a**). "Native and non-native perception of moraic phonemes in Japanese: Effect of identification training and exposure," in Proceedings of the 2003 Spring Meeting of the Acoustical Society of Japan, Tokyo, pp. 491–492.

Tajima, K., Kato, H., Rothwell, A., and Munhall, K. G. (**2003b**). "Perception of phonemic length contrasts in Japanese by native and non-native listeners," in Proceedings of the 15th International Congress of Phonetics Sciences, Barcelona, Spain, pp. 1585–1588.

Takeda, K., Sagisaka, Y., and Kuwabara, H. (**1989**). "On sentence-level factors governing segmental duration in Japanese," J. Acoust. Soc. Am. **86**, 2081–2087.

Toda, T. (**2003**). *Second Language Speech Perception and Production: Acquisition of Phonological Contrasts in Japanese* (University Press of America, Lanham, MD).

Uchida, T. (**1998**). "Categorical perception of relatively stead-state speech sound duration in Japanese moraic phoneme," Journal Phonetic Soc. of Japan **2**, 71–86 (in Japanese).

Wang, Y., Spence, M. V., Jongman, A., and Sereno, J. A. (**1999**). "Training American listeners to perceive Mandarin tones," J. Acoust. Soc. Am. **106**, 3649–3658.

Warner, N., and Arai, T. (**2001**). "The role of the mora in the timing of spontaneous Japanese speech," J. Acoust. Soc. Am. **109**, 1144–1156.

Watson, C., Kelly, W., and Wroton, H. (**1976**). "Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty," J. Acoust. Soc. Am. **60**, 1176–1186.

Yamada, R. A., (**1995**). "Age and acquisition of second language speech sounds: Perception of American English /r/ and /l/ by native speakers of Japanese," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York, Timonium, MD), pp. 305–320.

Yamada, T., Yamada, R. A., and Strange, W. (**1994**). "Perceptual learning of Japanese mora syllables by native speakers of American English: An analysis of acquisition processes of speech perception in second language learning," in Proceedings of the 1994 International Conference on Spoken Language Processing, (Yokohama, Japan), pp. 2007–2010.

The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception

Martin Cooke[a)]
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

M. L. Garcia Lecumberri
Department of English Philology, University of the Basque Country, Paseo de la Universidad 5, 01006, Vitoria, Spain

Jon Barker
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

(Received 8 March 2007; revised 8 October 2007; accepted 12 October 2007)

Studies comparing native and non-native listener performance on speech perception tasks can distinguish the roles of general auditory and language-independent processes from those involving prior knowledge of a given language. Previous experiments have demonstrated a performance disparity between native and non-native listeners on tasks involving sentence processing in noise. However, the effects of energetic and informational masking have not been explicitly distinguished. Here, English and Spanish listener groups identified keywords in English sentences in quiet and masked by either stationary noise or a competing utterance, conditions known to produce predominantly energetic and informational masking, respectively. In the stationary noise conditions, non-native talkers suffered more from increasing levels of noise for two of the three keywords scored. In the competing talker condition, the performance differential also increased with masker level. A computer model of energetic masking in the competing talker condition ruled out the possibility that the native advantage could be explained wholly by energetic masking. Both groups drew equal benefit from differences in mean F0 between target and masker, suggesting that processes which make use of this cue do not engage language-specific knowledge.
 [DOI: 10.1121/1.2804952]

PACS number(s): 43.71.Hw, 43.71.Es, 43.66.Dc [ARB] Pages: 414–427

I. INTRODUCTION

It is widely assumed that two kinds of processes play a part in decoding speech when other sound sources are present. First, general-purpose "signal-driven" processes are thought to help in creating an initial segregation of the auditory scene into components belonging to different acoustic sources (Bregman, 1990). Second, "knowledge-driven" processes which exploit prior knowledge of individual sources such as speech could be used to integrate the components into a coherent linguistic interpretation. The boundary between signal-driven and knowledge-driven processes is not clear, and there is some debate over the extent to which auditory scene analysis is capable of grouping the segregated components into coherent structures, or whether this is accomplished primarily by learned models of speech (Remez *et al.*, 1994).

Nearly all speech perception studies have focused on signal-driven processes. For example, in a competing talker experiment, listeners may be faced with sentence pairs whose mean fundamental frequency (F0) difference is the key variable. While these studies measure the effect of factors such as F0 differences on intelligibility, they say little about the role played by prior speech knowledge. One way to explore this latter factor is by comparing listener groups differing in the state of spoken language acquisition. Such nonhomogeneous listener groups are frequently chosen on the basis of native language (e.g., Florentine *et al.*, 1984) although it is also possible to vary prior linguistic experience by comparing perception in children and adults with the same native language (e.g., Hazan and Markham, 2004). If nonhomogeneous listener groups were to show similar intelligibility benefits of factors such as F0 differences, this would constitute strong evidence for the hypothesis that general auditory (or at least language-universal) processes are mainly responsible for F0-based source separation.

The presence of other sound sources results in masking of the "target" speech in ways summarized in Fig. 1. It is conventional to distinguish (i) energetic masking (EM), which occurs when components of the speech signal in some time-frequency region are rendered inaudible because of swamping by the masker, and (ii) informational masking (IM), which covers everything that reduces intelligibility

[a)]Author to whom correspondence should be addressed. Electronic mail: m.cooke@dcs.shef.ac.uk

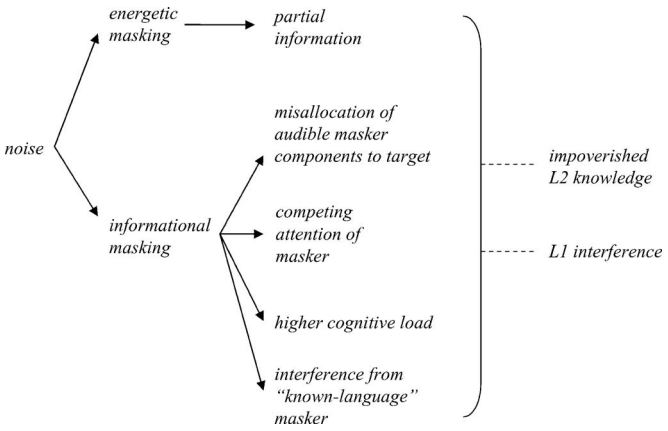


FIG. 1. Summary of potential masking effects for native and non-native listeners.

once energetic masking has been accounted for. Informational masking has multiple facets and is a catch-all term whose use reflects the current state of both conceptual and scientific uncertainty (Durlach, 2006).

The primary purpose of the current study was to investigate differential effects of energetic and informational masking on native and non-native speech perception and to relate the findings to the relative roles of signal- versus knowledge-driven processes in the perception of speech.

Energetic masking leads to a loss of signal components so that partial information has to be used to interpret the speech signal. Fortunately, speech is sufficiently redundant that, for many tasks, only a relatively small proportion of the time-frequency plane has to be "glimpsed" to support high levels of intelligibility (Cooke, 2006; Barker and Cooke, 2007).

Several studies have compared native and non-native or bilingual speech perception in stationary noise, which can be considered to be a pure energetic masker (Florentine *et al.*, 1984; Hazan and Simpson, 2000; Bradlow and Bent, 2002; van Wijngaarden *et al.*, 2002; Garcia Lecumberri and Cooke, 2006; Rogers *et al.*, 2006). While the languages used varied and the tasks ranged from consonant identification in short tokens to keyword identification in sentences, all these studies demonstrated that natives outperform non-natives in noisy conditions. However, for those studies which compared performance in quiet (or low noise) and high noise conditions, estimates of the relative size of the native advantage differed. While one study (Bradlow and Bent, 2002) found that the native advantage remained constant as noise level increased, another (Garcia Lecumberri and Cooke, 2006) demonstrated an increased native advantage in the high noise condition.

We now consider the various potential elements of informational masking listed in Fig. 1. One, misallocation, denotes situations where the listener uses audible elements from the masker to arrive at an incorrect identification of the target, or, equivalently, assigns target elements to the masker, again resulting in an error. Informational masking of speech has most often been studied using maskers which are themselves composed of speech material (Carhart *et al.*, 1969; Brungart, 2001; Freyman *et al.*, 2004). When a single competing talker is present, whole words from the masker can be reported as belonging to the target (Brungart, 2001). However, misallocation could, in principle, apply to units of any size. In particular, patches of acoustic information (e.g., bursts, formant transitions, or frication) might be wrongly attributed. For this reason, any maskers containing speech could produce informational masking through misallocation. In fact, significant IM effects have been demonstrated for *N*-talker babble over a wide range of values for *N* (Simpson and Cooke, 2005). Since misallocation can apply to units smaller than words or phonemes, it might also result in the reporting of a sound or word which is not present in either the target or masking speech. For example, the aspiration following a plosive could be interpreted as the voiceless glottal fricative "h."

A second component of IM comes from the higher cognitive load which results when processing a signal containing multiple components. If both target and masker might contain important information, it is reasonable to suppose that processing resources are allocated to both. A related facet of IM is the failure to attend to the target in the presence of a competing speech masker. The role of differences in fundamental frequency, vocal tract size, and spatial cues in determining which of two competing sentences is attended to has been studied (Darwin and Hukin, 2000). If attention is based on limited resources (e.g., Kahneman, 1973), then a higher cognitive load may well result in difficulties in tracking the target source.

A further aspect of informational masking is dependent on whether the language of the masking speaker is known to listeners. A number of recent studies have demonstrated that the language of the masker can affect the intelligibility of the target sentence (Rhebergen *et al.*, 2005; Garcia Lecumberri and Cooke, 2006; Van Engen and Bradlow, 2007). Rhebergen *et al.* found worse speech reception thresholds for Dutch sentences presented in competing Dutch speech than when the competitor material was Swedish. Using a consonant in vowel context identification task, Garcia Lecumberri and Cooke (2006) showed that monolingual English listeners performed better when the language of a competing speaker was Spanish rather than English, whereas Spanish listeners with English as their second language (L2) were equally affected by maskers in both languages. Van Engen and Bradlow (2007) demonstrated that for native English listeners, English sentence intelligibility was better when the noise consisted of two-talker Mandarin Chinese babble than when it was composed of two-talker English babble. These results suggest that a masking talker using a language known to the listener increases informational masking relative to one using an unknown language, perhaps due to the engagement of language-specific decoding processes for both masker and target, which in turns increases cognitive load.

In contrast to energetic masking, the role of informational masking in non-native speech processing has received little attention to date. A number of researchers have compared native and non-native or bilingual performance using maskers composed of speech babble (Mayo *et al.*, 1997; Cutler *et al.*, 2004; Garcia Lecumberri and Cooke, 2006), which

is known to be capable of inducing IM (Simpson and Cooke, 2005). While all three studies mentioned IM-related factors such as native language (L1) interference, none of them supported a quantitative assessment of the size of any native advantage in IM, and the tasks involved were not designed with IM in mind.

Consequently, a key goal of the current study was to compare native and non-native performance using a competing talker task which is known to induce extensive informational masking (Brungart, 2001). The competing talker situation arises frequently in speech communication, so it is of interest to determine whether informational masking effects cause significantly more problems for non-native listeners. Brungart presented listeners with pairs of sentences added at a range of target-to-masker ratios (TMRs). Sentences were drawn from the CRM corpus (Bolia *et al.*, 2000), which consists of sentences with a simple structure such as "ready baron go to green 4 now" or "ready charlie go to red 3 now." The call sign ("baron," "charlie") acts as a keyword to identify which of the sentences of the pair the listener is to attend to, and the task usually involves reporting the color and the digit of the attended sentence. Brungart varied the availability of potential cues for separating the sentence pairs and estimated the amount of informational masking in each case. In one condition, listeners were asked to identify keywords from a target talker when the same talker was used as a masker. This condition provides few cues to separate the two utterances and produced large amounts of informational masking, which was especially evident when target and masker sentences were mixed at the same rms level. Listeners were able to use a level difference cue even in the same talker condition to improve keyword identification scores. In other conditions, the target and masking talkers were of the same gender or of differing genders, providing cues such as voice quality and F0 differences which contributed to a reduction in informational masking.

The current study employed a task similar to that used by Brungart to explore the role of informational masking in non-native speech perception. Speech material was drawn from the "Grid" corpus (Cooke *et al.*, 2006), which consists of 1000 sentences from each of 34 talkers. Sentences have a particularly simple six-word form such as "place white at B 4 now" and "lay green with N 8 again." The availability of many utterances and individual talkers in the corpus also allowed the effect of factors such speech rate, F0 differences, and individual speaker intelligibility to be compared across the native and non-native groups.

Experiment 1 measured native and non-native listeners' keyword identification scores for Grid sentences in quiet and in three levels of speech-shaped noise (SSN). The primary goal was to derive an estimate of pure energetic masking to enable the contribution of EM in experiment 2 to be estimated in order to better assess the role of IM. A further goal was to determine whether an increasing native advantage with greater amounts of energetic masking found in our earlier study with VCV tokens (Garcia Lecumberri and Cooke, 2006) carries over to the sentence material of the Grid corpus. Sentences contain a wider range of phonetic realizations, including intersegmental effects, admit more variation in speaking rate and prosodic structure, and invoke a higher cognitive load than isolated VCVs. By using simple sentences, our aim was to introduce some natural variation while restricting the use of high-level knowledge. Although it is known that native listeners are better able to exploit higher-level knowledge such as syntactic context contained in sentence-level material in processing noisy speech (Mayo *et al.*, 1997; van Wijngaarden *et al.*, 2004), the Grid corpus was felt to minimize demands on higher-level processing since all utterances are syntactically, semantically, and pragmatically equal and sufficiently short to reduce memory loading. In addition, the total lexicon used in the corpus is a collection of 51 very common words, of which only 39 act as keywords, minimizing non-native listener disadvantage due to deficits in the L2 lexicon. Our goal in using Grid sentences was similar to that of Meador *et al.* (2000), who used semantically unpredictable sentences composed of words likely to be known to non-native listeners. A further aim was to compare the intelligibility of multiple talkers for the two listener groups. Grid contains 34 talkers of both genders with a variety of accents, and differences in the intelligibility ranking of talkers might be expected based on the native listeners' richer knowledge of talker differences.

Experiment 2 compared informational masking in natives and non-natives and measured keyword identification scores for target sentences in the presence of a competing sentence in conditions where the availability of cues to the separability of the sentence pair was systematically varied in the manner of Brungart (2001).

II. EXPERIMENT 1: SENTENCES IN STATIONARY NOISE

A. Methods

1. Participants

The non-native group consisted of 49 native speakers of (European) Spanish. All were students at the University of the Basque Country studying English as a foreign language (age range: 20–25, mean: 21.2 years). They were enrolled in a one-semester course in English Phonetics in the second year of a four-year BA degree in English Language and Literature. All students had attained the level of the Cambridge Advanced Examination. Students received course credit for participating in the listening tests. The results of 7 of the 49 Spanish listeners were excluded from the analysis of experiment 1 based on their performance in experiment 2 (see Sec. III A 1).

Results for native listeners were derived from an earlier study on speaker identification in noise (Barker and Cooke, 2007), which employed similar stimuli. Twenty English listeners took part in the Barker and Cooke study, 18 of which also participated in experiment 2 of the current study. For this reason, results from the common subset of 18 were extracted for comparison with non-natives in experiment 1.

2. Speech and noise materials

Speech material was drawn from the Grid corpus (Cooke *et al.*, 2006), which consists of six word sentences such as "lay red at H 3 now." In experiment 1, colors, letters, and digits acted as keywords. Four choices of color ("red," "white," "green," and "blue"), 25 letters of the English alphabet (excluding "W" due to multisyllabicity), and ten spoken digits ("one" to "nine" and "zero") were available. All 34 talkers (18 male, 16 female) who contributed to the Grid corpus were used.

Listeners heard utterances without noise and in three stationary noise conditions created by the addition of SSN whose long-term spectrum was the average of sentences in the Grid corpus. Noise was added at token-wise signal-to-noise ratios (SNRs) of 6, 0, and −6 dB. Spanish listeners identified 60 sentences in each of the four conditions while native listeners had been tested in an earlier study across a larger range of SNRs and heard 100 sentences in each condition (Barker and Cooke, 2007).

Two considerations led to the choice of SNRs for the non-native group. First, since one goal of the experiment was to provide a means of estimating the effect of energetic masking in experiment 2, the noise levels had to be chosen such that they would result in a similar degree of energetic masking as found in the competing talker conditions in that experiment. Since a competing talker provides around 6–8 dB less masking than does stationary noise when presented at the same SNR (Miller, 1947; Festen and Plomp, 1990), it was necessary to avoid the use of extremely low SNRs in experiment 1. A second reason for choosing these SNRs was on the basis of estimates of the expected non-native performance disadvantage, derived by extrapolating from our earlier study (Garcia Lecumberri and Cooke, 2006). There, native listener performance in stationary noise was at the same level as that of non-native listeners in a competing talker condition, suggesting a non-native deficit of around 6–8 dB. While Garcia Lecumberri and Cooke used VCVs rather than sentences, the lack of strong contextual cues in the current task suggests that a similar non-native deficit might result for both types of stimuli. By applying this estimate to the full SNR-intelligibility relation for natives in the current task provided by Barker and Cooke (2007), the SNR values of 6, 0, and −6 dB were predicted to provide a representative mapping of the non-native SNR-intelligibility relation.

To enable sufficient representation of speech material from different talkers, individual listeners heard different sets of sentences in each condition. For the natives, each listener heard a different set of sentence/talker combinations drawn at random from the Grid corpus. For the non-native group, ten different sets of sentences/talker combinations were extracted from the corpus at random, and each listener was randomly assigned to one of the ten sets.

3. Procedure

The non-native group was tested at the University of the Basque Country. Stimulus presentation and response collection was under computer control. Listeners were asked to identify the color, letter, and digit spoken and entered their results using a conventional computer keyboard in which four of the nonletter/digit keys were marked with colored stickers. Listeners were familiarized with the stimuli and the task by identifying an independent practice set of 60 sentences in quiet prior to the main set. Stimuli were blocked according to noise level, and the order of the blocks was randomized across listeners.

Native listeners had been tested individually in an IAC single-walled acoustically isolated booth using Sennheiser HD250 headphones at the University of Sheffield. At the University of the Basque Country, participants were tested in groups of 15–20 in a quiet laboratory using Plantronics Audio-90 headphones. The difference in the two stimulus presentation setups in the two countries was shown to be nonsignificant in a previous study involving the perception of VCV tokens in noise (Garcia Lecumberri and Cooke, 2006).

B. Results

Listener responses were scored separately for each of the three keywords. Due to near-ceiling performance in quiet and the low noise condition for the native group, scores were converted to rationalized arcsin units (RAU; Studebaker, 1985) for both statistical analyses and graphical displays. Figure 2 shows RAU-transformed keyword scores for the two groups together with the native advantage (N-NN), expressed as a difference in RAUs. As expected, native listeners performed better in all conditions. For the color and number keywords, the native advantage increased with background noise level (from 6 to 14 RAUs for the color keyword and from 12 to 27 RAUs for the number keyword). On average, the native advantage was least for the color keyword and greatest for the letter keyword. Separate repeated measures ANOVAs with one within-subjects factor (noise level) and one between-subjects factor (nativeness) were performed for each of the three keywords, confirming the clear effects of noise level and nativeness in each case. The interaction noise level × nativeness was significant for color $[F(3,56)=2.85,\ p<0.05,\ \eta^2=0.13]$ and number $[F(3,56)=12.6,\ p<0.001,\ \eta^2=0.40]$ but not for letter $(p=0.29)$.

These results suggest that non-native listeners suffer more from the effects of increasing stationary background noise when identifying certain keywords in simple sentences. Interestingly, the native advantage for the most difficult keyword (letter) did not increase with noise level.[1] It is not clear why non-native listeners did not suffer more for the highly confusable set of spoken letters, though it is notable that the native advantage in quiet was already large.

As mentioned earlier, the native listeners of the Barker and Cooke (2007) study were exposed to a larger range of SNRs and a greater number of tokens in each noise condition than the non-native group. To determine whether the greater exposure to the task was beneficial for the natives, the full range of conditions from the earlier study was analyzed for both within-condition learning effects and across-condition

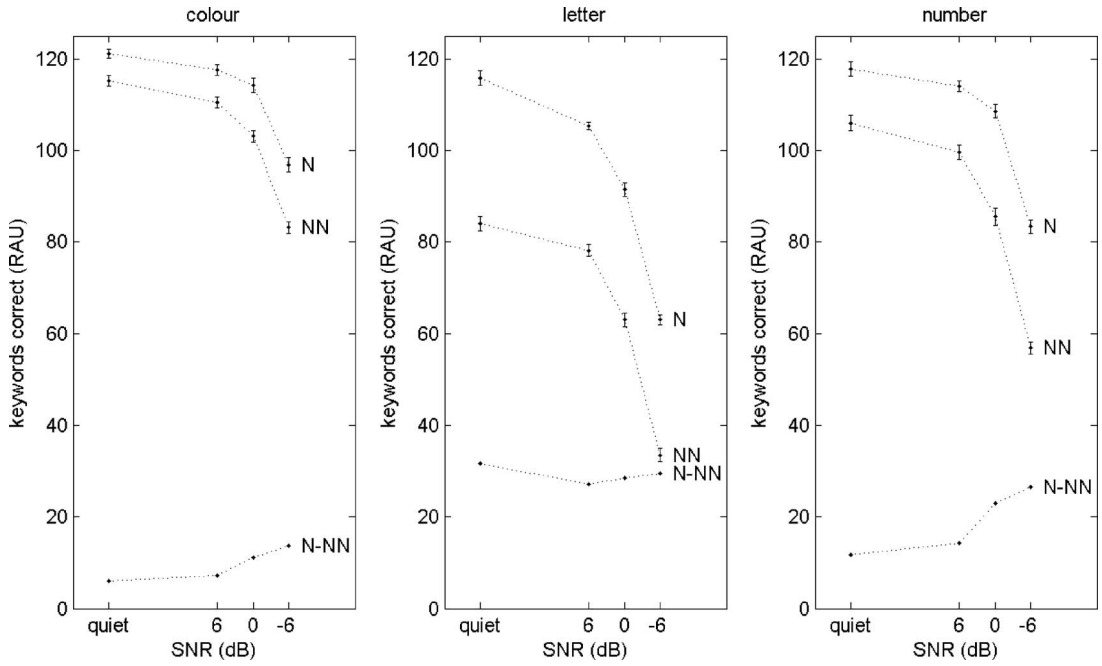


FIG. 2. Native (N) and non-native (NN) keyword identification scores in rationalized arcsine units (RAUs) for quiet and in three levels of speech-shaped noise for color, letter, and number keywords. The native advantage (N-NN in RAUs) is also shown. Error bars here and elsewhere denote ±1 standard errors.

order effects. No such learning effects were found,[2] suggesting that the differences reported here were due to differences in the two groups of listeners.

C. Acoustic and speaker analyses

To further explore the origins of energetic masking suffered by listeners, a number of additional analyses involving the factors of gender, speech rate, and talker were performed. For these analyses, results are presented as percentage keywords correct, transformed to RAUs.

1. Gender

The top row of Fig. 3 shows the breakdown of intelligibility by speaker gender for native and non-native listeners in the four conditions of experiment 1. Both listener groups showed a similar pattern in each condition. Utterances from neither gender proved more intelligible than the other in quiet and low levels of noise. However, female speakers were more intelligible by approximately equal amounts for both groups in higher levels of noise. A two-way ANOVA (gender × nativeness) at each SNR confirmed a significant intelligibility advantage for female talkers at both 0 dB[$F(1,58)=6.5$, $p<0.05$, $\eta^2=0.10$] and −6 dB[$F(1,58)=57.1$, $p<0.001$, $\eta^2=0.50$]. It also confirmed the lack of a gender by nativeness interaction in any of the four conditions, suggesting that both groups benefitted equally from the more intelligible female talkers.

2. Speech rate

The lower row of Fig. 3 illustrates the effect of speech rate on intelligibility. Since utterances have the same number of words (and nearly the same number of syllables), speech rate is approximately in inverse proportion to utterance duration. Utterances were split into three equal-sized groups based on duration, and keyword identification scores for the fastest and slowest groups were compared. Overall, listeners scored significantly better for the slower utterances ($p<0.001$) in quiet and at all SNRs. In the quiet condition, a small but significant interaction between duration and nativeness was present [$F(1,58)=5.0$, $p<0.05$, $\eta^2=0.08$]. Here, the native group showed no benefit of increased duration, probably because their performance was near ceiling.

3. Individual talkers

Figure 4 shows scatterplots of mean talker intelligibility scores for the native versus non-native groups in quiet and each of the three levels of speech-shaped noise. Correlations between native and non-native scores are also shown. Apart from quiet, all correlations are statistically significant ($p<0.01$ for the 6 dB condition, $p<0.001$ for the 0 and −6 dB conditions).

In the quiet condition, most of the native scores are at or near "ceiling" levels (no listener errors are made for 21 of the 34 talkers). However, the non-native group found certain talkers more difficult than others. For example, talkers m28 and f33 are outliers (defined here as lying more than 2 s.d. from the mean). These talkers are the only two with a Scottish accent in the Grid corpus, and it is perhaps not surprising that nonstandard accents are problematic for non-native listeners, who lack the exposure to a range of regional variation. Indeed, dialectal/idiolectal variation can contribute to poor non-native perception (Strange, 1995) and the intelligibility of unfamiliar accents is correlated with the nativeness of the listener (Ikeno and Hansen, 2006).

As conditions become increasingly adverse, native and non-native judgments of talker difficulty converge, as demonstrated by the increase in correlation from 0.44 at 6 dB SNR to 0.80 at −6 dB SNR. For example, in the latter con-

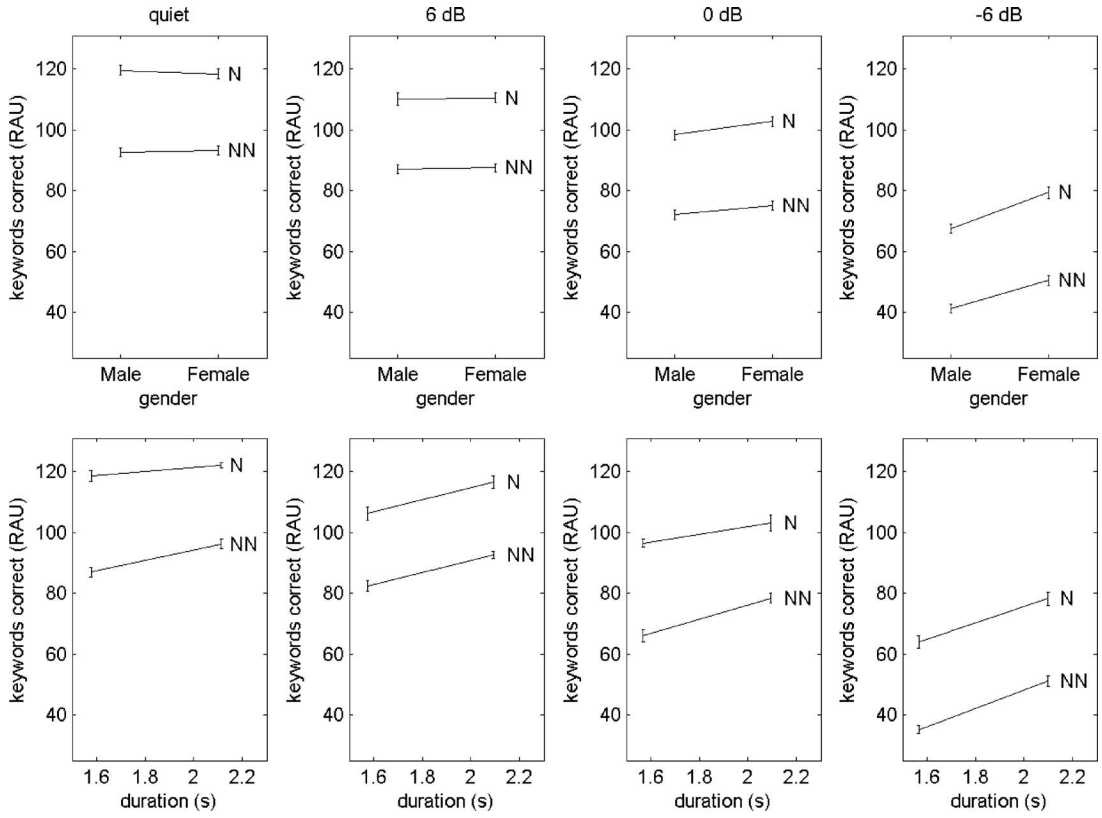


FIG. 3. Effect of speaker gender (top) and speech rate (bottom) on intelligibility for native and non-native listeners in the four conditions of experiment 1. Scores are averaged over color, letter, and number keywords and transformed to RAUs. For speech rate, the ordinate represents the mean duration of the fastest and slowest thirds of the utterance set.

dition, while talkers such as f7 and fl8 are highly intelligible for natives and non-natives, talker m1 is problematic for both groups. The increasing similarity for the two listener groups in noise suggests that accent and other idiosyncratic speaker-related information is rendered less salient by energetic masking, while talker characteristics which promote robust speech cues that resist masking are useful to native and non-native listeners, although not necessarily equally so.

D. Summary and discussion

Experiment 1 measured the effect of pure energetic masking on the two listener groups and revealed that the non-native group suffered more from increasing levels of noise for two of the three keywords in the simple sentences used here. An analysis of gender effects in experiment 1 demonstrated a similar pattern of increasing intelligibility of female talkers in the high noise conditions for both listener groups. The factors that underpin the higher mean female intelligibility in noise for this corpus are not known but appear to be equally useful to both native and non-native listeners, suggesting that it is relatively low-level acoustic differences such as higher formant frequencies or language-independent differences in speaking style rather than language-specific factors which govern the female intelligibility advantage. Bradlow *et al.* (1996) and Hazan and Markham (2004) also found that females were more intelligible for speech presented in quiet conditions. In the current study, the masker had a long-term spectrum derived from both male and female talkers, so it is possible that the higher center of gravity of female speech (caused both by a higher mean F0 and higher formant frequencies) led to some release from masking relative to male talkers.

Non-native listeners benefitted from a slower speech rate in quiet and all noise levels, while natives benefitted in all but the quiet condition. A slower speaking rate can help in two ways. First, in absolute terms, it leads to a greater "visibility" of the target speech. If an informative acoustic feature is masked at one instant, it may be "glimpsed" at a later instance with a probability inversely proportional to the speaking rate. Of course, slower speaking rates do not lengthen all sounds equally, so the increased glimpsing opportunities may not reduce misidentifications across all phonemes. Second, a slower speaking rate results in a slower information rate and thus reduced attentional load for higher-level tasks such as lexical retrieval. While the increased visibility of the target will help both natives and non-natives in noise, it seems plausible that non-natives will benefit from anything which increases the available processing time due to the greater complexity of speech perception in a second language (Gass, 1997).

One striking finding of experiment 1 was that while native and non-natives found different individual talkers more or less intelligible in quiet and low noise conditions, both groups tended to agree on an intelligibility ordering of talkers in noisier conditions. A similar finding was reported for automatic speech recognition scores in noise for this corpus (Barker and Cooke, 2007). This suggests that in situations of low noise, where detailed acoustic information pertaining to

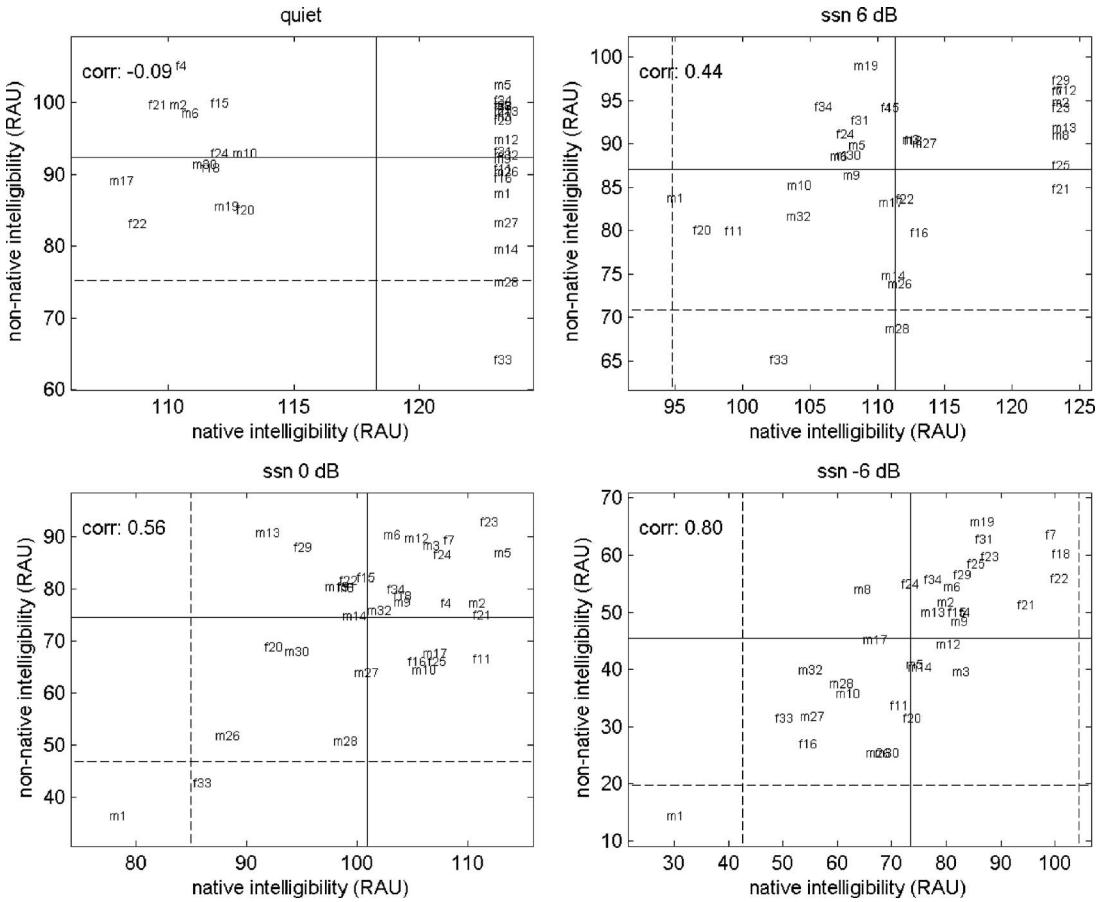


FIG. 4. Intelligibility of individual talkers for native and non-native listeners in the four conditions of experiment 1. Solid lines indicate mean intelligibilities across all talkers for the two listener groups, while dotted lines locate ±2 s.d. from the mean (these are plotted where the value is within the range of talker scores). Male and female talkers are identified by m and f, respectively, and the numbers distinguish different talkers in the Grid corpus. Scores computed as for Fig. 3.

individual talkers is available, previous experience of speech variation produced by factors such as differing accents are dominant in producing the native advantage. Native listeners are able to draw upon a richer knowledge base in interpreting the signal. However, in the presence of high levels of noise, these knowledge-driven factors appear to give way to more general acoustic factors that make individual talkers more resistant to noise, since both groups find the same talkers difficult or easy to recognize. This might be seen as a generalization of the result for different genders. A talker with a "peaky" spectrum (i.e., with energy concentrated at the information-bearing formant frequencies) will resist energetic masking more than one whose spectral profile is more diffuse. It appears that both native and non-native listeners benefit from the more informative distribution of glimpses of the target which result.

III. EXPERIMENT 2: COMPETING TALKERS

A. Methods

1. Participants

The same listeners (47 non-native, 18 native) who took part in experiment 1 participated in experiment 2. However, 7 of the Spanish participants were deemed to be responding randomly in the most difficult conditions (mean keyword identification of 2% in the most difficult condition compared with a mean of 47% for the rest of the non-native group). Consequently, their results were excluded from the analysis of both experiments.

2. Speech materials

As in experiment 1, utterances were drawn from the Grid corpus (Cooke *et al.*, 2006). Sentences were paired to be approximately equal in duration and added at six different TMRs: 6, 3, 0, −3, −6, −9 dB, chosen on the basis of Brungart (2001). The target sentence always contained the keyword "white." The letter and digit keywords always differed in the target and masker. Following Brungart (2001), sentence pairs were split into three subconditions: "same talker," "same gender," and "different gender." There were 20 sentence pairs in each of the three subconditions, leading to a total of 60 pairs at each of the 6 TMRs.

3. Procedure

Stimulus presentation was as described for experiment 1. Listeners reported the letter and digit spoken by the target, identified as the speaker producing the keyword "white." Listeners were familiarized with the task through an independent practice set of 60 utterance pairs. In the main part of the experiment, each TMR constituted a block. Presentation order of the six blocks was randomly chosen for each lis-

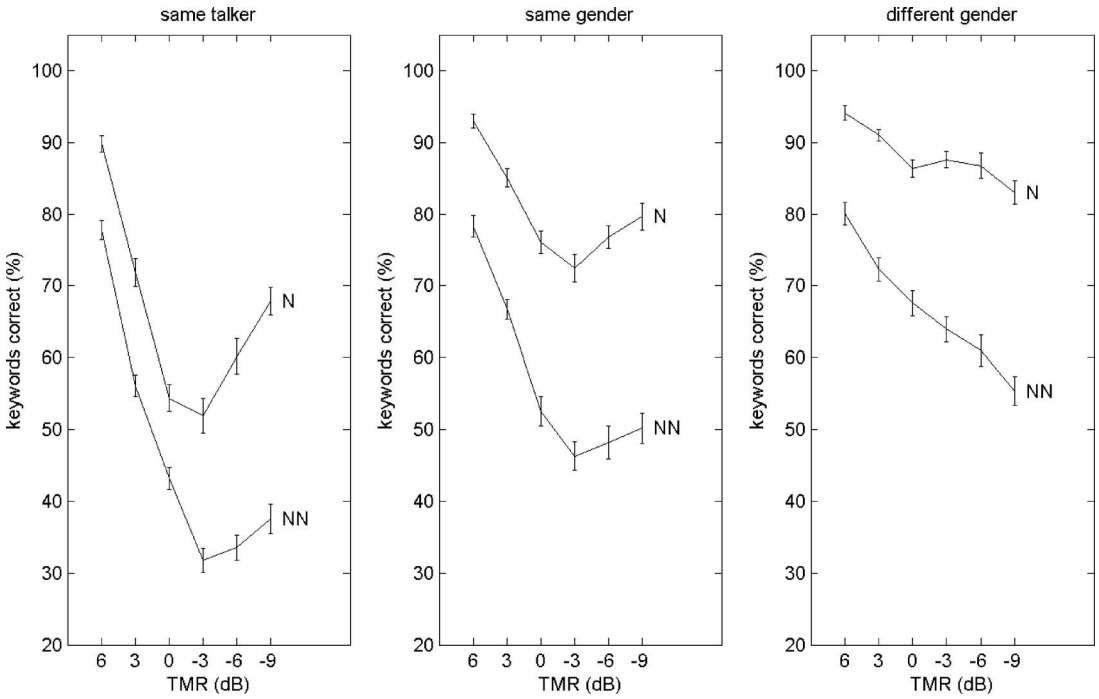


FIG. 5. Native and non-native keyword identification scores in the two-talker conditions.

tener. Within each block, the 60 utterance pairs were presented in a random order. Consequently, the same talker, same gender, and different gender utterance pairings were mixed within blocks. To prevent listeners from using absolute level as a cue to the target utterance, presentation level was randomly roved over a 9.5 dB range from stimulus to stimulus.

B. Results

Results are presented as percentage keywords correct scored for the (letter, digit) keyword pair. As for experiment 1, percentages were converted to RAUs for statistical analysis. However, this produced essentially identical outcomes as for raw percentages, so results are displayed in terms of percentages for ease of interpretation.

Figure 5 presents native and non-native keyword identification performance as a function of TMR in the three simultaneous talker subconditions. The pattern of results for native listeners was similar to that found by Brungart (2001). Listeners had least difficulty in identifying target keywords when the masking talker was of a different gender, and had most difficulty when the same talker was used for target and masker. The strongly nonmonotonic pattern as a function of TMR in the same talker and same gender conditions was also found by Brungart, and demonstrates the beneficial effect of level differences between target and masker in helping to assign keywords to the target speaker. In the different gender case, other cues are sufficiently strong to render level differences unnecessary. These results confirm the usefulness of the Grid corpus in studies of informational masking in speech. Overall, scores for the non-native group followed the same pattern.

For the current study, whose focus is on differences in native and non-native performance in a task designed to produce large amounts of informational masking, the main feature of interest in the results is the large native advantage at all TMRs and in all subconditions. The native advantage ranges from 12 to 15 percentage points at the most favorable TMR and reaches 30 percentage points at the least favorable TMR.

A repeated measures ANOVA with two within-subjects factors (TMR and sentence pairing condition) and one between-subjects factor (nativeness) showed that the three-way interaction of TMR × nativeness × sentence pairing was not significant ($p>0.5$). However, all two-way interactions were significant [TMR × nativeness: $F(5,54)=8.7$, $p<0.001$, $\eta^2=0.45$; sentence pairing × nativeness: $F(2,57)=5.4$, $p<0.01$, $\eta^2=0.16$; sentence pairing × TMR: $F(10,49)=30.2$, $p<0.001$, $\eta^2=0.86$]. The first of these (TMR × nativeness) confirms that non-native listeners are more seriously disadvantaged at adverse TMRs than natives. The second interaction (sentence pairing × nativeness) suggest that the pattern of native advantage is different for each sentence pair type (same talker, same gender, different gender), although the effect is small. The remaining interaction (sentence pairing × TMR) arises from the differing non-monotonic behavior with TMR across the three conditions. The analysis also confirmed highly significant effects of TMR [$F(5,54)=127$, $p<0.001$, $\eta^2=0.92$], speaker pairing condition [$F(2,57)=369$, $p<0.001$, $\eta^2=0.93$], and nativeness [$F(1,58)=131$, $p<0.001$, $\eta^2=0.69$].

C. Acoustic analyses

As for experiment 1, analyses were performed to determine how the two listener groups responded to low-level factors. Here, fundamental frequency differences between the target and masker sentences and absolute duration were examined.

1. Fundamental frequency differences

A difference in fundamental frequency (F0) between two simultaneous sentences has long been known to improve intelligibility (Brokx and Nooteboom, 1982; Bird and Darwin, 1998). F0 difference is considered a primitive source separation cue which is independent of the type of source and hence ought to be equally beneficial to listeners, regardless of their first language.

For each sentence pair employed in experiment 2, the mean instantaneous difference in F0 (measured in semitones) was computed for all frames where both utterances were voiced. F0 information and binary voicing decisions were computed at 10 ms intervals automatically using an autocorrelation approach implemented in PRAAT (Boersma and Weenink, 2005). Using the same approach as was applied for duration in experiment 1 (Sec. II C 2), sentence pairs were split into three equal-sized groups based on F0 difference and the lower and upper groups compared. The three subconditions (same talker, same gender, and different gender) were analyzed separately to prevent the differing extent of mean F0 differences in the three subconditions from masking any differences in the lower and higher terciles.

The effect of F0 differences on keyword identification in each of the two-talker conditions is shown in the upper panel of Fig. 6. Significant effects were found in all three subconditions [same talker: $F(1,58)=7.79$, $p<0.01$, $\eta^2=0.12$; same gender: $F(1,58)=40.4$, $p<0.001$, $\eta^2=0.41$; different gender: $F(1,58)=6.51$, $p<0.05$, $\eta^2=0.10$]. None of the interactions with nativeness were significant, suggesting that F0 differences were equally beneficial for the two groups. The smallest effect of F0 differences was found in the different gender condition. All utterance pairs in the different gender condition had large F0 differences, ranging from a mean of around 5 semitones in the lower tercile to nearly an octave in the higher tercile. These findings suggest that, for both listener groups, the effects of F0 differences reach a ceiling at around 5 semitones, a result in line with the findings (for native listeners) of Darwin *et al.* (2003), who employed similar sentences and near-identical competing talker conditions.

2. Speech rate

To determine the effect of speech rate on keyword identification in the two-talker conditions, an analysis similar to that described in Sec. II C 2 was performed. Since utterances were paired by similar duration, this analysis is not of speech rate differences between the utterances in a pair, but rather examines the effect of absolute speech rate. The results shown in the lower panel of Fig. 6 combine durational information across the three two-talker subconditions since very similar patterns were observed in analyses of each subcondition. The interaction between duration and nativeness was significant [$F(1,58)=7.2$, $p<0.01$, $\eta^2=0.11$]. Non-natives benefitted significantly from a slower overall speaking rate [$F(1,58)=46.5$, $p<0.001$, $\eta^2=0.45$] while natives did not ($p>0.2$).

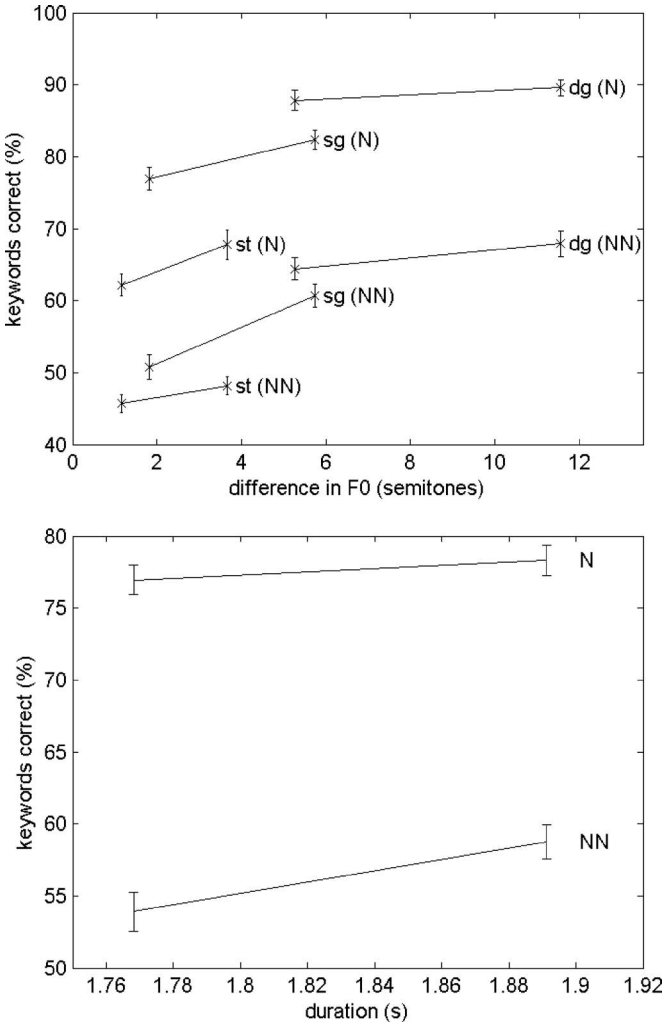


FIG. 6. Upper panel: Effect of fundamental frequency differences in the two-talker conditions. Keyword identification scores for natives and non-natives in the three subconditions (sg=same gender, st=same talker, dg=different gender) are presented for subsets of utterance pairs in the lower and upper tercile of F0 differences. Lower panel: Effect of absolute duration on keyword identification in the two-talker conditions.

D. Quantifying the degree of informational and energetic masking in the two-talker situation

At issue in the current study is the origin of the native advantage in the two-talker case. A competing talker provides relatively little energetic masking at the TMRs used here (Brungart *et al.*, 2006) and informal listening suggests that the letter-digit pair from both target and masker are usually clearly audible in the two-talker signals. Following Brungart (2001), one way to assess the extent of informational masking in the two-talker task is to examine the proportion of listener responses which were present in the masker rather than the target utterance. These errors might be considered to result from informational masking. Figure 7 partitions listener responses into three categories: keywords correctly reported, i.e., present in the target (black), errors where the keywords reported were from the masking talker (midgray) and keywords not contained in either utterance of the pair (light gray). In general, native listeners make fewer "masker confusions" than non-native listeners in the three speaker-pairing subconditions. At the higher TMRs, the na-

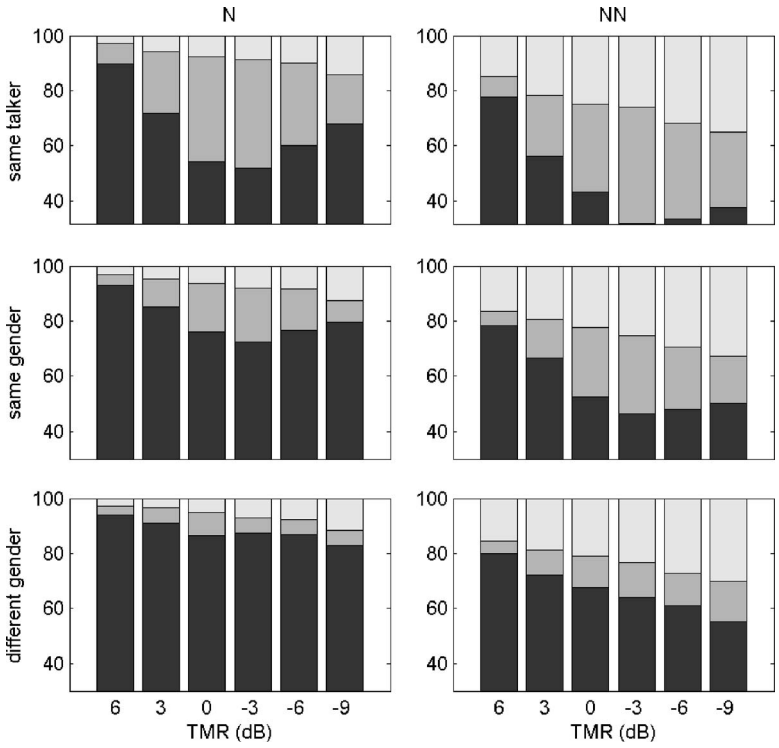


FIG. 7. Proportions of keywords from the target utterance (black) and from the masker (mid gray). The residual (light gray) shows the proportion of responses which were not part of the target or masker.

tive advantage is slight but increases to around 9 percentage points at the lowest TMR. The three subconditions show different degrees of native advantage. The same talker condition shows the least difference between the two groups while the greatest difference occurs in the same gender subcondition. This may reflect the relative ease of the different gender task for both listener groups, reducing the scope for any native advantage.

The difference in the proportion of keywords reported that were present in neither the target nor masker increases monotonically with decreasing TMR, from approximately 12 to 20 percentage points, with a similar increase in each of the three speaker-pairing subconditions. It is tempting to ascribe this type of error solely to energetic masking, but, while it is likely that many of these errors do originate in EM, it is also possible that listeners sometimes combine acoustic cues from the target and masker to "invent" a third sound. This is a form of informational masking by misallocation. It is more likely to occur in the Grid corpus, which contains the highly confusable spoken alphabetic letters, than in corpora such as CRM (Bolia *et al.*, 2000; Brungart, 2001), which used a restricted range of color keywords in the equivalent position. In support of this notion, listeners make 5.2 times as many errors in reporting spoken letters not present as in reporting digits, twice as many as would be expected on the basis of the relative number of response alternatives. Consequently, the proportion of keywords present in neither target nor masker cannot be seen as a reliable measure of energetic masking in the competing talker situation.

An alternative approach to quantifying the extent of energetic masking in experiment 2 is to extrapolate from the results of the pure energetic masking conditions of experiment 1. The technique is based on a glimpsing model (Cooke, 2006), which assumes that intelligibility in a pure energetic masking situation is a function of both prior speech knowledge and the availability of glimpses of the target speech in regions not dominated by the masker. For each listener group, prior speech knowledge is the same in both experiments. Consequently, it is possible to use the pure energetic masking conditions of experiment 1 to estimate the relationship between intelligibility and the proportion of the spectrotemporal plane glimpsed, and to use this relationship to quantify energetic masking effects in experiment 2 by measuring the glimpse proportion at each TMR.

One potential problem with the use of the glimpse proportion metric is that it ignores the distributional characteristics of glimpses in the time-frequency plane. One would expect different glimpse distributions to result from the stationary and competing talker maskers of experiments 1 and 2, although to some extent the distribution of foreground speech glimpses will be similar regardless of the masker type due to the relatively sparse concentration of energy in harmonics and formants. To determine whether glimpse proportion is a good predictor of intelligibility for the different glimpse distributions resulting from the maskers of experiments 1 and 2, a "missing-data" automatic speech recognition system modified to handle glimpses (Cooke *et al.*, 1994; 2001) was used to identify keywords in the stimuli of both experiments. The automatic speech recognizer (ASR) scores at the four SNRs of experiment 1 (measured as raw percentages) and the six TMRs of experiment 2 are shown in the upper panel of Fig. 8. Since only four noise levels were used in experiment 1, piecewise linear interpolation was used to estimate the glimpse-intelligibility relation across the entire range of glimpse proportions. Figure 8 shows that the ASR scores in the two experiments are very similar at the same glimpse proportions, being within 1 percentage point at those glimpse proportions where they overlap, apart from the low-

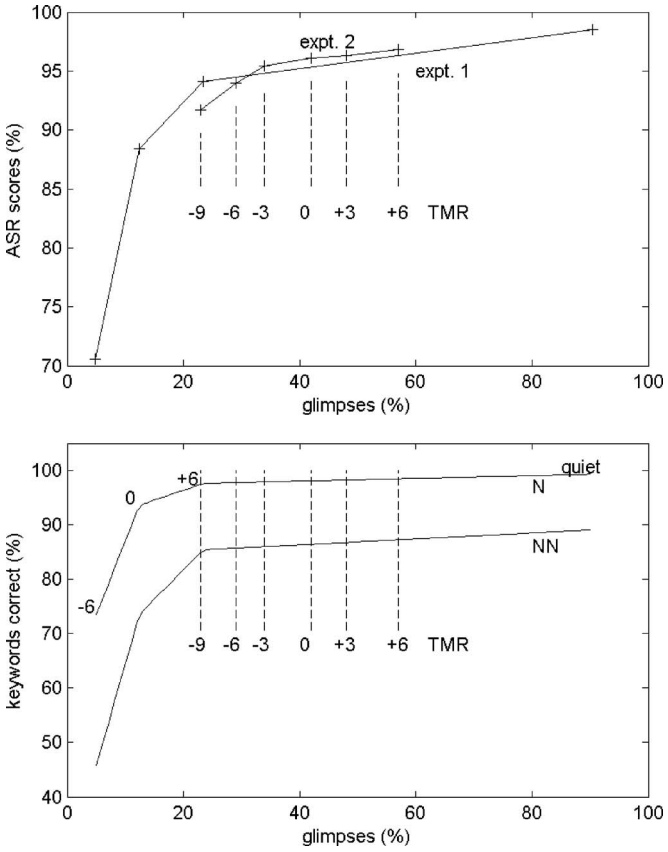


FIG. 8. Upper panel: Automatic speech recognition scores based on glimpse recognition for the stimuli of experiments 1 and 2. Rather than being expressed in terms of SNR (experiment 1) or TMR (experiment 2), recognition scores are plotted as a function of the mean glimpse percentage on which recognition was based. Lower panel: Solid lines (after pairwise linear interpolation between the four SNRs indicated) depict glimpse proportion vs intelligibility (scored as the mean identification rate of the letter and digit keywords) derived from experiment 1, for native and non-native listeners. Vertical lines indicate the measured glimpse percentages for the six TMR conditions of experiment 2.

est TMR value of −9 dB where the scores differ by 2.5%. Consequently, it appears that glimpse proportion alone can be used to mediate between intelligibility reductions due to energetic masking in the two experiments.

The lower panel of Fig. 8 plots intelligibility versus glimpse percentage for the native and non-native groups based on the results of experiment 1. To permit comparison of the two experiments, experiment 1 was rescored based on the two keywords (letter and number) identified in experiment 2. Also plotted in Fig. 8 are the glimpse percentages at the six TMRs used in experiment 2. The intelligibilities at the points where the vertical lines intersect the two curves are the estimates of the energetic masking effect for natives and non-natives in the competing talker experiment. These estimates suggest that the energetic masking effect of a competing talker is rather small, even at the lowest TMR used, and varies over a narrow range across TMRs. Indeed, the predicted keyword identification scores, if energetic masking were the only factor operating, range from 97.5 to 98.5 for natives and from 85.0 to 87.3 for non-natives.

A further estimate of the effect of energetic masking is provided by the ASR scores shown in the upper panel of Fig. 8. A comparison of the ASR scores with native listeners' identification rates in experiment 1 (Fig. 8, lower panel) suggests that the glimpsing model applied in the ASR system is about 3 percentage points worse than listeners on average. Consequently, the energetic masking effect suggested by ASR scores almost certainly overestimates the masking effect on listeners.

A promising alternative approach to isolating the energetic masking component was recently developed by Brungart *et al.* (2006). Like the approach described here, their technique uses a model of energetic masking to determine which portions of the target are audible, but instead of using glimpse counts or ASR scores based on glimpses, a resynthesis technique was used to reconstruct those parts of the signal which resist energetic masking, and the resulting signal is scored by listeners. They also found that energetic masking plays a relatively small role in the overall masking effect in the two-talker situation.

E. Summary and discussion

The native listener results of experiment 2 confirm the findings of Brungart (2001) and extend them to a more extensive and challenging corpus. More important, experiment 2 provides what we believe to be the first test of the "non-native cocktail party," where listeners had to identify keywords spoken by the target talker in the presence of a competing talker uttering very similar material. The speech-on-speech masking condition is known to provoke large amounts of informational masking. Here, native listeners scored between 10 and 30 percentage points better than non-natives. While it is difficult to determine precisely how much of this deficit was due to energetic masking, several analyses suggested that informational masking played by far the dominant role. Even after accounting for the effect of energetic masking on the two groups, there remains a native advantage of up to 10 percentage points due to reduced informational masking. In fact, analyses based on a computational model of energetic masking estimate a higher deficit of perhaps 20 percentage points. The results of experiment 2 suggest that non-native listeners are more adversely affected than natives by informational masking in multiple talker situations, and that the native advantage increases as the relative level of the masking talker increases.

Further acoustic analyses of the two-talker experiment demonstrated that both groups drew equivalent benefits from differences in the mean fundamental frequencies of the two simultaneous sentences. This is a very strong indication that the processes which lead to intelligibility improvements with increases in F0 difference precede the engagement of native-language-specific speech processes, since if the latter were involved in exploiting F0 differences (for example, by taking advantage of the more "visible" target speech harmonics which might results from F0 differences), one would expect to see a greater benefit for native talkers in conditions of large F0 difference.

Speech rate also played a part in the two-talker conditions. Non-natives identified substantially more keywords in slower utterances than in the more rapid utterances, while the

effect of speech rate was marginal for native listeners. As in the pure energetic masking case, this probably reflects the advantages of a slower information rate in a task which makes great attentional and cognitive demands.

The analysis of keyword confusions (Fig. 7), where listeners reported tokens from the masking source, are of particular interest. Non-natives found the same gender condition particularly confusing but, intriguingly, reported similar numbers of confusions as natives in the two positive TMRs of the same talker condition, and indeed reported fewer confusions than the natives at a TMR of 0 dB in that condition. Results in the different gender condition were intermediate between the other two conditions. To make sense of these findings, it is helpful to consider the cues which listeners might use to separate speakers in the three conditions.

First, in the same talker condition, cues such as differences in level and F0 as well as continuity of formants and harmonics are available. It seems that when the target is least masked, native and non-native listeners misallocate masker components to the target at about equal rates, suggesting that both groups are equally able to exploit level and F0 differences. In the same gender condition, additional cues are available. These fall into two classes: those that are language-universal (e.g., differences in voice quality, vocal tract length) and those which are language-specific (e.g., differences in accent and other speaker idiosyncrasies). Since the native group is best placed to take advantage of the latter type of cue [e.g., Ikeno and Hansen (2006) demonstrated that native listeners are better at detecting and classifying accents], it is not surprising that this group makes fewer background confusions. In the different gender condition, speaker differences are more extreme. Consequently, the native advantage seen in the same gender condition is somewhat reduced. It is unlikely that native listeners benefitted from the fact that multiple talkers were presented in a mixed order. Bradlow and Pisoni (1999) showed that native and non-native listeners drew similar advantages from having a consistent talker.

The differences discussed above apply to the situations where the target speaker is dominant. However, non-natives are much more likely to report keywords from the masker when the masker is dominant. In principle, there are at least two strategies that listeners could use to solve the two-talker problem. One would involve the use of cues such as F0 differences to "track" the separate talkers through time from the color keyword to the appropriate letter-digit combination. An alternative approach is to extract speaker "tags" from the color keyword and match these against the appropriate letter-digit keywords. Indeed, informal listening to two-talker utterances suggests that a more sophisticated form of tagging is possible, whereby listeners use "tags" from the color keyword belonging to the masker in order to eliminate letter-digit keywords produced by the masker. Such a strategy is the obvious one to use at lower TMRs when the masker is dominant.

If tracking were the dominant method, one might expect similar scores for the two listener groups, since both show similar benefits of F0 differences and tracking would seem to be a linguistically universal process. Since the difference in confusion scores in the negative TMR conditions is so great, it is more likely that a tagging strategy is dominant in this task. Given that in solving the two-talker problem listeners have to not only pick out the weaker target keywords but also to detect speaker cues (such as gender, mean F0, voice quality, accent) based on the color keyword in order to decide which letter-digit keywords to report, it is clear that in a tagging approach the non-native group has a double disadvantage because of their less-rich models of language variation.

Finally, caution is required in interpreting the target/masker confusions as wholly the result of informational masking. Since energetic masking plays a dual role in the two-talker case (color identification followed by letter-digit identification), it is difficult to ascribe all of the target/masker confusions to informational masking. It may well be that both letters and digits are audible and that some combination is reported so that the results appear to favor informational masking, but if the decision on which combination to report is based on a partially audible color keyword, then energetic masking is partly responsible for the results.

IV. GENERAL DISCUSSION

The two experiments reported in this paper demonstrate that in a task involving the identification of keywords in simple sentences spoken by native English talkers, Spanish listeners are more adversely affected than English listeners by increases in masker level for both stationary noise and competing speech maskers.

In principle, non-native listeners suffer both because of impoverished knowledge of the second language, and due to interference from their first language (Trubetzkoy, 1939; Strange, 1995). It is of interest to consider how these factors interact with effects of masking as listed in Fig. 1 to determine possible origins of the non-native deficit in noise.

In the case of energetic masking, native listeners presumably perform well due to their extensive experience of speech and in particular the effects of masking on the signal. Since there are multiple redundant cues to important phonetic distinctions such as voicing (Lisker, 1986), native listeners may have learned which cues survive in different noise conditions. On the other hand, non-native listeners have far less exposure to the second language and indeed may have virtually no experience of hearing the L2 in noisy conditions, so one might expect to see a differential effect of energetic masking, with non-native listeners suffering more in adverse conditions.

Of the multiple causes of informational masking, non-native listeners might be expected to suffer more than natives from target/masker misallocation. The accuracy of "sorting" audible components into speech hypotheses is likely to be higher for listeners with richer knowledge of the target language, which can be used to prevent false rejections and acceptances. For example, a listener whose knowledge of English is restricted to one specific accent may be less able to assign speech sounds from other accents to the target or

the background source (McAllister, 1997; Strange, 1995). Similarly, influences from the non-native L1 may create further difficulties in allocation of sound components, particularly at the phonemic level required to report the spoken letters in experiment 2. Cues in the target which do not conform to L1 categories may be wrongly allocated to the masker.

The other facets of IM listed in Fig. 1 may also be responsible for reduced performance amongst non-native listeners. The two-talker task creates a high cognitive load even for native listeners, and there is evidence that some aspects of processing a foreign language are slower than in processing a native language (Callan *et al.*, 2004; Mueller, 2005; Clahsen and Felser, 2006). Speech segregation processes requiring tracking and focus of attention may also affect non-natives adversely. For example, if listeners lack knowledge about English stress-timed rhythm, they are missing what may be a useful cue in segregating and tracking competing speech sources. Likewise, interference from L1 expectations of intonational contours might affect tracking.

V. CONCLUSIONS

The two experiments reported in this paper attempted to quantify the effect of energetic and informational masking on native and non-native listeners. English and Spanish listener groups identified keywords in simple sentences presented in stationary speech-shaped noise and in the presence of a competing talker speaking a similar sentence. Both conditions induced significantly more errors in the non-native group. A computer model suggested that the effect of energetic masking on the two groups could not account for the large native advantage in the competing talker conditions. It can be concluded that non-native listeners suffer a large performance deficit due to informational masking relative to native listeners.

Just as comparisons involving speech and nonspeech (e.g., music) sources can be used to distinguish the roles of general auditory from speech-specific processes, studies comparing listener populations with different native languages can be used to distinguish those parts of the speech interpretation process which make use of language-specific prior knowledge from those which are speech specific but language independent. In the current study, both groups derived equal benefit from differences in mean fundamental frequency between the target and masking talker, suggesting that segregation of speech using fundamental frequency cues has no language-specific component. Further studies will determine which other potential cues for understanding speech in noise act independently of prior linguistic knowledge.

ACKNOWLEDGMENTS

This work was supported by grants from the Spanish Ministry of Science and Technology, the Basque Government (9/UPV 00103.130-13578/2001) and the University of Sheffield Research Fund. We thank Jonny Laidler and Balakrishnan Kolluru for useful comments on the manuscript as well as three anonymous reviewers and Ann Bradlow for their insightful comments.

[1]In raw percentage terms, the native advantage increased for all three keywords (colors: 1% in quiet to 10% at −6 dB; letters: 16% to 31%; numbers: 4% to 25%).

[2]Within-condition learning effects: The mean difference between second and first half token presentations across conditions and listeners was −0.01 percentage points. Across-condition: The correlation between listeners' standardized (z) scores and condition order was insignificant (correlation =0.009; $p=0.89$).

Barker, J., and Cooke, M. P. (**2007**). "Modelling speaker intelligibility in noise," Speech Commun. **49**, 402–417.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Boersma, P., and Weenink, D. (**2005**). "Praat: Doing phonetic by computer," version 4.3.14 (computer program), http://www.praat.org. Last accessed 10 August 2007.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Bradlow, A. R., and Pisoni, D. B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074–2085.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. I. Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organisation of Sound* (MIT Press, Cambridge, MA).

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

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., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Callan, D. E., Jones, J. A., Callan, A. M., and Akahane-Yamada, R. (**2004**). "Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models," Neuroimage **22**, 1182–1194.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.

Clahsen, H., and Felser, S. (**2006**). "How native-like is non-native language processing?," Trends Cogn. Sci. **10**, 564–570.

Cooke, M. P. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Cooke, M. P., Barker, J., Cunningham, S. P., and Shao, X. (**2006**). "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am. **120**, 2421–2424.

Cooke, M. P., Green, P. D., and Crawford, M. D. (**1994**). "Handling missing data in speech recognition," *Proceedings of the Third International Conference on Spoken Language Processing*, Yokohama, Japan, pp. 1555–1558.

Cooke, M. P., Green, P. D., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and uncertain acoustic data," Speech Commun. **34**, 267–285.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," J. Acoust. Soc. Am. **116**, 3668–3678.

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.

Durlach, N. (**2006**). "Auditory masking: Need for improved conceptual structure," J. Acoust. Soc. Am. **120**, 1787–1790.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Florentine, M., Buus, S., Scharf, B., and Canevet, G. (**1984**). "Speech reception thresholds in noise for native and non-native listeners," J. Acoust. Soc. Am. **75**, s84 (abstract).
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.
Garcia Lecumberri, M. L., and Cooke, M. P. (**2006**). "Effect of masker type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. **119**, 2445–2454.
Gass, S. M. (**1997**). *Input, Interaction and the Second Language Learner* (Lawrence Erlbaum Associates, Mahwah, NJ).
Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility in adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.
Hazan, V., and Simpson, A. (**2000**). "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects," Lang Speech **43**, 273–294.
Ikeno, A., and Hansen, H. L. (**2006**). "Perceptual recognition cues in native English accent variation: Listener accent, perceived accent, and comprehension," *International Conference on Acoustics Speech and Signal Processing*, Toulouse, France, pp. 401–404.
Kahneman, D. (**1973**). *Attention and Effort* (Prentice-Hall, Englewood Cliffs, NJ).
Lisker, L. (**1986**). "Voicing in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Lang Speech **29**, 3–11.
Mayo, L. H., Florentine, M., and Buus, S. (**1997**). "Age of second-language acquisition and perception of speech in noise," J. Speech Lang. Hear. Res. **40**, 686–693.
McAllister, R. (**1997**). "Perceptual foreign accent: L2 users' comprehension ability," in *Second Language Speech: Structure and Process*, edited by A. James and J. Leather (Mouton de Gruyter, New York).
Meador, D., Flege, J. E., and MacKay, I. R. (**2000**). "Factors affecting the recognition of words in a second language," Bilingualism: Lang. Cognit. **3**, 55–67.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Mueller, J. L. (**2005**). "Electrophysiological correlates of second language processing," Second Lang. Res. **21**, 152–174.
Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (**1994**). "On the perceptual organization of speech," Psychol. Rev. **101**, 129–156.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., and Abrams, H. B. (**2006**). "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," Appl. Psycholinguist. **27**, 465–485.
Simpson, S., and Cooke, M. P. (**2005**). "Consonant identification in N-talker babble is a non-monotonic function of N," J. Acoust. Soc. Am. **118**, 2775–2778.
Strange, W. (**1995**). *Speech Perception and Linguistic Experience* (York, Timonium, MD).
Studebaker, G. A. (**1985**). "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Trubetzkoy, N. (**1939**). *Principles of Phonology (Grundzüge der Phonologie)* (University of California Press, Berkeley).
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.
van Wijngaarden, S. J., Bronkhorst, A. W., Houtgast, T., and Steeneken, H. J. M. (**2004**). "Using the Speech Transmission Index for predicting non-native speech intelligibility," J. Acoust. Soc. Am. **115**, 1281–1291.
van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (**2002**). "Quantifying the intelligibility of speech in noise for non-native listeners," J. Acoust. Soc. Am. **111**, 1906–1916.

Auditory-visual speech perception in normal-hearing and cochlear-implant listeners[a)]

Sheetal Desai, Ginger Stickney, and Fan-Gang Zeng[b)]
Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences and Otolaryngology – Head and Neck Surgery, 364 Medical Surgery II, University of California, Irvine, California 92697-1275

(Received 1 September 2006; revised 30 October 2007; accepted 31 October 2007)

The present study evaluated auditory-visual speech perception in cochlear-implant users as well as normal-hearing and simulated-implant controls to delineate relative contributions of sensory experience and cues. Auditory-only, visual-only, or auditory-visual speech perception was examined in the context of categorical perception, in which an animated face mouthing /ba/, /da/, or /ga/ was paired with synthesized phonemes from an 11-token auditory continuum. A three-alternative, forced-choice method was used to yield percent identification scores. Normal-hearing listeners showed sharp phoneme boundaries and strong reliance on the auditory cue, whereas actual and simulated implant listeners showed much weaker categorical perception but stronger dependence on the visual cue. The implant users were able to integrate both congruent and incongruent acoustic and optical cues to derive relatively weak but significant auditory-visual integration. This auditory-visual integration was correlated with the duration of the implant experience but not the duration of deafness. Compared with the actual implant performance, acoustic simulations of the cochlear implant could predict the auditory-only performance but not the auditory-visual integration. These results suggest that both altered sensory experience and improvised acoustic cues contribute to the auditory-visual speech perception in cochlear-implant users.
 [DOI: 10.1121/1.2816573]

PACS number(s): 43.71.Ky, 43.71.Es, 43.66.Ts [PEI] Pages: 428–440

I. INTRODUCTION

Multisensory integration provides a natural and important means for communication. The benefit of integrating auditory and visual (AV) cues in speech perception has been well documented, particularly in difficult listening situations and for hearing-impaired listeners (Sumby and Pollack, 1954; Erber, 1972; Binnie *et al.*, 1974; Dodd, 1977; Summerfield, 1979; Easton and Basala, 1982; Walden *et al.*, 1993; Grant *et al.*, 1998; Massaro, 1998). The benefit derived from speechreading can be substantial, allowing unintelligible speech to become comprehensive, or even exceeding the benefit derived from the use of assistive listening devices, counseling, or training (Sumby and Pollack, 1954; Walden *et al.*, 1981; Montgomery *et al.*, 1984; Grant and Braida, 1991; Grant and Walden, 1996). Two fundamental questions arise naturally concerning this AV integration in speech recognition: (1) what acoustic and optical cues are integrated? and (2) how and where are they integrated in the brain? (Rosen *et al.*, 1981; Braida, 1991; Massaro and Cohen, 2000; Bernstein *et al.*, 2002; De Gelder and Bertelson, 2003).

Acoustic and optical cues can be complementary to enhance speech perception. On the one hand, some speech sounds can be more easily distinguished in the visual modality than in the auditory modality, e.g., the bilabial /ba/ versus the alveolar /da/ stop consonant (Binnie *et al.*, 1974; Dodd, 1977; Summerfield, 1979; Walden *et al.*, 1990). On the other hand, other speech sounds, called visemes, such as /b/, /p/, and /m/, rely on the acoustic cues to differentiate from each other because they are visually indistinguishable (Fisher, 1968; Binnie *et al.*, 1974). While normal-hearing listeners have little trouble doing so, hearing-impaired listeners, including cochlear-implant listeners, often have great difficulty differentiating these phonemes due to reduced auditory temporal and spectral resolution (Miller and Nicely, 1955; Turner *et al.*, 1997; Munson *et al.*, 2003).

Because the optical cues provided by a speaker's facial and lip movements are not affected by the presence of noise, they are particularly useful at relatively low signal-to-noise ratios (SNRs). Sumby and Pollack (1954) demonstrated robust and relatively constant visual contribution to AV speech perception at SNRs over a 30-dB range. A recent study, however, showed the maximal integration efficiency at an intermediate SNR of about −12 dB (Ross *et al.*, 2007).

There is also evidence that different acoustic cues contribute differently to the amount of auditory and visual integration. When presented acoustically, voice pitch is virtually unintelligible but can significantly improve speechreading. The reason for this improvement is that the voice pitch cue provides important segmental and suprasegmental information that is usually invisible (Rosen *et al.*, 1981; Grant, 1987). Similarly, the temporal wave form envelope cue im-

[a)]Portions of this work were presented at the Conference on Implantable Auditory Prostheses, Asilomar Conference Center, Pacific Grove, California, 2003.

[b)]Author to whom correspondence should be addressed. Presently at: University of California, Irvine, 364 Med Surge II, Irvine, CA 92697. Electronic mail: fzeng@uci.edu

proves speechreading as long as the periodicity information is included (50–500 Hz) (Rosen, 1992). If only the envelope information (<25 Hz) is included, then this temporal envelope cue produces little or no effect on speechreading (Grant *et al.*, 1991; 1994). Overall, these studies suggest that it is not necessary to provide accurate information in both modalities, rather complementary acoustic and optical cues are sufficient to support high-level AV speech perception (e.g., Van Tasell *et al.*, 1987).

To understand how and where these acoustic and optical cues are integrated in the brain, researchers have used incongruent cues from auditory and visual modalities (e.g., Calvert and Campbell, 2003; De Gelder and Bertelson, 2003; van Wassenhove *et al.*, 2007). A compelling example showing interactions between incongruent auditory and visual cues in speech perception is the McGurk effect (McGurk and MacDonald, 1976). The McGurk effect is evoked by dubbing the audio recording of one sound (e.g., /ba/) onto the visual recording of a different sound (e.g., /ga/), obligating many listeners to report hearing an illusive sound (e.g., /da/ in this case). The McGurk effect has been extended to sentences, different languages, children, hearing-impaired listeners, and special patient populations (Green *et al.*, 1991; Sams *et al.*, 1998; Cienkowski and Carney, 2002; Burnham and Dodd, 2004). It is believed that AV speech integration occurs in a relatively early, prelexical integration stage (e.g., Calvert *et al.*, 1997; Reale *et al.*, 2007).

Recently, there has been an intensified interest in using the cochlear implant to study AV speech perception and integration. There are at least two reasons for this intensified interest. First, because the present implant extracts and delivers only the temporal envelope cue, lacking access to fine structure including the low-frequency voice pitch that is typically accessible to a hearing aid user, the implant users are particularly susceptible to noise (e.g., Stickney *et al.*, 2004; Kong *et al.*, 2005; Zeng *et al.*, 2005). The optical cue, when available, is essentially unaffected by noise. Therefore, the implant users rely more than normal listeners on the visual cue, forcing them to become not only better speechreaders but also better multisensory integrators (Goh *et al.*, 2001; Clark, 2003; Schorr *et al.*, 2005; Rouger *et al.*, 2007). Indeed, some cochlear-implant users can integrate AV cues to increase the functional SNR in noise (Lachs *et al.*, 2001; Bergeson *et al.*, 2005; Hay-McCutcheon *et al.*, 2005; Moody-Antonio *et al.*, 2005).

Second, the dramatic auditory experience and intervention with the cochlear implant provide a unique tool to study brain plasticity in multiple ways. For example, Schorr *et al.* (2005) demonstrated a critical period in developing AV integration, with the critical age of implantation being at about 2.5 years old. On the other hand, brain imaging studies have shown a profound cortical reorganization in cochlear implant users, with good users being able to recruit a larger cortical area, even the visual cortex, than poor users to perform an auditory task (Giraud *et al.*, 2001b; Lee *et al.*, 2001; Doucet *et al.*, 2006).

At present it remains unclear how much the cochlear-implant users can integrate auditory and visual information and whether this integration is related to stimulus and subject variables. The primary goal of the present study was to address the following two questions: (1) Do postlinguistically deafened persons fitted with a cochlear implant really integrate auditory and visual information? (2) How will altered stimuli and sensory experience affect AV integration? We first quantified the degree of AV integration by measuring performance in normal-hearing listeners, actual implant listeners, and simulated-implant listeners. We then delineated the relative contributions of stimulus and subject variables to AV integration by relating the degree of the AV integration to the duration of deafness and the duration of the implant experience.

II. METHODS

A. Subjects

1. Normal-hearing listeners

A total of 14 young, normal-hearing listeners participated in this study. All subjects were native English speakers with reported normal hearing. These young subjects ranged in age from 18 to 36 years. Subjects reported normal or corrected-to-normal vision.

Because of the large age difference between normal-hearing and cochlear-implant listeners as well as the known cognitive differences in sensory and cross-modality processing (Walden *et al.*, 1993; Gordon-Salant and Fitzgibbons, 1997; Humes, 2002; Hay-McCutcheon *et al.*, 2005), three elderly, nearly normal-hearing listeners (average age=77) were recruited to evaluate whether age is a significant factor in the present study. Pure-tone averages (across 500, 1000, and 2000 Hz) of two subjects were below 15 dB HL, and the third subject had a pure-tone average of 22 dB HL. Pure-tone averages for all three subjects were taken from the right ear. These elderly subjects reported normal or corrected-to-normal vision. Because of time limitation, they only participated in the experiment with the original unprocessed stimuli.

2. Cochlear-implant listeners

A total of eight postlingually deafened, adult cochlear-implant listeners were evaluated in this experiment. These subjects were recruited locally from the Southern California area. They had a mean age of 66 years old, duration of deafness of 18 years, and >1 year of experience with their device. Table I shows additional information on the individual cochlear-implant listeners evaluated in this study, including consonant and vowel identification scores. Significant correlation was observed between duration of deafness and consonant ($r=-0.96$) and vowel ($r=-0.85$) recognition, as well as between implant experience and consonant ($r=-0.70$) and vowel ($r=-0.91$) recognition. All cochlear-implant listeners were native English speakers and were postlingually deafened. All cochlear-implant listeners reported normal or corrected-to-normal vision.[1]

TABLE I. Demographics of the cochlear-implant listeners, including age, implant type, etiology, duration of deafness, years of experience with the cochlear implant, and percent correct scores in consonant and vowel recognition. Normal-hearing listeners typically score >90% on these consonant and vowel tests (Hillenbrand *et al.*, 1995; Shannon *et al.*, 1999).

Subject	Age (years)	Implant	Etiology	Duration of deafness (years)	Implant Experience (years)	Percent consonants (%)	Percent vowels (%)
1	61	Nucleus 22	Genetic	9	5	72	59
2	78	Nucleus 24	Unknown	12	1	44	31
3	70	Nucleus 24	Unknown	11	3	54	51
4	80	Med-El	Genetic	49	1	8	25
5	58	Nucleus 24	Genetic	44	1	12	22
6	46	Nucleus 22	Trauma	1	11	71	79
7	68	Clarion II	Genetic	16	2	59	51
8	69	Nucleus 24	Virus	1	6	77	63

B. Stimuli

1. Unprocessed stimuli

The auditory stimuli were created using a web-based Klatt Synthesizer (1980), developed by Bunnell (1996) and colleagues at the Speech Research Laboratory, A.I. duPont Hospital for Children (1996). Eleven consonant-vowel (CV) tokens were synthesized to represent an auditory continuum along /ba/, /da/, and /ga/. The continuum was created by varying the starting frequency of the *F*2 or *F*3 formants in 200-Hz steps while keeping the other formant frequencies constant (see Table II). The formant frequencies for the /a/ sound paired with each consonant were also kept constant. Reference tokens for /ba/, /da/, and /ga/ are highlighted in gray. The formant values for these reference tokens were adopted from Turner and Robb (1987).

To test the effects of formant-transition-duration (i.e., from the onset of the consonant sound to the onset of the steady state /a/ sound), two continuums were created: one with a formant-transition-duration of 20 ms and a second with a formant-transition-duration of 40 ms. The fundamental frequency for the consonant sounds started at 150 Hz and changed after 20 or 40 ms (depending on the type of stimulus) and then decreased to 100 Hz for the vowel sound over the remaining duration.

The total duration of each CV stimulus token was kept constant at 305 ms for both 20- and 40-ms stimuli (i.e., 20 or 40 ms was allotted to the respective consonant sounds and the remaining portion of the 305 ms was designated as the vowel sound). The auditory stimuli were calibrated using the Bruel & Kjaer sound level meter (Model No. 2260). A calibration tone, created by using a 1000-Hz sinusoid matched to the same rms level as the synthesized speech sounds, was used to adjust the level of the auditory stimuli to a 70 dB SPL presentation level.

TABLE II. Starting formant frequencies for each consonant token as well as steady-state formant frequencies for the vowel (/a/).

Stimulus token		*F*1 (Hz)	*F*2 (Hz)	*F*3 (Hz)
/b/	**1**	**300**	**700**	**2800**
	2	300	900	2800
	3	300	1100	2800
	4	300	1300	2800
	5	300	1500	2800
/d/	**6**	**300**	**1700**	**2800**
	7	300	1700	2600
	8	300	1700	2400
	9	300	1700	2200
	10	300	1700	2000
/g/	**11**	**300**	**1700**	**1800**
/a/		720	1250	2500

Normal-hearing and cochlear-implant listeners were seated in a sound-attenuated booth during the experiments. Normal-hearing listeners listened to auditory stimuli monaurally through the right ear with Sennheiser HDA 200 headphones. Seven of the cochlear-implant listeners were presented with stimuli through a direct audio input connection to their speech processor and one cochlear-implant listener was presented with stimuli through a speaker because her ear-level device did not allow a direct audio connection.

Visual stimuli from an animated face ("Baldi"), which corresponded to the /ba/, /da/, and /ga/ sounds, were created using the Center for Spoken Language Understanding (CSLU) Speech Toolkit (Barnard *et al.*, 2000; Massaro *et al.* 2000). The animated face was temporally aligned with each auditory stimulus token to represent the initial consonant position, the transition (20 or 40 ms), and the final vowel position. The computer monitor window displaying the animated face was modified so that the lips were 2 in. in width and 1 in. in height.

The synthetic sound and the animated face, instead of a natural sound and a human face, were used for the following two reasons. First, they can rule out a possible confounding factor of idiosyncratic acoustic and optical cues in a small set of stimuli as used in the present study. Second, the "Baldi" program allowed accurate temporal alignment between the congruent and incongruent acoustic and optical cues. An apparent weakness of using these synthetic stimuli was their relatively weaker signal strength compared with natural stimuli, because the synthetic stimuli were limited to the number of variables in the models, resulting in lower accuracy and poorer resolution than the natural stimuli (Massaro *et al.*, 2000). This weak signal strength could have contributed to relatively low-level AV integration (e.g., the McGurk effect) found in the present study.

2. Four- and eight-channel processed stimuli

The cochlear-implant simulation was generated using an algorithm developed by Shannon *et al.* (1995). The original stimuli from the 11-token continuum were band-pass filtered (sixth-order elliptical IIR filters) by either four or eight bands using the Greenwood map (Greenwood, 1990). The envelope from each band was extracted through full-wave rectification followed by a 500-Hz low-pass filter (first-order Bessel IIR filter). The envelope was then used to modulate a sinusoidal carrier set to the center frequency of the narrowband. The outputs of the narrowband signals were combined to create the four- or eight-channel cochlear-implant simulations.

C. Procedures

We adopted a classical categorical perception paradigm similar to that used by Walden *et al.* (1990) in hearing-impaired listeners and by Clark and colleagues (Clark, 2003) in prelingually deafened pediatric cochlear-implant users. A three-alternative-forced-choice procedure was used in three experimental conditions: auditory-alone (*A*), visual-alone (*V*), and AV (AV). In the AV condition, an animated face mouthing /ba/, /da/, or /ga/ was paired with each speech sound from the auditory continuum, creating both congruent and incongruent AV combinations. The *V* condition evaluated the subjects' ability to lipread the mouthed phonemes. Each condition was tested with both 20- and 40-ms formant-transition-durations. Subjects were given both verbal and written instructions to click buttons labeled /ba/, /da/, and /ga/ on a computer-based interface that corresponded to what they thought they perceived. The three-alternative forced-choice protocol allowed quantification of perceptual boundaries but could be a liability if the subject perceived a sound that was different from the three choices, e.g., a /bg/ response to an auditory /g/ and visual /b/ stimuli (McGurk and MacDonald, 1976).

Prior to the test session, practice sessions were given with feedback for the *A*, *V*, and AV conditions. For the *A* and AV conditions, the reference tokens were presented a total of 20 times, resulting in 60 presentations for one complete practice session. For the *V* practice sessions, each reference token was mouthed a total of ten times, resulting in 30 presentations for one complete practice session. Normal-hearing listeners were required to achieve 80% correct identification of all three unprocessed CV pairs in the *A* practice session (at a 40-ms formant-transition-duration) to continue with the experiment. All of the normal-hearing listeners met this criterion.

Following the practice sessions, the test sessions were given without feedback. The test order of these conditions was randomized for each subject. For the *A* condition, each of the 11 continuum tokens was presented randomly a total of 20 times, resulting in 220 presentations for one complete test session. For the AV condition, each of the 11 continuum tokens was paired with a /ba/, /da/, and /ga/ face and were presented randomly a total of ten times, resulting in 330 presentations. For the *V* condition, each /ba/, /da/, and /ga/ face was presented randomly a total of 20 times, resulting in 60 presentations for one complete test session. The scores were calculated as the number of tokens correctly identified (in % correct) or as a distribution in response to each of the three tokens (in % identification).

D. Data analysis

The phonemic boundaries were estimated with a four-parameter sigmoidal function fit to each function:[2]

$$f(x) = y_0 + \frac{a}{1 + e^{-(x - x_0/b)}}$$

where y_0=minimum y value; x_0=x value corresponding to peak y; a=50% point; and b=slope.

A repeated-measures analysis of variance (ANOVA) was performed on both within- and between-subjects factors to examine the main effects for each condition and stimulus. Chance performance was considered to be 33% identification for consonants. If an interaction was found between any of the main effects, a simple effects analysis was carried out followed by planned comparisons with a Bonferroni correction between the conditions or stimuli in question. A significant effect in a Bonferroni analysis was calculated by dividing the p value of 0.05 by $n-1$ where n was the number of conditions or stimuli in question.

III. RESULTS

A. Categorical perception

1. Normal-hearing listeners

Since no main effect was found for the formant-transition-duration factor, the data for all conditions were averaged across 20- and 40-ms stimuli. The left panels in Fig. 1 show perceptual labeling (% identification) by young normal-hearing listeners, who listened to the unprocessed auditory continuum along /ba/-/da/-/ga/. The top panel shows the results for the *A* condition, while the bottom three panels show results for the auditory continuum simultaneously presented with a visual /ba/ (second panel), a visual /da/ (third panel), or a visual /ga/ stimulus (bottom panel). Open circles, filled squares, and open triangles represent the subjects' percent identification for /ba/, /da/, and /ga/, respectively. The curves represent the best fit of the four-parameter sigmoidal function to the data. The left vertical dashed line in each panel represents the estimated phonemic boundary between /ba/ and /da/, while the right dashed line represents the boundary between /da/ and /ga/. The asterisk symbol in the visual /ga/ condition (bottom panel) placed above the /da/ response (filled square) at the token 1 location represents one of the commonly observed McGurk effects (auditory /ba/ + visual/ga/ = perceived/da/).

The *A* condition (left-top panel) shows the classical pattern of categorical perception, with a significant main effect being observed for the three responses [$F_{(2,12)}=5.2$; $p<0.05$]. *Post hoc* analysis with a Bonferroni correction revealed that tokens 1–3 were primarily labeled as /ba/ [$F_{(10,4)}=1805.3$; $p<0.017$], tokens 4–8 labeled as /da/ [$F_{(10,4)}=551.7$; $p<0.017$], and tokens 9–11 labeled as /ga/ [$F_{(10,4)}=963.3$; $p<0.017$]. The estimated /ba-da/ boundary

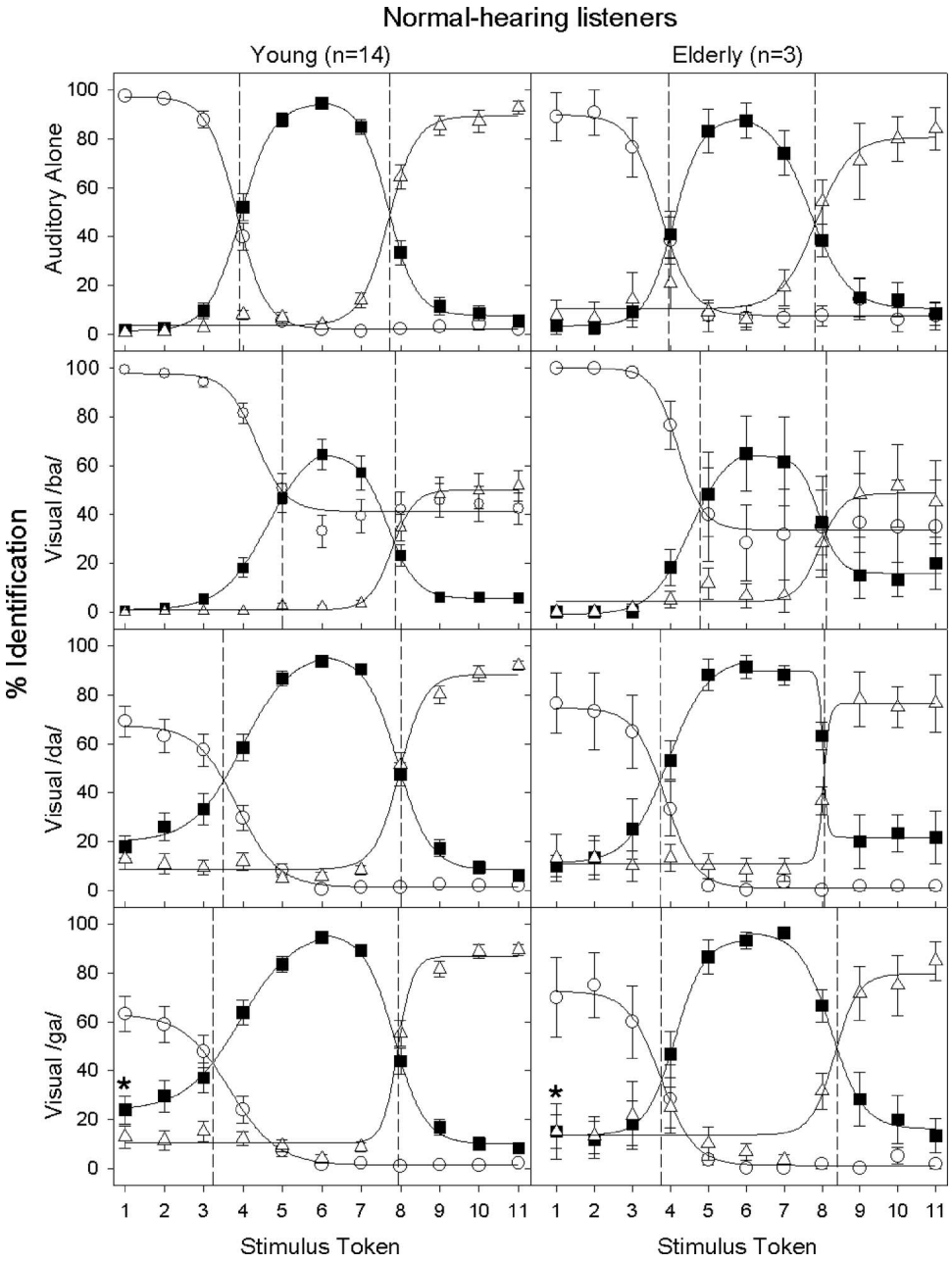


FIG. 1. Percent identification as a function of consonant continua in young (left column) and elderly (right column) normal-hearing listeners. The top panel shows the results for the auditory-alone continuum. The data for the three AV conditions are shown in separate panels (visual /b/: second row; visual /d/: third row; and visual /g/: bottom row). Open circles (○), filled squares (■), and open triangles (△) represent the percentage response to /b/, /d/, and /g/, respectively. Error bars represent the standard error of the mean. Sigmoidal four-parameter functions were fitted to the data to reveal /b/-/d/ and /d/-/g/ boundaries. Vertical dashed lines show where these boundaries occur along the continuum. An asterisk (*) denotes one of the commonly observed McGurk effects, i.e., when subjects responded /da/ when a visual /ga/ face was paired with the reference auditory /ba/ sound (token 1).

was located at about token 4 and the /da-ga/ boundary was at token 8.

The simultaneous presentation of the visual stimulus (bottom three panels) generally increased the subjects' overall percent identification towards the visual stimulus while decreasing the response to the incompatible auditory stimuli. For example, the visual /ba/ stimulus (second panel) increased the response to /ba/ from 0% to about 40% for tokens between 5 and 11, whereby the peak /da/ response decreased to about 60% and the peak /ga/ response decreased to about 50%. Compared with the effect of the visual /ba/ stimulus, the visual /da/ and /ga/ stimuli produced a similar but slightly smaller effect on the subjects' responses.

The simultaneous presentation of the visual /ba/ stimulus significantly shifted the /ba/-/da/ boundary but not the /da/-/ga/ boundary [$F_{(3,12)}=11.3$; $p<0.05$]. Compared to the *A* condition, the visual /ba/ shifted the /ba/-/da/ boundary rightward from the 3.9-token position to the 5-token position, indicating that subjects tended to respond /ba/ to more stimulus tokens when the corresponding visual cue was present. This shift in the /ba/-/da/ boundary was not significant when the *A* condition was compared to the visual /da/ and /ga/ conditions. Additionally, none of the visual stimuli produced any significant shift for the /da/-/ga/ boundary.

The right panels of Fig. 1 show the results obtained under the same conditions from three elderly listeners. Except for the larger error bars, reflecting the smaller sample size ($n=3$) than the young normal-hearing listeners ($n=14$), the elderly listeners produced essentially the same results. For example, the categorical boundary was at about token 4 between /ba/ and /da/, and at about 8 between /da/ and /ga/. Similarly, the visual /ba/ shifted the /ba/-/da/ boundary rightward to about token 5, the visual /da/ and /ga/ slightly shifted

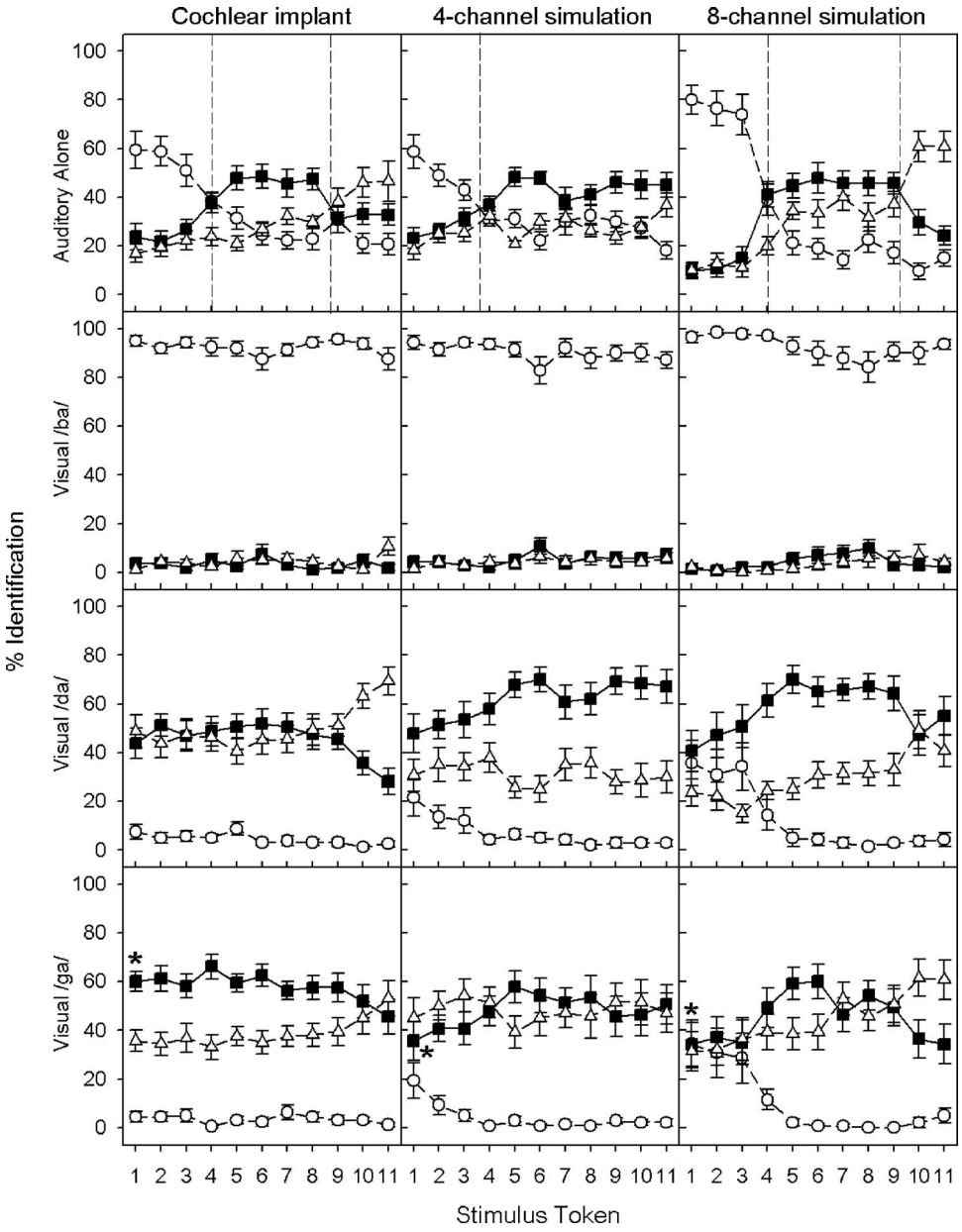


FIG. 2. Percent identification as a function of consonant continuum in the cochlear-implant listeners (first column), and the four-channel (middle column) and eight-channel (right column) simulated implant listeners. The top panels show the results for the auditory-alone continuum. The data for the three AV conditions are shown in separate panels (visual /b/: second row; visual /d/: third row; and visual /g/: bottom row). Open circles, filled squares, and open triangles represent the percentage response to /b/, /d/, and /g/, respectively. Error bars represent the standard error of the mean. An asterisk (*) denotes the McGurk effect, i.e., when subjects responded /da/ when a visual /ga/ face was paired with the reference auditory /ba/ sound (token 1).

the /ba/-/da/ boundary leftward, whereas the visual stimuli had minimal effect on the /da/-/ga/ boundary. This small set of data from elderly listeners suggests that age *per se* plays a negligible role in the present AV tasks (see also Walden *et al.*, 1993; Helfer, 1998; Cienkowski and Carney, 2002; Sommers *et al.*, 2005).

2. Cochlear-implant listeners

Figure 2 shows perceptual labeling by cochlear-implant listeners (left column). The figure configuration and symbol conventions are the same as Fig. 1. Several differences are apparent between the normal and implant listeners. First, cochlear-implant listeners produced an insignificant main effect for the three responses in the *A* condition [$F_{(2,6)}=3.9$; $p>0.05$]. Because of a significant interaction between responses and stimulus tokens [$F_{(20,140)}=4.2$; $p<0.05$], *post hoc* analysis with a Bonferroni correction was conducted to show that only tokens from 1 to 3 produced significantly higher responses to /ba/ than to /da/ or /ga/ ($p<0.025$). Second, different from the nearly perfect response to the reference tokens by the normal-hearing listeners, the highest /ba/ response obtained by the implant listeners was about 60% to tokens 1–2, followed by 50% /da/ to tokens from 5–8 and 30% /ga/ to tokens 10–11. Although their overall categorical responses were much weaker, the implant listeners produced a proper /ba/-/da/ boundary at token 4 and a slightly rightward shifted /da/-/ga/ boundary at token 9.

Third, a totally different pattern emerged for the visual effect on categorical perception in cochlear-implant listeners than in normal-hearing listeners. Independent of the auditory stimulus, the cochlear-implant listeners showed an almost total bias toward the visual /ba/ cue (second panel), and to a lesser extent, toward the visual /da/ (third panel) and /ga/ cue (bottom panel). Except for the separate /da/ and /ga/ labeling at high tokens (10 and 11) in the visual /da/ condition, the dominant visual cue wiped out the relatively weak categories that existed in the *A* condition. On the surface, the McGurk

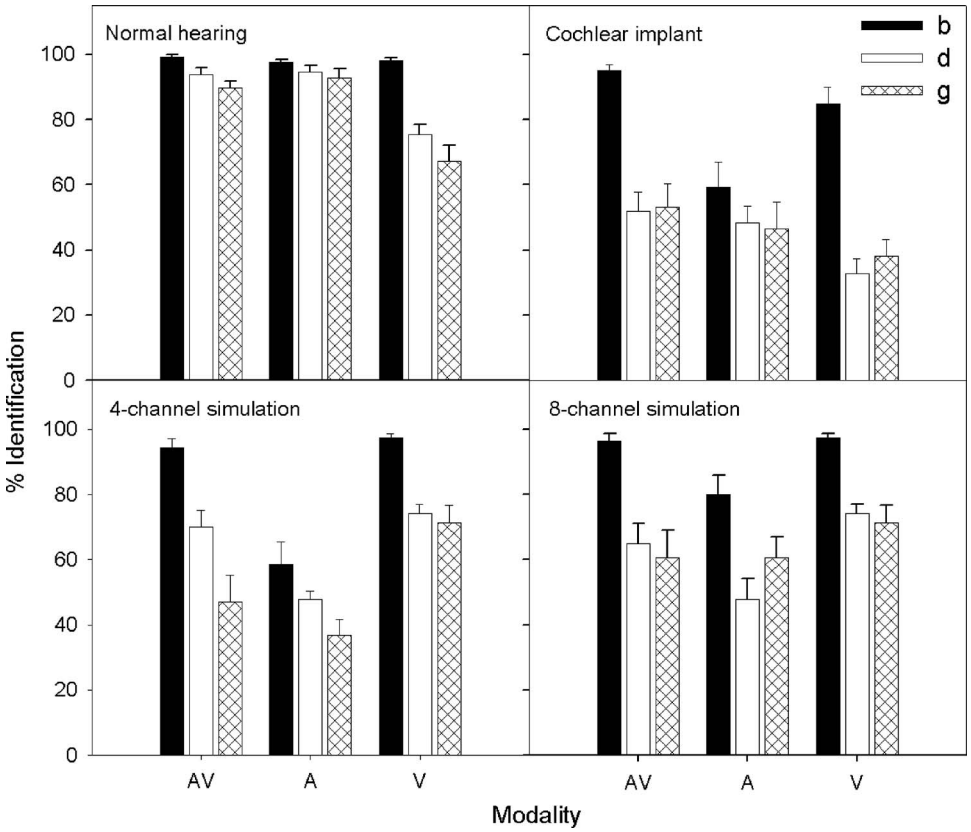


FIG. 3. Performance of all four groups of subjects in the congruent AV, *A*, and *V* conditions. Error bars represent the standard error of the mean.

effect (the asterisk symbol in the bottom panel) appeared to be much stronger in the implant listeners (60% labeling to the combined auditory /ba/ and visual /ga/ stimuli) than in the normal listeners (24%). However, note that the overall baseline response to /da/ was also much higher in implant listeners (e.g., 46% to the combined auditory /ga/, i.e., token 11, and visual /ga/ stimuli) than in normal listeners (8%). We shall return to this point in Sec. III.

3. Cochlear-implant simulations

Figure 2 also shows perceptual labeling by young normal-hearing listeners attending to four- and eight-channel cochlear-implant simulations (middle and right columns). Like cochlear-implant listeners, normal-hearing listeners presented with cochlear-implant simulations produced relatively weak categorical perception and showed a strong bias toward the visual cue, particularly to the visual /ba/. Moreover, the simulated implant listeners produced a characteristic, albeit relatively weak, categorical boundary for the /ba/-/da/ pair at the token 4 position and /da/-/ga/ boundary close to the token 9 position for the eight-channel simulation condition only. Finally, the simulated implant listeners produced a McGurk effect (~40%, represented by asterisk symbols in the middle and right bottom panels) that was greater than normal-hearing listeners attending to unprocessed stimuli (~20%) but smaller than the actual implant listeners (~60%). Does this mean that actual and simulated implant listeners are more susceptible to the McGurk illusion than normal-hearing listeners? The following sections analyze congruent and incongruent conditions in detail.

B. Congruent AV perception (AV benefit)

Figure 3 shows perceptual labeling results in response to *A*, *V*, and congruent AV stimuli in normal-hearing listeners, cochlear-implant listeners, and four- or eight-channel cochlear-implant simulations. The filled, unfilled, and hatched bars correspond to the percent identification of /b/, /d/, and /g/, respectively. In this case, only the reference tokens were included (i.e., token 1 for /ba/, token 6 for /da/, and token 11 for /ga/).

We also used two measures to define the amount of relative AV benefit that can be derived relative to either the *A* condition or the *V* condition. The first relative AV benefit measures the amount of phoneme recognition improvement relative to the *A* baseline and is defined as (AV-*A*)/(100-*A*) with AV and *A* scores expressed as percent scores (Grant and Seitz, 1998). Similarly, the second AV benefit measures the improvement relative to the *V* baseline and is defined as (AV-*V*)/(100-*V*) with AV and *V* expressed as percent scores. These relative measures, as opposed to the absolute differences (i.e., AV-*A* or AV-*V*), were adopted to avoid potential biases because a wide range of *A* and *V* scores occurred in the present diverse group of subjects. For example, a bias occurs because a high *A* score will certainly produce a low AV benefit score with the absolute measure but not necessarily with the relative measure.

1. Normal-hearing listeners

Normal-hearing listeners demonstrated a significant main effect on modality [$F_{(2,12)}=9.0$; $p<0.05$], showing nearly perfect performance for the identification of the refer-

ence stimuli in the *A* and AV conditions, but lower and more variable performance in the *V* condition. On average, the proper labeling with the original unprocessed stimuli was 95% for the *A* condition, and 94% for the AV condition, but only 80% for the *V* condition.

Normal-hearing listeners also demonstrated a significant main effect on phoneme identification [$F_{(2,12)}=23.0$; $p<0.05$], showing an averaged labeling of /ba/ 98% of the time, /da/ 88% of the time, and /ga/ 83% of the time. There was a significant interaction between modality and consonant identification [$F_{(4,10)}=7.0$; $p<0.05$]. The interaction was due to the fact that the normal-hearing listeners confuse the phonemes /da/ and /ga/ only in the *V* condition. This result is not too surprising because the voiced bilabial stop consonant, /b/, was more visually salient than either /d/ or /g/.

Averaged across all three phonemes, the normal-hearing listeners produced an AV benefit score of −0.14 relative to the *A* baseline and +0.71 relative to the *V* baseline. The negative score was due to the fact that the 94% AV percent score was slightly lower than the 95% *A*-alone score. These relative AV benefit scores suggest that the relative signal strength in the auditory domain is strong, producing a ceiling effect.

2. Cochlear-implant listeners

In contrast to the normal control, the implant listeners produced much lower overall percent correct scores in all conditions. Average performance collapsed across all conditions was 67% for the AV condition, 51% for the *A* condition, and 52% for the *V* condition. A significant main effect was found for modality [$F(2,6)=15.7$; $p<0.05$] in cochlear-implant listeners. A significant main effect was also found for phoneme identification [$F_{(2,6)}=15.7$; $p<0.05$]. Average performance collapsed across all modalities was significantly higher for /b/ (80%), but no difference was shown between /d/ (44%) and /g/ (46%). A significant interaction was found between modality and phoneme identification [$F_{(4,4)}=41.5$; $p<0.05$]. This interaction was attributed to the larger difference in performance between /b/ and /d/ in AV (43 percentage points) and *V* (52 percentage points) conditions, as compared to the auditory-only condition (11 percentage points).

The present cochlear-implant listeners were able to integrate the auditory and visual cues to significantly improve the *A* or the *V* performance by 15–16 percentage points. The corresponding relative AV benefit score was 0.30 relative to the *A* baseline and 0.29 relative to the *V* baseline.

3. Cochlear-implant simulations

First, a significant main effect was found for modality [$F_{(2,5)}=22.8$; $p<0.05$]. Like actual implant listeners, the simulated implant listeners appeared to benefit from the additional lipreading cue when presented with the degraded auditory stimuli. The performance was improved from 48% with *A* stimuli to 70% with AV stimuli in the four-channel condition, and from 63% to 74% in the eight-channel condition. However, unlike actual implant listeners, the *V* condition produced the highest performance of about 80% in both cochlear-implant simulations, suggesting no integration between the auditory and visual cues.

Second, a significant main effect was found for number of channels [$F_{(1,6)}=37.5$; $p<0.05$], with the eight-channel condition producing 72% performance and the four-channel condition producing significantly lower performance at 66%. A significant interaction between channel and modality was also observed [$F_{(2,5)}=15.6$; $p<0.05$], reflecting lower *A* performance for the four-channel condition than the eight-channel condition. This result was expected because the greater spectral resolution associated with the eight-channel would produce better *A* performance than the four-channel condition.

Third, a significant main effect was found for phoneme identification [$F_{(2,5)}=24.3$; $p<0.05$]. In the four-channel condition, the mean performance was 83%, 64%, and 52% for /b/, /d/, and /g/, respectively. In the eight-channel condition, the mean performance was 91%, 62%, and 64% for /b/, /d/, and /g/, respectively. Overall, the simulated implant listeners performed poorer than the normal-hearing listeners when listening to the unprocessed stimuli, but more similarly to the actual cochlear-implant listeners.

Finally, the simulated four-channel implant listeners produced the following percent correct scores: 50% for *A*, 81% for *V*, and 68% for AV; the simulated eight-channel implant listeners produced the following percent correct scores: 63% for *A*, 81% for *V*, and 74% for AV. The corresponding relative AV benefit score was 0.36 relative to the *A* baseline and −0.70 relative to the *V* baseline in the four-channel simulation, and was 0.30 and −0.37, respectively, in the eight-channel simulation. Similar to the actual implant listeners, the simulated implant listeners produced similar auditory-only percent scores and AV benefit scores relative to the auditory baseline. In contrast to the actual implant listeners, the simulated listeners produced higher visual-only scores and negative AV benefit scores relative to the visual baseline.

C. Incongruent AV perception (McGurk effect)

Figure 4 shows perceptual labeling in response to the incongruent condition (the reference auditory /ba/ paired with the visual /ga/) in normal-hearing listeners listening to the unprocessed auditory stimuli, cochlear-implant listeners, four-channel, and eight-channel cochlear-implant simulations. The filled, unfilled, and hatched bars represent the percentage of the /ba/, /da/, and /ga/ responses, respectively. While a /ba/ response indicates a bias towards the auditory cue and a /ga/ response indicates a bias towards the visual cue, a /da/ response represents integration of the auditory and visual cues, also known as the McGurk effect in this case.

We used two different measures to estimate the size of the McGurk effect. First, hearing-impaired listeners are prone to numerous auditory errors, with some of the errors being consistent with McGurk-like responses, making it difficult to distinguish between these auditory errors and a true McGurk effect (Grant and Seitz, 1998). To overcome this auditory error problem, we followed Grant and Seitz's corrective method (Grant and Seitz, 1998) and adjusted the size of the McGurk effect by subtracting the subject's /da/ response to visual /ga/ and auditory /ba/ (i.e., the solid square data point above token 1 in the bottom panels of Figs. 1 and

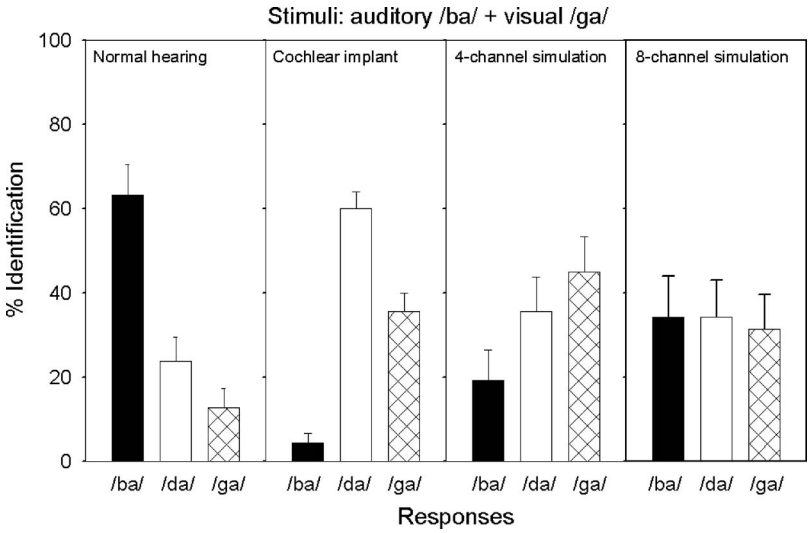


FIG. 4. Percent identification of /ba/, /da/, or /ga/ in all four groups of subjects for the incongruent AV condition, in which an auditory /ba/ cue was paired with a visual /ga/ cue. Error bars represent the standard error of the mean.

2) from the subject's averaged /da/ response to the auditory /ba/ token (i.e., the solid square data point above token 1 in the top panels of Figs. 1 and 2) and the auditory /ga/ token (i.e., the solid square data point above token 11 in the top panels of Figs. 1 and 2). We will refer to this measure as the error-adjusted McGurk effect.

Second, there may be a general subject response bias. If the bias happens to be a McGurk-type response, it will inflate the size of the McGurk effect. To overcome this response bias problem, we adjusted the size of McGurk effect by subtracting the subject's /da/ response to the incongruent stimulus from the subject's /da/ response to the congruent /ga/ stimulus (i.e., the solid square data point above token 11 in the bottom panels of Figs. 1 and 2). We will refer to this measure as the bias-adjusted McGurk effect.

1. Normal-hearing listeners

A significant effect was found for consonant identification [$F_{(2,12)}=6.5$; $p<0.05$]. The normal-hearing listeners responded 63% of the time to /ba/, 24% to /da/, and 13% to /ga/, implicating a strong auditory bias. The 24% McGurk effect was relatively weak, possibly due to the weak signal strength of the present synthetic stimuli. Interestingly, six out of 14 normal-hearing listeners did not experience any McGurk effect, responding to the auditory /ba/ cue 100% of the time. The remaining eight subjects who experienced the McGurk effect had a mean /da/ response of 42% (SD=23, with a range from 15% to 85%). The unadjusted 24% McGurk effect was significantly higher than the averaged 3% /da/ response to the auditory /ba/ and /ga/ tokens (paired-t test, $p<0.05$), resulting in an error-adjusted McGurk effect of 21%. On the other hand, the normal subjects produced an 8% /da/ response to the congruent /ga/ token, which was significantly lower than the 24% unadjusted McGurk effect (paired-t test, $p<0.05$). The bias-adjusted McGurk effect size is therefore 16%.

2. Cochlear-implant listeners

A significant effect was found for consonant identification [$F_{(2,6)}=54.1$; $p<0.05$]. The implant listeners responded 4% of the time to /ba/, 60% to /da/, and 36% to /ga/ with the incongruent auditory /ba/ and visual /ga/ stimulus. The implant response pattern was significantly different from the normal control pattern [$F_{(2,19)}=8.6$; $p<0.05$]. For example, the response to the auditory /ba/ cue was greatly reduced from 63% in the normal control to 4% in the implant listeners, suggesting that the auditory signal strength via a cochlear implant was much weaker than the visual signal strength.

On the surface, the McGurk effect appeared to be much stronger in implant listeners, producing a fused /da/ response 60% of the time, compared with the /da/ response 24% of the time in normal control. The 33% error-adjusted McGurk effect was significant (paired t-test, $p<0.05$) but the 14% bias-adjusted McGurk effect just missed the predefined significance criterion (paired t-test, $p=0.08$). Overall, the 33% error-adjusted and the 14% bias-adjusted McGurk effects in implant subjects were statistically indistinguishable from their 21% and 16% counterparts in the normal control (t-test with two-sample unequal variance; $p>0.2$). Overall, these adjusted measures suggest that although there is some evidence for integration between incongruent auditory and visual cues, but this integration is weak, if present at all, in postlingually deafened, adult implant listeners.

3. Cochlear-implant simulations

No significant main effect was found for consonant identification for either the four-channel [$F_{(2,5)}=0.9$; $p>0.05$] or eight-channel [$F_{(2,5)}=1.0$; $p>0.05$] condition. In the four-channel condition, the response was the least to the auditory /ba/ cue (19%), but was increasingly biased toward the fused /da/ cue (36%) and the visual /ga/ cue (45%). In the eight-channel condition, the response was evenly distributed across the auditory /ba/, the fused /da/, and the visual /ga/ cues at near chance performance. Neither adjusted McGurk effect showed any indication of integration between the incongruent auditory and visual cues in these simulated implant listeners.

IV. DISCUSSION

A. Categorical perception

The present results show sharp categories in young and elderly normal-hearing listeners, but greatly weakened categories in actual and simulated cochlear-implant listeners. This finding is consistent with previous studies on hearing-impaired listeners (Walden *et al.*, 1990) and pediatric cochlear-implant users (Clark, 2003) who also showed broader than normal boundaries in a similar categorical perception task. Together, these data suggest that categorical perception is affected by hearing loss or reduced spectral resolution, but not by age.

B. Congruent auditory and visual benefit

The present AV benefit in normal-hearing subjects was confounded by the ceiling effect. The actual implant data showed similar 0.30 AV benefit scores relative to either the auditory or the visual baseline. This relative AV benefit score was about half of the 0.67 AV benefit score obtained by hearing-impaired listeners in the Grant and Seitz study (Grant and Seitz, 1998). This discrepancy might be due to either the specific and limited set of stimuli, or the perceptual difference between hearing-impaired and cochlear-implant subjects, or both. The simulated implant data suggest that current acoustic simulation models of the cochlear implant (Shannon *et al.*, 1995) can simulate auditory perception of degraded acoustic stimuli but cannot simulate the *V* and AV perception in actual cochlear-implant users.

C. Incongruent auditory and visual integration

When incongruent acoustic and optic cues are present, listeners may be biased toward either auditory or visual cues. Consistent with previous studies (Easton and Basala, 1982; Massaro and Cohen, 1983; Bernstein *et al.*, 2000; Cienkowski and Carney, 2002; Clark, 2003; Schorr *et al.*, 2005), the present results show that normal-hearing listeners were biased toward the auditory cue (i.e., greater /ba/ response in Fig. 4) while cochlear-implant listeners were biased toward the visual cue (i.e., greater /da/ and /ga/ response in Fig. 4).

This bias appears to depend on the relative signal strength between acoustic and optical stimuli and is likely to influence the degree of integration (e.g., Green *et al.*, 1991; Green and Norrix, 1997). In the normal-hearing subjects, the optical signal strength from the animated "Baldi" was relatively weaker than the synthetic acoustic signal strength. In actual and simulated implant listeners, however, the same optical signal became relatively stronger than the improvised acoustic signal. While the unadjusted McGurk effect was much greater in implant subjects than normal subjects, this difference could be totally accounted for by the McGurk-like auditory error responses and response biases (Grant and Seitz, 1998). Overall, the present data showed that the adjusted McGurk effect was weak in normal-hearing and cochlear-implant subjects but totally absent in simulated-implant subjects.

D. Experience and performance

Typically before receiving their devices, postlingually deafened adult-implant users experience a period of deafness from several months to decades and during which they rely on lipreading. After implantation, users usually need several months to years to achieve asymptotic performance (e.g., Tyler and Summerfield, 1996; Skinner *et al.*, 2002). This unique experience may allow them to use the visual cue and integrate the auditory and visual cues to a greater extent than the normal-hearing listeners (Schorr *et al.*, 2005; Rouger *et al.*, 2007).

Recent brain imaging studies showed strong association between cortical plasticity and cochlear-implant performance: In general, good performance was correlated with the amount of overall cortical activation, not only in the auditory cortex but also in the visual cortex when using the implant only (Giraud *et al.*, 2001a; Lee *et al.*, 2001; Green *et al.*, 2005). Because good performance is correlated with the duration of implant experience, we would expect that both variables are correlated with AV integration. On the other hand, Doucet *et al.* (2006) found an intramodal reorganization in good implant performers but a profound cross-modality reorganization in poor performers, suggesting that the duration of deafness is correlated with the AV integration.

To address this experience and performance issue, we performed correlational analysis between two implant variables (duration of deafness and duration of implant usage) and seven implant performance measures (Table III). The implant performance included three direct measures in response to *A*, *V*, and AV stimuli and four derived measures: the AV benefit score relative to *A* baseline (AV_*A*), the AV benefit score relative to *V* baseline (AV_*V*), the error-adjusted McGurk effect (*M*_error), and the bias-adjusted McGurk effect (*M*_bias).

Consistent with previous studies (e.g., Blamey *et al.*, 1996; van Dijk *et al.*, 1999; Gomaa *et al.*, 2003), the duration of deafness is negatively correlated with the duration of the implant usage and the auditory-only performance. However, we found that the duration of deafness is not correlated with any other measures, including the visual-only performance. On the other hand, we found that the duration of implant experience is directly correlated with the *A* and AV performance, as well as the AV benefit relative to the visual baseline and the McGurk effect. Because visual-only performance is not correlated to any implant performance, the present data suggest, at least in the present postlingually deafened implant users, that implant experience, rather than auditory deprivation, determines the implant users' ability to integrate both congruent AV cues (i.e., the AV benefit) and incongruent AV cues (i.e., the McGurk effect).

V. CONCLUSIONS

Auditory-only, visual-only, and AV speech perception was conducted in normal-hearing listeners, postlingually deafened actual implant users, and acoustically simulated cochlear-implant listeners. Given the limitations of using the synthetic acoustic stimuli, the animated face, and the three-alternative, forced-choice method, we can reach the follow-

TABLE III. Correlational analysis between implant variables and implant performance in eight postlingually deafened cochlear-implant listeners. The implant variables included the duration of deafness (Deaf) and duration of implant usage (CI). The implant performance included three direct measures in response to congruent auditory and visual cues: the auditory-only score (*A*), the visual-only score (*V*), the AV score (AV), as well as four derived measures (see text for details): the AV benefit score relative to *A* baseline (AV_*A*), the AV benefit score relative to *V* baseline (AV_*V*), the error-adjusted McGurk effect (*M*_error), and the bias-adjusted McGurk effect (*M*_bias). Pearson correlation coefficient was used.

	Deaf	CI	A	V	AV	AV_*A*	AV_*V*	*M*_error	*M*_bias
Deaf	1	−0.67[a]	−0.65[a]	0.06	−0.38	0.20	−0.41	−0.44	−0.23
CI		1	0.87[a]	−0.09	0.82[a]	0.13	0.87[a]	0.77[a]	0.63[a]
A			1	0.02	0.84[a]	0.14	0.80[a]	0.90[a]	0.83[a]
V				1	0.24	0.48	−0.27	0.41	0.23
AV					1	0.37	0.87[a]	0.77[a]	0.89[a]
AV_*A*						1	0.17	0.14	−0.11
AV_*V*							1	0.63[a]	0.65[a]
*M*_error								1	0.87[a]
*M*_bias									1

[a]Significant correlation at the 0.05 level.

ing conclusions:

(1) Normal-hearing listeners show not only sharp categorical perception of place of articulation but also strong reliance on the auditory cue;
(2) Cochlear-implant listeners show weak categorical perception of place of articulation but strong reliance on the visual cue;
(3) The implant listeners can derive significant AV benefit from the congruent acoustic and optical cues;
(4) Both normal and implant listeners produced a relatively weak McGurk effect in response to the incongruent acoustic and optical cues, possibly due to the weak signal strength in the synthetic stimuli;
(5) It is the duration of the implant experience, rather than the duration of deafness, that correlates with the amount of AV integration; and
(6) The present acoustic simulation model of the cochlear implant can predict the actual implant auditory-only performance but not the AV speech integration.

ACKNOWLEDGMENTS

We are very grateful for the time and dedication that our cochlear-implant and normal-hearing listeners have offered for this study. We thank Dominic Massaro, Jacques de Villiers, and Paul Hosom at the Center for Spoken Language Understanding, UC Santa Cruz, for their help and for the use of their Baldi software. We also thank Abby Copeland, Jeff Carroll, Hongbin Chen, Sara Ghazi, Jeremy Liu, Sheng Liu, and Kaibao Nie for their assistance in technical support and helpful comments on the manuscript. This work was supported in part by a grant from the National Institutes of Health (Grant No. RO1 DC002267).

[1]One of the cochlear-implant listeners tested reported that she had corrective cataract surgery and that the vision in the right eye was still somewhat limited. However, she also reported that glasses are sufficient for viewing the television and objects at close range. She was also able to drive during daylight hours.

[2]Occasionally there were functions that did not optimally fit the parameters of the sigmoid four-parameter equation. In these cases, alternative methods were used to find the *x* intercepts: (1) if a five-parameter sigmoidal function was able to fit a boundary instead of a four-parameter sigmoidal function, then the values for a, b, x_0, and y_0 were taken from calculations made by the graphing software (SigmaPlot); (2) if two boundary lines were linear functions, the *x* intercept was determined from the equation of the intersecting lines. When none of these rules applied, for instance when one function was linear and the second sigmoidal, the point that most closely estimated the *x* intercept was taken upon inspection of the graph. These exceptions only occurred in nine of 672 cases.

Barnard, E., Bertrant, J., Rundlem, B., Cole, R., Cosi, P., Cronk, A., de Villiers, J., Fanty, M., Hosom, P., Kain, A., Kaiser, E., Macon, M., Rozman, R., Shobaki, K., Schalkwyk, J., Sutton, S., Tadrist, A., van Vuuren, S., Vermelen, P., Wouters, J., Yan, Y. (**2000**). ⟨http://cslu.cse.ogi.edu/⟩, Oregon Graduate Institute of Science and Technology, (Last viewed 10/16/2007).

Bergeson, T. R., Pisoni, D. B., and Davis, R. A. (**2005**). "Development of audiovisual comprehension skills in prelingually deaf children with cochlear implants." Ear Hear. **26**, 149–164.

Bernstein, L. E., Auer, E. T., Jr., Moore, J. K., Ponton, C. W., Don, M. and Singh, M. (**2002**). "Visual speech perception without primary auditory cortex activation." NeuroReport **13**, 311–315.

Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (**2000**). "Speech perception without hearing." Percept. Psychophys. **62**, 233–252.

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (**1974**). "Auditory and visual contributions to the perception of consonants." J. Speech Hear. Res. **17**, 619–630.

Blamey, P., Arndt, P., Bergeron, F., Bredberg, G., Brimacombe, J., Facer, G., Larky, J., Lindstrom, B., Nedzelski, J., Peterson, A., Shipp, D., Staller, S., and Whitford, L. (**1996**). "Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants." Audiol. Neuro-Otol. **1**, 293–306.

Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments." Q. J. Exp. Psychol. A **43**, 647–677.

Bunnell, H. (**1996**). ⟨http://wagstaff.asel.udel.edu/speech/tutorials/synthesis/⟩, Speech Research Laboratory, A.I. duPont Hospital for Children, (Last viewed 10/16/2007).

Burnham, D., and Dodd, B. (**2004**). "AV speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect." Dev. Psychobiol. **45**, 204–220.

Calvert, A. A., Bullmore, T. T., Brammer, J. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (**1997**). "Activation of auditory cortex during silent lipreading." Science **276**, 593–596.

Calvert, G. A., and Campbell, R. (**2003**). "Reading speech from still and moving faces: the neural substrates of visible speech." J. Cogn Neurosci. **15**, 57–70.

Cienkowski, K. M., and Carney, A. E. (**2002**). "AV speech perception and aging." Ear Hear. **23**, 439–449.

Clark, G. (**2003**). "Cochlear implants in children: safety as well as speech and language." Int. J. Pediatr. Otorhinolaryngol. **67**, S7–20.

De Gelder, B., and Bertelson, P. (**2003**). "Multisensory integration, perception and ecological validity." Trends Cogn. Sci. **7**, 460–467.

Dodd, B. (**1977**). "The role of vision in the perception of speech." Percep-

tion **6**, 31–40.
Doucet, M. E., Bergeron, F., Lassonde, M., Ferron, P., and Lepore, F. (**2006**). "Cross-modal reorganization and speech perception in cochlear implant users." Brain **129**, 3376–3383.
Easton, R. D., and Basala, M. (**1982**). "Perceptual dominance during lipreading." Percept. Psychophys. **32**, 562–570.
Erber, N. P. (**1972**). "Auditory, visual, and AV recognition of consonants by children with normal and impaired hearing." J. Speech Hear. Res. **15**, 413–422.
Fisher, C. G. (**1968**). "Confusions among visually perceived consonants." J. Speech Hear. Res. **11**, 796–804.
Giraud, A. L., Price, C. J., Graham, J. M., and Frackowiak, R. S. (**2001a**). "Functional plasticity of language-related brain areas after cochlear implantation." Brain **124**, 1307–1316.
Giraud, A. L., Price, C. J., Graham, J. M., Truy, E., and Frackowiak, R. S. (**2001b**). "Cross-modal plasticity underpins language recovery after cochlear implantation." Neuron **30**, 657–663.
Goh, W. D., Pisoni, D. B., Kirk, K. I., and Remez, R. E. (**2001**). "Audiovisual perception of sinewave speech in an adult cochlear implant user: a case study." Ear Hear. **22**, 412–419.
Gomaa, N. A., Rubinstein, J. T., Lowder, M. W., Tyler, R. S., and Gantz, B. J. (**2003**). "Residual speech perception and cochlear implant performance in postlingually deafened adults." Ear Hear. **24**, 539–544.
Gordon-Salant, S., and Fitzgibbons, P. J. (**1997**). "Selected cognitive factors and speech recognition performance among young and elderly listeners." J. Speech Lang. Hear. Res. **40**, 423–431.
Grant, K. W. (**1987**). "Encoding voice pitch for profoundly hearing-impaired listeners." J. Acoust. Soc. Am. **82**, 423–432.
Grant, K. W., and Braida, L. D. (**1991**). "Evaluating the articulation index for AV input." J. Acoust. Soc. Am. **89**, 2952–2960.
Grant, K. W., Braida, L. D., and Renn, R. J. (**1991**). "Single band amplitude envelope cues as an aid to speechreading." Q. J. Exp. Psychol. A **43**, 621–645.
Grant, K. W., Braida, L. D., and Renn, R. J. (**1994**). "Auditory supplements to speechreading: combining amplitude envelope cues from different spectral regions of speech." J. Acoust. Soc. Am. **95**, 1065–1073.
Grant, K. W., and Seitz, P. F. (**1998**). "Measures of AV integration in nonsense syllables and sentences." J. Acoust. Soc. Am. **104**, 2438–2450.
Grant, K. W., and Walden, B. E. (**1996**). "Evaluating the articulation index for AV consonant recognition." J. Acoust. Soc. Am. **100**, 2415–2424.
Grant, K. W., Walden, B. E., and Seitz, P. F. S. (**1998**). "AV speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and AV integration." J. Acoust. Soc. Am. **103**, 2677–2690.
Green, K. M., Julyan, P. J., Hastings, D. L., and Ramsden, R. T. (**2005**). "Auditory cortical activation and speech perception in cochlear implant users: effects of implant experience and duration of deafness." Hear. Res. **205**, 184–192.
Green, K. P., Kuhl, P. K., Meltzoff, A. N., and Stevens, E. B. (**1991**). "Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect." Percept. Psychophys. **50**, 524–536.
Green, K. P., and Norrix, L. W. (**1997**). "Acoustic cues to place of articulation and the McGurk Effect: the role of release bursts, aspiration, and formant transitions." J. Speech Lang. Hear. Res. **40**, 646–665.
Greenwood, D. (**1990**). "A cochlear frequency-position function for several species - 29 years later." J. Acoust. Soc. Am. **87**, 2592–2605.
Hay-McCutcheon, M. J., Pisoni, D. B., and Kirk, K. I. (**2005**). "Audiovisual speech perception in elderly cochlear implant recipients." Laryngoscope **115**, 1887–1894.
Helfer, K. S. (**1998**). "Auditory and AV recognition of clear and conversational speech by older adults." J. Am. Acad. Audiol **9**, 234–242.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels." J. Acoust. Soc. Am. **97**, 3099–3111.
Humes, L. E. (**2002**). "Factors underlying the speech-recognition performance of elderly hearing-aid wearers." J. Acoust. Soc. Am. **112**, 1112–1132.
Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer." J. Acoust. Soc. Am. **67**, 971–995.
Kong, Y. Y., Stickney, G. S., and Zeng, F. G., (**2005**). "Speech and melody recognition in binaurally combined acoustic and electric hearing." J. Acoust. Soc. Am. **117**, 1351–1361.
Lachs, L., Pisoni, D. B., and Kirk, K. I. (**2001**). "Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report." Ear Hear. **22**, 236–251.
Lee, D. S., Lee, J. S., Oh, S. H., Kim, S. K., Kim, J. W., Chung, J. K., Lee, M. C., and Kim, C. S., (**2001**). "Cross-modal plasticity and cochlear implants." Nature (London) **409**, 149–150.
Massaro, D. W. (**1998**). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press, Cambridge, MA).
Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception." J. Exp. Psychol. Hum. Percept. Perform. **9**, 753–771.
Massaro, D. W., and Cohen, M. M., (**2000**). "Tests of AV integration efficiency within the framework of the fuzzy logical model of perception." J. Acoust. Soc. Am. **108**, 784–789.
Massaro, D. W., Cohen, M. M., Beskow, J., and Cole, R. (**2000**). "Developing and evaluating conversational agents." in *Embodied Conversational Agents*, edited by J. Chassell, J. Sullivan, S. Prevost, and E. Churchill (MIT Press, Cambridge, MA), pp. 287–318.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices." Nature (London) **264**, 746–748.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants." J. Acoust. Soc. Am. **27**, 338–352.
Montgomery, A. A., Walden, B. E., Schwartz, D. M., and Prosek, R. A. (**1984**). "Training AV speech reception in adults with moderate sensorineural hearing loss." Ear Hear. **5**, 30–36.
Moody-Antonio, S., Takayanagi, S., Masuda, A., Auer, T. T., Jr., Fisher, L., and Bernstein, L. E. (**2005**). "Improved speech perception in adult congenitally deafened cochlear implant recipients." Otol. Neurotol. **26**, 649–654.
Munson, B., Donaldson, S. S., Allen, S. L., Collison, E. A., and Nelson, D. A. (**2003**). "Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability." J. Acoust. Soc. Am. **113**, 925–935.
Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., Howard, M. A., and Brugge, J. F. (**2007**). "AV processing represented in the human superior temporal gyrus." Neuroscience **145**, 162–184.
Rosen, S. M. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects." Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.
Rosen, S. M., Fourcin, A. J., and Moore, B. C. (**1981**). "Voice pitch as an aid to lipreading." Nature (London) **291**, 150–152.
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (**2007**). "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments." Cereb. Cortex **17**, 1147–1153.
Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., and Barone, P. (**2007**). "Evidence that cochlear-implanted deaf patients are better multisensory integrators." Proc. Natl. Acad. Sci. U.S.A. **104**, 7295–7300.
Sams, M., Manninen, P., Surakka, P., Helin, P., and Katto, R. (**1998**). "McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context." Speech Commun. **26**, 75–87.
Schorr, E. A., Fox, N. A., van Wassenhove, V., and Knudsen, E. I. (**2005**). "AV fusion in speech perception in children with cochlear implants." Proc. Natl. Acad. Sci. U.S.A. **102**, 18748–18750.
Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing." J. Acoust. Soc. Am. **106**, L71–74.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues.." Science **270**, 303–304.
Skinner, M. W., Holden, L. K., Whitford, L. A., Plant, K. L., Psarros, C., and Holden, T. A. (**2002**). "Speech recognition with the nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults." Ear Hear. **23**, 207–223.
Sommers, M. S., Tye-Murray, N., and Spehar, B. (**2005**). "AV speech perception and AV enhancement in normal-hearing younger and older adults." Ear Hear. **26**, 263–275.
Stickney, G. S., Zeng, F. G., Litovsky, R. Y., and Assmann, P. F. (**2004**). "Cochlear implant speech recognition with speech masker." J. Acoust. Soc. Am. **116**, 1081–1091.
Sumby, W., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise." J. Acoust. Soc. Am. **26**, 212–215.
Summerfield, Q. (**1979**). "Use of visual information for phonetic perception." Phonetica **36**, 314–331.
Turner, C. W., and Robb, M. P. (**1987**). "Audibility and recognition of stop

consonants in normal and hearing-impaired subjects." J. Acoust. Soc. Am. **81**, 1566–1573.
Turner, C. W., Smith, S. J., Aldridge, P. L., and Stewart, S. L. (**1997**). "Formant transition duration and speech recognition in normal and hearing-impaired listeners." J. Acoust. Soc. Am. **101**, 2822–2825.
Tyler, R. S.,and Summerfield, A. Q. (**1996**). "Cochlear implantation: Relationships with research on auditory deprivation and acclimatization." Ear Hear. **17**, 38S–50S.
van Dijk, J. E., van Olphen, A. F., Langereis, M. C., Mens, H. H., Brokx, J. P., and Smoorenburg, G. F. (**1999**). "Predictors of cochlear implant performance." Audiology **38**, 109–116.
Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition." J. Acoust. Soc. Am. **82**, 1152–1161.
van Wassenhove, V., Grant, K. W., and Poeppel, D. (**2007**). "Temporal window of integration in AV speech perception." Neuropsychologia **45**, 598–607.
Walden, B. E., Busacco, D. A., and Montgomery, A. A. (**1993**). "Benefit from visual cues in AV speech recognition by middle-aged and elderly persons." J. Speech Hear. Res. **36**, 431–436.
Walden, B. E., Erdman, S. A., Montgomery, A. A., Schwartz, D. M., and Prosek, R. A. (**1981**). "Some effects of training on speech recognition by hearing-impaired adults." J. Speech Hear. Res. **24**, 207–216.
Walden, B. E., Montgomery, A. A., Prosek, R. A., and Hawkins, D. B. (**1990**). "Visual biasing of normal and impaired auditory speech perception." J. Speech Hear. Res. **33**, 163–173.
Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations." Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.

Perceptual coherence in listeners having longstanding childhood hearing losses, listeners with adult-onset hearing losses, and listeners with normal hearing

Andrea Pittman[a)]
Department of Speech and Hearing Sciences, P.O. Box 870102, Arizona State University, Tempe, Arizona, USA 85287-0102

(Received 29 December 2006; revised 11 October 2007; accepted 12 October 2007)

Perceptual coherence, the process by which the individual elements of complex sounds are bound together, was examined in adult listeners with longstanding childhood hearing losses, listeners with adult-onset hearing losses, and listeners with normal hearing. It was hypothesized that perceptual coherence would vary in strength between the groups due to their substantial differences in hearing history. Bisyllabic words produced by three talkers as well as comodulated three-tone complexes served as stimuli. In the first task, the second formant of each word was isolated and presented for recognition. In the second task, an isolated formant was paired with an intact word and listeners indicated whether or not the isolated second formant was a component of the intact word. In the third task, the middle component of the three-tone complex was presented in the same manner. For the speech stimuli, results indicate normal perceptual coherence in the listeners with adult-onset hearing loss but significantly weaker coherence in the listeners with childhood hearing losses. No differences were observed across groups for the nonspeech stimuli. These results suggest that perceptual coherence is relatively unaffected by hearing loss acquired during adulthood but appears to be impaired when hearing loss is present in early childhood. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2804953]

PACS number(s): 43.71.Ky, 43.71.Ft, 43.71.An [ARB] Pages: 441–449

I. INTRODUCTION

Speech is a stream of complex, dissimilar acoustic elements that are interpreted as orderly and meaningful events by experienced listeners. The process by which these disparate elements are bound together and perceived as speech has been given the term "perceptual coherence." Both phonemic and auditory cues within the speech signal have been shown to promote perceptual coherence. At the level of the phoneme, the listener appears to bind (or cohere) the various acoustic components within the phoneme during perception (Best *et al.*, 1989; Remez and Rubin, 1990; Carrell and Opie, 1992; Barker and Cooke, 1999). Remez *et al.* (2001) demonstrated the coherence of acoustic components using simple sine-wave speech. In sine-wave speech, the formants that comprise a word or phrase are replicated with pure tones that vary in frequency and amplitude over time. Listeners in this study were first asked to detect differences between pairs of isolated second formants and then to indicate if the isolated formants occurred within synthetic words. Although the listeners were able to differentiate the isolated formants well ($d' \sim 3$), they were unable to detect the presence of a specific formant within a word ($d' \sim 0$). The results suggest that the process of binding the acoustic elements together resulted in the percept of one event rather than several individual events.

Perceptual coherence appears to operate in the same manner as comodulation masking release (Hall *et al.*, 1984; McFadden, 1987; Grose and Hall, 1993). During comodulation masking release, the threshold of detection for a tone presented in an amplitude modulated band of noise is improved (i.e., threshold is reduced) by adding another amplitude modulated band of noise in a different frequency region, provided that the modulation rates are the same. It is thought that the listener perceives the two noise bands as one auditory event and thus perceives the tone separately, improving the threshold. During speech perception, the components of speech are thought to group together as one auditory event and stand apart from other simultaneous acoustic events (Bregman, 1990).

Gordon (1997a, b) demonstrated this aspect of coherence using synthetic representations of the vowels /ɛ/ and /ɪ/. The first formants were centered at 625 and 375 Hz, respectively. The second and third formants were identical for both vowels and occurred at 2200 and 2900 Hz, respectively. All of the formants were amplitude modulated (or comodulated) at 125 Hz. Because the two high-frequency formants were the same for both vowels, the low-frequency formant provided the only unique information with which to identify either vowel. The stimuli were mixed with low-pass noise (corner frequency: 1000 Hz) that overlapped the first formant only. The listeners in this study were asked to indicate which of the two vowels was presented in each trial. The signal-to-noise ratio necessary to distinguish the two vowels was measured with and without the addition of the ambiguous higher-frequency formants. The results showed a small

[a)]A portion of the data was collected while the author was affiliated with Boys Town National Research Hospital. Electronic mail: andrea.pittman@asu.edu

but significant reduction in signal-to-noise ratio (3 dB) when the high-frequency formants were added to the speech signal. These results suggest that the comodulated formants cohered to become one auditory event, improving the threshold of detection in noise.

Similarly, Carrell and Opie (1992) showed that intelligibility of sine-wave speech increased when the sinusoids were amplitude comodulated at rates similar to the fundamental frequency of adult talkers. In the third of three experiments, they comodulated the sinusoidal formants of four sentences at 50, 100, and 200 Hz and presented them for recognition to normally hearing listeners. They also presented the sentences with no comodulation. Because comodulation masking release is strongest at low modulation rates and weakest at high modulation rates, the authors hypothesized that intelligibility of the sentences would be best for the 50 Hz modulation rate and poorest for the 200 Hz modulation rate. The intelligibility results were consistent with their hypothesis and suggested that the principles that govern comodulation masking release are also involved in speech perception. That is, the talker's fundamental frequency contributes to perceptual coherence and that lower fundamental frequencies may promote stronger coherence.

Although studies in normally hearing adults have provided important evidence regarding the nature of perceptual coherence and the principles that govern it, additional insight may be gained from examination of the underdeveloped perceptual system (children) or the disordered system (listeners with hearing loss). In the only study of perceptual coherence in children to date, Nittrouer and Crowther (2001) used a categorical perception paradigm to examine the developmental course of phonemic coherence. Adults and children (5 and 7 years of age) were asked to indicate if two stimuli were the same or different. The second of the two stimuli contained gap duration and formant frequency cues that were consistent with naturally produced speech or conflicted with it. The consistent and conflicting cues were intended to influence the listener's decision regarding the similarity between the two stimuli. The responses of the youngest children were least influenced by the consistent and conflicting cues contained in the stimuli suggesting stronger perceptual coherence in the 5-year-old children compared to the 7-year-olds and adults. Their results suggest that young children may learn to perceive speech by overcoming the effects of coherence rather than developing the ability to bind acoustic elements together. That is, children may learn to detect the subtle, but important, acoustic cues that lead to speech perception (e.g., the third formant distinction between /la/ and /ra/) rather than learning which cues to group together.

Unfortunately, there are no data regarding perceptual coherence in the impaired auditory system. However, several studies have examined the effect of hearing loss for a number of related auditory processes. For example, comodulation masking release has been shown to be significantly reduced in listeners with hearing loss relative to listeners with normal hearing (Hall and Grose, 1989; 1994; Eisenberg *et al.*, 1995). Also, listeners with hearing loss have been shown to have difficulty organizing nonspeech stimuli for the purpose of analytic listening (i.e., hearing out individual parts of a complex sound) (Grose and Hall, 1996; Rose and Moore, 1997; Kidd *et al.*, 2002). Although the poor performance of listeners with hearing loss has been attributed to peripheral factors associated with hearing loss (e.g., abnormally wide auditory filters, reduced temporal resolution, poor frequency selectivity), the contribution of these factors does not appear to be substantial and suggests that other factors may also contribute to coherence. Therefore, it may be informative to examine the perceptual coherence of listeners with adult-onset hearing losses relative to those having longstanding childhood hearing losses. The ability to organize the acoustic components of speech and nonspeech stimuli may differ across listeners for whom perceptual coherence developed in the presence of normal hearing or in the presence of hearing loss.

The purpose of the current investigation was to examine the perceptual coherence of listeners with childhood hearing losses and listeners with adult-onset hearing losses relative to that of listeners with normal hearing. For this study, a version of the paradigm employed in Remez *et al.* (2001) was used (described earlier). This paradigm required the listener to detect individual acoustic components within speech and nonspeech stimuli. Poor performance on this task would suggest relatively strong perceptual coherence whereas good performance would suggest relatively weak perceptual coherence. It was hypothesized that for both types of stimuli, the perceptual coherence of the listeners with adult-onset hearing losses would be stronger than normal due to their inability to perceive the subtle acoustic cues of speech. It was also hypothesized that the perceptual coherence of the listeners with childhood hearing loss would be greater than that of the listeners with adult-onset hearing loss due to deficits associated with life-long hearing loss as well as their inability to perceive the subtle acoustic cues of speech.

II. METHODS

A. Participants

Ten normal-hearing listeners, ten listeners with adult-onset hearing losses (A-HL), and ten listeners with longstanding, childhood hearing losses (C-HL) participated in this study. All of the listeners with hearing loss had bilateral, symmetrical, sensorineural hearing losses. The age and hearing levels for each of the listeners with hearing loss is given in Table I. The average hearing levels (and standard error) of the listeners with normal hearing are also given. On average, the thresholds of the listeners with A-HL were within the range of normal at 125 and 250 Hz and increased to ~70 dB SPL at frequencies >1000 Hz. The listeners with C-HL had equally poor thresholds across frequency (~60 dB SPL).

B. Stimuli

Three different stimuli were used in this study: isolated second formants, naturally produced words, and time-varying sinusoids. The stimuli were fashioned after those used in Remez *et al.* (2001) with the exception that the speech stimuli were naturally produced and all of the stimuli were longer in duration. The speech stimuli were selected from an original stimulus set containing digital recordings of

TABLE I. Age and hearing thresholds for the listeners with adult-onset hearing losses (A-HL) and childhood hearing losses (C-HL).

ID No.	Age (years)	Ear	Hearing thresholds (dB SPL) 0.125	0.25	0.5	1	2	4	8
A-HL1	67	L	27	23	15	12	60	65	67
		R	32	33	20	13	67	60	58
A-HL2	60	L	43	38	38	33	67	70	73
		R	35	37	33	43	58	73	87
A-HL3	57	L	33	17	23	18	37	63	43
		R	33	17	23	23	48	70	47
A-HL4	69	L	27	5	7	7	15	60	38
		R	17	15	15	8	37	67	50
A-HL5	68	L	30	22	33	30	30	60	73
		R	30	13	13	13	35	60	73
A-HL6	66	L	33	22	13	13	57	65	77
		R	27	23	13	23	67	67	77
A-HL7	66	L	30	17	17	20	15	83	87
		R	32	26	17	8	15	63	70
A-HL8	65	L	37	33	27	37	67	67	93
		R	52	43	30	37	53	63	77
A-HL9	58	L	57	66	61	69	73	65	46
		R	69	54	56	61	45	57	58
A-HL10	64	L	58	50	57	53	46	51	82
		R	48	48	58	50	48	55	68
Avg	64	L	38 (3)	29 (4)	29 (4)	29 (4)	47 (5)	65 (3)	68 (4)
(1 SE)		R	38 (4)	31 (4)	28 (4)	28 (4)	47 (4)	64 (2)	67 (4)
C-HL11	39	L	55	53	53	53	63	53	83
		R	58	58	58	58	58	58	83
C-HL12	35	L	73	83	93	92	83	70	68
		R	83	78	88	93	87	68	68
C-HL13	19	L	53	37	38	38	23	13	33
		R	33	33	38	47	38	5	13
C-HL14	14	L	61	63	71	68	63	63	59
		R	67	73	68	68	63	63	62
C-HL15	54	L	63	63	78	78	83	73	88
		R	53	43	53	68	68	68	78
C-HL16	35	L	66	72	76	74	68	67	81
		R	62	63	63	68	67	62	73
C-HL17	36	L	38	43	53	78	79	66	73
		R	43	43	56	88	83	71	75
C-HL18	41	L	36	33	33	37	43	57	33
		R	43	38	36	28	33	48	36
C-HL19	44	L	57	48	51	58	78	108	130
		R	78	81	78	77	68	91	107
C-HL20	43	L	35	48	62	68	78	83	78
		R	43	48	63	73	83	98	130
Avg	36	L	54 (4)	54 (4)	61 (4)	64 (4)	66 (4)	65 (5)	73 (5)
(1 SE)		R	56 (4)	56 (4)	60 (4)	67 (4)	65 (4)	63 (5)	73 (6)
Normal hearing									
Avg	25	L	33 (2)	21 (3)	17 (2)	8 (3)	6 (3)	11 (3)	14 (4)
(1 SE)		R	34 (3)	25 (2)	16 (2)	12 (2)	10 (2)	8 (3)	17 (3)

30 bisyllabic words produced by an adult male, an adult female, and a 7-year-old child. The words consisted of voiced consonants and vowels (sonorants). The words produced by each talker were concatenated and saved to separate files (one per talker). The stimuli were then subjected to a pitch extraction algorithm in MATLAB (YIN: de Cheveigne and Kawahara, 2002). On average, the male, female, and child talkers had fundamental frequencies of 96, 180, and 227 Hz, respectively. These stimuli provided a range of fundamental frequencies which was expected to influence perceptual coherence in a systematic fashion. Specifically, coherence was expected to be best for the male talker (lowest fundamental frequency) and poorest for the child talker (highest fundamental frequency).

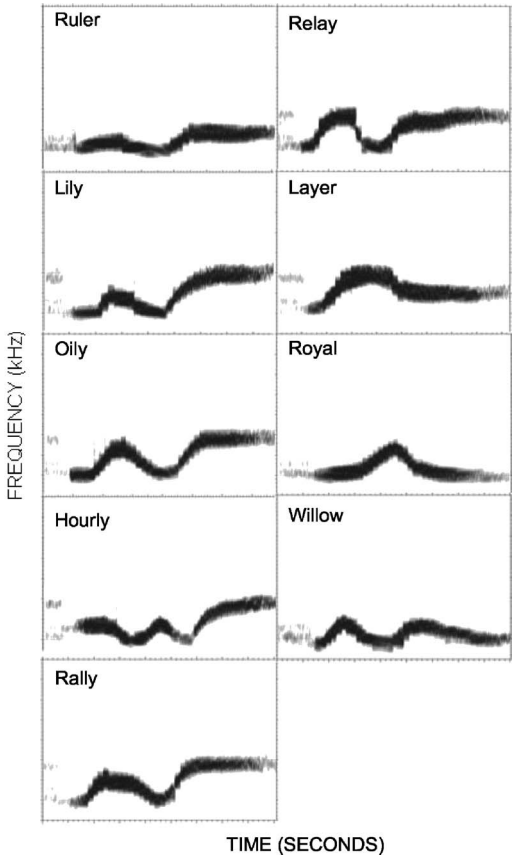


FIG. 1. Spectrograms of the second formants for the female talker. The abscissa and ordinate of each spectrogram extend to 8000 Hz and 1000 ms, respectively.

The second formant in each word was isolated using custom laboratory software written in MATLAB. The program allowed the user to visually inspect the spectrogram and manually trace the second formant. The trace was then converted into a bandpass filter having a selectable bandwidth. For these stimuli, a filter with a bandwidth of 250 Hz (at the 3 dB down point) was applied throughout the length of the formant. The stimuli were presented for identification to a separate group of ten normal-hearing adults. The words *ruler, lily, oily, hourly, rally, relay, layer, royal*, and *willow* could not be identified by any of the listeners based on the acoustic information provided by the second formant alone and were selected as stimuli for the study. Figure 1 shows the spectrograms for the second formants extracted from the nine stimulus words produced by the female talker. The isolated formants varied in frequency over time between 750 and 3500 Hz across talkers.

A series of simple time-varying sinusoids also were created. The sinusoids were amplitude modulated at 100 Hz using a triangular wave with a 50% duty cycle. The sinusoids were arranged into three-tone complexes having a high-, mid-, and low-frequency component. The on- and offset frequencies of each component in the complexes are listed in Table II. Note that the high-frequency component was the same for all stimuli and consisted of on- and offset frequencies of 3500 and 4000 Hz, respectively. Likewise, the low-frequency component was the same for all stimuli and consisted of on- and offset frequencies of 1000 and 500 Hz, respectively. The on- and offset frequencies of the midfrequency component varied from 1500 to 3000 Hz in six steps of 200–500 Hz.

TABLE II. On- and offset frequency of each component in the time-varying sinusoid stimuli.

Stim No.	Component	Onset (Hz)	Offset (Hz)
1	Low	1000	500
	Mid	3000	1500
	High	3500	4000
2	Low	1000	500
	Mid	2800	1800
	High	3500	4000
3	Low	1000	500
	Mid	2500	2000
	High	3500	4000
4	Low	1000	500
	Mid	2000	2500
	High	3500	4000
5	Low	1000	500
	Mid	1800	2800
	High	3500	4000
6	Low	1000	500
	Mid	1500	3000
	High	3500	4000

C. Amplification parameters

Prior to testing, hearing thresholds were obtained from each listener. The stimuli were frequency shaped for the listeners with hearing loss according to the Desired Sensation Level fitting algorithm (DSL i/o version 4.1) (Seewald *et al.*, 1997). The amplification parameters provided levels sufficient to perceive average conversational speech (65 dB SPL). An estimate of stimulus audibility was calculated for each listener using the speech intelligibility index (SII). The importance function for nonsense syllables was used (ANSI 1997). On average, the SII for the listeners with A-HL and C-HL was 0.84 (s.d.=0.05) and 0.73 (s.d.=0.23), respectively, which is sufficient for maximum perception of speech. Figure 2 shows the average long-term speech spectra (dashed lines) relative to the average hearing levels (solid lines) for the listeners with A-HL and C-HL in the upper and lower panels, respectively. Note that sufficient amplification was provided to ensure audibility of formants occurring at or below 4000 Hz for all listeners. The hatched area in each panel represents the range of thresholds for the listeners with normal hearing as defined by +1 s.d. around the mean for both the right and left ears. All measures were referenced to a 6 cm^3 coupler.

D. Procedures

Each listener participated in three tasks. In the first task, the listeners' ability to perceive a word from the second formant alone was determined. This task was necessary to confirm that the listeners perceived the second formants as acoustic stimuli rather than as meaningful speech. On each trial, an isolated second formant was presented and the lis-

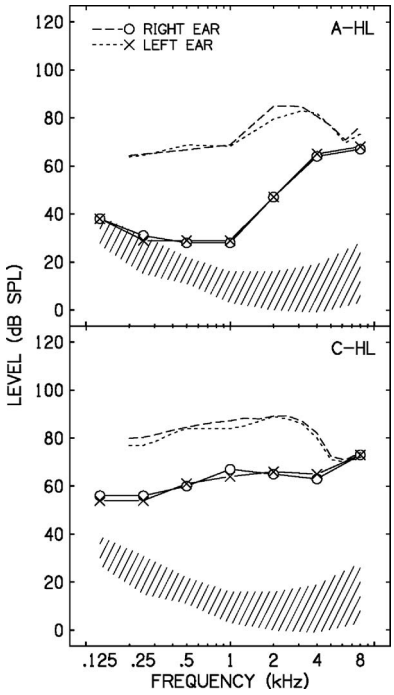


FIG. 2. Long-term average speech spectra (dashed lines) and hearing thresholds (solid lines) for the left and right ears of the listeners with adult onset hearing loss (A-HL) in the upper panel and with longstanding, childhood hearing loss (C-HL) in the lower panel. The hatched area in each panel represented the range of hearing thresholds of the listeners with normal hearing as defined by 1 s.d. around the mean.

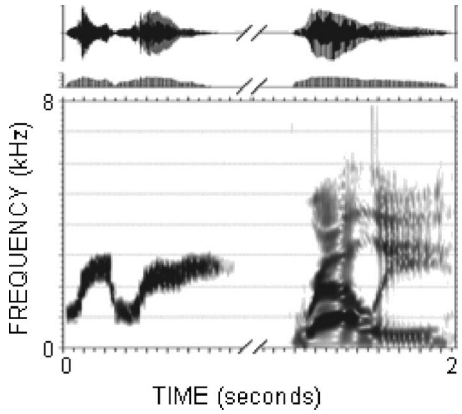


FIG. 3. An example of a stimulus set used in the primary task. A second formant was paired with an intact, neighboring word separated by 800 ms. In this example of a "no" trial, the second formant *relay* was paired with the word *rally*.

tener verbally responded with the word he/she perceived, if any. The listeners were not provided with a list of words from which to choose their responses nor were they given any feedback. Each second formant was presented once for a total of 27 trials (9 formants × 3 talkers).

The second (and primary) task assessed the listener's ability to perceive a specific auditory object (the second formant) within a speech stimulus (the intact word). Each trial consisted of an isolated formant, an 800 ms silent gap, and an intact word. Figure 3 shows the wave form and spectrogram of one stimulus produced by the female talker. In this example, the second formant was extracted from the word *relay* and paired with the word *rally*. For each trial the listener reported whether or not the formant was contained within the word. The listeners indicated their response by selecting "yes" or "no" buttons on a touch-screen monitor. Feedback was provided for correct responses.

To minimize the number of stimuli while maximizing the difficulty of the task, the isolated formants were paired with words having similar second formant characteristics. To achieve this, the formants were ordered according to offset frequency and overall morphology. This order is given in Table III (as well as in Fig. 1) and shows the manner in which the second formants (rows) were paired with the words (columns). Half of the trials contained a second formant and intact word which neighbored the original word ("no" trials). The remaining trials contained stimuli with the same second formants ("yes" trials). The words *ruler* and *willow* were not paired with neighboring formants because they comprised the end points of the continuum. Admittedly, the progression from one formant to the next could have been accomplished through several different arrangements; however the perceptual differences between these formants was quite subtle. This task consisted of a total of 168 trials (3 talkers × 2 repetitions of the 14 no stimuli + 4 repetitions of the 7 yes stimuli).

The third task assessed the listener's ability to perceive a single amplitude modulated sinusoid within a three-tone complex of sinusoids. Because this task contained nonspeech stimuli, coherence under these conditions may suggest that

TABLE III. Pairing of the isolated second formants with the intact words. "Yes" trials contained a second formant and the intact word from which it originated. "No" trials contained a second formant and an intact word neighboring the word from which the second formant was obtained.

	Intact words								
Second formant	ruler	lily	oily	hourly	rally	relay	layer	royal	willow
Lily	No	Yes	No						
Oily		No	Yes	No					
Hourly			No	Yes	No				
Rally				No	Yes	No			
Relay					No	Yes	No		
Layer						No	Yes	No	
Royal							No	Yes	No

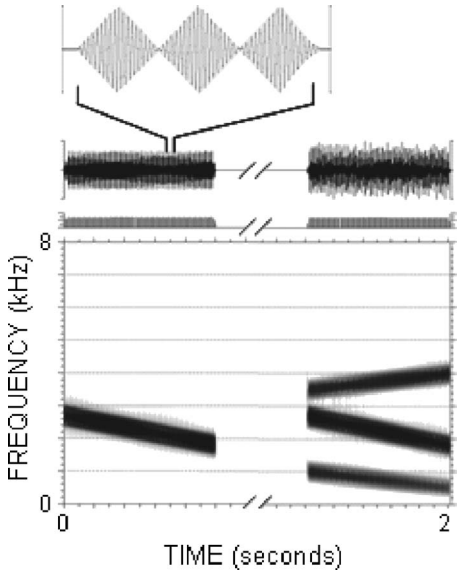


FIG. 4. An example of a stimulus set used in the final task. A midfrequency amplitude modulated sinusoid was paired with a three-tone complex. In this example of a "yes" trial, the midfrequency sinusoid was contained within the three-tone complex. The upper wave form shows three modulation cycles of the midfrequency sinusoid.

the fundamental acoustic characteristic of frequency modulation is a significant contributor to the coherence of speech. As with the speech stimuli, a midfrequency sinusoid was paired with a three-tone complex containing the same midfrequency sinusoid or that of a neighboring complex. Figure 4 shows the wave form and spectrogram of one stimulus pair. Table IV shows the stimulus pairs in their respective yes and no categories. The procedures for this task were the same as those for the previous task. After hearing an isolated sinusoid followed by a three-tone complex, the listener indicated whether or not the isolated tone was contained in the complex. This task consisted of 80 trials (5 repetitions of the 8 no stimuli+10 repetitions of the 4 yes stimuli).

All testing took place in an audiometric test suite. The stimuli were presented binaurally in quiet under earphones (Sennheiser, HD25). Testing required approximately 2 h and each listener was paid for his/her time.

III. RESULTS

A. Task 1: Perception of words from isolated second formants

In general the listeners did not perceive the isolated second formants as speech. Recall that this task consisted of 27 trials (1 repetition × 9 formants × 3 talkers). On average, the listeners with normal hearing, A-HL, and C-HL correctly identified the word from which the isolated formant originated only 4%, 1%, and 3% of the trials, respectively. The highest score achieved was 11% (3 of 27 trials) by two normally hearing listeners and two listeners with C-HL. However, most of the other listeners (17 of 30) did not identify any of the isolated second formants as the words from which they originated. These results suggest that the listeners perceived the isolated formants as acoustic stimuli rather than speech. This result was important because it demonstrated that in the primary task the listeners were not simply comparing two speech stimuli.

B. Task 2: Coherence of real speech

Because the average performance of a listener may be biased toward one response in a two-alternative forced choice paradigm (yes/no), each listener's sensitivity to the yes and no trials was calculated and represented as a value of d' (Marshall and Jesteadt, 1986). Specifically, a listener's correct responses on yes trials (hits) were subtracted from the portion of time he/she responded incorrectly to no trials (false alarms). The d' value was then used to calculate the listener's maximized performance in the absence of a response bias. On average the maximized performance for each group was adjusted by less than 1 percentage point suggesting that these listeners demonstrated little or no response bias.

Figure 5 displays the average maximized performance (bars re: left axis) and d' (open circles re: right axis) for the primary task as a function of group. The error bars represent +1 SE. In general, the performance and d' values for the listeners with C-HL were higher than that of the listeners with normal hearing or with A-HL. A repeated measures ANOVA of the d' data revealed a significant main effect of group ($F_{2,27}=4.862$; $p=0.016$). Tukey's HSD post-hoc analyses revealed that, on average, the d' values of the listeners with normal hearing and with A-HL did not differ from one another. However, the d' values of the listeners with C-HL were significantly higher than that of the other groups. These results indicate that the listeners with C-HL were better able to hear the isolated second formants within the words. The lower d' values for the listeners with NH and with A-HL suggest that they incorrectly matched the isolated formants to intact words more often.

The ability of the listeners with C-HL to match the isolated formants to the words from which they originated is

TABLE IV. Pairing of the midfrequency components with the three-tone complexes. "Yes" trials contained a midfrequency component and the three-tone complex from which it originated. "No" trials contained a midfrequency component and a neighboring three-tone complex.

Midfrequency component	Three-tone complex					
	1	2	3	4	5	6
2	No	Yes	No			
3		No	Yes	No		
4			No	Yes	No	
5				No	Yes	No

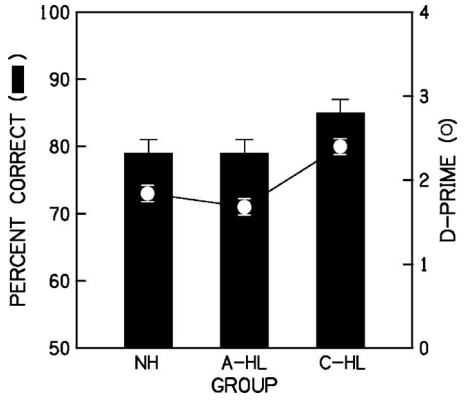


FIG. 5. Average maximized performance and d' values for the speech task as a function of group. Performance (bars) is referenced to the left axis and d' (circles) is referenced to the right axis. All error bars are +1 SE. Groups are listeners with normal hearing (NH), adult-onset hearing loss (A-HL), and childhood hearing loss (C-HL).

more apparent from the results for each talker. Figure 6 shows d' and percent correct performance as a function of group. The parameter in this figure is talker. These data show that performance was lower for the child talker relative to the adult talkers. A repeated measures ANOVA confirmed a significant main effect of talker ($F_{2,54}=29.364$; $p<0.001$); but no group × talker interaction was revealed ($F_{4,54}=1.299$; $p=0.282$). Recall that perceptual coherence was expected to be strongest for the male talker and poorest for the child talker based on the differences in their fundamental frequencies. However, this was not the case. Instead, the perceptual coherence of all three groups was greatest for the child talker.

Finally, several Pearson correlation coefficients were calculated to determine the relation between a listener's ability to match formants to the words from which they originated (d') and his/her age, hearing threshold, and sensation level. Significance was defined as $p<0.01$. Although a significant correlation was observed between age and hearing threshold at 4000 Hz ($r=0.66$ left ear, $r=0.68$ right ear) and at 8000 Hz ($r=0.66$ left ear, $r=0.60$ right ear), no significant correlation was observed between d' and age ($r=-0.30$, $p=0.102$). These results are consistent with decreasing hearing

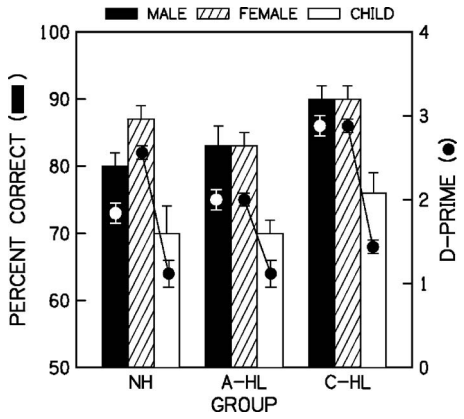


FIG. 6. Same convention as in Fig. 5 but with the parameter of talker. Closed, hatched, and open bars represent the male, female, and child talkers, respectively.

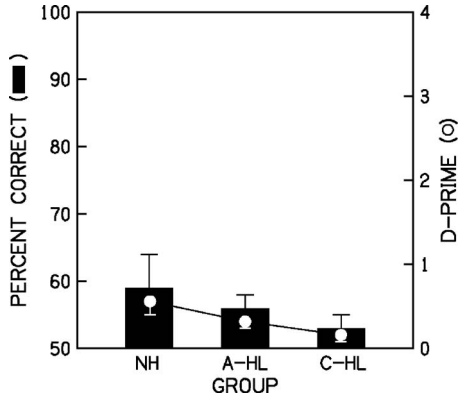


FIG. 7. Same convention as Fig. 5 but for the amplitude modulated sinusoid stimuli.

sensitivity with increasing age and confirm that the rather large difference in ages across the groups (nearly 40 years) did not contribute significantly to performance. A significant correlation was also observed between d' and hearing threshold in left and right ears at 1000 Hz but the relation was relatively weak ($r=0.47$ left ear, $r=0.48$ right ear). These results suggest that the listener's level of performance increased somewhat as hearing loss increased at 1000 Hz. However, because the thresholds at this frequency differed most across groups, the significant correlation likely reflects the overall effect of group rather than hearing sensitivity. Finally, the relation between d' and stimulus sensation level at each of the audiometric frequencies (250 through 8000 Hz) was calculated for the listeners with hearing loss only. No significant correlation was found for any frequency in either ear. These results confirm that the difference in d' between the listeners with A-HL and C-HL groups was not due to the amplification parameters they received.

C. Task 3: Coherence of amplitude modulated sinusoids

d' and maximized performance was also calculated for the amplitude modulated sinusoids. Average maximized performance (bars re: left axis) and d' (open circles re: right axis) are displayed as a function of group in Fig. 7. The error bars represent ±1 SE. On average, all three groups performed at chance levels. A one-way ANOVA of the d' data revealed no significant difference between the groups ($F_{2,29}=1.132$; $p=0.337$). These data suggest strong perceptual coherence for these stimuli in that none of the listeners were able to hear the midfrequency component in the same manner as the second formant of the speech stimuli.

IV. DISCUSSION

The purpose of this study was to examine the perceptual coherence for both speech and nonspeech stimuli in listeners with hearing loss and in listeners with normal hearing. Overall, the results suggested that perceptual coherence differs for speech and nonspeech stimuli and is influenced by the onset of hearing loss more so than the hearing loss itself. Specifically, the perceptual coherence for the speech stimuli in listeners with adult-onset hearing losses was equivalent to that

of the listeners with normal hearing. These results suggest that perceptual coherence remains even in the presence of a condition that is associated with significant peripheral abnormalities (e.g., wide auditory filters, reduced temporal resolution, poor frequency selectivity). In contrast, the listeners with childhood hearing losses showed significantly weaker perceptual coherence for speech (they were better able to hear isolated formants) suggesting that the presence of hearing loss early in life affected their performance. The results for the nonspeech stimuli, on the other hand, revealed similarly strong perceptual coherence across groups. That is, all three groups were equally poor at identifying a single acoustic component within a three-tone complex.

These results are difficult to reconcile with those of previous studies for two reasons. First, a consistent theme throughout much of the research in this area is the idiosyncratic performance of listeners with hearing loss (Grose and Hall, 1996; Rose and Moore, 1997). In the present study, the only significant departure from the norm was observed for the listeners with childhood hearing losses. It is worth noting that similar variation in performance would have been observed if the hearing-impaired listeners had been combined into a single group as is often the convention for studies in this population. Second, attempts have been made to relate the performance of listeners with hearing loss to the peripheral abnormalities that accompany cochlear lesions (e.g., wide auditory filters, reduced temporal resolution, poor frequency selectivity). However, no direct relationship has been found (Hall and Grose, 1989; 1994; Kidd *et al.*, 2002). As for the present study, the results for the listeners with childhood hearing losses are even more difficult to explain on the basis of peripheral abnormalities. That is, it is unlikely that these abnormalities served to improve their ability to discriminate the individual acoustic components within the speech stimuli. Although one might speculate that the peripheral abnormalities associated with childhood hearing loss differ from those of acquired hearing loss, a direct examination of the psychophysical characteristics of listeners with early versus late onset hearing loss would provide valuable insight into this issue.

It is also possible that experience with speech perception contributed substantially to each group's performance. Although listeners with adult-onset hearing loss no longer receive a fully intact speech signal, their lengthy experience with speech processing may allow them to compensate for signal degradation, particularly in highly contextual situations. The same may be true for perceptual coherence. As for the listeners with childhood hearing losses, they were expected to demonstrate significantly stronger perceptual coherence based on the assumption that their ability to utilize the subtle acoustic elements of speech is underdeveloped. This hypothesis was motivated by the findings of Nittrouer and Crowther (2001), who reported stronger perceptual coherence in normally hearing children than in adults. For the present study, however, the listeners with childhood hearing losses were better able to hear specific acoustic elements within the speech signal than either of their normal-hearing and hearing-impaired counterparts. These results may reflect delayed, arrested, or impaired perceptual development. However, this interpretation should be considered speculative until the characteristics and development of perceptual coherence in children are further defined.

Recall that differences in perceptual coherence were expected across talkers on the basis of fundamental frequency (96, 180, and 227 Hz for the male, female, and child talkers, respectively). This expectation was based on the results of Carrell and Opie (1992), who reported significantly improved intelligibility for sine-wave sentences that were amplitude comodulated at low frequencies (50 and 100 Hz) relative to the same sentences comodulated at a higher frequency (200 Hz). This aspect of the study was motivated by a desire to determine whether or not the effects observed using synthetic replicas of speech may be generalized to naturally produced speech. The results of this study suggest that they do not generalize well. By way of example, had fundamental frequency played a substantial role in perceptual coherence, then the results for the female and child talkers should have been more similar given their proximity in frequency (within 50 Hz). Yet the poorest perceptual coherence occurred for the female talker while the strongest perceptual coherence occurred for the child talker. These results suggest that other acoustic/phonetic characteristics likely contribute to perceptual coherence and that fundamental frequency may only play a minor role.

Finally, because little evidence is available regarding the practical implications of perceptual coherence on a listener's ability to perceive speech, the effects of relatively weak or strong coherence can only be speculated. Some have argued that perceptual coherence may assist a listener to perceive speech in noise (Carrell and Opie, 1992; Gordon, 1997a, b). If so, the listener with childhood hearing loss demonstrating weak perceptual coherence may then be expected to have more difficulty perceiving speech in noise than listeners with adult-onset hearing losses. However, the latter group is better known for their complaints regarding difficulty in noise. To complicate the argument, the listeners with adult-onset losses demonstrated perceptual coherence similar to that of the normal-hearing listeners who generally have no complaints regarding noise. Even so, it may be the case that listeners with childhood hearing losses experience more difficulty in noise but they simply have no reference to normal hearing and therefore have no reason to suspect that what they hear is any different than what others hear. We do know that children with even the mildest hearing losses are at risk for poor speech perception, academic performance, and social development in typical classroom noise (Davis *et al.*, 1986; Crandell, 1993; Briscoe *et al.*, 2001). We also know that the vocabulary development of children with hearing loss tends to be delayed relative to that of children with normal hearing by as much as 2 years, and that the delay persists throughout childhood and increases with increasing severity of hearing loss (Briscoe *et al.*, 2001; Pittman *et al.*, 2005). Unfortunately, there are no data regarding the outcomes of children with hearing loss when they mature and are absorbed into the much larger population of adults with acquired hearing losses. Because the results of the present study suggest that perceptual coherence in adults is detrimentally affected when hearing loss is present early in life, it is possible that other

perceptual processes are also affected over the long term. Therefore, it may be wise to examine directly the effects of childhood hearing losses in adults and to consider hearing history as a potential source of variance in all studies involving adults with hearing loss.

ACKNOWLEDGMENTS

Gratitude is extended to Christina Sergi, Ann Hickox, Dawna Lewis, and Brenda Hoover for their help with data collection, Chad Rotolo for the computer software, and Pat Stelmachowicz for her input during the early development of this project. Also, two anonymous reviewers provided many substantive and editorial comments that served to improve the paper substantially. This work was supported by a Grant from NIDCD (No. RO3DC06573).

American National Standards Institute (**1997**). "Methods for calculation of the speech intelligibility index," (ANSI S3.5-1997), New York.

Barker, J., and Cooke, M. (**1999**). "Is the sine-wave speech cocktail party worth attending?," Speech Commun. **27**, 159–174.

Best, C. T., Studdert-Kennedy, M., Manuel, S., and Rubin-Spitz, J. (**1989**). "Discovering phonetic coherence in acoustic patterns," Percept. Psychophys. **45**, 237–250.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Briscoe, J., Bishop, D. V., and Norbury, C. F. (**2001**). "Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment," J. Child Psychol. Psychiatry **42**, 329–340.

Carrell, T. D., and Opie, J. M. (**1992**). "The effect of amplitude comodulation on auditory object formation in sentence perception," Percept. Psychophys. **52**, 437–445.

Crandell, C. C. (**1993**). "Speech recognition in noise by children with minimal degrees of sensorineural hearing loss," Ear Hear. **14**, 210–216.

Davis, J. M., Elfenbein, J., Schum, R., and Bentler, R. A. (**1986**). "Effects of mild and moderate hearing impairments on language, educational, and psychosocial behavior of children," J. Speech Hear Disord. **51**, 53–62.

de Cheveigne, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.

Eisenberg, L. S., Dirks, D. D., and Bell, T. S. (**1995**). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," J. Speech Hear. Res. **38**, 222–233.

Gordon, P. C. (**1997a**). "Coherence masking protection in speech sounds: The role of formant synchrony," Percept. Psychophys. **59**, 232–242.

Gordon, P. C. (**1997b**). "Coherence masking protection in brief noise complexes: Effects of temporal patterns," J. Acoust. Soc. Am. **102**, 2276–2283.

Grose, J. H., and Hall, J. W. (**1993**). "Comodulation masking release: Is comodulation sufficient?," J. Acoust. Soc. Am. **93**, 2896–2902.

Grose, J. H., and Hall, J. W. (**1996**). "Cochlear hearing loss and the processing of modulation: Effects of temporal asynchrony," J. Acoust. Soc. Am. **100**, 519–527.

Hall, J. W., and Grose, J. H. (**1989**). "Spectrotemporal analysis and cochlear hearing impairment: Effects of frequency selectivity, temporal resolution, signal frequency, and rate of modulation," J. Acoust. Soc. Am. **85**, 2550–2562.

Hall, J. W., and Grose, J. H. (**1994**). "Signal detection in complex comodulated backgrounds by normal-hearing and cochlear-impaired listeners," J. Acoust. Soc. Am. **95**, 435–443.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.

Kidd, G., Arbogast, T. L., Mason, C. R., and Walsh, M. (**2002**). "Informational masking in listeners with sensorineural hearing loss," J. Assoc. Res. Otolaryngol. **3**, 107–119.

Marshall, L., and Jesteadt, W. (**1986**). "Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures," J. Speech Hear. Res. **29**, 82–91.

McFadden, D. (**1987**). "Comodulation detection differences using noise-band signals," J. Acoust. Soc. Am. **81**, 1519–1527.

Nittrouer, S., and Crowther, C. S. (**2001**). "Coherence in children's speech perception," J. Acoust. Soc. Am. **110**, 2129–2140.

Pittman, A. L., Lewis, D. E., Hoover, B. M., and Stelmachowicz, P. G. (**2005**). "Rapid word-learning in normal-hearing and hearing-impaired children: Effects of age, receptive vocabulary, and high-frequency amplification," Ear Hear. **26**, 619–629.

Remez, R., and Rubin, P. E. (**1990**). "On the perception of speech from time-varying acoustic information: Contributions of amplitude variation," Percept. Psychophys. **48**, 313–325.

Remez, R. E., Pardo, J. S., Piorkowski, R. L., and Rubin, P. E. (**2001**). "On the bistability of sine wave analogues of speech," Psychol. Sci. **12**, 24–29.

Rose, M. M., and Moore, B. C. (**1997**). "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **102**, 1768–1778.

Seewald, R. C., Cornelisse, L. E., Ramji, K. V., Sinclair, S. T., Moodie, K. S., and Jamieson, D. G. (**1997**). *DSL v4.1 for Windows: A Software Implementation of the Desired Sensation Level (DSL[i/o]) Method for Fitting Linear Gain and Wide-Dynamic-Range Compression Hearing Instruments. User's Manual*, Hearing Healthcare Research Unit, University of Western Ontario, London, Ontario, Canada.

Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects[a)]

Helen E. Cullington[b)] and Fan-Gang Zeng
Hearing and Speech Laboratory, University of California, Irvine, 364 Med Surge II, Room 315, Irvine, California 92697

(Received 30 June 2006; revised 12 October 2007; accepted 12 October 2007)

Cochlear-implant users perform far below normal-hearing subjects in background noise. Speech recognition with varying numbers of competing female, male, and child talkers was evaluated in normal-hearing subjects, cochlear-implant users, and normal-hearing subjects utilizing an eight-channel sine-carrier cochlear-implant simulation. Target sentences were spoken by a male. Normal-hearing subjects obtained considerably better speech reception thresholds than cochlear-implant subjects; the largest discrepancy was 24 dB with a female masker. Evaluation of one implant subject with normal hearing in the contralateral ear suggested that this difference is not caused by age-related disparities between the subject groups. Normal-hearing subjects showed a significant advantage with fewer competing talkers, obtaining release from masking with up to three talker maskers. Cochlear-implant and simulation subjects showed little such effect, although there was a substantial difference between the implant and simulation results with talker maskers. All three groups benefited from a voice pitch difference between target and masker, with the female talker providing significantly less masking than the male. Child talkers produced more masking than expected, given their fundamental frequency, syllabic rate, and temporal modulation characteristics. Neither a simulation nor testing in steady-state noise predicts the difficulties cochlear-implant users experience in real-life noisy situations.
 [DOI: 10.1121/1.2805617]

PACS number(s): 43.71.Ky, 43.71.Gv, 43.66.Dc, 43.66.Sr [KWG] Pages: 450–461

I. INTRODUCTION

Speech recognition in background noise depends on the properties of the interfering sounds. It is usually characterized in terms of the subject's speech reception threshold (SRT): signal-to-noise ratio (SNR) at which they score 50% correct. In test situations, a steady-state masking noise is often used. However, fluctuating backgrounds are much more common in real life, and most often speech is heard against a background of other speech. Although in normal-hearing subjects the SRT decreases when the masker has temporal fluctuations, most hearing-impaired subjects show very little difference for a fluctuating and steady-state masker (Drullman and Bronkhorst, 2004; Duquesnoy, 1983; Festen and Plomp, 1990; Hawley *et al.*, 2004; Miller, 1947; Peters *et al.*, 1998; Summers and Molis, 2004; Wagener and Brand, 2005), including those using a cochlear implant (Zeng *et al.*, 2005). Cochlear-implant users may even show negative effects of modulated maskers (Nelson *et al.*, 2003). Normal-hearing subjects appear to be able to take advantage of listening in the gaps which occur when the level of the competing speech is low, for example in pauses between words, or during the production of low-energy sounds like m, n, k, or p (Peters *et al.*, 1998). This allows brief glimpses of the target speech and leads to improved SRTs. Hearing-impaired subjects are usually unable to utilize glimpsing; this discrepancy is not due to inaudibility (Summers and Molis, 2004) or subject age (Festen and Plomp, 1990). Suprathreshold differences like reduced frequency selectivity may be involved (Peters *et al.*, 1998). It has been suggested that cochlear-implant users' difficulty understanding speech in modulated noise may be related to reduced spectral information (Fu *et al.*, 1998) and problems fusing auditory information across temporal gaps (Nelson and Jin, 2004).

Normal-hearing subjects can also use differences in the voice fundamental frequency (F0) between target and masker to help segregate competing voices, resulting in better speech recognition when the F0 of the target voice differs from that of the masker voice (Brokx and Nooteboom, 1982; Brungart, 2001; Brungart *et al.*, 2001; Drullman and Bronkhorst, 2004). No such effect has been seen in cochlear-implant users (Stickney *et al.*, 2007; Stickney *et al.*, 2004) or normal-hearing subjects using a cochlear-implant simulation (Qin and Oxenham, 2003; Qin and Oxenham, 2005; Stickney *et al.*, 2007; Stickney *et al.*, 2004). The speech processing method only weakly encodes the F0, so it may be difficult to segregate voices on this basis, despite reasonably good F0 difference limens (around one semitone or less) in cochlear-

[a)]Portions of this work were presented in "Two's company; three's a crowd. Speech recognition with competing talkers: normally-hearing, cochlear implant and CI simulation subjects," American Auditory Society meeting, Arizona, March 2006.

[b)]Author to whom correspondence should be addressed. Electronic mail: hculling@uci.edu

implant simulation subjects with access to eight or more spectral bands (Carroll and Zeng, 2007; Qin and Oxenham, 2005).

Masking can be broadly divided into two types. *Energetic masking* results from competition between target and masker at the auditory periphery, i.e., overlapping excitation patterns in the cochlea or auditory nerve. *Informational masking* can be defined as the elevation of threshold caused by stimulus uncertainty (Durlach *et al.*, 2003). In the case of a speech target, this would suggest that the interfering talker is intelligible and so similar to the target speech that it becomes difficult for the subject to disentangle target and interfering speech. Energetic masking is believed to be a purely peripheral phenomenon; informational masking is thought to be related to central or attentional mechanisms (Durlach *et al.*, 2003; Watson and Kelly, 1981). The effects of purely energetic masking are well documented, and can be predicted using models such as the Speech Intelligibility Index or Articulation Index (French and Steinberg, 1947).

Informational masking is difficult to predict and document. When other talkers mask speech, there is probably a combination of energetic and informational masking occurring. One method that has been used to separate the two types of masking is reversed speech. When speech is time reversed, its long-term spectral content and the F0 remain unchanged; however, it contains no linguistic information above the phoneme level and thus should cause limited informational masking. Reversal of the temporal envelope though may increase forward masking due to the abrupt offsets (Rhebergen *et al.*, 2005). Some studies have shown that speech recognition is better in the presence of reversed compared with forward maskers (Rhebergen *et al.*, 2005; Summers and Molis, 2004; Trammell and Speaks, 1970). Duquesnoy (1983), however, found negligible difference. Another method to study informational masking is to minimize spectral overlap between the signal and masker, thus eliminating energetic masking. This can be done by presenting speech stimuli into nonoverlapping bands (Arbogast *et al.*, 2002). Brungart and colleagues specifically examined the role of informational masking in speech recognition in normal-hearing subjects. Significant differences in performance were found between talker maskers and noise maskers leading to the conclusion that, although energetic masking occurred, informational masking dominated performance (Brungart, 2001; Brungart *et al.*, 2001). Drullman and Bronkhorst (2004) had assumed that informational masking would reduce with more interfering talkers, until the SRT approached that for steady-state noise. This hypothesis was based on the idea that the spectral and temporal modulations in the masking signal would diminish with increasing numbers of talkers, and eventually approach the dynamics of steady-state noise. However, even with eight interfering talkers, they found poorer SRTs than for steady-state noise. Carhart *et al.* (1975) found that even 64 competing talkers gave more masking that steady-state noise, although informational masking was at its maximum with three competing talkers and thereafter decreased.

The aim of the current research was to investigate the performance of cochlear-implant users in real-life listening situations, in comparison to normal-hearing subjects. Speech recognition was measured in the presence of background talkers as a function of the number and characteristics of the competing voices. Target and maskers originated from the same location so that spatial release from masking was not considered; Arbogast *et al.* (2005) are among several researchers who have conducted work in this field. In addition, most cochlear-implant users listen with just one ear and therefore may be unable to exploit spatial release from masking. Three experiments were performed. The aim of the first experiment was to assess to what extent cochlear-implant users can obtain release from masking due to temporal and spectral fluctuations in the masker. This was done by examining the influence of masker type on the SRT using combinations of female, male, and child talkers, and steady-state noise as maskers. Normal-hearing and cochlear-implant simulation subjects were also evaluated as a control. The simulation subjects were included in an attempt to compensate for the disparity in age and other characteristics between the normal-hearing and cochlear-implant subjects. It is acknowledged, however, that a simulation does not exactly mimic the performance of cochlear-implant subjects, due to inherent differences between acoustic and electric stimulation. Results therefore should be viewed in terms of trends, rather than a quantitative estimate of cochlear-implant performance (Throckmorton and Collins, 2002). Additional results were collected on one implant subject who has virtually normal hearing in the contralateral ear. Comparison of his results between ears reflects only hearing capabilities and removes the effect of subject characteristics. In order to assess the influence of informational masking on the SRT, a second experiment was performed whereby normal-hearing subjects were tested with one and two talker maskers using both forward and time-reversed masker sentences. This was done in an attempt to resolve the conflicting results obtained by previous authors (Duquesnoy, 1983; Rhebergen *et al.*, 2005; Summers and Molis, 2004; Trammell and Speaks, 1970). The third experiment investigated further the masking effectiveness of a child's voice in normal-hearing subjects. Although previous research has used children as subjects in informational masking of speech experiments (Hall *et al.*, 2002; Johnstone and Litovsky, 2006), results have not been reported using children's voices as maskers. Results were examined in relation to F0, syllabic rate, and temporal modulation rate of the talkers.

II. EXPERIMENT 1: EFFECT OF MASKER TYPE ON THE SRT IN NORMAL-HEARING, COCHLEAR-IMPLANT, AND COCHLEAR-IMPLANT SIMULATION SUBJECTS

A. Methods

1. Test material

In all three experiments, the target material consisted of sentences drawn from the HINT database, spoken by a male talker. These comprise 25 phonemically balanced lists of ten sentences, with each sentence containing between three and seven words (mean=5.3, mode=5 words) (Nilsson *et al.*, 1994). The HINT sentences were designed to be scored as

correct if the subject repeats all of the words exactly correct, with the exceptions of article confusion (a/the/an) and tense for the verbs "to be" or "to have" (is/was, has/had, etc.). However, preliminary investigation with cochlear-implant subjects showed that most did not repeat the exact sentence word for word, even if they appeared to have clearly understood it. This may be a function of age or hearing impairment. Therefore, a loose keyword scoring method was adopted. The HINT sentences were developed from the Bamford-Kowal-Bench (BKB) sentences (Bench *et al.*, 1979), and most of the sentences are almost identical between the two databases. The BKB sentences are commonly scored for keywords (Bench and Bamford, 1979); it was therefore relatively easy to designate keywords for the HINT sentences. Three to five keywords (mean=3.3, mode=3 keywords) were identified for each sentence. In common with criteria often used for BKB sentences, if two or more keywords were repeated correctly, the sentence was considered correct (Blandy and Lutman, 2005). Loose keyword scoring was used, meaning that if the subject repeated the root of the keyword correctly, this would be considered correct; precise inflexion or word ending were not required. Loose keyword scoring is easier to apply, especially if there are difficulties understanding precisely the speech of the test subject (Foster *et al.*, 1993). No target sentence lists were repeated during the test session, as recommended to avoid familiarity (Foster *et al.*, 1993; Wagener and Brand, 2005). All test material was digitized with a sampling rate of 44.1 kHz, and comprised mono 16 bit resolution wav files.

Twenty different maskers were available to compete with the target talker; they were selected to represent various real-life competing talker situations. These are shown in Table I. The names of the masker conditions were abbreviated to represent the constituent talkers, for example, "m2f2" represented two males and two females. The abbreviations are listed in Table I. Different combinations of maskers were used in each experiment. The talker maskers comprised various combinations of ten different voices: two females, two males, and six children. The first female talker and both the male talkers were obtained from the IEEE sentence material (IEEE, 1969) (used with permission from the Sensory Communication Group of the Research Laboratory of Electronics at MIT). Each spoke 40 different sentences. The IEEE sentences are typically longer, and use more complex language than the HINT sentences. The second female talker spoke 30 of the IEEE sentences; this recording was obtained from and used with permission from Ruth Litovsky at University of Wisconsin, Madison. The third female and male talkers (used only for condition m3f3) were the same as the first female and male talkers; as sentence choice was randomized, they were very unlikely to speak the same sentence.

The child talkers were obtained from the Carnegie Mellon University (CMU) Kids Corpus (Eskenazi, 1996; Eskenazi and Mostow, 1997); this is a large database of sentences read aloud by children. Six child talkers were included; they were labeled child-A to child-F. The details of the children used are shown in Table II. Although the database contains hundreds of sentences, only those spoken fluently, without hesitation, mistakes, or extraneous noise were included. This meant that for some children very few sentences were available.

TABLE I. Masking material used in the three experiments, including the abbreviations utilized in this paper. Each experiment used only a subset of the maskers, due to the limited target material available; the conditions used are indicated by 'X'. All talker maskers spoke sentences. The adults spoke IEEE sentences; the children spoke sentences from the CMU Kids Corpus.

No. of talkers	Masker	Abbreviation	Expt 1	Expt 2	Expt 3
1	female	f	×	×	
	6 different children	child-A to child-F	child-E only		×
	male	m	×	×	
2	2 females	f2	×	×	
	1 male and 1 female	m1f1	×	×	
	2 males	m2	×	×	
	2 children	2ch			×
3	1 male and 2 females	m1f2	×		
	2 males and 1 female	m2f1	×		
	3 children	3ch			×
4	2 males and 2 females	m2f2	×		
	4 children	4ch			×
6	3 males and 3 females	m3f3			×
	6 children	6ch			×
—	steady-state noise	noise	×		

The sentences used for the single-talker maskers were selected such that they would have greater duration than the longest HINT sentence, ensuring that no part of the target sentence would be presented in quiet. All sentence material (including target HINT sentences) was edited digitally so that there were minimal silent periods at the start and end of each sentence.

An eight-channel sine-carrier cochlear-implant simulation was implemented in MATLAB® (The MathWorks, Inc.). The signal was first split into eight logarithmically spaced frequency bands from 80 to 8000 Hz, using eighth-order Butterworth bandpass filters. The amplitude envelope was extracted from each band by full-wave rectification and low-pass filtering with a cutoff frequency of 160 Hz. The enve-

TABLE II. Details of child maskers. Child talkers were obtained from the CMU Kids Corpus. They read aloud from grade-appropriate Weekly Reader Stories. Only sentences spoken fluently without mistakes, hesitation, or background noise were included.

Masker	Sex	Age (years)	School grade	No. sentences used
child-A	female	8	3	8
child-B	male	8	2	41
child-C	female	8	2	9
child-D	male	8	2	15
child-E	female	9	3	13
child-F	female	7	1	10

lope in each band was used to modulate a sine wave carrier whose frequency was equal to the band's center frequency. The modulated signal was filtered again using the original analysis filters to ensure that the amplitude-modulated signal had the same bandwidth. The bands were then summed to produce an eight-channel cochlear-implant simulation (Shannon *et al.*, 1995).

In Experiment 1, ten masker conditions were used: f, child-E, m, f2, m1f1, m2, m1f2, m2f1, m2f2, and steady-state noise. The steady-state noise was a 3 s sample spectrally matched to the average long-term spectrum of the HINT sentences (Nilsson *et al.*, 1994), ensuring that on average the SNR was approximately equal at all frequencies. For the cochlear-implant simulation testing, all maskers were preprocessed with the same eight-channel sine-carrier simulation program used for the target sentences. The target and masker materials were individually processed and then added, to allow real-time variation of signal-to-noise ratio during the test. Although in real-life situations the target and noise mix and are then processed together by the cochlear implant, this method was not used with the simulation as it would have introduced a few seconds processing delay before each stimulus.

2. Subjects

Normal-hearing subjects were undergraduates at UC Irvine. They chose to participate in the experiment in order to receive course credit. All subjects had hearing threshold levels within normal limits (≤20 dB HL re ANSI-1996 for octave frequencies between 0.25 and 8 kHz), reported no history of hearing problems, and stated that English was their native language. The subjects were naïve subjects for the HINT sentences. Each subject was included in only one experiment; a total of 38 normal-hearing people participated in the three experiments. All subjects signed an informed consent. The study protocol was approved by the UC Irvine Institutional Review Board.

a. Normal-hearing subjects. Six females and one male with ages ranging from 18 to 21 years (mean=20 years) participated in Experiment 1.

b. Cochlear-implant subjects. Five females and two males with ages ranging from 49 to 80 years (mean =69 years) participated. They were all regular participants in experiments in our laboratory and others; they had all most likely been exposed to the HINT sentences on some or many occasions. They received payment for their participation and had their travel expenses reimbursed. All subjects were postlingually deafened and were experienced users of the Nucleus® or Advanced Bionics Clarion® cochlear-implant device (five had Nucleus 24, one had Nucleus 22, one had CII). They listened with their usual speech processor (SPrint, ESPrit 3G, Spectra 22, or Auria), without a contralateral hearing aid.

c. Simulation subjects. Four female and three male normal-hearing subjects with ages ranging from 18 to 22 years (mean=20 years) participated.

d. Subject CINH001. This subject has a Clarion® HiRes 90k cochlear-implant in his right ear, and virtually normal hearing in his left ear (≤20 dB HL re ANSI-1996 for octave frequencies between 0.25 and 8 kHz, except 35 dB HL at 4 kHz). He received an implant due to intractable tinnitus and is 46 years old.

3. Procedure

A MATLAB® program, developed by the first author, was used to present and score the sentences. Testing was done in quiet or in noise, with a choice of 20 maskers (as shown in Table I). The target and masker were added digitally, and the root mean square (rms) level in dB sound pressure level was adjusted so that all maskers were at a constant intensity regardless of the number of talkers involved. Testing took place in a sound-treated audiometric booth, with the subject sitting approximately 1 m from a loudspeaker placed at 0° azimuth. The operator was also inside the booth at a computer terminal, scoring the subject's responses and running the test. All test material was presented in the sound field except for testing for subject CINH001, as described later. Testing took approximately 1 h, with an option for a break if required.

Testing began with at least one list of HINT sentences presented in quiet at a rms level of 60 dB(A). The quiet testing allowed the subject to become accustomed to the sound of the target talker's voice. The rms level of the target remained at 60 dB(A) throughout the testing; the masker intensity was adjusted to create the appropriate SNR. Using a fixed target level avoids presenting target stimuli at intensities where compression occurs; the cochlear-implant device has a limited input dynamic range (Stickney *et al.*, 2004). Wagener and Brand (2005) found no significant difference in the SRT for normal-hearing or hearing-impaired subjects whether the target level was held constant and the masker level varied or vice versa, although their experiment used only steady-state noise. Two sentence lists (20 sentences total) were used for each masking condition; the pair of lists used was selected at random.

A one-up, one-down adaptive procedure was used to estimate the subject's SRT. The initial SNR was −5 dB for normal-hearing subjects, and +5 dB for cochlear-implant users. This procedure, first described by Levitt and Rabiner (1967), is commonly used to ensure observations are concentrated in the region of interest. Initially, the same target sentence was presented repeatedly and the SNR was increased by 8 dB until the subject correctly repeated the sentence; this allowed the program to quickly find the approximate SRT. Once this occurred, the step size was reduced to 4 dB, and the adaptive procedure began, with the SNR decreasing by 4 dB when the subject answered correctly, and increasing by 4 dB when the response was erroneous. The SRT (in dB) was calculated as the mean of the last six reversals. Although the usual HINT step size is 2 dB, it was found that with only 20 sentences presented, cochlear-implant users would produce insufficient reversals with this step size. The masker segment was demonstrated to the subject at the beginning of each condition; they were told to ignore this voice or these voices and listen only to the target male talker. The masker sentence began approximately 0.45 s before the target; both target and masker were presented from the same speaker. An onset difference between masker and target has been used by other

authors (Drullman and Bronkhorst, 2004; Festen and Plomp, 1990; Freyman *et al.*, 2004; Wagener and Brand, 2005); it provides a basis for attending to the target, although in this experiment the subjects were not instructed as such.

Due to the relatively small sample size, the univariate (mixed-model) approach was used for statistical analyses in Experiments 1 and 3, using the Greenhouse–Geiser ε adjustment to control for Type I errors. Experiment 2 had more subjects and therefore used multivariate analysis of variance. In the case of multiple planned comparisons, the observed p value was compared to a critical p value of 0.05/C (where C is the number of planned comparisons) in order to maintain the familywise alpha level at 0.05 using the Bonferroni approach.

Subject CINH001 was tested in two ways: listening with his normal-hearing ear in the sound field while not wearing his cochlear implant, and by direct connection to his cochlear-implant speech processor (to prohibit the use of his normal-hearing ear). For the direct cochlear-implant connection, the level was adjusted to a comfortable listening level.

B. Results and discussion

Five normal-hearing subjects scored 100% for sentences in quiet; two missed one word. Sentence scores in quiet for cochlear-implant users varied from 65 to 100%, with a mean of 84%. Although initially it was considered appropriate only to test in noise if scores in quiet exceeded 90%, the score in quiet was not found to be a good predictor of performance in noise, so noise testing was performed on all subjects. The simulation subjects obtained scores in quiet ranging from 80 to 100% correct, with a mean of 90%. Subject CINH001 scored 100% correct for sentences in quiet with his near normal-hearing ear and 70% correct with his cochlear-implant ear.

1. Effect of masker type

Figure 1 shows the mean SRT for each masker condition for the three groups of subjects and for subject CINH001. Lines joining the points are purely for clarity, and are not suggesting a functional relation, due to the maskers being categorically different. Three initial observations are noteworthy. First, normal-hearing subjects performed vastly better than the implant users on all conditions. The mean discrepancy was 8 dB for steady-state noise, but as much as 24 dB difference with a female masker. Second, the standard deviation across the individual cochlear-implant users' SRTs was generally higher than those for the normal-hearing subjects, reflecting a larger variation in listening performance. Third, the simulation provides a very comparable result to cochlear-implant users for steady-state noise, but there is a large discrepancy for the talker maskers. This eight-channel simulation may not provide a fair representation of implant users' performance in the presence of competing talkers.

A repeated-measures analysis of variance (ANOVA) (with Greenhouse–Geiser adjustment where appropriate) showed a highly significant main effect of group (normal-hearing, cochlear-implant, and cochlear-implant simulation) ($F(2,18)=102.8$, $p<0.0005$), and a highly significant main

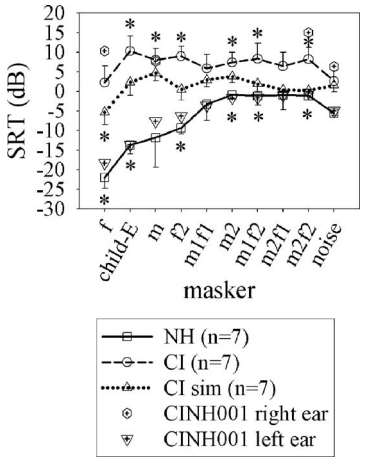


FIG. 1. Mean SRT as a function of masker type in seven normal-hearing, seven cochlear-implant subjects, and seven normal-hearing subjects using an eight-channel sine-carrier cochlear-implant simulation. Target material was HINT sentences spoken by a male. The squares represent the mean SRT for normal-hearing subjects. The circles represent the mean SRT for cochlear-implant subjects. The triangles represent the mean SRT for cochlear-implant simulation subjects. The hexagons and inverted triangles represent the SRTs for the right and left ears, respectively, for subject CINH001, who has a cochlear implant in the right ear and virtually normal hearing in the left ear. Error bars represent one standard deviation. For clarity, only the upward bar is shown for the cochlear-implant users, and only the downward bar for the normal-hearing and simulation subjects. Lines joining the points are purely for clarity, and are not suggesting a functional relation, due to the maskers being categorically different. The asterisks represent a SRT value significantly different from the SRT with a steady-state noise masker.

effect of masker type ($F(4.2,75.8)=36.3$, $p<0.0005$). There was also a highly significant interaction between group and masker type ($F(8.4,75.8)=16.3$, $p<0.0005$), suggesting that the effect of masker type differs in the three groups. Further analysis was therefore performed separately for the three subject groups.

Subject CINH001 was tested over two sessions separated by several months to avoid duplication of target material. At the time of the second test, he had not been using his speech processor for several days and obtained only 30% correct in quiet. At this session, SRTs from 15 to 25 dB were obtained with his implanted right ear for maskers child-E, f2, m1f1, m2, m1f2, and m2f1. Therefore for his implanted ear, only those masker conditions tested in the first session were plotted due to the device nonuse prior to the second session. The SRT results from his normal-hearing ear are almost all within one standard deviation of the mean of those from the normal-hearing young adults. Limited results from his implanted ear fall close to or outside one standard deviation of the mean for the implant subjects. These results suggest that age-related cognitive differences between the normal-hearing and cochlear-implant subject groups are not responsible for the vast differences in the SRT. It is acknowledged, however, that cognitive effects may play a part in some elderly subjects; previous work demonstrated that performance in noise worsened significantly with increasing age (Souza *et al.*, 2007).

A repeated-measures ANOVA was performed to assess nine planned comparisons: the difference between the SRT for the nine masker conditions (f, child-E, m, f2, m1f1, m2, m1f2, m2f1, m2f2) and the SRT for steady-state noise. The

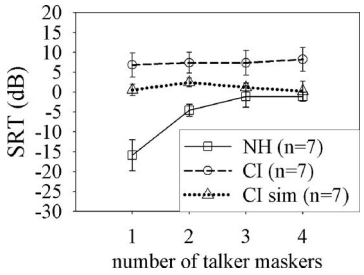


FIG. 2. Mean SRT as a function of number of talker maskers in seven normal-hearing, seven cochlear-implant subjects, and seven normal-hearing subjects using an eight-channel sine-carrier cochlear-implant simulation. Target material was HINT sentences spoken by a male. The squares represent the mean SRT for normal-hearing subjects. The circles represent the mean SRT for cochlear-implant subjects. The triangles represent the mean SRT for cochlear-implant simulation subjects. Error bars represent ± one standard deviation.

analysis was performed separately for each subject group. Those results found to be significantly different from steady-state noise are indicated with an asterisk in Fig. 1. A significant p value was considered to be 0.006, using the Bonferroni correction for multiple comparisons. In addition, for each subject, the SRTs were averaged across conditions f, child-E, and m to obtain a measure of the SRT for a one-talker masker; over conditions f2, m1f1, and m2 for a two-talker condition; over conditions m1f2 and m2f1 for a three-talker condition; and condition m2f2 was used for the four-talker condition. Figure 2 shows mean SRT results as a function of number of interfering talkers. Three planned comparisons were performed to assess the effect of changing from one to two, two to three, and three to four interfering talkers. The p values were compared to 0.017, using the Bonferroni correction for multiple planned comparisons.

a. Normal-Hearing Subjects. All masker conditions gave significantly different SRTs from steady-state noise, except conditions m, m1f1, and m2f1. The SRT for these conditions had the largest standard deviations, probably due to the subject inadvertently following the wrong male talker. This confusion usually happened when these conditions occurred near the beginning of the test, and the subject was unaccustomed to the target talker's voice. The single-talker conditions female and child produced significantly better SRT results than steady-state noise. For two interfering talkers, only f2 gave better SRT results, while m1f1 provided a comparable SRT to steady-state noise, and m2 was worse. With more than two interfering talkers, the SRT was significantly worse than for steady-state noise, except for m2f1 where the SRT was comparable. Considering Fig. 2, there was a significant increase in the mean SRT as the number of interfering talkers changed from one to two ($F(1,6)=45.6$, $p=0.001$) and from two to three ($F(1,6)=18.5$, $p=0.005$). However, as the number of interfering talkers increased from three to four, there was not a significant change in the mean SRT ($F(1,6)<0.0005$, $p=0.991$). This suggests that once there are three interfering talkers, the inclusion of one additional talker did not influence speech recognition, although, as discussed later, there must eventually be a drop in the SRT to the level of that with a noise masker, as more and more talkers are included.

The normal-hearing subjects in this study appeared to be able to use temporal and spectral fluctuations in the interferers to obtain release from masking with one or two interfering talkers. With fewer maskers, the subject can take advantage of favorable SNRs in the temporal gaps; as more maskers are introduced, these gaps are filled in and energetic masking increases. However, as the number of talkers increases, informational masking also increases; with three or four interfering talkers, the SRT was generally worse than for steady-state noise. These results agree with those found by Brungart *et al.* (2001) who showed that, at a negative SNR, performance is worse for two or three interfering talkers than for only one. In common with Drullman and Bronkhorst (2004) the authors believe that when there are only one or two interfering talkers, the interfering speech is still intelligible, so grammatical and semantic information in the masker help the subject to pick out the target speech. However, when there are multiple interfering talkers, the subject is able to hear words in the maskers, but because they cannot fully perceive the linguistic structure of the masker, they are unable to take advantage of this to decide whether the words were spoken by target or masker. The SRTs seem to reach a plateau at three talker maskers, suggesting that additional background talkers would not affect the SRT. An early study by Miller (1947) using target words against continuous discourse maskers and very different methodology had shown an increase in masking from two to four masker voices, but no further increase from four to six or six to eight. However, in this study the SRT for four interfering talkers (m2f2) is significantly worse than that for steady-state noise, and if steady-state noise is considered to be the sum of an infinite number of talkers, it appears that at some point the effect of informational masking would decline, and the SRT would decrease again. Carhart *et al.* (1975) though still demonstrated informational masking with 64 competing talkers.

b. Cochlear-Implant Subjects. The maskers f, m1f1, and m2f1 produced comparable SRTs to steady-state noise; all other maskers gave significantly higher SRTs. As shown in Fig. 2, there was no significant difference in the mean SRT related to number of interfering talkers ($F(2.6,15.4)=1.6$, $p=0.213$). The best SRTs were obtained for one female masker and for steady-state noise. Most of the multiple talker maskers gave worse SRTs than steady-state noise. This may be a result of informational masking, or some kind of modulation interference. The cochlear implant users are clearly unable to take advantage of temporal glimpsing. Assessing whether the subjects confused the masker words with the target would indicate the extent of informational masking; error analysis was not done in this study. In common with Qin and Oxenham (2003) these results suggest that testing implant users in steady-state noise may underestimate the difficulties they experience in everyday life.

c. Simulation Subjects. The only masker condition that gave a significantly different SRT from steady-state noise was the female talker; its mean SRT was lower than that of steady-state noise ($F(1,6)=21.4$, $p=0.004$). Considering Fig. 2, as the number of interfering talkers increased from one to two, there was a significant increase in the mean SRT ($F(1,6)=26.2$, $p=0.002$). Significant changes were not seen for two vs three talkers ($F(1,6)=7.0$, $p=0.038$), or three vs four talkers ($F(1,6)=0.9$, $p=0.377$).

Simulation subjects showed little difference in the mean SRT across the masker types except a significantly better result with a female masker. They did not show a pronounced effect of informational masking and seemed unable to make much use of amplitude minima in the temporal

structure of the maskers. This does not agree with simulation results from Qin and Oxenham (2003) who found that simulation subjects performed significantly worse with male or female single-talker maskers than steady-state noise. This simulation used narrowband noise carriers, and the present study used sine carriers, so it is possible that this caused the discrepancy. The current data do show that the SRT for the male masker was worse than that for steady-state noise, and if a p value of 0.05 were used (as used by Qin and Oxenham), then this difference would be significant. As discussed later, the female masker used in Qin and Oxenham's study had an abnormally low F0, so cannot be compared to that used in the current research.

Five subjects from each group were asked to identify the number and gender of talkers in each of the maskers played separately without the target. Although the normal-hearing subjects were able to accurately specify the one, two, and three talker conditions, none was correct for the four-talker condition. All stated that they heard three voices: two male and one female, when in fact there were four voices (two male and two female). (One subject stated that he heard one male talker as two males; this is believed to have been a concentration error). The plateau for masking effectiveness coincided with the limit of talker number that the subjects were able to identify. The cochlear-implant and simulation subject results were variable, but worse than those for normal-hearing subjects. Although all implant and most simulation subjects were able to identify one female or one male talker, only one implant user and three simulation subjects could identify the child's voice. The others reported the child talker to be the female.

Cochlear-implant users clearly performed much worse than normal-hearing subjects when identifying talkers. Voice gender perception is dependent on accurate pitch information, which is lacking in cochlear-implant speech processing. The implant users were able to accurately identify one female or one male talker as found by Fu *et al.* (2005), but identification of the child's voice was very poor.

2. Effect of voice pitch on the SRT

Further analysis was conducted using only the single-talker maskers, in order to evaluate the effect of voice pitch on the SRT. The hypothesis was that more separation between talker and masker F0 would lead to lower (better) SRTs. A modification of the MATLAB® program STRAIGHT was used to extract the F0 for voiced parts of the speech (Kawahara *et al.*, 1999); these were averaged across each sentence. The mean for each sentence was then averaged over all the sentences (250 for the target, 40 for the adult maskers, 13 for the child) providing a single mean value for the target and each of the maskers. These are shown in Fig. 3. The child's voice clearly has the highest F0, followed by the female voice, followed by the two male voices, as expected. Three planned paired comparisons were assessed to evaluate differences in the SRTs for the single-talker maskers, and the observed p value was compared to 0.017 using the Bonferroni correction for multiple comparisons.

For all three groups (normal-hearing, cochlear-implant, and simulation), the mean SRT for the female masker was significantly better than the mean SRT for the child masker, and significantly better than the mean SRT for the male masker. There was not a significant difference between the mean SRT for the child and male maskers. Table III contains the t and p values for the comparisons.

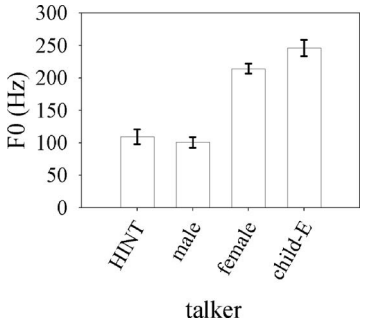


FIG. 3. Mean fundamental frequency (F0) for the target and each of the masker voices. The target HINT sentences were spoken by a male. Error bars represent ± one standard deviation. As expected, the child has the highest F0, followed by the female, then the two males.

The three subject groups behaved in a similar manner on this comparison; they were all able to take advantage of differences in voice F0 to separate the female from the male and child talkers. This result does not agree with other researchers who found that implant (Stickney *et al.*, 2007; Stickney *et al.*, 2004) and simulation users (Qin and Oxenham, 2003; Stickney *et al.*, 2007; Stickney *et al.*, 2004) showed no difference in speech recognition when the gender of the masker was changed. The simulations used were different in two of the studies (Qin and Oxenham, 2003; Stickney *et al.*, 2004): these used noise carriers, but this should not affect the result. The previous studies had more generous low-pass cutoff frequencies (300 and 500 Hz, respectively, for Qin and Oxenham and Stickney *et al.*) compared to the current study's 160 Hz. Qin and Oxenham (2003) felt that one likely reason for their null effect was that their female voice had an atypically low mean F0 of 129 Hz. The current study uses a female masker with F0 of 214 Hz, which is much closer to the average value of 220 Hz (Hillenbrand *et al.*, 1995). In Stickney *et al.*'s 2004 study, although the cochlear-implant subjects performed well in quiet (78%–92% on IEEE sentences), their performance in noise was poor, with the average reaching only around 50% even at 20 dB SNR. The performance of the subjects in the later

TABLE III. t and p values from the paired t tests performed on the mean SRT with female, child, and male masker in normal-hearing, cochlear-implant, and cochlear-implant simulation subjects. Values shown are two tailed, with six degrees of freedom. The p values were compared to 0.017 using the Bonferroni correction for multiple comparisons. All three groups showed a significantly better mean SRT for the female masker than both the child and male maskers. There was no significant difference between the mean SRT for the child and male maskers.

	Normal-hearing	Cochlear-implant	Cochlear-implant simulation
Female/child	t=−21.8	t=−6.0	t=−4.3
	p<0.0005	p=0.001	p=0.005
Female/male	t=−4.8	t=−4.0	*t*=−6.6
	p=0.003	p=0.007	p=0.001
Child/male	t=−8	t=1.7	t=−1.4
	p=0.465	p=0.135	p=0.204

study was even worse (Stickney *et al.*, 2007). This may explain their null finding. The cochlear-implant users in this study performed much better in noise. It is possible that the subjects in the current study are using some characteristic other than the F0 in order to segregate the voices, for example, speaking rate or coarse spectral differences. However, it is shown later that syllabic rate is similar between the female and male maskers.

Considering that previous research has shown that voices are easier to segregate if there is a larger separation between their pitches, one would expect that the child's voice would be the least effective masker. This was not found. In fact no significant difference was seen in the masking effectiveness between the child and the male masker, although their F0 difference is 145 Hz. Brungart *et al.* (2001) suggested that a particularly salient masker could cause the subject's attention to be drawn away from the target phrase; this appears to be the situation here. The characteristics of the child's voice make it more difficult to ignore than would be expected given its spectral qualities; this is further explored in Experiment 3.

III. EXPERIMENT 2: EFFECT OF REVERSED SPEECH MASKERS ON THE SRT IN NORMAL-HEARING SUBJECTS

A. Methods

1. Test material

The HINT sentences were the target material, as described in Experiment 1. Only the single- and two-talker adult maskers were used: f, m, f2, m1f1, and m2. They were played either forward or reversed, making ten different maskers. In order for target material not to be repeated, only these conditions were involved.

2. Subjects

The normal-hearing subjects were as described previously; results were obtained from 12 females and four males, with ages ranging from 18 to 36 years (mean=21 years). None of these subjects had taken part in Experiment 1.

3. Procedure

The procedure was as described in Experiment 1. A within-subjects crossed design was used, with each subject acting as his or her own control; this had the advantage of requiring fewer subjects, although there was the possible disadvantage of differential carryover effects. Differential carryover effects occur when a subject's participation in one part of the experiment affects their performance on a later condition one way, and on a different condition in another way. In contrast, practice effects affect all treatment conditions equally. Test order was randomized for each subject across both factors in an attempt to avoid differential carryover effects.

B. Results and discussion

All subjects scored 100% for sentences in quiet. All 16 subjects were tested in the single-talker masker-forward and masker-reversed conditions. The final seven (six female, one male, age range 18–23 years, mean=20 years) were also tested in the two-talker masker-forward and masker-reversed conditions.

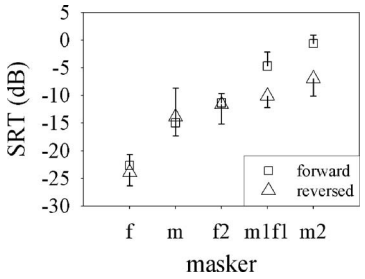


FIG. 4. Mean SRT as a function of masker type and masker reversal in normal-hearing subjects. Target material was HINT sentences spoken by a male. Maskers were one or two female or male talkers played either forward or reversed. The squares represent the mean SRT when masker sentence was played forward. The triangles represent the mean SRT when the masker segment was reversed. Sixteen subjects were evaluated for the single-talker maskers (f and m), seven subjects for the two-talker maskers (f2, m1f1, and m2). Error bars represent one standard deviation. For clarity, only the upward bar is shown for the masker-forward condition, and only the downward bar for the masker-reversed condition.

Mean SRT results are shown in Fig. 4 for the single-talker (16 subjects) and two-talker (seven subjects) masker conditions. For the single-talker maskers no difference is observed between the masker-forward and masker-reversed conditions. (Multivariate analysis of variance $F(1,15)=0.02$, $p=0.896$). For the two-talker conditions, the reversed masker produced better SRTs in conditions m1f1 and m2. A repeated-measures ANOVA showed a significant main effect of reversal ($F(1,6)=25.7$, $p=0.002$). After examining the data, the significance of the difference between m1f1 and m2 for masker-forward or masker-reversed conditions was tested. As this was a post hoc comparison, the critical value used was a multivariate extension of Scheffé's method developed by Roy and Bose (1953). Both comparisons were statistically significant (m1f1: $t=4.6$, $df=6$, $p=0.004$; m2: $t=4.7$, $df=6$, $p=0.003$; critical $t=2.97$). The results in Fig. 4 again show a much larger variance for masker condition m, when the masker was forward. This was presumably because subjects were more likely to pay attention to the wrong talker in this condition.

A significant effect of reversal was only seen for conditions m1f1 and m2: the two-talker conditions that involved a male voice. These results coincided with those of Hawley *et al.* (2004), who used a male target and the same male interferers. Their results did not show a large effect of reversed speech when there was only one interfering talker; however, for two interferers, reversed speech gave consistently better SRTs than forward speech. In common with Drullman and Bronkhorst (2004), it is believed that this occurs because there is minimal informational masking with only one interfering talker as the subject can use the grammatical and semantic information in the masker to segregate it from the target. However, with several interfering talkers, the sentences are not individually intelligible, so the subject cannot take advantage of the grammatical and semantic content. Rhebergen *et al.* (2005) did, however, show a signifi-

cant improvement in the SRT when using a male target and a time-reversed female masker. This result was found in Dutch listeners using Dutch target and masker; differences in the dynamics of this language compared to English may contribute to the discrepancy.

Although time-reversed speech has unchanged spectral content, the temporal envelope is reversed. In forward speech, words usually begin with plosives: quick onset and slow decay. When speech is reversed, it contains abrupt offsets. The auditory system cannot follow these abrupt offsets so accurately, so soft sounds can be masked by a preceding strong signal: forward masking (Rhebergen *et al.*, 2005). When a speech masker is time reversed, one may expect an improvement in the SRT due to the release from informational masking. However, there may also be a decrease in performance due to increased forward masking. The two effects act in opposition, therefore in order to show a release from informational masking using reversed speech, this effect must exceed the opposing increase in forward masking. Rhebergen *et al.* (2005) found the increase in the SRT due to forward masking to be approximately 2 dB. Results from the current study therefore suggest that any release from informational masking that is present with reversed speech in conditions f, m, and f2 is less than or around 2 dB. The release from informational masking for conditions m1f1 and m2 may be approximately 7–8 dB, assuming the 2 dB of forward masking still applies with a multitalker background. The two effects can be separated using a speech masker in a foreign language, thus offering release from informational masking, but not altering the amount of forward masking (Rhebergen *et al.*, 2005).

IV. EXPERIMENT 3: FURTHER INVESTIGATION OF CHILD MASKERS

A. Methods

1. Test material

The target material was the HINT sentences as in Experiments 1 and 2. Subjects were tested with six different child maskers (child-A to child-F), combinations of two, three, four, and six children (2ch, 3ch, 4ch, and 6ch), and six adults (m3f3). Results from Experiment 1 showed that a child's voice had a greater masking effect than expected, given its fundamental frequency. The purpose of Experiment 3 was to further investigate this finding. Six different child talkers were used, in order to rule out the hypothesis that there was an anomalous feature associated with the child's voice used in Experiment 1. Combinations of child maskers were used to replicate and extend the adult talker masker findings shown in Fig. 2. The masking effect of a babble of six child talkers was examined and compared to that of a babble of six adults. It was hypothesized that whatever feature of a child masker that increased its masking effectiveness would disappear once the voices were not individually distinguishable. Although previous work has used children as subjects in masking experiments (Hall *et al.*, 2002), children's voices have not been examined as maskers.

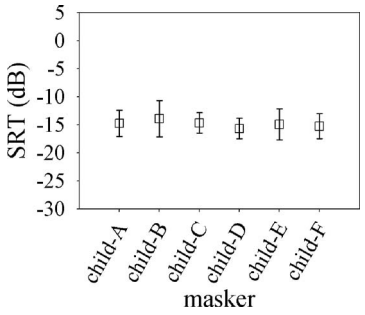


FIG. 5. Mean SRT for six different child maskers in eight normal-hearing subjects. Target material was HINT sentences spoken by a male. Error bars represent ± one standard deviation. Four female and two male children were used, with ages ranging from seven to nine years, with voice fundamental frequencies from 215 to 281 Hz. The mean SRT, however, was approximately the same for all child maskers.

2. Subjects

Eight normal-hearing subjects (five female, three male), with ages ranging from 18 to 23 years (mean=21 years) participated. None of these subjects had taken part in Experiments 1 or 2.

3. Procedure

The procedure was as used in Experiment 2.

B. Results and discussion

Seven subjects scored 100% correct for sentences in quiet; one scored 90% correct for sentences, 98% correct for words.

1. Comparison of the SRT in Six Different Child Maskers

Figure 5 shows the mean SRT for each of the child maskers in eight normal-hearing subjects. A repeated-measures ANOVA using the Greenhouse–Geiser correction showed no significant difference ($F(2.9,20.3)=0.655$, $p=0.585$). This confirms that the finding in Experiment 1 (that the child masker produced more masking than expected) was not simply a result of characteristics of that particular child's voice.

2. Effect of number of interfering child talkers

Figure 6 shows the mean SRT for one, two, three, four, and six child maskers. A repeated-measures ANOVA using

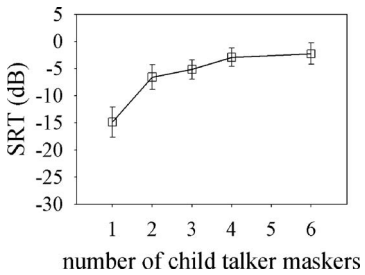


FIG. 6. Mean SRT for one, two, three, four, and six child maskers in eight normal-hearing subjects. Target material was HINT sentences spoken by a male. Error bars represent ± one standard deviation. The mean SRT gradually increases as the number of interfering talkers increases.

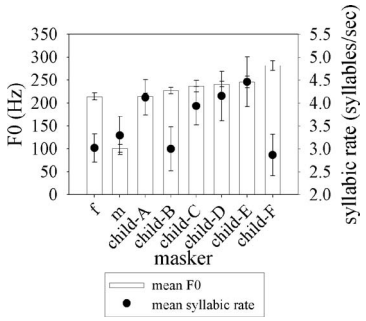


FIG. 7. Mean F0 and syllabic rate for all the single-talker maskers. The bars represent the mean F0 for each talker, using the left-hand axis. The circles represent the means syllabic rate for each talker, using the right-hand axis. Error bars represent ± one standard deviation. The male F0 is significantly lower than all the other maskers. All child maskers except child-A have a significantly higher F0 than the female. There is much variation in syllabic rate, with some children having higher rate than the female, and some comparable.

the Greenhouse–Geiser correction showed a significant effect of the number of talkers ($F(2.8,19.8)=47.3$, $p<0.0005$). Four planned paired comparisons were evaluated to assess the change from one to two, two to three, three to four, and four to six interfering child talkers. The p value was compared to 0.01 using the Bonferroni correction for multiple planned comparisons. The only significant change was from one to two interfering child talkers ($F(1,7)=38.3$, $p<0.0005$); the other changes were too gradual to reach significance. A comparison of the child six-talker SRT to that obtained using steady-state noise in Experiment 1, showed that the child six-talker masker had a significantly greater masking effect than steady-state noise (two-tailed $t=4.3$, $df=13$, $p<0.0005$). This reflects informational masking. The mean SRT with six adults as the masker was −2.7 dB; mean SRT with six child maskers was −2.3 dB. A paired t test showed that this difference was not significant (two-tailed $t=0.424$, $df=7$, $p=0.685$). This demonstrates that the characteristic of a child masker that is providing more masking than expected is not apparent when a babble of child voices is used. A babble of child voices produces essentially the same masking as a babble of adult voices.

3. Effect of child voice pitch, syllabic rate, and temporal modulation on the SRT

The program STRAIGHT implemented in MATLAB® was used to calculate the mean F0 for the six child maskers (Kawahara *et al.*, 1999). The rate of the talkers was also examined. This was simply the mean number of syllables spoken per second; it was evaluated for 15 sentences for female, male, child-B, and child-D, and fewer sentences for the other children (see Table II). Figure 7 shows mean F0 and syllabic rate for all the one-talker maskers. The male F0 is significantly lower than the female (two-tailed $t=63.4$, $df=78$, $p<0.0005$), and all the child maskers except child-A have a significantly higher F0 than the female masker ($p<0.008$ using the Bonferroni correction for multiple comparisons). Even though F0 of the child maskers is higher than the female, the SRT is still significantly worse; this suggests that frequency separation alone is not responsible for the SRT difference. Figure 7 also shows the variation in the syllabic rate of the child talkers. Compared to the female masker, child-A, child-C, child-D, and child-E have a significantly faster syllabic rate ($p<0.008$ using the Bonferroni correction for multiple comparisons). Child-B and child-F have comparable speaking rates to the female talker. It seems that talking rate is not a consistent factor, and therefore is not responsible for the increased masking effectiveness of a child talker.

The authors observed that children's voices often have a "sing-song" characteristic: voice pitch appears to rise and fall more during a sentence than for adult talkers. The F0 calculation using STRAIGHT involves averaging instantaneous F0 across a sentence. In order to evaluate whether the variation in F0 was similar for female, male, and child maskers, the coefficient of variation (standard deviation/mean) was evaluated in each case. The values were 0.16, 0.32, and 0.21 for female, male, and child, respectively. The child's value is between that for the female and male; this rejects the hypothesis that the child's voice is more distracting because of the variation in the F0 across a sentence.

Spectra of the octave band envelope modulations were examined for the target HINT sentences, female, male, and child maskers. Ten sentences of each were concatenated to simulate running speech; no gaps were inserted between the sentences. Following a method described by Payton and Braida (1999) and implemented in MATLAB®, the speech material was first bandpass filtered using octave bandwidth digital Butterworth filters with center frequencies from 125 to 8000 Hz (the 8000 Hz filter was a high-pass filter). The samples were squared, low-pass filtered to extract the intensity envelope, and power spectra were computed. Average spectra were summed to third octave band representations. The square root of this sum was the one third octave modulation spectrum. The modulation index represents depth of modulation. Greater informational masking may occur if temporal modulation properties are similar between target and masker. Figure 8 shows the modulation index difference between each masker and the target as a function of third octave band modulation frequency for female masker, male masker, and child-E. The measurement was made for seven octave bands (125 Hz through 8000 Hz), although the child results at 125 Hz were artifactual and were excluded. Temporal envelope modulations play an important role in speech intelligibility (Shannon *et al.*, 1995). It is difficult to visually assess from these graphs whether any real differences exist in terms of envelope modulation depth, although in the 1000 Hz band it appears that the child masker had slightly more modulation than the other talkers from 3 to 8 Hz. Correlation coefficients between the target and the female, male, and two child maskers (child-B was included because it has a similar syllabic rate to the female) were calculated for each octave band. The values were averaged across frequencies from 250 to 8000 Hz. The mean correlation coefficients were as follows: female 0.709, male 0.670, child-E 0.653, child-B 0.670. These numbers describe how similar the modulation indices of the masker are to the target. There is no consistent trend in these numbers; they do not explain the masking properties of the child maskers. However, this

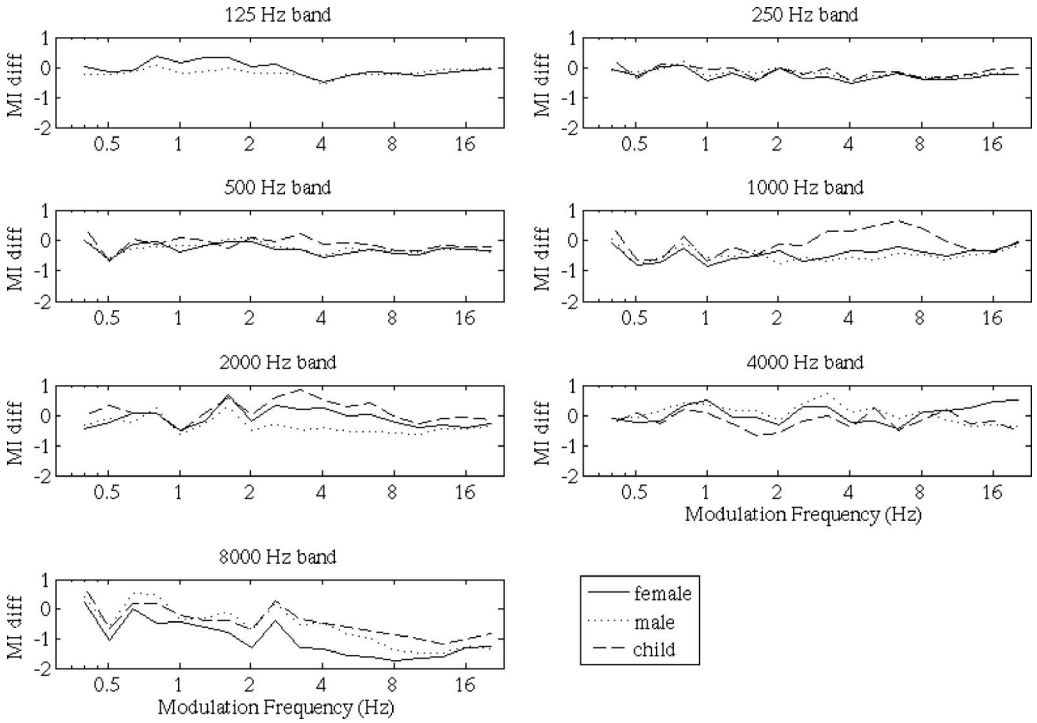


FIG. 8. Temporal envelope modulation data. Graphs show envelope spectra differences between masker and target for seven octave bands, depicted by modulation index difference (MI diff) as a function of third octave band modulation frequency. Data are shown for three maskers: female, male, and child (child-E).

analysis is preliminary, and further research is required. Future work should also examine the speech intelligibility index (SII) of speech masked by a child's voice, using the modification to the SII proposed to predict intelligibility in the presence of a fluctuating masker (Rhebergen and Versfeld, 2005).

V. SUMMARY AND CONCLUSIONS

This research investigated speech recognition in normal-hearing and cochlear-implant subjects, under a variety of talker and noise masker conditions. Normal-hearing subjects performed vastly better than implant users on all conditions; the largest mean discrepancy in the SRT was 24 dB with a female masker. These differences were not caused by age-related cognitive differences in the subject groups. Although an eight-channel sine-carrier cochlear-implant simulation provided an almost identical SRT to cochlear-implant users with a steady-state noise masker, there was a large discrepancy for talker maskers.

Normal-hearing subjects used temporal fluctuations in interferers to obtain release from masking. Cochlear-implant and simulation subjects made much less use of temporal fluctuations. A talker background provides a combination of energetic and informational masking. Results from masker reversal suggested that single-talker maskers produce little informational masking. As the number of talkers increase, both energetic and informational masking increase. Normal-hearing, cochlear-implant, and simulation subjects all showed a significantly better SRT for a female than male masker. Despite the weak representation of voice fundamental frequency in their coding scheme, implant users appeared to use spectral differences in the talkers to segregate the voices.

Although the child maskers had higher voice pitch, all subject groups showed no difference between the mean SRT for the male and child maskers, and a significantly better SRT for the female compared to the child masker. The child maskers possessed greater masking ability than suggested by their spectral qualities; this did not seem to be related to talking rate, variation in the F0 within a sentence, or temporal envelope modulation characteristics.

Clinical cochlear-implant testing generally uses steady-state noise as a masker. The current research suggests that this does not reflect the vast discrepancy between normal-hearing and cochlear-implant subjects in real-life situations with competing talkers. Caution must be exercised when a cochlear-implant simulation is used, as results may reflect implant users' performance in a steady-state noise background, but are discrepant in more realistic listening situations.

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.

Bench, J., and Bamford, J. (**1979**). *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children* (Academic Press, London).

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Ophthamol. **13**, 108–112.

Blandy, S., and Lutman, M. (**2005**). "Hearing threshold levels and speech recognition in noise in 7 year olds," Int. J. Audiol. **44**, 435–443.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.
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.
Carhart, R., Johnson, C., and Goodman, J. (**1975**). "Perceptual masking of spondees by combinations of talkers," J. Acoust. Soc. Am. **58**, S35.
Carroll, J., and Zeng, F. G. (**2007**). "Fundamental frequency discrimination and speech perception in noise in cochlear implant simulations," Hear. Res. **231**, 42–53.
Drullman, R., and Bronkhorst, A. W. (**2004**). "Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers," J. Acoust. Soc. Am. **116**, 3090–3098.
Duquesnoy, A. J. (**1983**). "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," J. Acoust. Soc. Am. **74** 739–743.
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.
Eskenazi, M. (**1996**). "KIDS: A database of children's speech," J. Acoust. Soc. Am. **100**(4), 2759.
Eskenazi, M., and Mostow, J. (**1997**). "The CMU KIDS Speech Corpus (LDC97S63)," Linguistic Data Consortium (http://www.ldc.upenn.edu), University of Pennsylvania (viewed 8-27-07).
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Foster, J. R., Summerfield, A. Q., Marshall, D. H., Palmer, L., Ball, V., and Rosen, S. (**1993**). "Lip-reading the BKB sentence lists: Corrections for list and practice effects," Br. J. Ophthamol. **27**, 233–246.
French, N., and Steinberg, J. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
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.
Fu, Q. J., Chinchilla, S., Nogaki, G., and Galvin, J. J., III (**2005**). "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution," J. Acoust. Soc. Am. **118**, 1711–1718.
Fu, Q. J., Shannon, R. V., and Wang, X. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.
Hall, J. W., III, Grose, J. H., Buss, E., and Dev, M. B. (**2002**). "Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children," Ear Hear. **23**, 159–165.
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.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU-17**, 225–246.
Johnstone, P. M., and Litovsky, R. Y. (**2006**). "Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults," J. Acoust. Soc. Am. **120**, 2177–2189.
Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, 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.
Levitt, H., and Rabiner, L. R. (**1967**). "Use of a sequential strategy in intelligibility testing," J. Acoust. Soc. Am. **42**, 609–612.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.
Nelson, P. B., and Jin, S. H. (**2004**). "Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **115**, 2286–2294.
Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A. (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Payton, K. L., and Braida, L. D. (**1999**). "A method to determine the speech transmission index from speech waveforms," J. Acoust. Soc. Am. **106**, 3637–3648.
Peters, R. W., Moore, B. C., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Qin, M. K., and Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on F0 discrimination and concurrent-vowel identification," Ear Hear. **26**, 451–460.
Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Roy, S. N., and Bose, R. C. (**1953**). "Simultaneous confidence interval estimation," Ann. Math. Stat. **24**, 513–536.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Souza, P. E., Boike, K. T., Witherell, K., and Tremblay, K. (**2007**). "Prediction of speech recognition from audibility in older listeners with hearing loss: Effects of age, amplification, and background noise," J. Am. Acad. Audiol **18**, 54–65.
Stickney, G. S., Assmann, P. F., Chang, J., and Zeng, F. G. (**2007**). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078.
Stickney, G. S., Zeng, F. G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Summers, V., and Molis, M. R. (**2004**). "Speech recognition in fluctuating and continuous maskers: Effects of hearing loss and presentation level," J. Speech Lang. Hear. Res. **47**, 245–256.
Throckmorton, C. S., and Collins, L. M. (**2002**). "The effect of channel interactions on speech recognition in cochlear implant subjects: Predictions from an acoustic model," J. Acoust. Soc. Am. **112**, 285–296.
Trammell, J. L., and Speaks, C. (**1970**). "On the distracting properties of competing speech," J. Speech Hear. Res. **13**, 442–445.
Wagener, K. C., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters," Int. J. Audiol. **44**, 144–156.
Watson, C. S., and Kelly, W. J. (**1981**). In *Auditory and Visual Pattern recognition*, D. J. Getty and J. H. Howard, eds. (Erlbaum, Hillsdale, NJ), pp. 37–59.
Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.

Longitudinal changes in speech recognition in older persons

Judy R. Dubno,[a] Fu-Shing Lee, Lois J. Matthews, Jayne B. Ahlstrom, Amy R. Horwitz, and John H. Mills

Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, 135 Rutledge Avenue, P.O. Box 250550, Charleston, South Carolina 29425

(Received 29 March 2007; revised 29 October 2007; accepted 4 November 2007)

Recognition of isolated monosyllabic words in quiet and recognition of key words in low- and high-context sentences in babble were measured in a large sample of older persons enrolled in a longitudinal study of age-related hearing loss. Repeated measures were obtained yearly or every 2 to 3 years. To control for concurrent changes in pure-tone thresholds and speech levels, speech-recognition scores were adjusted using an importance-weighted speech-audibility metric (AI). Linear-regression slope estimated the rate of change in adjusted speech-recognition scores. Recognition of words in quiet declined significantly faster with age than predicted by declines in speech audibility. As subjects aged, observed scores deviated increasingly from AI-predicted scores, but this effect did not accelerate with age. Rate of decline in word recognition was significantly faster for females than males and for females with high serum progesterone levels, whereas noise history had no effect. Rate of decline did not accelerate with age but increased with degree of hearing loss, suggesting that with more severe injury to the auditory system, impairments to auditory function other than reduced audibility resulted in faster declines in word recognition as subjects aged. Recognition of key words in low- and high-context sentences in babble did not decline significantly with age. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2817362]

PACS number(s): 43.71.Lz, 43.71.Es, 43.71.Gv, 43.71.Ky [MSS] Pages: 462–475

I. INTRODUCTION

A. Cross-sectional studies of age-related differences in speech recognition

Many studies of speech recognition in older adults report age-related differences in performance. However, the interpretation of these results is complicated by pure-tone thresholds that change with increasing age and rates of change that vary among individuals. One of the largest reports of speech recognition among older persons is from Jerger (1973), who obtained scores from clinical records of 2162 patients. With subjects grouped according to age and degree of hearing loss, results suggested that speech recognition (defined as the maximum score obtained using a monosyllabic word list) differed among age groups, but differences were dependent on degree of loss. That is, for subjects with mild loss (<30 dB HL), differences in scores with increasing age were measurable but small through age 70. However, for subjects with moderate-to-severe hearing loss (40–69 dB HL), larger differences in scores with increasing age were observed, particularly for persons between the ages of 45 and 85. Thus, with hearing loss held constant, age had little effect on speech recognition for individuals with mild hearing loss but greater effect for those with more loss.

In another study of speech recognition by older persons, Jerger (1990, 1992) compared scores for (1) PB word lists (PAL-PB50 developed by Egan, 1948); (2) key words from low- and high-context sentences of the Speech Perception in Noise (SPIN) test (Kalikow *et al.*, 1977); and (3) the Synthetic Sentence Identification (SSI) test (Speaks and Jerger, 1965). Subjects ranging in age from 50 to 90 years were assigned to four age groups. To minimize the confounding effects of hearing loss and age, subjects were selected such that average thresholds at a particular frequency across the four groups were within 6 dB. Although scores for all tests differed somewhat from the youngest to the oldest age group, only the difference in scores for the SSI reached statistical significance. Thus, when thresholds across age groups were equated, speech recognition did not differ from age 50 to age 90, except for the SSI (consistent with Jerger and Hayes, 1977). Using a similar approach, Dubno *et al.* (1997) found no significant differences in speech recognition for individuals in three age groups (55–64, 65–74, and 75–84 years) who were selected so that average pure-tone thresholds for the three groups were within 5 dB. Speech-recognition measures included word recognition and maximum word recognition in quiet, key word recognition of low- and high-context sentences in babble (SPIN), and maximum SSI sentence recognition with an ipsilateral competing message.

In a population-based study, word-recognition scores obtained with CID W-22 lists (Hirsh *et al.*, 1952) were reported for subjects in the Framingham Heart Study (Mościcki *et al.*, 1985; Gates *et al.*, 1990), organized into 5-year age groups. Although changes in word recognition with increasing age group were larger for males than females, differences were attributed to gender-related differences in hearing levels. That is, when subjects were grouped by audiometric configuration ("flat, gradual, sharp"), only very small gender-related differences were observed. This result was also consistent with Articulation Index (AI) values computed from subjects' better-ear thresholds and speech spectra, using the procedure of Popelka and Mason (1987). Mean AI values were signifi-

[a]Electronic mail: dubnojr@musc.edu

cantly higher for females than males; however, the change in AI values with increasing age group did not differ for males and females.

Another population-based study (Wiley *et al.*, 1998) compared age-related changes in word recognition in quiet and with a single-talker competing message for subjects in the Epidemiology of Hearing Loss study from Beaver Dam, WI (Cruickshanks *et al.*, 1998). Subjects were organized into four age groups (48–59, 60–69, 70–79, and 80–92 years). Word recognition in quiet and with the competing message was poorer for older age groups and poorer for males than females, even when results were stratified by degree of hearing loss (pure-tone average of 0.5, 1.0, 2.0, 4.0 kHz). Using analysis of covariance (ANCOVA), degree of hearing loss accounted for the largest portion of the variance in speech recognition in quiet and in the competing message; age and gender were much smaller, but significant, contributors. Another recent population-based study of age-related hearing loss, the Blue Mountains Hearing Study of older Australians, included measures of speech recognition but outcomes focused on evidence of "central auditory processing abnormalities" and, therefore, had little relevance to the results of the current study (Golding *et al.*, 2004, 2005, 2006).

A weakness of retrospective analyses of individuals grouped to equate average thresholds is that effects on speech recognition of variables other than age and thresholds (such as gender) are difficult to assess. Moreover, age and thresholds are continuous variables and groupings such as described earlier are necessarily arbitrary. Therefore, Dubno *et al.* (1997) used partial correlations to adjust both score and age for their associations with average thresholds. For males, significant differences in scores with increasing age were observed in several speech-recognition measures, after adjusting for threshold differences; for females, no significant age-related changes in any speech-recognition measures were observed.

Thus, results of several cross-sectional studies consistently report differences in speech recognition between age groups. However, results are inconsistent with regard to the contributions of hearing loss, age, and gender to these differences. These inconsistencies may be attributed to differences in studies' sample sizes, sampling methods, speech materials, procedures, or statistical methods. Contradictory findings may also relate to cohort differences, which are known to confound group differences in cross-sectional studies. Most important, without controls for age-related threshold changes, it is not possible to determine the extent to which age-related differences in speech recognition relate to differences in speech audibility or to other auditory or cognitive factors.

Other types of cross-sectional analyses are reported in experimental studies of speech recognition in older persons in which background noises, reverberation, or other forms of degradation were used to increase the difficulty of the task, or in which binaural or sound-field listening were required. Some show age-related differences in auditory behavior, whereas others do not [see Humes (1996) and Gordon-Salant (2005) for reviews]. These studies by design have small sample sizes and relatively narrow ranges of age, hearing loss, and speech-recognition scores. Moreover, because threshold-related differences in speech audibility are the primary contributors to individual differences in speech recognition, effects of age or gender on speech recognition may be relatively small and, therefore, difficult to detect.

B. Longitudinal studies of age-related declines in speech recognition

In a longitudinal study, subjects serve as their own controls, thus minimizing effects of uncontrollable factors, such as noise history, health history, and occupation. In contrast to cross-sectional studies, which typically focus on group effects, longitudinal studies can measure age-related changes in hearing levels and speech recognition for groups and individuals. Thus, longitudinal studies provide a method for studying the effects of age on speech recognition in groups or individuals with fewer confounding factors. A disadvantage of longitudinal studies of older persons is that data collection takes many years, making it difficult to retain subjects in good general health over long periods of time. This raises concerns about "selective attrition," wherein healthier older persons (who may also be higher performing subjects) remain in the study for longer periods of time.

Only a few large-scale longitudinal studies of hearing have been conducted, including the Baltimore Longitudinal Study of Aging (Brant and Fozard, 1990; Pearson *et al.*, 1995), the British Medical Research Council's epidemiologic studies in the United Kingdom and Denmark (Davis *et al.*, 1991), the Framingham Heart Study (Gates and Cooper, 1991), and the Beaver Dam Epidemiology of Hearing Loss study (Cruickshanks *et al.*, 2003). These studies report rates of change in pure-tone thresholds for males and females derived from two to six repeated measurements. Rate of change in hearing was estimated either by fitting a regression line to thresholds for a group of subjects or by taking the difference between two threshold measurements for each individual. Fitting a regression line to group data provides an estimate of the rate of change for a group but does not describe the change for each individual. Calculating the rate of threshold change from two measurements is prone to error because the rate of change is small relative to the measurement error.

In the longitudinal study of age-related hearing loss at the Medical University of South Carolina (MUSC), pure-tone thresholds for conventional and extended high frequencies were analyzed for longitudinal changes and to determine effects of initial thresholds, age, gender, and noise history on these changes (Lee *et al.*, 2005). Subjects had between 2 and 21 threshold measures over a period of 3–11.5 years. The slope of a linear regression was used to estimate the rate of change in pure-tone thresholds for each ear. The average rate of change in thresholds was 0.7 dB/year at 0.25 kHz, increasing gradually to 1.2 dB/year at 8.0 and 12.0 kHz. Rate of threshold change increased significantly with age at 0.25 to 3.0, 10.0, and 11.0 kHz for females and at 6.0 kHz for males. After adjusting for age, females had a significantly slower rate of change than males at 1.0 kHz but a significantly faster rate of change than males from 6.0 to 12.0 kHz.

Many of the large-scale longitudinal studies of hearing mentioned earlier included measures of speech recognition

but none have reported longitudinal changes. The few studies that have assessed longitudinal changes were limited by small sample sizes of older subjects, only one to two repeated measurements, and short time spans. Most important, none of the studies has assessed changes in speech recognition over time while controlling for concurrent changes in pure-tone thresholds. Briefly, in a study of 80-year-old residents of central Finland, pure-tone thresholds and monosyllabic word recognition were measured three times over a 10 year period (Hietanen *et al.*, 2004). Results showed significant declines in word recognition, which did not differ for males and females. However, the number of subjects tested over the 10 year period was relatively small (18 males, 62 females) and age-related changes in pure-tone thresholds likely confounded the results. Møller (1981) had a relatively large subject sample (124 males and 137 females) but only one repeated measure and a 5 year time span. Scores were reported as distributions, making it difficult to determine the amount of change in scores for individuals. The longitudinal study of Bergman *et al.* (1976) measured speech recognition two or three times over a 3 or 7 year period in a small group of older subjects (eight to ten subjects age 60 years and older) and showed significant declines in some measures of speech recognition. However, given that pure-tone thresholds were not measured (hearing was screened at 35–40 dB HL at baseline only), the contribution of age-related changes in hearing levels was unknown. Pedersen *et al.* (1991) had a relatively large sample size and long (11-year) time span, but relationships between word-recognition scores and average pure-tone thresholds were assessed only by correlations for different age groups. The longitudinal study of Divenyi *et al.* (2005) included 29 older subjects with mild-to-moderate hearing loss who were tested two times within ~5 years; measurements included pure-tone thresholds and a battery of speech-recognition tests (Divenyi and Haupt, 1997). Comparing results to a group of younger subjects with normal hearing, these authors concluded that speech-recognition measures in older subjects declined with increasing age more rapidly than their pure-tone thresholds.

The current study differs from previous assessments of longitudinal changes in speech recognition in several respects. Recognition of isolated monosyllabic words in quiet and recognition of key words in low- and high-context sentences in babble were measured in 256 and 85 older persons, respectively; a minimum of three scores was obtained from each ear of each subject over a range of 3–15 years. As noted earlier, assessing longitudinal changes in speech recognition of older subjects is not straightforward because pure-tone thresholds change with increasing age and rates of change vary among subjects. In addition, if degree of hearing loss changed from measurement to measurement, speech presentation levels may have also changed. Because speech audibility is the primary contributor to individual differences in speech recognition, it was critical that changes in word recognition over time were measured independently of changes in speech audibility over time.

To determine the extent to which changes in pure-tone thresholds and speech levels explained changes in word recognition, speech-recognition scores were adjusted using an importance-weighted speech-audibility metric (AI, ANSI, 1969a; 1997). Linear-regression slope was then used to estimate the rate of change in adjusted speech-recognition scores. That is, an AI value and predicted score was computed for each observed score, using pure-tone thresholds, speech level, and babble level (if present) measured at the same time as the speech-recognition score. The rationale for comparing observed and predicted scores was as follows. Increasing age results in higher pure-tone thresholds, which lowers speech audibility and the computed AI, corresponding to a lower predicted score. If declines in observed scores with increasing age were determined entirely by reduced audibility and were independent of age and age-related factors, observed scores should decline as predicted. Thus, the difference between observed and predicted scores at different time points was used to determine how word recognition changed with increasing age while accounting for changes in speech audibility. Finally, using the procedures developed for estimating rates of change in pure-tone thresholds, longitudinal changes in speech recognition were analyzed to determine effects of gender, initial thresholds, age, and noise history on these longitudinal changes.

II. METHODS

A. General methods

In the longitudinal study of age-related hearing loss at MUSC, subjects 55 years of age[1] and older and in good general health were recruited through advertisements and subject referral. Subjects were excluded if there was evidence of conductive hearing loss or active otologic or neurologic diseases. All subjects were screened with the Short Portable Mental Status Questionnaire (a ten-item memory test; Pfeiffer, 1975). Subjects were scheduled approximately once per month for a total of three to six visits to complete a test battery that included the following: (1) pure-tone air-conduction thresholds at conventional frequencies (repeated at every visit) and extended high frequencies; (2) speech-recognition thresholds (SRT) using the Auditec recording of the CID W-1 spondaic word lists (Hirsh *et al.*, 1952); (3) several measures of speech recognition in quiet and in noise using a variety of recorded test materials; (4) middle-ear measurements; (5) otoacoustic emissions; (6) upward and downward spread of masking; (7) auditory brainstem responses; (8) an otologic examination; and (9) blood draws for clinical chemistries, including serum estradiol and progesterone levels for female subjects, and to extract and store DNA. In addition, subjects completed questionnaires on hearing and medical history, tinnitus, medication use, smoking, occupational and nonoccupational noise history, hearing-aid use, family history, and self-evaluation of hearing handicap.

After completion of the test battery, subjects were scheduled annually to update their contact information, to update medical and hearing histories and prescription drug information, and for measurement of thresholds at conventional frequencies and monosyllabic word-recognition scores. To obtain longitudinal data, the entire test battery was repeated

every 2 to 3 years. Measurement of the SPIN test was considered part of the test battery and was therefore obtained once every 2 to 3 years.

To retain a nearly constant number of subjects actively involved in the longitudinal study, approximately 50 new subjects are enrolled each year since the start of the study in 1987; this enrollment is based on average attrition rates. Subjects have voluntarily withdrawn from the study due to nonstudy-related illnesses or death, moving from the area, no longer perceiving a benefit of participation, and increased time constraints. Subjects have been discontinued from the study due to difficulty scheduling, nonaging-related changes in hearing or otologic or neurologic conditions, and poor test reliability.

Nearly 840 subjects have participated in the program; of these, some longitudinal data covering at least a 3-year period are available from nearly 380 subjects. The current study reports results for all subjects with longitudinal measures of: (1) recognition of isolated monosyllabic words in quiet from the Northwestern University Auditory Test No. 6 (NU#6; Tillman and Carhart, 1996) or (2) key-word recognition of low- and high-context sentences in multitalker babble from the SPIN test.

B. Subjects

Word-recognition (NU#6) scores were obtained from both ears of 256 subjects (128 males and 128 females). At the time of their first measure, these subjects ranged in age from 50 to 82 years (mean=67.6 years). At the time of their last measure, their ages ranged from 60 to 91 years (mean =75.0 years). Scores for low- and high-context sentences in babble (SPIN) were obtained from both ears of 85 subjects (39 males and 46 females). At the time of their first measure, these subjects ranged in age from 56 to 81 years (mean =67.0 years). At the time of their last measure, their ages ranged from 63 to 90 years (mean=77.2 years).

As described earlier, the protocol of the longitudinal study called for yearly measures of pure-tone thresholds and word-recognition scores in quiet (NU#6 at 30 dB above the SRT). In addition, scores for the NU#6 and SPIN tests were obtained at 2 to 3 year intervals (see the procedures to follow). This protocol yields many more word-recognition (NU#6) scores than sentence-recognition (SPIN) scores over a given period of time. For NU#6, subjects had between 3 and 18 scores (mean=7.2 scores) over a period of 3–15 years (mean=7.3 years). A total of 3683 scores from 512 ears were analyzed in the current study. For SPIN, subjects had 4 scores for low-context sentences and 4 scores for high-context sentences over a period of 7–13 years (mean=10.3 years). A total of 1360 scores (680 low-context and 680 high-context) from 170 ears were analyzed in the current study.

Figure 1 shows mean (±1 standard error, SE) pure-tone thresholds for 512 ears from female and male subjects at the time of their initial NU#6 test. Mean thresholds were very similar for the 170 ears from female and male subjects at the time of their initial SPIN test. For comparison, Fig. 1 also includes mean thresholds for female and male subjects of similar ages from the Framingham Heart Study cohort (Gates *et al.*, 1990) and the Beaver Dam cohort (Cruickshanks *et al.*, 1998). Mean thresholds of subjects in the MUSC study are generally similar to those of both community-based, epidemiological studies of age-related hearing loss.

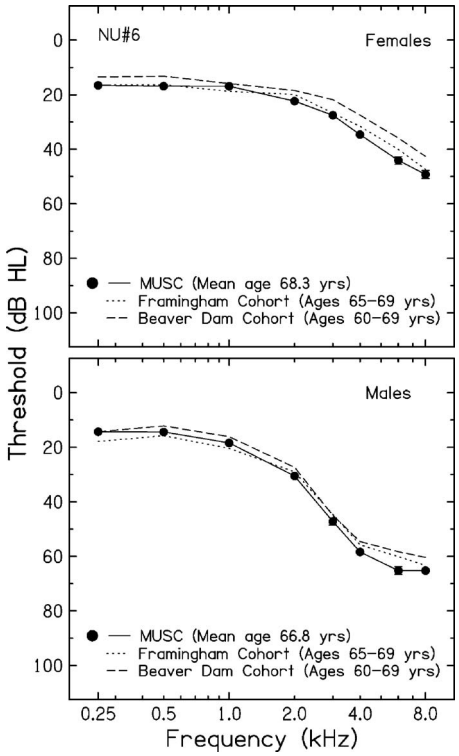


FIG. 1. Mean (±1 standard error, SE) initial pure-tone thresholds in dB HL for females (top panel, circles) and males (bottom panel, circles) enrolled in the Medical University of South Carolina (MUSC) longitudinal study of age-related hearing loss. This group (512 ears) was followed longitudinally using monosyllabic word-recognition scores in quiet (NU#6). SE ranges are smaller than data points at some frequencies. For comparison, mean thresholds are shown for females (top panel) and males (bottom panel) of similar ages from the Framingham Heart Study (dotted line) and Epidemiology of Hearing Loss Study of Beaver Dam, WI (dashed line).

C. Stimuli and procedures

Conventional pure-tone thresholds were measured with a Madsen OB822 or Orbiter 922 clinical audiometer calibrated to appropriate ANSI standards (1969b, 1989, 1996, 2004) and equipped with TDH-39 headphones mounted in MX-41/AR cushions. Thresholds were measured in 5 dB steps at octave and half octave intervals (0.25, 0.5, 1.0, 2.0, 3.0, 4.0, 6.0, and 8.0 kHz).

For measures of speech recognition, the following test lists were used: (1) Auditec tape-recorded NU#6 25-item word lists and (2) the revised version of the SPIN test consisting of eight 50-sentence lists (Bilger, 1984). There were four NU#6 lists and four randomizations of each list; a different list or randomization was used for each ear at each visit. SPIN sentences were presented at 50 dB above a babble threshold, which was estimated from pure-tone thresholds, specifically the best threshold from 0.25–4.0 kHz (according to Bilger, 1984). Signal-to-babble ratio was fixed at +8 dB. If a particular speech level was uncomfortably loud, speech level was reduced to a comfortable level and babble level was reduced to maintain a +8 dB signal-to-

babble ratio. Scores were obtained for recognition of the final words of high-context and low-context sentences. High-context sentences ("The watchdog gave a warning growl") contain both semantic and syntactic context. Low-context sentences ("I had not thought about the growl") contain only syntactic context. One SPIN list was used for each ear at each visit, with different lists used at subsequent visits.

In all cases, testing began with the better-hearing ear; if hearing loss was equivalent in the two ears, testing began with the right ear. Masking was introduced to the nontest ear when that ear may have contributed to the observed response. Additional details of subject selection and test administration are included in previous publications reporting results from the MUSC longitudinal study of age-related hearing loss (Dubno *et al.*, 1995, 1997; Lee *et al.*, 2005).

AI values and predicted scores for isolated words in quiet were computed from the spectra and levels of the NU#6 words and subjects' quiet thresholds using procedures similar to ANSI (1969a, 1997) and the frequency-importance function and AI-recognition transfer function for the Auditec of St. Louis recordings of the NU#6 word test (Studebaker *et al.*, 1993). To maintain consistency with our previous studies, the AI (ANSI, 1969a) was computed and used to determine predicted scores, rather than the Speech Intelligibility Index (SII; ANSI, 1997). According to Pavlovic (2006), the key difference between the standards describing the AI and the SII is that the SII allows for the use of frequency-importance functions for specific speech materials. Thus, given that the frequency-importance functions developed for NU#6 and SPIN were used in the current study, our procedures were consistent with the standard that describes the SII. In the implementation of the AI used in this study, no speech or masker level corrections were included.

Using similar procedures, AI values and predicted scores for key words in low- and high-context sentences were computed from the spectra and levels of the final words of the SPIN sentences, the spectra and levels of the SPIN babble, and subjects' quiet thresholds, using the frequency-importance function and AI-recognition transfer function for the SPIN materials (Bell *et al.*, 1992). Pure-tone thresholds measured at octave or half octave intervals were interpolated to obtain thresholds in one-third octave intervals; speech spectra were measured in one-third octave bands.

III. RESULTS AND DISCUSSION

A. Longitudinal changes in word recognition in quiet

Figure 2 includes mean (±1 SE) presentation levels for the NU#6 words in dB HL (closed circles, left ordinate) for laboratory visits at which NU#6 scores were obtained plotted against mean age during those visits. Also shown are mean (±1 SE) pure-tone averages (PTA) in dB HL measured during the same visits (open circles, right ordinate). The PTA is the average thresholds at 0.5, 1.0, 2.0, and 4.0 kHz. Over the 14 year time span (i.e., difference in mean age from initial visit to last visit), mean PTA increased by 15.5 dB, an average of 1.1 dB/year, consistent with longitudinal changes in pure-tone thresholds reported previously from the MUSC study (Lee *et al.*, 2005). Because the presentation level of the

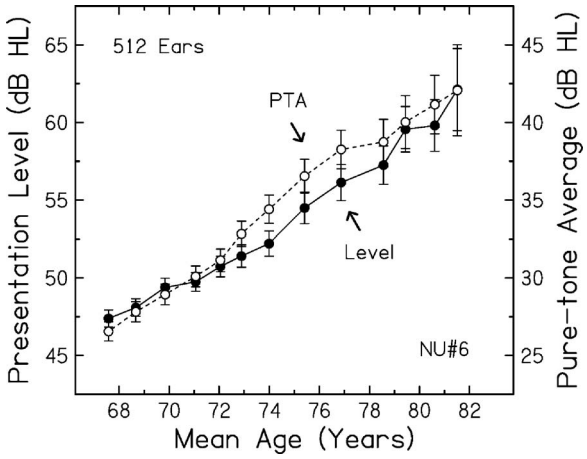


FIG. 2. Mean (±1 SE) speech presentation levels in dB HL for NU#6 words (closed, left ordinate) for the laboratory visits at which NU#6 scores were obtained plotted against mean age during those visits. Also shown are mean (±1 SE) pure-tone averages (PTA) in dB HL measured during the same visits (open, right ordinate). PTA is the average thresholds at 0.5, 1.0, 2.0, and 4.0 kHz.

NU#6 items for each subject was set relative to the subject's SRT, the average speech level generally increased in a similar manner as the PTA, but the match was not perfect. This illustrates the importance of accounting for age-related changes in speech audibility due to changes in pure-tone thresholds and speech levels when evaluating longitudinal changes in speech recognition.

For individual ears, the slope of a linear regression relating NU#6 scores to age was used to calculate the rate of change in word-recognition scores and adjusted word-recognition scores (i.e., rate of change in observed minus predicted scores, as described earlier). By way of example, Fig. 3 (top) shows word-recognition scores plotted as a function of age for two subjects. Each set of results was fit with a linear regression; the slope of the function represents the estimated rate of change in score with increasing age. The negative slope of the linear regression functions indicates that word recognition declined as subjects got older. The variance in scores was quite large because, as subjects got older, pure-tone thresholds and presentation levels were also changing at different rates. The bottom panel of Fig. 3 shows the adjusted word-recognition score (i.e., the difference between the observed score and the score predicted by the AI) plotted against age for the same two subjects. The adjusted word-recognition scores take into account any changes in speech audibility due to increasing thresholds and changes in speech levels. A negative adjusted word-recognition score means that the observed score was poorer than the AI-predicted score. Fitting the linear regression function to adjusted word-recognition scores reduced the variance in scores to some degree and changed the estimated rate of decline. The linear regression functions with negative slopes for adjusted word recognition scores mean that observed scores declined faster than predicted scores as subjects got older.

The average rate of change in NU#6 scores for 512 ears was -1.04%/year ($p < 0.0001$). Although this suggests that word-recognition scores declined significantly over time,

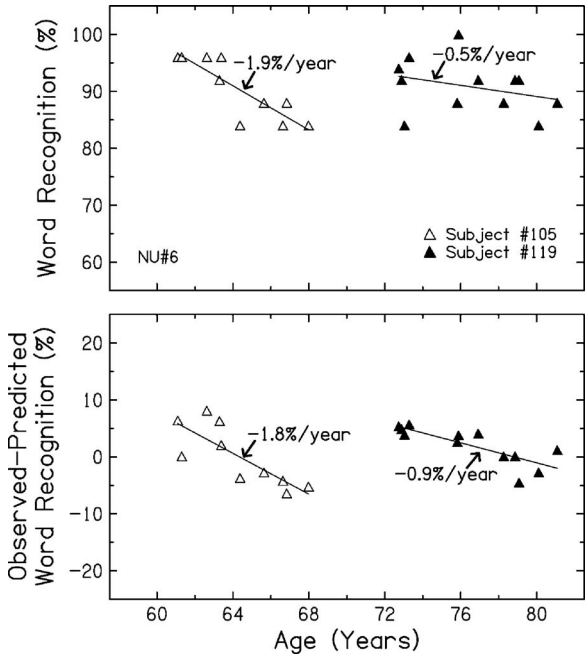


FIG. 3. Top panel: Word-recognition scores in percent correct (NU#6) for two subjects (open and closed triangles) plotted as a function of age. Linear regression functions and slopes (%/year) show the rate of change in word recognition as subjects aged. Bottom panel: For the same two subjects and time points, differences between observed word-recognition scores and scores predicted using the articulation index (i.e., adjusted word recognition) plotted as a function of age (see the text for details). Linear regression functions and slopes (%/year) show the rate of change in adjusted word recognition as subjects aged.

some of the change may be attributed to changes in speech audibility resulting from threshold or speech-level changes occurring during the same time span. The average rate of change in *adjusted* NU#6 scores was −0.74%/year ($p < 0.0001$). Thus, word recognition declined significantly with increasing age even when accounting for age-related changes in speech audibility.

The current results are consistent with the small longitudinal study of Divenyi *et al.* (2005), which found that speech-recognition measures in older subjects declined with increasing age more rapidly than their pure-tone thresholds. However, they are in contrast to previous cross-sectional studies reporting that age-related differences in speech recognition were accounted for by grouping subjects according to audiometric configuration or degree of hearing loss (e.g., Gates *et al.*, 1990; Wiley *et al.*, 1998). The current results also differ from studies that used multivariate approaches to identify "hearing loss" or "audibility" as the primary factor contributing to individual differences in speech recognition of older adults, with age and auditory and cognitive function accounting for only small portions of the variance (e.g., van Rooij *et al.*, 1989; van Rooij and Plomp, 1990, 1992; Jerger *et al.*, 1991; Humes *et al.*, 1994). However, studies of factors contributing to speech recognition by older adults rarely measure or carefully control for age- and threshold-related differences in speech audibility, as in the current study, which may explain differences in conclusions. Finally, inconsistencies with conclusions of the current longitudinal study may also be attributed to limitations of the previous studies' cross-sectional designs.

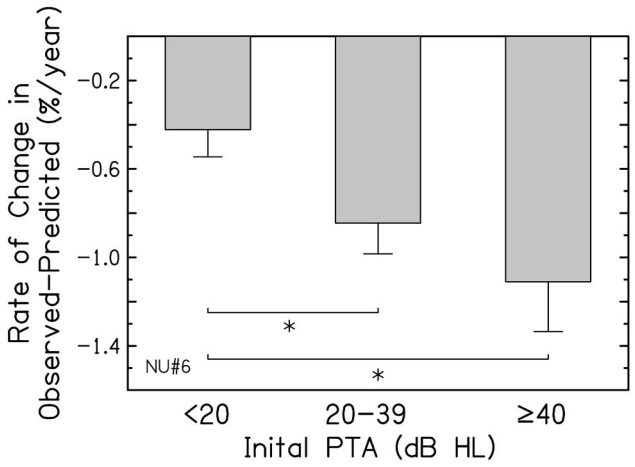


FIG. 4. Mean (±1 SE) rate of change (in %/year) in observed−predicted differences for subjects grouped according to their initial PTA (<20 dB HL, 20−39 dB HL, ≥40 dB HL). Asterisks indicate statistically significant differences between PTA groups ($p < 0.05$).

1. Effects of initial hearing loss and age

Because initial thresholds were higher for older males than older females (see Fig. 1) and subjects entered the study at different ages, effects of gender, initial hearing loss, and initial age were confounded, making it difficult to assess these effects on declines in word recognition. To determine the gender effect without the confounded threshold and age effects, an ANCOVA of the rate of change in adjusted word recognition was performed, with gender as a grouping factor and initial hearing loss (PTA) and age as covariates. Results of the ANCOVA showed significant effects of gender ($p = 0.002$) and initial PTA ($p = 0.0002$), and nonsignificant effects of age ($p = 0.321$). Effects of initial PTA and age are discussed now; the gender effect is discussed in Sec. III A 2.

Rate of decline in word recognition increased by a small, but significant, amount as initial hearing loss increased, even while taking into account hearing-level-related differences in speech audibility. Figure 4 illustrates this effect by grouping subjects into three PTA categories; the interaction between initial PTA and gender was not statistically significant. Rate of decline in word recognition was significantly faster for individuals with more severe hearing loss. This effect of degree of hearing loss suggests that with more severe injury to the peripheral and/or central auditory system, impairments to auditory and/or cognitive function other than elevated thresholds (reduced audibility) resulted in faster declines in word recognition as subjects aged. In the current study, reduction in simple audibility due to elevated thresholds was eliminated as a factor by evaluating differences between observed scores and scores predicted by importance-weighted speech audibility (AI).

The absence of an effect of initial age suggests that, among older subjects, the rate of decline in word recognition did not accelerate with increasing age. That is, the rate of decline in word recognition was not faster for subjects in their 70s or 80s than for subjects in their 60s. This finding, together with the significant effect of initial hearing loss, suggests that the changes in function that accompanied higher thresholds and resulted in faster declines in word rec-

ognition were not increasing with age. Taken together, these age-related declines in word recognition were more consistent with underlying changes in auditory, rather than cognitive, function resulting from peripheral, rather than central, auditory system pathology. Nevertheless, it is also possible that a single nervous-system factor underlies observed age-related changes and accounts for all results. This "common cause" hypothesis was put forth by Lindenberger and Baltes (1994) and Baltes and Lindenberger (1997) to interpret the relatively high correlations observed among sensory and cognitive factors in older adults. These authors proposed that the link between sensory and cognitive measures increased with age because both functions reflect the same anatomic and physiologic changes in the aging brain. In the current study, a single underlying neural mechanism could underlie both hearing levels and cognitive function, such that subjects with poorer initial PTA (as shown in Fig. 4) also had greater cognitive impairment. The combined effects of auditory and cognitive impairments could have resulted in faster rates of decline in word recognition for subjects with poorer hearing.

2. Effects of gender and serum hormone levels

As noted earlier, results of ANCOVA revealed a significant effect of gender. Specifically, declines in adjusted word recognition were significantly faster for females (−0.92%/year) than for males (−0.57%/year), even while taking into account gender-related differences in speech audibility due to their threshold differences. Some longitudinal studies of pure-tone thresholds in older persons have reported faster rates of threshold increases for females than males (Møller, 1981; Gates and Cooper, 1991; Lee *et al.*, 2005 for higher frequencies only). However, other studies found equivalent rates (Cruickshanks *et al.*, 2003 for PTA) or faster rates of increases for males than females (Pearson *et al.*, 1995). Regardless of these threshold findings, the gender difference observed in the current study in the rate of decline in word recognition was independent of gender-related threshold and speech-audibility differences.

Given the limited information on longitudinal changes in speech recognition in older persons, gender-related differences in longitudinal effects are unknown. From cross-sectional studies, changes in word recognition with increasing age group appeared larger for males than females, but differences were often attributed to gender-related differences in hearing levels (e.g., Gates *et al.*, 1990). In contrast, Wiley *et al.* (1998) reported a small, but significant gender effect (males worse than females) in results from the Beaver Dam study that remained after stratifying for hearing loss. An earlier cross-sectional analysis from the MUSC study reported age-related differences in speech recognition in males but not in females, when scores were adjusted for their association with average thresholds (Dubno *et al.*, 1997). Thus, there is little evidence of faster declines in word recognition for females than males. However, as noted previously, differences among results of cross-sectional studies may occur due to differences in sampling or statistical methods, procedures, controls for age-related threshold changes, or cohort effects.

It is possible that different etiologies underlie the hearing loss observed in older females and males. For example, a higher percentage of males in the MUSC longitudinal study reported a significant history of noise exposure than females. Although presbyacusis in humans likely has multiple causes, threshold elevation in males may result from combined effects of noise and aging (plus other exogenous factors) whereas threshold elevation in females may have a smaller noise component (see Sec. III A 3 for effects of noise history on rate of change in word recognition). Physiologic evidence from noise-aged gerbils (Schmiedt *et al.*, 1990) are consistent with the hypothesis that noise exposure alters the micromechanics of the cochlea and results in dysfunction or loss of sensory cells. In contrast, quiet-aged gerbils exhibit a degeneration of the lateral wall of the cochlea, including capillary beds of the stria vascularis (Gratton *et al.*, 1996). A consequence of this "metabolic presbyacusis" is an age-related decrease in the endocochlear potential (Schulte and Schmiedt, 1992), which reduces the gain of the cochlear amplifier in the apex by as much as 20 dB and in the base by as much as 60 dB (Cooper and Rhode, 1997; Schmiedt *et al.*, 2002), yielding higher thresholds. This gives rise to the typical audiometric configuration of older adults as seen in the top panel of Fig. 1, including a flat 10–40 dB low-frequency hearing loss, with thresholds at higher frequencies gradually increasing to ~60 dB. Thresholds in the higher frequencies for males (bottom panel of Fig. 1) reflect the additional effects of sensory-cell loss due to noise exposure (Schmiedt *et al.*, 2002). These patterns are consistent with faster rates of pure-tone threshold changes in the higher frequencies for females than males (Lee *et al.*, 2005) and may relate to faster declines in adjusted word recognition seen in the current study. Thus, gender-related differences in the etiology of cochlear injury in older persons, and mechanisms underlying presybacusis, could have implications for age-related changes in auditory function, such as speech recognition.

There may also be other gender-related factors that covary with age and explain the faster decline in word recognition for females than males. For example, calcium channel blockers, beta (adrenergic) blockers, antihistamines, and high levels of cholesterol were shown to affect hearing levels of females but not males in the MUSC study (Lee *et al.*, 1998a, b). In addition, older females who were on estrogen therapy had slightly better hearing than those not taking hormone supplements (Lee *et al.*, 1998b; Kilicdag *et al.*, 2004; Hultcrantz *et al.*, 2006), but results are inconsistent (Kim *et al.*, 2002; Helfer, 2004). Guimaraes *et al.* (2006) found no protective effect of estrogen therapy alone on hearing but a negative effect of estrogen combined with progestin; in the Guimaraes study, serum levels were not reported.

To explore effects of hormones on longitudinal changes in speech recognition in females, an ANCOVA on the rate of change in adjusted word recognition was performed, with serum estradiol and serum progesterone levels as grouping factors, and initial hearing loss (PTA) and age as covariates. Subjects were 108 females for whom serum hormone levels and rates of change in word recognition were available.[2] Serum estradiol and progesterone were treated as grouping factors because of their skewed distributions. Serum levels were organized into three groups with similar sample sizes. Results of the ANCOVA again showed a significant effect of

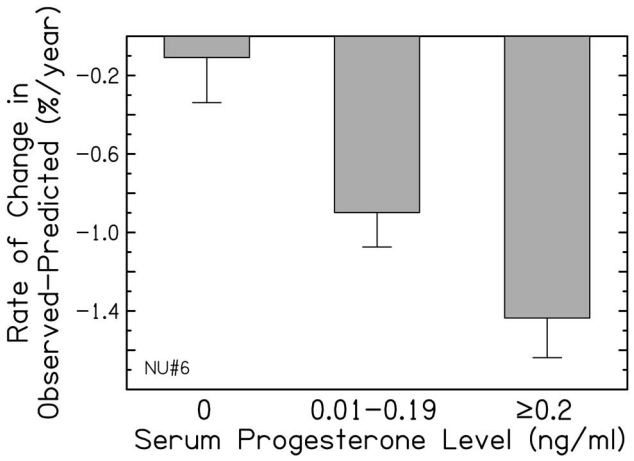


FIG. 5. Mean (±1 SE) rate of change (in %/year) in observed−predicted differences for female subjects grouped according to their serum progesterone level (0 ng/ml, 0.01–0.19 ng/ml, ⩾0.2 ng/ml). Rates of change differed significantly between all progesterone groups ($p<0.05$).

initial PTA ($p=0.025$) and a nonsignificant effect of age ($p=0.335$); note that serum estradiol and progesterone levels did not vary significantly with age. The ANCOVA also revealed a significant effect of serum progesterone level ($p=0.002$) and a nonsignificant effect of serum estradiol level ($p=0.264$) on rate of change in adjusted word recognition. As illustrated in Fig. 5, females with higher levels of progesterone in their blood had faster declines in word recognition than females with lower levels of progesterone. This result is consistent with the negative effect of hormone therapy that includes progestin reported by Guimaraes *et al.* (2006) and a biochemical mechanism that relates progesterone to activation of inhibitory neurotransmittors, such as γ-aminobutyric acid (GABA) in the aging auditory system. Nevertheless, with the exception of exogenous ototraumatic factors (such as noise exposure) and potential effects of hormone replacement therapy and serum hormone levels, biological explanations for gender differences in presbyacusis remain unclear. The results obtained here suggest that further investigation of gender differences in presbyacusis is warranted.

3. Effects of noise history

It was also of interest to determine if noise history had a significant effect on the rate of change in adjusted word recognition. In an analysis of pure-tone thresholds from the Framingham Heart Study, Gates *et al.* (2000) reported that age-related increases in thresholds for older subjects whose audiograms had a noise "notch" were faster at frequencies adjacent to the notch than for older subjects whose audiograms had no notch. Because the notched audiograms were attributed to prior noise exposure, the authors concluded that the effect of noise on pure-tone thresholds could continue long after the noise exposure has stopped. To test this hypothesis, a similar analysis was conducted on pure-tone thresholds from the MUSC longitudinal study. Rates of threshold change for subjects with a positive noise history did not differ significantly from those with a negative noise history [Lee *et al.* (2005); see also Gates (2006) and Lee *et al.* (2006) for further discussion].

To assess the effect of noise exposure on the rate of change in word recognition, the results of a seven-item noise history questionnaire on occupational and nonoccupational noise exposure were used as an index of noise exposure. Subjects answered yes or no to questions related to noisy work environments (including the military), and exposure to noise from guns, music, power tools, and farm machinery. In the current study, 111 of the 256 subjects reported a positive noise history (related primarily to occupational noise exposure). Although noise history had a significant effect on absolute word-recognition scores (due to higher pure-tone thresholds), noise history did not have a significant effect on the rate of change of adjusted word-recognition scores ($p=0.372$) and the noise-history effect did not differ significantly between genders ($p=0.427$). Thus, noise history had no effect on age-related declines in word recognition by older persons. Moreover, noise history does not explain the gender-related difference in declines in word recognition, as discussed in Sec. III A 2.

B. Longitudinal changes in observed and predicted word-recognition scores

Rates of change in word-recognition scores and adjusted word-recognition scores estimated from slopes of regression functions provide a method to quantify the longitudinal change in word recognition for each individual. However, rate-of-change values cannot reveal the relationship between observed and predicted scores through time. An alternative method of assessing changes in observed and predicted word-recognition scores over time is to group data according to laboratory visit and plot mean scores against subjects' mean ages at each visit (as shown in Fig. 2 for speech presentation level and PTA). Figure 6 (top) presents mean (±1 SE) observed word-recognition scores (closed) and mean (±1 SE) predicted scores (open) for laboratory visits 1–13 during which NU#6 scores were obtained plotted against mean subject age during those visits (21 scores obtained during laboratory visits 14–18 are not shown). The larger variance in the observed scores and the flattening of the predicted scores at older ages were due to smaller sample sizes.

Mean observed scores declined from the first to last visit from 83.6% to 70.5%. Scores predicted by the AI declined less over the same time span, from 80.5% to 78.5%. That is, changes in speech audibility over time, independent of age, predicted only a 2% decline in mean word-recognition scores. Comparing observed and predicted changes with age in Fig. 6 (top) shows that mean observed scores at younger ages (⩽74 years) were better than predicted whereas mean observed scores at older ages (>74 years) were worse than predicted. The observed−predicted difference function (bottom panel, solid line) has a negative slope. That is, as subjects aged, their observed scores deviated increasingly from their predicted scores at a rate of −0.79%/year, which was significantly different from zero ($p<0.0001$) and similar to the mean rate of decline in adjusted word recognition reported in Sec. III A (−0.74%/year). The linear fit of the difference function further suggests that this decline did not

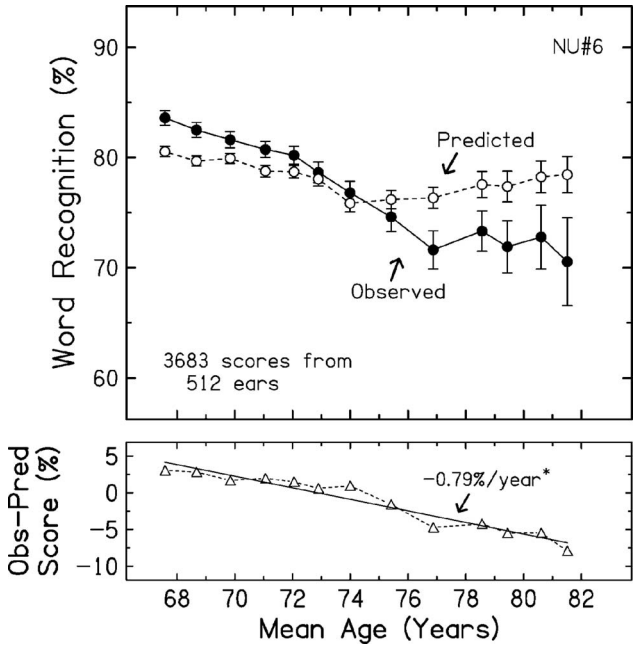


FIG. 6. Top panel: Mean (±1 SE) observed word-recognition scores (in percent correct, closed) and mean (±1 SE) scores predicted by the articulation index (in percent correct, open) for laboratory visits 1–13 at which NU#6 scores were obtained plotted as a function of mean age during those visits (21 scores obtained during laboratory visits 14–18 are not shown). Bottom panel: Mean differences between observed and predicted scores (triangles) for the same laboratory visits plotted as a function of mean age during those visits. A linear regression function fit to observed−predicted differences is shown (solid line), along with the slope (%/year). The asterisk indicates statistical significance.

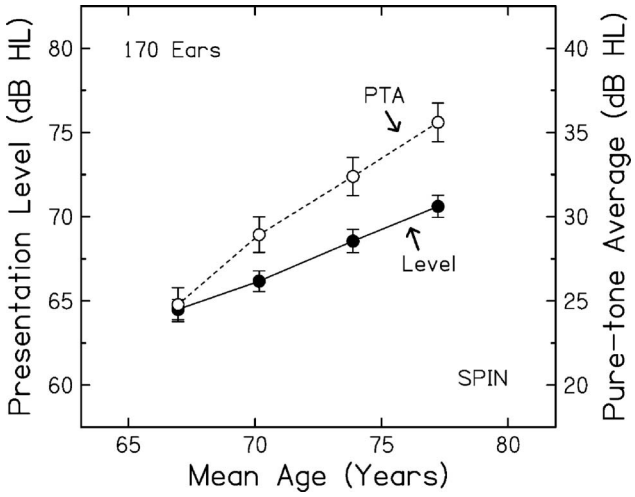


FIG. 7. Mean (±1 SE) speech presentation levels in dB HL for SPIN sentences (closed, left ordinate) for the four laboratory visits at which SPIN scores were obtained plotted against mean age during those visits. Also shown are mean (±1 SE) PTAs in dB HL measured during the same visits (open, right ordinate).

accelerate with age. Thus, word recognition in quiet declined significantly with age faster than predicted by declines in pure-tone thresholds.

The same analysis was conducted separately for females (1818 scores from 256 ears) and males (1865 scores from 256 ears); as in Fig. 6, functions for observed and predicted scores intersected near age 74 years. However, predicted scores for females were higher than for males, due to females' better pure-tone thresholds. Observed scores at younger ages remained better than predicted but more so for females than males. Observed scores at older ages were worse than predicted to a similar extent for females and males. Thus, although scores for both genders declined significantly ($p<0.0001$), the observed−predicted difference function was significantly more negative for females than for males (−0.97%/year vs −0.61%/year; $t_{22}=-2.87$, $p=0.009$), similar to the gender differences for mean rate of decline in adjusted word recognition reported in Sec. III A. Also as noted earlier, the relationship between observed and predicted scores with increasing age did not differ significantly for subjects with positive and negative noise histories.

C. Longitudinal changes in recognition of key words in sentences in babble

To determine if the decline in word recognition with age was a general trend or was specific to monosyllabic word recognition in quiet, observed and predicted recognition of key words in low-context sentences (PL) and high-context sentences (PH) in babble were analyzed using the same procedures as described earlier. Recall that there were fewer SPIN test scores than NU#6 test scores. SPIN scores were included for all subjects who had four repeated measures.

Figure 7 includes mean (±1 SE) presentation levels for the SPIN sentences in dB HL (closed circles, left ordinate) for the four laboratory visits at which SPIN scores were obtained plotted against mean subject age during those visits. Also shown are mean (±1 SE) PTAs in dB HL measured during the same visits (open circles, right ordinate). Over the 10.3 year average time span, the average PTA increased by 10.8 dB, or an average rate of increase in pure-tone thresholds of ~1.0 dB/year [similar to that reported by Lee *et al.* (2005)]. Over the same time span, average speech presentation levels for SPIN sentences increased by only 6.1 dB. Recall that the level of the SPIN sentences was set relative to a babble threshold that was estimated from pure-tone thresholds for each ear, specifically the best threshold from 0.25 to 4.0 kHz. Typically, the best threshold was at one of the lower frequencies, which increase at a slower rate than higher frequencies (e.g., Lee *et al.*, 2005). Thus, increases in average presentation level for the SPIN sentences did not keep up with increases in the four-frequency PTA for the time span over which SPIN scores were obtained. In contrast, average presentation level for NU#6 words generally increased in the same manner as the PTA (see Fig. 2), perhaps because levels were set relative to the measured SRT, which was similar in magnitude to the PTA.

As described earlier for NU#6 scores, the slope of a linear regression was used to estimate the rates of change in key-word recognition for low-context (PL) and high-context (PH) sentences for individual ears; the rate of change in adjusted word recognition (i.e., observed−predicted scores) was also determined. The average rate of change in scores for low-context sentences was −0.04%/year ($p=0.749$) and 0.31%/year ($p=0.008$) for adjusted scores. The comparable values for scores for high-context sentences were

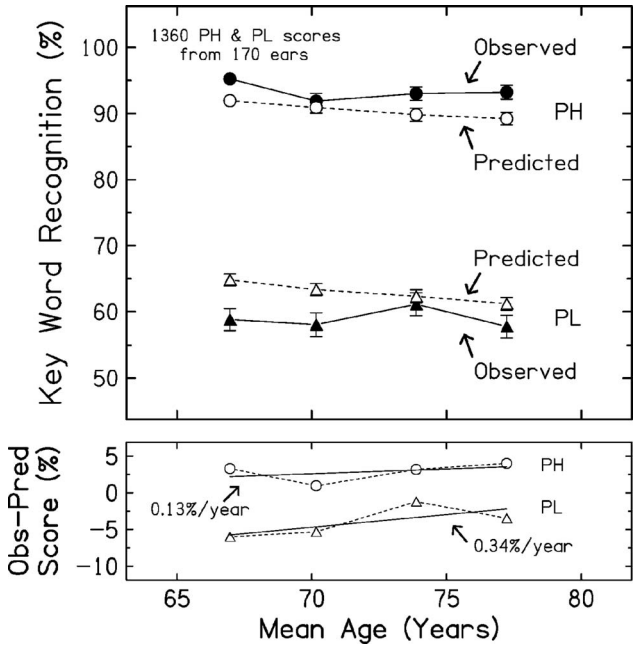


FIG. 8. Top panel: Mean (±1 SE) observed (closed) and predicted (open) recognition scores in percent correct for key words in high-context sentences (PH, circles) and low-context sentences (PL, triangles) for the four laboratory visits at which SPIN scores were obtained plotted as a function of mean age during those visits. Bottom panel: Mean differences between observed and predicted scores for high-context (PH, circles) and low-context (PL, triangles) sentences plotted as a function of age. Linear regression functions fit to observed−predicted differences are shown (solid lines), along with the slopes (%/year).

−0.14% / year ($p=0.037$) and 0.13%/year ($p=0.035$) for adjusted scores. Thus, in contrast to age-related declines in recognition of isolated monosyllabic words, key-word recognition in low- and high-context sentences did not decline with increasing age, but improved by a small, but significant amount.

An ANCOVA of the rate of change in adjusted recognition of key words in low- and high-context sentences was performed, with gender and noise history as grouping factors, and initial hearing loss (PTA) and age as covariates. Results of ANCOVA for both low- and high-context sentences showed nonsignificant effects of all factors. Finally, for females only, serum estradiol and progesterone levels were analyzed for possible hormonal effects on rate of change for word recognition in low- and high-context sentences; all effects were nonsignificant.

D. Longitudinal and cross-sectional changes in observed and predicted recognition of key words in sentences

As discussed earlier for NU#6 scores, rate-of-change values cannot reveal the relationship between observed and predicted scores through time. An alternative method is to group data according to laboratory visit and plot mean scores against subjects' mean ages at each visit. Figure 8 (top) includes mean (±1 SE) observed and predicted scores for high-context sentences (PH, circles) and low-context sentences (PL, triangles) for the four laboratory visits at which SPIN scores were obtained plotted against mean age during those visits. Observed PH scores initially declined, then recovered, and were consistently better than predicted, which is a different pattern than for observed NU#6 scores (Fig. 6). PH scores were predicted to decline with increasing age, based on changes in speech audibility, similar to predicted changes in NU#6 scores. The observed−predicted difference function (solid line, bottom panel) shows that the change with age was not significantly different from zero ($p=0.553$) and was similar in magnitude to the mean rate of change in adjusted PH scores reported in Sec. III C. Better-than-predicted PH scores are consistent with earlier findings suggesting that older subjects benefit from semantic context in sentences at least as well as younger subjects (Dubno *et al.*, 2000).

For sentences with syntactic but not semantic context (PL), observed scores improved and then declined with increasing age. With PL scores predicted to decline steadily with age, the deficit in subjects' performance appeared to decrease as subjects aged. The observed−predicted difference function for PL scores (solid line, bottom panel) showed a nonsignificant change of 0.34%/year ($p=0.285$), similar in magnitude to the mean rate of change in adjusted PL scores reported in Sec. III C. Age-related deficits that decreased over time were unexpected, based on the findings for monosyllabic word recognition in quiet and the assumption that changes with age would be even more pronounced for the more-complex sentence-recognition task in babble. In an effort to explain this result, effects of gender, initial PTA, and initial age on longitudinal changes in adjusted low- and high-context sentence scores were examined, and were found to be nonsignificant.

To explore additional reasons for the absence of significant age-related declines in recognition of key words in low- and high-context sentences in babble, effects of age on adjusted scores were examined cross sectionally. Figure 9 shows observed−predicted differences in scores for high-context (PH) sentences plotted against subject age for each of the four SPIN tests. Figure 9 includes the same scores from the same subjects as in the longitudinal analysis; note the increase in subjects' ages from Test 1 through Test 4. For recognition of words in sentences with context, adjusted scores did not change significantly with increasing age, except when subjects were at their most advanced age (PH Test 4, −0.32% / year, $p=0.039$).

Figure 10 shows the same results for low-context (PL) sentences. In each panel, the slope of the linear regression function is negative and becomes more negative from Test 1 to Test 4 (averaging approximately −0.75% / year across the four tests; p values ranged from 0.029 to <0.0001). These cross-sectional results suggest that scores were increasingly poorer than predicted as subjects aged, in contrast to conclusions based on longitudinal data. Among studies that report both longitudinal and cross-sectional analyses of pure-tone thresholds or speech-recognition scores, most reveal larger age-related changes from longitudinal than cross-sectional data (e.g., Bergman *et al.* 1976; Brant and Fozard, 1990; Hietanen *et al.*, 2004), although Pearson *et al.* (1995) described the differences as insignificant. In the current study, the longitudinal analyses of adjusted word recognition in low- and high-context sentences and the cross-sectional

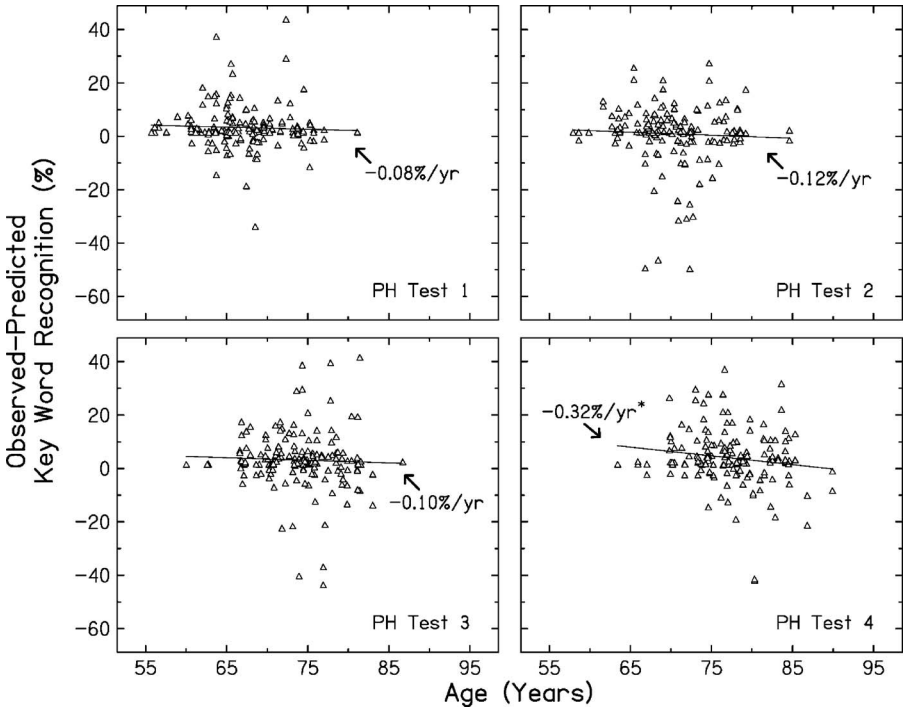


FIG. 9. Observed minus predicted recognition scores in percent correct for key words in high-context sentences (PH) plotted as a function of age. Scores were obtained from the same subjects at each of four time points (Test 1 through Test 4) and results are shown cross sectionally. A linear regression function (solid line) and slope (%/year) is included in each panel. Asterisk indicates statistical significance.

analyses in Figs. 9 and 10 were based on the same data set, but the results were inconsistent. Given differences in the assumptions of the statistical models, it is not surprising to see different results. That is, the longitudinal model assumes that trials were repeated measures from the same subjects whereas cross-sectional analyses assume all data were independent samples.

E. Comparing longitudinal changes in recognition of isolated words and words in sentences

There were several differences between the word-recognition (NU#6) and sentence-recognition (SPIN) measures, some of which may explain differences in longitudinal effects. For example, the NU#6 test measures recognition for words in isolation in quiet whereas the SPIN test measures recognition of words in sentences with a babble background. Speech levels for the NU#6 words ranged from 46 to 62 dB HL (20–25 dB above the PTA) and mean scores ranged from 70%–84%. Speech levels for low-context sentences ranged from 64 to 71 dB HL (40–45 dB above the PTA) and mean scores ranged from 58% to 61%. Each of these factors could have a differential effect on audibility estimates computed by the AI and determination of predicted scores. In addition, learning, practice, and list effects (if any) for repeated mea-

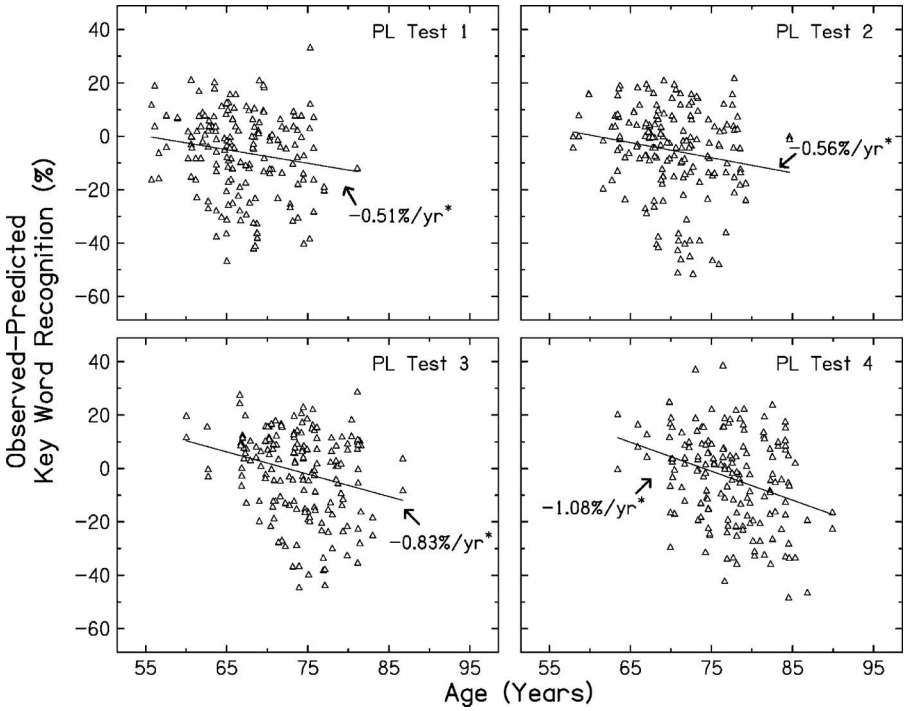


FIG. 10. Same as Fig. 9, but for recognition of key words in low-context sentences (PL).

surements of words in sentences may be different than for repeated measurements of isolated words, which could influence the magnitude of longitudinal changes in recognition. Finally, the size of the subject sample, the number of repeated measures, and the timing of the measures, were different for word-recognition and sentence-recognition scores, which may affect the accuracy of the estimated trends.

Recall that word recognition was measured annually and sentence recognition was measured every 2 to 3 years. One outcome of these different testing schedules was that subjects who provided longitudinal data for SPIN remained in the study 3 years longer (on average) than subjects who provided longitudinal data for NU#6. If higher performing or healthier older persons remain in longitudinal studies for longer periods of time, a potential concern is that the subjects who provided SPIN results may be a higher performing or healthier subset of the subjects who provided NU#6 results. To answer this question, the 256 subjects who provided longitudinal NU#6 scores were split into two subgroups: (1) those who completed four SPIN tests ($N=85$) and provided longitudinal SPIN scores and (2) those who completed less than four SPIN tests ($N=171$). The two groups did not differ significantly in initial age ($t_{254}=1.573$; $p=0.117$) and, after gender ratio was equated, did not differ significantly in initial PTA ($t_{466}=1.765$; $p=0.078$). However, adjusted NU#6 scores declined significantly faster for the subjects who completed four SPIN tests (−0.99%/year) than for subjects with less than four SPIN tests (−0.62%/year; $t_{510}=1.993$; $p=0.047$). These results do not support the assumption that the subjects who provided longitudinal SPIN results were higher performing or healthier older adults, at least in terms of their age, degree of hearing loss, and declines in recognition of isolated words. To assess selective attrition for the study as a whole, the effect of general health on subjects' time in the study was determined. For this analysis, general health was operationally defined as the number of chronic diseases reported on the initial medical history. For 465 subjects no longer enrolled, results of a one-way ANOVA revealed that time enrolled in the study did not vary significantly with subjects' general health, even after adjusting for differences due to gender and age [$F(1,461)=0.243$, $p=0.622$]. This result suggests that selective attrition may not be a major factor in the MUSC longitudinal study of age-related hearing loss.

Finally, it is also possible that mechanisms engaged in recognition of words presented with syntactic and semantic context (PL and PH sentences) may differ from those engaged in recognition of words in isolation (NU#6 words) and may be less susceptible to the effects of age. Evidence to support differing mechanisms or susceptibilities is the poor association between these measures, as shown by nonsignificant correlations between rate of change in adjusted recognition for isolated words and for words in low-context sentences ($r=0.073$, $p=0.347$) and high-context sentences ($r=0.038$, $p=0.624$).

IV. SUMMARY AND CONCLUSIONS

Recognition of isolated monosyllabic words (NU#6) in quiet and recognition of key words in low- and high-context sentences in babble (SPIN) were measured in a large sample of older persons enrolled in the MUSC longitudinal study of age-related hearing loss. Repeated measures were obtained yearly (NU#6) or every 2 to 3 years (SPIN). To control for age-related changes in pure-tone thresholds and speech levels, speech-recognition scores were adjusted using an importance-weighted speech-audibility metric (AI). Linear-regression slope was used to estimate the rate of change in adjusted speech-recognition scores. Slopes were analyzed for longitudinal changes and to determine effects of gender, initial thresholds, age, and noise history on these changes, and to assess differences between observed changes with age and changes predicted by the AI. Results may be summarized as follows.

(1) Monosyllabic word recognition in quiet (NU#6) declined with increasing age at a rate of 0.74%/year, even after accounting for reductions in speech audibility due to concurrent changes in hearing and speech levels. Thus, word recognition in quiet declined significantly faster with age than predicted by declines in speech audibility.
(2) Observed word-recognition scores at younger ages ($\leqslant$74 years) were better than predicted by the AI, whereas observed scores at older ages ($>$74 years) were worse than predicted. After age 74, observed scores deviated increasingly from predicted scores, but this effect did not accelerate with age.
(3) Rate of decline in word recognition in quiet was significantly faster for females than for males, even while taking into account gender-related differences in speech audibility due to threshold differences. Females with higher serum progesterone levels had faster declines in word recognition than females with lower levels of progesterone. Noise history had no effect on age-related declines in word recognition.
(4) Rate of decline in word recognition increased significantly as initial hearing levels increased, suggesting that with more severe injury to the auditory system, impairments in auditory function other than elevated thresholds resulted in faster declines in word recognition as subjects aged. However, the rate of decline in word recognition did not accelerate with age. Taken together, these age-related declines in word recognition were more consistent with underlying changes in auditory, rather than cognitive, function resulting from peripheral, rather than central, auditory system pathology. A "common cause" hypothesis relating age-related declines to a single neural mechanism also provides a reasonable account of the results.
(5) In contrast to age-related declines in recognition of monosyllabic words in quiet, recognition of key words in low- and high-context sentences in babble did not decline significantly with increasing age. Evidence of age-related changes in word recognition in low-context sentences was revealed when results were viewed cross sectionally.

ACKNOWLEDGMENTS

This work was supported (in part) by Grant Nos. P50 DC00422 and R01 DC00184 from NIH/NIDCD and the MUSC General Clinical Research Center (M01 RR 01070). This investigation was conducted in a facility constructed with support from Research Facilities Improvement Program Grant No. C06 RR14516 from the National Center for Research Resources, National Institutes of Health. The authors thank Sarah Ferguson, Tracy Fitzgerald, Kelly C. Harris, Dawn Konrad-Martin, Johanna Larsen, Elizabeth A. Poth, and Christine Strange for assistance with data collection. Valuable suggestions for improving the manuscript were provided by Associate Editor Mitchell Sommers and two anonymous reviewers.

[1]One subject was enrolled in the study at age 50.

[2]For subjects who had repeated measures of serum hormone levels, mean levels were used. Of the 108 female subjects, 56% reported a history of hormone replacement therapy. Of these subjects, 62% reported taking estrogen alone and 38% reported taking a combination of estrogen and progestin.

American National Standards Institute (**1969a**). "American National Standard methods for the calculation of the Articulation Index," ANSI S3.5–1969 (American National Standards Institute, Inc., New York).

American National Standards Institute (**1969b**). "American National Standard specification for audiometers," ANSI S3.6–1969 (American National Standards Institute, Inc., New York).

American National Standards Institute (**1989**). "American National Standard specification for audiometers," ANSI S3.6–1989 (American National Standards Institute, Inc., New York).

American National Standards Institute (**1996**). "American National Standard specification for audiometers," ANSI S3.6–1996 (American National Standards Institute, Inc., New York).

American National Standards Institute (**1997**). "American National Standard methods for the calculation of the Speech Intelligibility Index," ANSI S3.5–1997 (American National Standards Institute, Inc., New York).

American National Standards Institute (**2004**). "American National Standard specification for audiometers," ANSI S3.6–2004 (American National Standards Institute, Inc., New York).

Baltes, P. B., and Lindenberger, U. (**1997**). "Emergence of a powerful connection between sensory and cognitive functions across the adult life span: A new window to the study of cognitive aging?," Psychol. Aging **12**, 12–21.

Bell, T. S., Dirks, D. D., and Trine, T. D. (**1992**). "Frequency-importance functions for words in high- and low-context sentences," J. Speech Hear. Res. **35**, 950–959.

Bergman, M., Blumenfeld, V. G., Cascardo, D., Dash, B., Levitt, H., and Margulies, M. K. (**1976**). "Age-related decrement in hearing for speech," J. Gerontol. **31**, 533–538.

Bilger, R. (**1984**). *Manual for the Clinical Use of the Revised SPIN Test* (University of Illinois Press, Champaign, IL).

Brant, L. J., and Fozard, J. L. (**1990**). "Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging," J. Acoust. Soc. Am. **88**, 813–820.

Cooper, N. P., and Rhode, W. S. (**1997**). "Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea," J. Neurophysiol. **78**, 261–270.

Cruickshanks, K. J., Tweed, T. S., Wiley, T. L., Klein, B. E. K., Klein, R., Chappel, R., Nondahl, D. M., and Dalton, D. S. (**2003**). "The 5-year incidence and progression of hearing loss: The epidemiology of hearing loss study," Arch. Otolaryngol. Head Neck Surg. **129**, 1041–1046.

Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E. K., Klein, R., Mares-Perlman, J. A., and Nondahl, D. M. (**1998**). "Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin. The epidemiology of hearing loss study," Am. J. Epidemiol. **148**, 879–886.

Davis, A. C., Ostri, B., and Parving, A. (**1991**). "Longitudinal study of hearing," Acta Oto-Laryngol., Suppl. **476**, 12–22.

Divenyi, P. L., and Haupt, K. M. (**1997**). "Audiological correlates of speech understanding deficits in elderly listeners with mild-to-moderate hearing loss. I. Age and laterality effects," Ear Hear. **18**, 42–61.

Divenyi, P. L., Stark, P. B., and Haupt, K. M. (**2005**). "Decline of speech understanding and auditory thresholds in the elderly," J. Acoust. Soc. Am. **118**, 1089–1100.

Dubno, J. R., Ahlstrom, J. B., and Horwitz, A. R. (**2000**). "Use of context by younger and older adults with normal hearing," J. Acoust. Soc. Am. **107**, 538–546.

Dubno, J. R., Lee, F. S., Klein, A. J., Matthews, L. J., and Lam, C. (**1995**). "Confidence limits for maximum word-recognitions scores," J. Speech Hear. Res. **38**, 490–502.

Dubno, J. R., Lee, F. S., Matthews, L. J., and Mills, J. H. (**1997**). "Age-related and gender-related changes in monaural speech recognition," J. Speech Hear. Res. **40**, 444–452.

Egan, J. (**1948**). "Articulation testing methods," Laryngoscope **58**, 955–991.

Gates, G. A. (**2006**). "Letter to the editor," Ear Hear. **27**, 91.

Gates, G. A., and Cooper, J. C. (**1991**). "Incidence of hearing decline in the elderly," Acta Oto-Laryngol. **111**, 240–248.

Gates, G. A., Cooper, J. C., Kannel, W. B., and Miller, N. J. (**1990**). "Hearing in the elderly: The Framingham cohort, 1983–1985. I. Basic audiometric test results," Ear Hear. **4**, 247–256.

Gates, G. A., Schmid, P., Kujawa, S. G., Nam, B., and D'Agostino, R. (**2000**). "Longitudinal threshold changes in older males with audiometric notches," Hear. Res. **141**, 220–228.

Golding, M., Carter, N., Mitchell, P., and Hood, L. J. (**2004**). "Prevalence of central auditory processing (CAP) abnormality in an older Australian population: The Blue Mountains Hearing Study," J. Am. Acad. Audiol **15**, 633–642.

Golding, M., Mitchell, P., and Cupples, L. (**2005**). "Risk markers for the graded severity of auditory processing abnormality in an older Australian population: The Blue Mountains Hearing Study," J. Am. Acad. Audiol **16**, 348–356.

Golding, M., Taylor, A., Cupples, L., and Mitchell, P. (**2006**). "Odds of demonstrating auditory processing abnormality in the average older adult: The Blue Mountains Hearing Study," Ear Hear. **27**, 129–138.

Gordon-Salant, S. (**2005**). "Hearing loss and aging: New research findings and clinical implications," J. Rehabil. Res. Dev. Clin. Suppl. **42**, 9–24.

Gratton, M. A., Schmiedt, R. A., and Schulte, B. A. (**1996**). "Age-related decreases in endocochlear potential are associated with vascular abnormalities in the stria vascularis," Hear. Res. **94**, 116–124.

Guimaraes, P., Frisina, S. T., Mapes, F., Tadros, S. F., Frisina, D. R., and Frisina, R. D. (**2006**). "Progestin negatively affects hearing in aged women," Proc. Natl. Acad. Sci. U.S.A. **103**, 14246–14249.

Helfer, K. S. (**2004**). "Cross-sectional study of differences in speech understanding between users and nonusers of estrogen replacement therapy," Exp. Aging Res. **30**, 195–204.

Hietanen, A., Era, P., Sorri, M., and Heikkinen, E. (**2004**). "Changes in hearing in 80-year-old people: A 10-year follow-up study," Int. J. Audiol. **43**, 126–135.

Hirsh, I., Davis, H., Silverman, S., Reynolds, E., Eldert, E., and Benson, R. (**1952**). "Development of materials for speech audiometry," J. Speech Hear Disord. **17**, 726–735.

Hultcrantz, M., Simonoska, R., and Stenberg, A. E. (**2006**). "Estrogen and hearing: A summary of recent findings," Acta Oto-Laryngol. **126**, 10–14.

Humes, L. E. (**1996**). "Speech understanding in the elderly," J. Am. Acad. Audiol **7**, 161–167.

Humes, L. E., Watson, B. U., Christensen, L. A., Cokely, C. G., Halling, D. C., and Lee, L. (**1994**). "Factors associated with individual differences in clinical measures of speech recognition among the elderly," J. Speech Hear. Res. **37**, 465–474.

Jerger, J. (**1973**). "Audiological findings in aging," Adv. Oto-Rhino-Laryngol. **20**, 115–124.

Jerger, J. (**1990**). "Can age-related decline in speech recognition be explained by peripheral hearing loss?," in *Presbyacusis and Other Age Related Aspects—14th Danavox Symposium*, edited by J. H. Jensen (Danavox, Copenhagen), pp. 193–203.

Jerger, J. (**1992**). "Can age-related decline in speech understanding be explained by peripheral hearing loss?," J. Am. Acad. Audiol **3**, 33–38.

Jerger, J., and Hayes, D. (**1977**). "Diagnostic speech audiometry," Arch. Otolaryngol. **103**, 216–222.

Jerger, J., Jerger, S., and Pirozzolo, F. (**1991**). "Correlational analysis of speech audiometric scores, hearing loss, age, and cognitive abilities in the elderly," Ear Hear. **12**, 103–109.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of

a test of speech intelligibility in noise using test materials with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.
Kilicdag, E. B., Yavuz, H., Bagis, T., Tarim, E., Erkan, A. N., and Kazanci, F. (**2004**). "Effects of estrogen therapy on hearing in postmenopausal women," Am. J. Obstet. Gynecol. **190**, 77–82.
Kim, S. H., Kang, B. M., Chae, H. D., and Kim, C. H. (**2002**). "The association between serum estradiol level and hearing sensitivity in postmenopausal women," Obstet. Gynecol. (N.Y., NY, U. S.) **99**, 726–730.
Lee, F. S., Matthews, L. J., Dubno, J. R., and Mills, J. H. (**2005**). "Longitudinal study of pure-tone thresholds in older persons," Ear Hear. **26**, 1–11.
Lee, F. S., Matthews, L. J., Dubno, J. R., and Mills, J. H. (**2006**). "Thresholds of older persons: A reply to Gates (2006)," Ear Hear. **27**, 92.
Lee, F. S., Matthews, L. J., Mills, J. H., Dubno, J. R., and Adkins, W. Y. (**1998a**). "Analysis of blood chemistry and hearing levels in a sample of older persons," Ear Hear. **19**, 180–190.
Lee, F. S., Matthews, L. J., Mills, J. H., Dubno, J. R., and Adkins, W. Y. (**1998b**). "Gender-specific effects of medicinal drugs on hearing levels of older persons," Otolaryngol.-Head Neck Surg. **118**, 221–227.
Lindenberger, U., and Baltes, P. B. (**1994**). "Sensory functioning and intelligence in old age: A strong connection," Psychol. Aging **9**, 339–355.
Møller, M. B. (**1981**). "Hearing in 70 and 75 year old people: Results from a cross sectional and longitudinal population study," Am. J. Otol. **2**, 22–29.
Mościcki, E. K., Elkins, E. F., Baum, H. M., and McNamara, P. M. (**1985**). "Hearing loss in the elderly: An epidemiologic study of the Framingham heart study cohort," Ear Hear. **6**, 184–190.
Pavlovic, C. (**2006**). "The speech intelligibility index standard and its relationship to the articulation index, and the speech transmission index," J. Acoust. Soc. Am. **119**, 3326.
Pearson, J. D., Morrell, C. H., Gordon-Salant, S., Brant, L. J., Metter, E. J., Klein, L. L., and Fozard, J. L. (**1995**). "Gender differences in a longitudinal study of age-associated hearing loss," J. Acoust. Soc. Am. **97**, 1196–1205.
Pedersen, K. E., Rosenhall, U., and Møller, M. B. (**1991**). "Longitudinal study of changes in speech perception between 70 and 81 years of age," Audiology **30**, 201–211.
Pfeiffer, E. (**1975**). "A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients," J. Am. Geriatr. Soc. **23**, 433–441.
Popelka, G. R., and Mason, D. I. (**1987**). "Factors which affect measures of speech audibility with hearing aids," Ear Hear. **8**, 109–118.
Schmiedt, R. A., Lang, H., Okamura, H., and Schulte, B. A. (**2002**). "Effects of furosemide applied chronically to the round window: A model of metabolic presbyacusis," J. Neurosci. **22**, 9643–9650.
Schmiedt, R. A., Mills, J. H., and Adams, J. (**1990**). "Tuning and suppression in auditory nerve fibers of aged gerbils raised in quiet or noise," Hear. Res. **45**, 221–236.
Schulte, B. A., and Schmiedt, R. A. (**1992**). "Lateral wall Na,K-ATPase and endocochlear potentials decline with age in quiet-reared gerbils," Hear. Res. **61**, 35–46.
Speaks, C., and Jerger, J. (**1965**). "Method for measurement of speech identification," J. Speech Hear. Res. **8**, 185–194.
Studebaker, G. A., Sherbecoe, R. L., and Gilmore, C. (**1993**). "Frequency-importance and transfer functions for the Auditec of St. Louis recordings of the NU-6 word test," J. Speech Hear. Res. **36**, 799–807.
Tillman, T. W., and Carhart, R. (**1966**). "An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6," Technical Report No. SAM-TR-66-55, USAF School of Aerospace Medicine, Brooks Air Force Base, TX, pp. 1–12.
van Rooij, J. C. G. M., and Plomp, R. (**1990**). "Auditive and cognitive factors in speech perception by elderly listeners. II. Multivariate analyses," J. Acoust. Soc. Am. **88**, 2611–2624.
van Rooij, J. C. G. M., and Plomp, R. (**1992**). "Auditive and cognitive factors in speech perception by elderly listeners. III. Additional data and final discussion," J. Acoust. Soc. Am. **91**, 1028–1033.
van Rooij, J. C. G. M., Plomp, R., and Orlebke, J. F. (**1989**). "Auditive and cognitive factors in speech perception by elderly listeners. I. Development of test battery," J. Acoust. Soc. Am. **86**, 1294–1309.
Wiley, T. L., Cruickshanks, K. J., Nondahl, D. M., Tweed, T. S., Klein, R., and Klein, B. E. K. (**1998**). "Aging and word recognition in competing message," J. Am. Acad. Audiol **9**, 191–198.

Effect of age, presentation method, and learning on identification of noise-vocoded words

Signy Sheldon, M. Kathleen Pichora-Fuller,[a] and Bruce A. Schneider
Department of Psychology, University of Toronto, 3359 Mississauga Road North, Mississauga, Ontario L5L 1C6, Canada

(Received 31 October 2006; revised 2 October 2007; accepted 15 October 2007)

Noise vocoding was used to investigate the ability of younger and older adults with normal audiometric thresholds in the speech range to use amplitude envelope cues to identify words. In Experiment 1, four 50-word lists were tested, with each word presented initially with one frequency band and the number of bands being incremented until it was correctly identified by the listener. Both age groups required an average of 5.25 bands for 50% correct word identification and performance improved across the four lists. In Experiment 2, the same participants who completed Experiment 1 identified words in four blocked noise-vocoded conditions (16, 8, 4, 2 bands). Compared to Experiment 1, both age groups required more bands to reach the 50% correct word identification threshold in Experiment 2, 6.13, and 8.55 bands, respectively, with younger adults outperforming older adults. Experiment 3 was identical to Experiment 2 except the participants had no prior experience with noise-vocoded speech. Again, younger adults outperformed older adults, with thresholds of 6.67 and 8.97 bands, respectively. The finding of age effects in Experiments 2 and 3, but not in Experiment 1, seems more likely to be related to differences in the presentation methods than to experience with noise vocoding. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2805676]

PACS number(s): 43.71.Lz, 43.66.Sr, 43.71.Gv [MSS] Pages: 476–488

I. INTRODUCTION

Older adults often report more difficulties than younger adults in understanding spoken language, especially in adverse listening conditions (for reviews see CHABA, 1988; Divenyi and Simon, 1999; Pichora-Fuller and Souza, 2003). Even older adults with normal audiometric thresholds in the speech range have these difficulties, so it seems unlikely that loss of audibility can fully explain their poor comprehension. Given that auditory temporal processing is an integral part of spoken language comprehension (e.g., de Boer and Dreschler, 1987; van Tasell *et al.*, 1987; Rosen, 1992; Shannon *et al.*, 1995), many have suggested that age-related declines in auditory temporal processing may contribute to the comprehension difficulties that older adults often have in challenging conditions (for reviews see Fitzgibbons and Gordon-Salant, 1996; Schneider and Pichora-Fuller, 2001; Versfeld and Dreschler, 2002; Pichora-Fuller and Souza, 2003).

A. Effect of age on speech and temporal processing

Temporally coded cues relevant to speech processing have been described at three levels: subsegmental (voice), segmental (phonemic), and suprasegmental (syllabic and lexico-syntactic). Subsegmental fine structure cues include periodicity cues based on the fundamental frequency and harmonic structure of the voice. Segmental information is provided by local gap and duration cues in the envelope which contribute to phoneme identification (e.g., presence or absence of a stop consonant, voicing). Suprasegmental cues, such as amplitude fluctuations in the region of 3–20 Hz, convey prosodic information involving the rate and rhythm of speech and are used in lexical and syntactic processing (Rosen, 1992; Philips, 1995; Greenberg, 1996; Schneider and Pichora-Fuller, 2001; Shannon, 2002; Pichora-Fuller and Souza, 2003). It is important to know whether age affects the temporal processing of speech cues at one or more of these three levels within and across spectral regions [for a discussion see also Souza and Boike (2006)]. While there is strong evidence of age effects on aspects of auditory temporal processing that are relevant to speech processing at the subsegmental and segmental levels, less is known about how age affects the processing of suprasegmental speech cues.

With respect to the subsegmental level of temporal processing, physiological and behavioral studies suggest that there are age-related decrements in synchrony coding at various stages of auditory processing which could undermine the periodicity coding of speech and nonspeech signals (e.g., Frisina, 2001; Pichora-Fuller *et al.*, 2007). For example, the pattern of binaural masking-level differences in younger and older adults suggests that the precision of periodicity coding is reduced with age (Pichora-Fuller and Schneider, 1992). Monaurally, frequency difference limens at lower frequencies are larger for older than for younger adults, but this age-related difference is less marked at higher frequencies where periodicity coding does not play as significant a role (Abel *et al.*, 1990). Furthermore, older adults have more difficulty discriminating a mistuned harmonic in a harmonic complex, especially for short duration sounds (Alain *et al.*, 2001), and older adults have more problems than younger

[a]Electronic mail: k.pichora.fuller@utoronto.ca

adults in segregating concurrent vowels (Summers and Leek, 1998; Vongpaisal and Pichora-Fuller, 2007). These studies demonstrate the mounting evidence pointing to age-related deficits in the temporal processing of subsegmental or fine-structure cues that are believed to play a role in voice identification and segregation (e.g., de Cheveigné, 2003).

Other studies have demonstrated age-related differences in temporal processing relevant to the segmental level. The effects of age are seen clearly in a large number of gap detection experiments in which listeners are asked to detect the presence of a gap between sound markers. In general, older adults do not detect a gap until it is significantly longer than the smallest gap that can be detected by younger adults for either nonspeech or speech markers (e.g., Snell, 1996; Pichora-Fuller *et al.*, 2006). For example, the thresholds of older adults were approximately twice as large as those of younger adults when detecting a gap between two Gaussian-enveloped tone pips (Schneider *et al.*, 1994). Similarly, mean gap thresholds were significantly larger for older listeners compared to younger listeners for low-passed filtered noise bursts (Snell and Frisina, 2000). Age-related deficits in temporal processing of segmental-level speech cues may also arise from declines in duration discrimination. In duration discrimination studies, younger and older listeners are asked to identify the longer or shorter of two stimuli. Older adults have more difficulty with this task than younger adults for both nonspeech and speech stimuli (e.g., Abel *et al.*, 1990; Bergerson *et al.*, 2001; Fitzgibbons and Gordon-Salant, 1994, 1995; Gordon-Salant *et al.*, 2006).

Suprasegmental cues involving variations in pitch, loudness, and/or timing contribute to speech prosody (e.g., Cutler *et al.*, 1997). Variations in timing include changes in the duration of phonemes and words to alter the rate and rhythm of speech. Previous research provides evidence that older adults may be more disadvantaged than younger adults when prosodic cues are disrupted by various types of temporal distortion, but that they may benefit as much or more than younger adults when prosodic cues are available, especially in challenging listening conditions.

On the one hand, various methods of temporally distorting speech, including speeding or time compression, have a more deleterious effect on the spoken language understanding of older compared to younger listeners on word tests (e.g., Sticht and Gray, 1969; Konkle *et al.*, 1977; Stuart and Phillips, 1996), sentence tests (e.g., Gordon-Salant and Fitzgibbons, 1993, 1997), and discourse tests (e.g., Vaughan and Letowski, 1997; for a review see Wingfield, 1996). In addition to the possible cognitive strain introduced when speech is speeded, speeding speech may also have negative consequences on the comprehension abilities of older adults because their auditory systems are more susceptible than are those of younger adults to the acoustical temporal distortions that some methods introduce in the speech signal (e.g., Wingfield *et al.*, 1999; Schneider *et al.*, 2005). Since time compression alters the envelope of speech, it may be that older adults are more affected than younger adults by distortions in the shape of the envelope that might compromise the auditory processing of cues used at the phonemic, lexical, and/or syntactic levels.

On the other hand, evidence that the use of envelope cues is preserved in older adults comes from studies showing that younger and older adults both benefit from duration and envelope cues to identify words (e.g., Wingfield *et al.*, 2000), and that both age groups benefit from the insertion of pauses to understand sentences (e.g., Wingfield *et al.*, 1999). There is also evidence that older adults benefit more than younger adults from prosodic cuing to understand speeded sentences (e.g., Wingfield *et al.*, 1992), or short passages (e.g., Stine and Wingfield, 1987). Furthermore, in conversational discourse older adults demonstrate preserved use of prosody to understand socio-emotional relational information even when hearing loss impedes the understanding of content because of poor ability to identify phonemes and words (Villaume *et al.*, 1994).

Since temporal amplitude envelope cues are among the suprasegmental cues that contribute to speech prosody and such cues can be used even by those with significant hearing loss (Turner *et al.*, 1995), it seems likely that the findings of preserved ability in older adults to use prosody and their greater reliance on it in challenging conditions may be, at least partially, attributable to their good ability to use envelope information when they are listening to words, sentences, or discourse. Souza and Boike (2006) suggest a disassociation between auditory cue use and age—younger listeners tend to use both temporal envelope and fine structure cues, whereas older listeners seem to rely more heavily on suprasegmental envelope cues than on spectral or temporal fine structure cues. Thus, one hypothesis is that use of these suprasegmental cues does not decline with age, but rather that the use of suprasegmental cues may even compensate for the reduced ability of older listeners to use subsegmental and segmental speech cues.

The divergent theories concerning the abilities of older adults to use temporal amplitude envelope cues make it an issue that warrants further investigation. Therefore, the primary goal of this study is to examine the effect of age on the ability of listeners to use envelope cues when fine structure cues are minimized. To this end, we tested word identification in younger and older listeners using noise-vocoded speech.

B. Noise vocoding

Noise vocoding is a form of speech distortion that involves dividing a speech signal into specific frequency bands, and then, within each band, extracting the temporal amplitude envelope and using it to modulate noise of the same bandwidth. In effect, the fine structure of the signal is replaced with noise. Thus, noise vocoding preserves temporal amplitude envelope cues within specific frequency bands and eliminates fine structure cues, including periodicity cues (see Fig. 1). As the number of bands is increased, more band-specific envelope information becomes available. It has been shown that the intelligibility of noise-vocoded speech stimuli is dependent on the number of frequency bands used in voc-

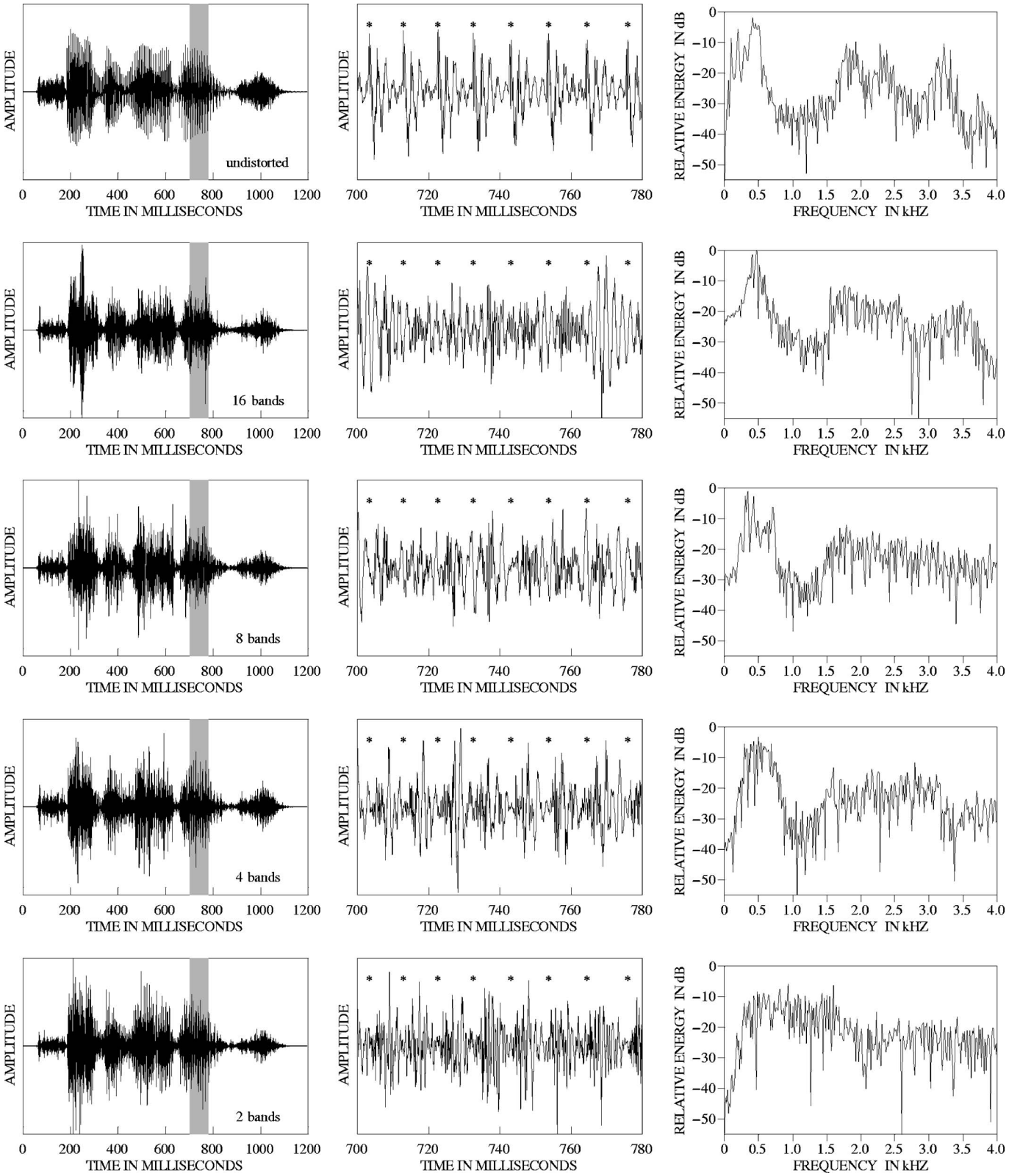


FIG. 1. Left panels, from top to bottom: The time wave forms of the undistorted followed by the vocoded versions (16, 8, 4, and 2 bands, respectively) of the sentence "Say the word ace." The shaded portion identifies an 80-ms segment of the vowel in the word ace. The middle panel, from top to bottom, shows the time wave forms of the undistorted (top) followed by the vocoded versions of the 80-ms segment of the vowel. The periodicity of the vowel in the undistorted case is indicated by the asterisks in the top middle panel. Note that the periodicity is virtually eliminated during vocoding. Right panels, from top to bottom: The energy spectra for the undistorted (top) followed by the vocoded versions of the 80-ms segment of the vowel.

oding (e.g., Loizou *et al.*, 1999; Shannon *et al.*, 1995). In a pivotal study by Shannon and colleagues (1995), near perfect levels of speech recognition were achieved with as few as four frequency bands for young adult listeners with good hearing. Not only did this study provide convincing evidence of the importance of envelope cues for spoken language

TABLE I. Mean (s.d.) of audiometric air-conducted pure-tone thresholds (dB HL) of the test ears for the younger and older participants in Experiments 1, 2, and 3.

	Frequency (kHz)							
	0.25	0.5	1	2	3	4	6	8
Younger participants (Experiments 1 and 2; $N=12$)								
Mean dB HL	6.25	4.58	1.67	0.833	0.833	1.67	7.98	5.42
s.d.	(4.33)	(8.11)	(3.89)	(3.60)	(5.57)	(6.51)	(7.82)	(6.90)
Younger participants (Experiment 3; $N=12$)								
Mean dB HL	6.25	4.17	0.00	3.33	−0.42	−0.42	3.75	0.42
s.d.	(6.78)	(5.57)	(5.22)	(4.92)	(5.82)	(6.20)	(7.42)	(7.21)
Older participants (Experiments 1 and 2; $N=12$)								
Mean dB HL	8.33	5.42	6.25	11.67	10.83	18.75	27.08	27.50
s.d.	(6.62)	(4.96)	(2.61)	(8.91)	(7.06)	(10.03)	(14.37)	(19.48)
Older participants (Experiment 3; $N=12$)								
Mean dB HL	8.33	5.42	6.25	11.67	10.83	22.50	31.25	45.42
s.d.	(6.62)	(4.96)	(2.61)	(8.91)	(7.06)	(9.68)	(14.16)	(19.63)

comprehension, but it also provided an important new method for studying the contribution of envelope cues to speech processing.

C. Perceptual learning

Noise-vocoded speech is not experienced outside of the lab; therefore, it is important to determine how perceptual learning might influence age-related differences in performance. The effect of perceptual learning on speech perception has been shown in behavioral (e.g., Davis *et al.*, 2005) and physiological studies (e.g., Tremblay *et al.*, 1997, 2001; Callan *et al.*, 2003). Furthermore, the effect of perceptual learning is seen for various forms of distorted speech, such as synthetic speech (Greenspan *et al.*, 1988), time-compressed speech (Dupoux and Green, 1997; Peelle and Wingfield, 2005), and noise-vocoded speech (Davis *et al.*, 2005). Specific to noise-vocoded speech, a study by Davis and colleagues (2005) showed that noise-vocoded sentences that were initially unintelligible to participants became markedly more intelligible as listeners gained experience with noise vocoding, implicating the role of perceptual learning in listeners' understanding of noise-vocoded speech.

Few studies have examined how younger and older listeners differ in terms of auditory learning. Among these studies, Peelle and Wingfield (2005) found that perceptual learning for noise-vocoded paragraphs was comparable for older and younger adults. However, in a subsequent experiment, they found an age-related deficit in transferring perceptual learning from one stimulus set to another when the speech was time compressed. Specifically, older adults were worse than younger adults in adapting to one rate of time-compressed speech after having adapted to another speech rate. Therefore, a secondary goal of our study was to investigate how age may affect perceptual learning as listeners acquire and consolidate learning gained from listening to noise-vocoded speech.

II. EXPERIMENT 1

The first goal of Experiment 1 was to determine the number of bands required for younger and older adults to correctly identify noise-vocoded words in a carrier phrase. The second goal was to determine whether or not participants in the two age groups showed improvement in word identification performance as they gained experience listening to noise-vocoded speech.

A. Method

1. Participants

Twelve younger adults (mean age=22.3 years, s.d.=2.2, range=19–25) and twelve older adults (mean age =70.2 years, s.d.=3.1, range=66–74) participated in the experiment. All participants in both age groups had pure-tone air-conduction thresholds in the test ear less than or equal to 25 dB HL from 0.25 to 3 kHz, i.e., they had audiometric thresholds that were clinically normal in the speech range (Table I). All participants had learned English before the age of 5 years and had been educated in English in a country where it is the dominant language. To measure verbal knowledge, each participant was given the Mill-Hill Vocabulary Scale (Raven, 1965). On average, the older group had better vocabulary scores than the younger group [mean for the younger group=12.5/20 and s.d.=2.0; mean for the older group=15.3/20 and s.d.=2.9; $t(22)=-2.68$, $p<0.05$]. On average, the younger group had 1.8 years more education than the older group [mean for the younger group =15.4 years of education and s.d.=1.6; mean for the older group=13.6, s.d.=2.9; $t(22)=1.84$; $p<0.05$]. All of the participants were paid volunteers recruited from the local community who gave informed consent in compliance with the protocol approved by the university's ethics review board. No participant had previously heard noise-vocoded speech.

2. Stimuli and apparatus

The test stimuli were the digital recordings of the four Northwestern University Auditory Test Number 6 (NU-6) word lists that have been standardized for use in clinical speech audiometry (see Penrod, 1985) and distributed on compact disk by Auditec of St. Louis. Each list consists of 50 monosyllabic words spoken by a male talker preceded by the carrier phrase "*Say the word*…." For each word, 16 different

TABLE II. Boundary frequencies for the 1 to 16 band-processed noise-vocoding.

1 band	300	6000															
2 band	300	1528	6000														
3 band	300	814	1528	6000													
4 band	300	722	1528	3066	6000												
5 band	300	546	994	1528	3296	6000											
6 band	300	494	814	1528	2210	3642	6000										
7 band	300	460	706	1083	1528	2549	3911	6000									
8 band	300	477	722	1061	1528	2174	3066	4298	6000								
9 band	300	418	584	814	1136	1528	2210	3083	4301	6000							
10 band	300	405	546	737	994	1528	1810	2443	3296	4447	6000						
11 band	300	394	517	679	892	1171	1528	2019	2650	3480	4570	6000					
12 band	300	385	494	634	814	1045	1528	1722	2210	2837	3642	4674	6000				
13 band	300	378	476	599	754	950	1196	1528	1896	2387	3005	3784	4765	6000			
14 band	300	372	460	570	706	875	1083	1342	1528	2058	2549	3158	3911	4844	6000		
15 band	300	366	447	546	667	814	994	1214	1528	1810	2210	2699	3296	4024	4914	6000	
16 band	300	382	477	590	722	878	1061	1276	1528	1825	2174	2584	3066	3632	4298	5080	6000

noise-vocoded band conditions were created, beginning with a one-band condition and increasing the number of bands, by one, up to a 16-band condition. That is, for every word, there were 17 files: one file for the intact condition and one for each of the 16 conditions differing in the number of bands used during noise vocoding.

To create noise-vocoded stimuli, we followed the procedure described in detail by Eisenberg *et al.* (2000). First, the speech stimuli were converted with the Goldwave digital audio editor to binary files with a sampling rate of 20 kHz. Using MATLAB software, stimuli were processed through a pre-emphasis filter (a high-pass first-order Butterworth infinite impulse response (IIR) filter with a cut-off frequency of 1.2 kHz and a roll-off of 6 dB/octave). The signal was split into a varying number of frequency bands ($n=1-16$) using fourth-order elliptical IIR bandpass filters with a maximum peak-to-peak ripple of 0.5 dB in the passband and a minimum attenuation of 40 dB in the stop band. The passband used to split the signal into frequency bands spanned a frequency range from 0.3 to 6 kHz for all conditions. The frequency spacing of the filter banks was based on the work of Greenwood (1990). The boundary frequencies for the band-processed conditions are shown in Table II. To extract the envelopes, the magnitude of the Hilbert transform was computed and passed through a low-pass filter (second-order Butterworth IIR with cut-off frequency of 0.1 kHz). One difference in procedures was that whereas Eisenberg *et al.* (2000) rectified and then low-pass filtered the filter bank outputs, we extracted the envelope using the magnitude of the Hilbert transform followed by a low-pass filter similar to that used by Eisenberg *et al.* (2000). Narrow-band noise was generated by passing a Gaussian white noise signal through the same Butterworth and elliptical filters. The envelopes extracted in the previous step were then used to modulate the corresponding band of noise. The bands of modulated noise were then summed together. Finally, the stimuli were converted to wav format with a sampling rate of 24.414 kHz using the Goldwave digital audio editor.

During the testing session, the participant was seated comfortably inside an Industrial Acoustics Company double-walled sound-attenuating booth. The stimulus files were played using a Tucker Davis Technologies System III and were presented monaurally to the participant's better ear over a Sennheiser HD 265 headphone. All stimuli were presented at 70 dB SPL for both age groups.

3. Procedure

The procedure for Experiment 1 was adapted from the gating paradigm (Grosjean, 1996). In the original gating procedure, the word is broken into a sequence of time gates. During the first trial, only one gate is presented and the listener attempts to identify the word. On each subsequent trial the number of gates is increased until the listener reliably identifies the word. Thus, the gating procedure involves multiple presentations of portions of the same stimulus with gradual increments in the amount of signal delivered in order to determine how much of the signal must be heard for the word to be correctly identified. The gating procedure was adapted for the present study such that, rather than incrementing the number of time gates presented to the listener, we incremented the number of bands that were presented. Specifically, a word was first presented in the one-band noise-vocoded condition. The participant was asked to identify the word or to respond "I don't know." If the listener did not correctly identify the word, it was presented again in the 2-band condition. The number of bands continued to be incremented by one until the word was correctly identified. If the word was not correctly identified by the 16-band condition, then the word was presented in the intact condition. Feedback was given via a computer monitor. Guessing was strongly encouraged.

Every participant heard every word in each of the four lists; however, individual participants heard a varying number of band conditions for each word depending on the band condition in which he or she correctly identified the word. The words were presented in random order within each list and the list order was counterbalanced across participants so that each list was presented an equal number of times in each order position for each age group.

Each participant completed the four word lists in a single session, with each word list taking approximately

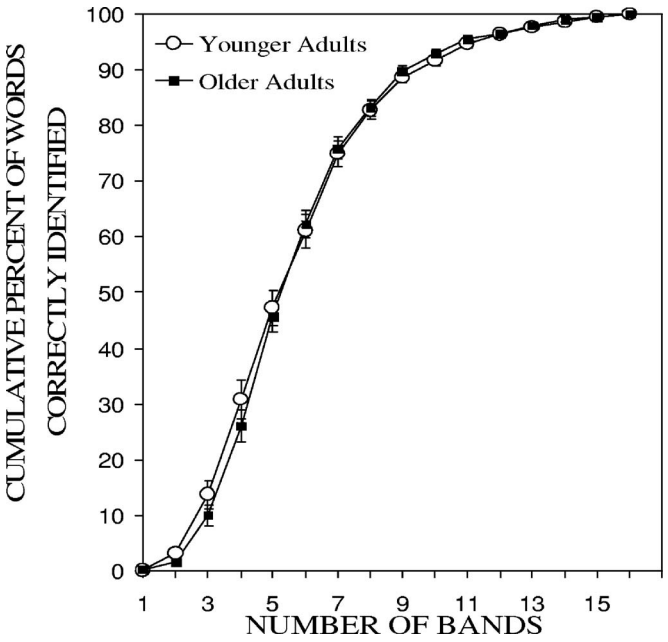


FIG. 2. The cumulative percentage of words correctly identified, averaged across participants, as a function of the number of bands for younger (open circles) and older (closed squares) participants in Experiment 1. Standard error bars are shown.

25 min to complete. Participants were given a 10-min break between each list. The experimenter, who was outside the booth, listened to and immediately scored each response. If the experimenter was uncertain about any response, the participant was asked to repeat and spell the word aloud. Any difference between response and target was marked as an error. All sessions were audiotaped to enable subsequent verification of the scoring.

B. Results and discussion

Words that were only correctly identified in the intact condition or never correctly identified were excluded from the analyses. In total, only 178 out of 4800 responses ($<4\%$) were excluded, of which 70 responses were from younger participants and 108 were from older participants. The cumulative percentage of words identified correctly in each band condition was calculated for each participant for each list. The mean of these values for each group is plotted in Fig. 2, illustrating that word identification performance was virtually identical for both age groups. Indeed, no statistically significant differences between age groups were found in the cumulative percentage of words correctly identified at any band value (for all independent samples t-tests, $p>0.4$).

For each list, we calculated each individual's threshold (the number of bands at which the cumulative percentage of correctly identified words was 50%). For both age groups, word identification improved similarly across the four lists, with mean band thresholds of 5.8, 5.3, 5.0, and 4.9 (overall mean threshold=5.25 bands) for the younger group, and 5.5, 5.3, 5.1, and 5.0 (overall mean threshold=5.25 bands) for the older group. This description was confirmed by an analysis of variance (ANOVA) with age as a between-subjects factor and list order as a within-subjects factor that revealed no significant effect of age, $F(1,22)=0.000$, $p>0.9$, but a significant effect of list order, $F(3,66)=14.53$, $p<0.001$, with no significant interaction between age and list order, $F(3,66)=1.65$, $p>0.1$. A Tukey test of multiple comparisons confirmed that performance did not differ significantly between Lists 1 and 2 ($p>0.1$), but that there was significant improvement between List 1 and List 3 ($p<0.01$) and between List 1 and List 4 ($p<0.002$), although performance on Lists 3 and 4 was not significantly different ($p>0.5$). Thus, there is an overall improvement in performance with increasing exposure to noise-vocoded stimuli across lists, but there does not seem to be an age-related difference in word identification or an age-related difference in the degree of improvement across lists.

III. EXPERIMENT 2

The goal of Experiment 2 was to investigate if the older and younger participants who had gained experience with noise-vocoded speech in Experiment 1 would differ in their ability to identify noise-vocoded words using a more common method of presentation in which the number of bands used in vocoding was blocked rather than gated as it had been in Experiment 1.

A. Method

1. Participants

The participants were those who completed Experiment 1.

2. Stimuli and apparatus

The test stimuli were the digital recordings of the four W-22 word lists that have been standardized for use in clinical speech audiometry (see Penrod, 1985) and distributed on compact disk by Auditec of St. Louis. Like the NU-6 lists, each W-22 list consists of 50 monosyllabic words spoken by a male talker preceded with the carrier phrase "*Say the word*…." The NU-6 lists were designed after the W-22 lists to be a more sensitive test for individuals with predominantly high-frequency hearing loss. The main differences between the NU-6 and W-22 lists are that the word frequency profile is lower and the occurrence of high-frequency consonants is higher in the NU-6 lists than in the W-22 lists. The digitized recordings were subjected to the same noise-vocoding procedure used by Eisenberg *et al.* (2000) and described earlier. For each list, the words were noise-vocoded using 16, 8, 4, and 2 bands. Stimuli were delivered using the same apparatus as was used in Experiment 1.

3. Procedure

Immediately following Experiment 1, in which the participants were familiarized with noise-vocoded speech, the participants completed Experiment 2. In contrast to Experiment 1, the presentation of stimulus conditions was blocked in Experiment 2, with the order of conditions progressing from easiest to hardest. First, participants heard a word list in the 16-band condition. Next, they heard a word list in the 8-band condition, then a word list in the 4-band condition,

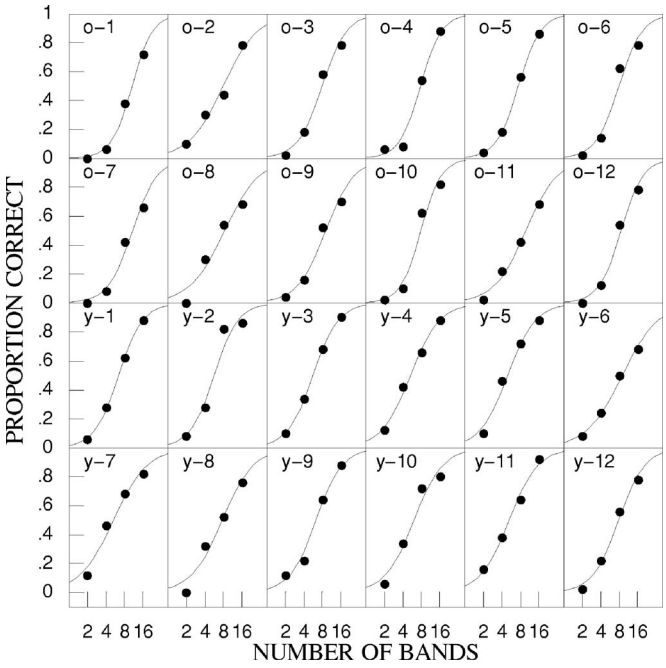


FIG. 3. The proportion of words correctly identified as a function of the number of bands for individual younger (y) and older (o) participants in Experiment 2.

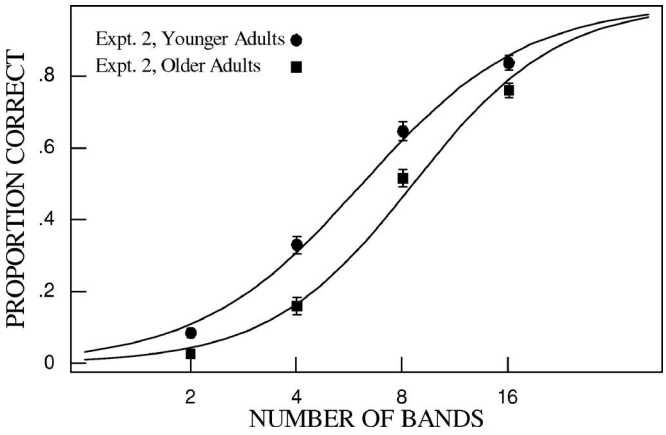


FIG. 4. The proportion of words correctly identified, averaged across participants, as a function of the number of bands for younger (circles) and older (squares) participants in Experiment 2. Standard error bars are shown.

and finally a word list in the 2-band condition. Words within a list were presented in random order and the list order was counterbalanced across participants so that each list was presented an equal number of times in each order position for each group.

For each word, the participant was asked to identify the word or respond "I don't know." Guessing was strongly encouraged. The experimenter, who was outside the booth, listened to and immediately scored each response. If the experimenter was uncertain about any response, the participant was asked to repeat and spell the word aloud. Any difference between response and target was marked as an error. No feedback was given after a response. All sessions were audiotaped. The testing session lasted approximately 25 min.

B. Results and discussion

Figure 3 plots the percentage of words that were correctly identified as a function of the number of bands for each participant. Logistic functions of the form

$$y = \frac{1}{1 + e^{-\sigma[(\log_{10} x) - \mu]}}$$

were fit to the data of each participant, where y is the proportion of words correctly identified, x is the number of bands, μ is the threshold (the value of $\log x$ resulting in 50% correct identification of the words), and σ is the slope parameter of the psychometric function (see the Appendix). It is apparent in Fig. 3 that a logistic function provides a good fit to the data of each individual.

Figure 4 plots the average proportion of words correctly identified by the younger and older groups by band condition. Figure 4 suggests that the psychometric functions for younger and older adults have similar slopes (σ), but different threshold values (μ). That is, the psychometric function for younger adults is the same as that for older adults, except that it is shifted to the left by 0.14 log units. To verify this, we conducted two ANOVAs with pure-tone thresholds (at all eight audiometric test frequencies) as covariates: one for the values of slopes and another for the band thresholds obtained from the individual data of younger and older adults (Fig. 3). These tests indicate that younger and older adults differed with respect to band thresholds, with a significant main effect of age, $F(1,14)=5.104$, $p<0.05$, but not with respect to slopes, $F(1,14)=0.814$, $p>0.1$. As shown in Fig. 4, older adults needed an average of 8.55 bands to correctly identify 50% of the noise-vocoded words, whereas younger adults needed an average of only 6.13 bands to achieve the same level of performance.

It is interesting that the band thresholds calculated in Experiment 2 based on 50% correct word identification were larger for both groups (older=8.55; younger=6.13), compared to the band thresholds calculated in Experiment 1 based on 50% cumulative correct word identification (5.25 for both age groups). Experiment 1 should have been more challenging than Experiment 2 insofar as the participants were less experienced with noise-vocoded speech in Experiment 1 than in Experiment 2, and the NU-6 word lists used in Experiment 1 could have been more difficult than the W-22 word lists used in Experiment 2, especially for the older listeners with high-frequency hearing loss because of the emphasis in the NU-6 lists on high-frequency consonants. The differences in the size of the band thresholds in the two experiments seem more likely to be explained by the way in which the functions were measured than by the stimuli that were used.

More important, the finding of a significant effect of age on the number of bands needed to correctly identify 50% of the noise-vocoded words differs from the finding of no effect of age on performance in Experiment 1. It is possible that the discrepancy between the finding of no age-related difference in Experiment 1 but a significant age-related difference in Experiment 2 could be explained by methodological differences between the two experiments, including differences associated with varying the number of bands using gating

versus blocked presentation methods, or differences related to whether or not feedback was provided. Another possibility is that there were age-related differences in the carry-over of learning from Experiment 1 to Experiment 2. Age-related differences in carry-over of learning would be consistent with the findings of previous studies showing that perceptual learning gained from a training session was more beneficial to younger than to older adults in a subsequent testing session (e.g., Peelle and Wingfield, 2005; Sommers, 1997). According to this explanation, the reason that an age effect was found in Experiment 2, but not in Experiment 1, would be that the younger adults were better able than the older adults to generalize or to use the experience they had gained during Experiment 1 to aid their performance in Experiment 2. In other words, there might be no age-related difference in temporal processing of envelope cues per se, but only a difference in the ability of younger and older adults to generalize from the training set to a testing set.

IV. EXPERIMENT 3

In Experiment 3, we tested new groups of younger and older listeners using the same procedure as in Experiment 2; however, the participants in Experiment 3 had never heard noise-vocoded stimuli prior to the experiment. To evaluate the possibility that the age-related differences found in Experiment 2 were due to age-related differences in the carry-over of learning from Experiment 1, the performance of the younger and older listeners in Experiment 2 (who had experience with noise-vocoded stimuli in Experiment 1) was compared to that of the younger and older listeners in Experiment 3 (who had not been previously exposed to noise vocoding).

A. Method

1. Participants

Twelve younger adults (mean age=21.0, s.d.=2.9, range=17–25 years) and twelve older adults (mean age =67.4, s.d.=2.8, range=64–72 years) participated in this experiment. Both groups had pure-tone air-conduction thresholds in the test ear that were less than or equal to 25 dB HL from 0.25 to 3 kHz (Table I). All participants had learned English before the age of 5 years and had been educated in English in a country where it is the dominant language. Mean scores on the Mill-Hill Vocabulary Scale (Raven, 1965) were better for older than for younger participants [mean for the younger group=12.7, s.d.=2.3; mean for the older group=15.0, s.d.=1.7; $t(22)=-2.85$, $p<0.05$]. On average, the younger adults had 1.2 years more education than the older adults [mean for younger participants=15.3 years of education, s.d.=3.1; mean for older participants=14.1, s.d.=2.2 years); $t(22)=1.06$, $p>0.10$]. All of the participants were paid volunteers recruited from the local community who gave informed consent in compliance with the protocol approved by the university's ethics review board. No participant had previously listened to noise-vocoded speech.

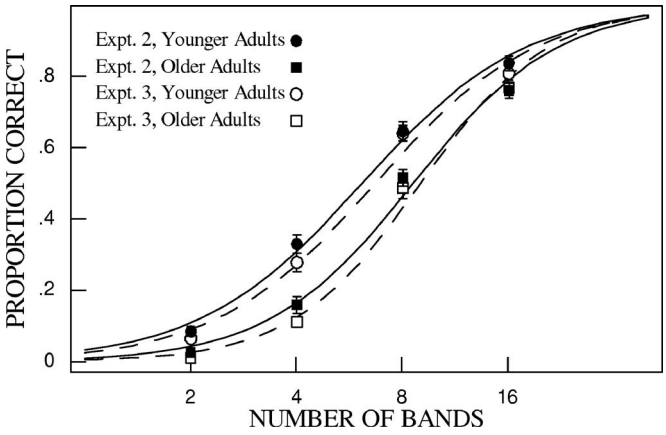


FIG. 5. The proportion of words correctly identified, averaged across participants, as a function of the number of bands for younger (circles) and older (squares). Closed symbols and solid lines are for Experiment 2. Open symbols and dashed lines are for Experiment 3. Standard error bars are shown.

2. Stimuli and procedure

The same materials and procedure used in Experiment 2 were used in Experiment 3.

B. Results and discussion

Figure 5 plots the average proportion of words correctly identified by the younger and older participants in Experiment 3 compared to the data obtained in Experiment 2. As in Experiment 2, the psychometric functions for younger and older adults have similar slopes (σ) and threshold values (μ). To verify this pattern for Experiment 3, we conducted two ANOVAs with pure-tone thresholds (at all eight audiometric test frequencies) as covariates: one for the values of slopes and another for the band thresholds obtained from the individual data of younger and older adults. These tests indicate that younger and older adults differed with respect to band thresholds, with a significant main effect of age, $F(1,14)=9.074$, $p<0.01$, but not with respect to slopes, $F(1,14)=1.295$, $p>0.1$. Older adults needed an average of 8.97 bands to correctly identify 50% of the noise-vocoded words, whereas younger adults needed an average of only 6.67 bands to achieve the same level of performance.

Furthermore, the similarity of the functions in Fig. 5 suggests that previous experience with noise-vocoded words had little effect on the threshold values (μ) for either age group. We conducted an ANOVA on the threshold and slope values with experiment (Experiment 2 versus Experiment 3) and age (younger versus older) as between-subjects factors. For the band threshold values, there was a significant difference between younger and older adults, $F(1,44)=47.47$, $p<0.0001$, but no significant difference between experiments, $F(1,44)=2.07$, $p>0.05$, and no significant interaction between age and experiment, $F(1,44)=0.15$, $p>0.05$. For the slope values, there was a significant difference between age groups, $F(1,44)=20.53$, $p<0.001$, and a borderline significant difference between experiments, $F(1,44)=4.32$, $p<0.05$, but no significant interaction between age and experiment, $F(1,44)=2.07$, $p>0.05$. Crucially, prior exposure and learning did not significantly alter the band threshold

values for either the younger or the older adults and the effect of age on word identification performance was observed in both Experiments 2 and 3. Therefore, we cannot attribute the results found in Experiment 2 to an age-related decline in the ability to carry-over learning gained in Experiment 1.

V. GENERAL DISCUSSION

The primary goal of the present study was to examine the effect of age on the ability of listeners to use envelope cues to identify words spoken in a carrier phrase when fine structure cues are minimized in noise-vocoded speech. A secondary goal was to investigate how age may affect perceptual learning as listeners acquire and consolidate learning gained from experience listening to noise-vocoded speech. In Experiment 1, we found no significant difference between the word identification accuracy of younger and older listeners when noise-vocoded stimuli were presented using an adapted gating procedure in which each target word was presented with an increasing number of bands until it was correctly identified. Not only was the mean band threshold for 50% cumulative correct word identification identical for both age groups, but the entire functions of the cumulative percentage of correctly identified words by band condition were also nearly identical for the two age groups. Both age groups also demonstrated similar perceptual learning in Experiment 1. In contrast, when the same listeners were subsequently tested in Experiment 2, there was a significant age-related difference in band threshold when each target word was presented in a blocked design. Similar to the results of the participants in Experiment 2 who had been exposed to noise-vocoded speech in Experiment 1, for the participants in Experiment 3 who had no prior exposure to noise-vocoded speech, older adults performed significantly worse than younger adults. Thus, age-related differences in the carry-over of learning from Experiment 1 do not seem to account for the discrepancy between the finding of an age effect in Experiment 2 but not in Experiment 1.

In this discussion, we will compare our results to those reported in previous research. Then we will consider if the discrepancy between Experiments 1 and 2 could be explained by methodological differences between the two experiments, namely differences associated with varying the number of bands when using gating versus blocked presentation methods or differences related to whether or not feedback was provided. Finally, we will discuss the absence of the carry-over of learning from Experiment 1 to Experiment 2.

A. Identification of noise-vocoded speech

To our knowledge, no other studies have adapted the gating paradigm to noise vocoding as we did in Experiment 1. Previous noise-vocoding studies typically used a blocked design similar to the one we used in Experiment 2 (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1998; Loziou *et al.*, 1999; Eisenberg *et al.*, 2000; Fu and Nogaki, 2005; Souza and Boike, 2006), or they used stimuli in only one selected band-processed condition (e.g., Davis *et al.*, 2005; Trout, 2005). In previous noise-vocoding studies, the accuracy with which the vocoded words were identified by younger adults has often been higher than was found in our Experiments 2 and 3. This difference may best be explained by differences in training and the type of speech material. We provided either minimal training (Experiment 2) or no training (Experiment 3) with noise-vocoded stimuli, whereas others have provided extensive training. For example, to achieve near-perfect levels of word identification with only four bands, Shannon and colleagues (1995) provided 8–10 h of training. Furthermore, the open-set test of monosyllablic words we used was more demanding than the closed-set recognition tests used in many previous studies (e.g., van Tasell *et al.*, 1992; Souza and Boike, 2006). In addition, other studies likely achieved higher levels of performance in low band-processed conditions because they used sentences in which lexical knowledge could be used to help decipher the degraded signal (e.g., Fishman *et al.*, 1997), whereas the carrier phrase we used did not provide listeners with any opportunity to use context to advantage. Results more similar to ours have been reported in studies in which the stimuli and response alternatives were also more similar to those of our study. In studies that tested monosyllablic word identification using an open set, results have ranged from about 55% correct in a 4-band processed condition (Friesen *et al.*, 2001) down to only 9.9% accuracy for 5-band noise-vocoded words that varied in lexical difficulty and were presented with no carrier phrase (Trout, 2005). Importantly, although identification of noise-vocoded words may have been harder in the present experiments than in less challenging tests used in prior studies, we were able to detect both similarities and differences between younger and older listeners.

One prior study has examined the relative effects of age and degree of hearing loss on the ability of adults to process noise-vocoded speech (Souza and Boike, 2006); however, their sample did not include older adults with good audiograms. Nevertheless, our finding of an age effect in Experiment 2 is in line with the results of Souza and Boike (2006) who found that age, but not degree of hearing loss, was a significant predictor of the ability of listeners to identify noise-vocoded /aCa/ nonsense bisyllables in a 16-alternative closed-choice task. They interpreted their findings for adults ranging in age from 23 to 80 years and degree of hearing loss from mild to severe as evidence for an age-related deficit in the use of temporal envelope information across all band conditions (1-, 2-, 4-, and 8-band conditions). Both age groups in our present study had clinically normal audiograms in the speech range, although the mean thresholds of the older adults were higher than those of the younger adults, especially at the highest frequencies (4–8 kHz). Nevertheless, consistent with the conclusion of Souza and Boike (2006) that there is an age-related deficit that is not explained by degree of hearing loss, the smaller audiometric threshold differences between the younger and older adults in our study could not explain why we found significant group differences in word identification performance in Experiments 2 and 3. It also seems unlikely that audiometric threshold differences would have affected performance in Experiments 2 and 3, but not in Experiment 1. If high-frequency audiomet-

ric loss at frequencies of 4 kHz and higher contributed to the problems of the older adults then age-related differences should have been more pronounced in Experiment 1 than in Experiments 2 and 3 because the NU-6 word lists used in Experiment 1 were designed to be more challenging for people with high-frequency hearing loss. In fact, using the proportion of words correctly identified in Experiment 1 as a baseline measure of word identification, both age groups achieved near-ceiling performance and the mean proportion correct for younger adults (0.97, s.d.=0.023) and older adults (0.96, s.d.=0.042) did not differ significantly, $t(22)=1.25$, $p>0.10$. It is also worth noting that the closeness of the cumulative percent correct functions for the two age groups shown in Fig. 2 seems inconsistent with the possibility that age-related differences were found only in Experiment 2 because older adults were less willing than younger adults to guess the first time that a word was presented. Nonaudiometric age-related differences may better explain the pattern of results.

B. Age-related differences in temporal processing

The main finding from Experiments 2 and 3 is that there is an age-related reduction in ability to use envelope cues to identify noise-vocoded words, as illustrated by the similar slopes (σ), but significantly different band thresholds (μ) of the younger and older adults. Although much prior research has been interpreted as suggesting that the processing of prosody is largely preserved with age, a closer examination of the pattern of results of key studies demonstrates that although older adults are able to benefit from prosodic cuing, including envelope duration and shape cues, it is not the case that older adults achieve the same level of performance as is achieved by younger adults. For example, in one study of speech prosody, the standard time-gating technique was used to investigate age-related differences in the use of envelope cues for word identification in three experimental conditions: in one condition, only the onset of the word was presented; in another condition, the onset of the word was presented and noise was used to terminate the word, adding information about the duration of the word; in the remaining condition, the onset of the word was provided plus a noise shaped by the speech envelope that provided duration and additional prosodic information (Wingfield *et al.*, 2000). Although older and younger listeners benefited similarly from the addition of duration and envelope cues, the younger adults outperformed the older adults even in the onset plus envelope-shaped noise condition. Thus, the age-related difference in overall performance that was observed seems to be consistent with our finding of an age-related difference in word identification based on the use of envelope cues in noise-vocoded speech.

Previous research has found evidence of age-related temporal processing deficits related to subsegmental and segmental speech cues. Our results suggest that there are also age-related differences in the use of envelope cues that could be relevant for processing suprasegmental speech information. Whether or how the age-related differences in auditory temporal processing relevant to the different levels of speech information are related has yet to be determined. Importantly, the present study establishes the possibility that age-related differences in auditory temporal processing exist at the suprasegmental level even when audiometric thresholds in the speech range are within clinically normal limits.

C. Benefit from repetition and feedback

Contrary to the findings from Experiments 2 and 3, in Experiment 1 we found no age-related differences in ability to use envelope cues when identifying noise-vocoded words. Nonetheless, the absence of an effect of age on word identification in Experiment 1 seems to be consistent with the idea that the processing of prosodic cues is relatively preserved in older adults. Clearly, it is important to consider why the results for Experiments 1 and 2 are discrepant. A number of methodological differences, including differences in the stimuli, may provide an explanation for the discrepancies. We will consider repetition and feedback as two such possibilities.

When words are presented using the band-gating procedure, as they were in Experiment 1, there is an opportunity for the listener to benefit from the summing of information that is incremented as the number of bands is increased over sequential presentations. No benefit from this type of summed information is available when a stimulus is presented only a single time, as was done in Experiment 2. The use of summed information has been cited in psychophysical studies to explain improvements in performance on tasks such as detecting interaural differences in intensity for trains of clicks (Hafter and Dye, 1983; Hafter *et al.*, 1983). Furthermore, it seems possible that summation of information over repeated utterances could be at play in everyday conversational behavior. Repetition, especially repetition with clearer pronunciation, is the most common conversational repair strategy (Drew, 1997). For a listener in a conversation, the information about the identity of a word provided by its first presentation may be used in conjunction with the information gained from later repetition to achieve correct word identification. That is, it could be possible to use the first degraded exposure to the utterance to narrow the set of possible lexical alternatives and to focus listening on the critical but missed portions of the utterance when it is repeated. Future experiments could directly explore the possible contribution of stimulus repetition.

If listeners can sum information to aid word identification, since everyday listening environments are more challenging for older adults than for younger adults, older adults are likely to have considerable experience in using this type of compensatory mechanism. Age-related differences may be found in Experiment 2 because there is no opportunity to compensate by summing information. The notion that older adults may sum information to support word identification in adverse listening conditions is consistent with other research indicating that older adults are better able to use other compensatory mechanisms, such as phonological knowledge (e.g., Pichora-Fuller *et al.*, 2006) or sentence context to identify speech in adverse listening situations (e.g., Pichora-Fuller *et al.*, 1995; Wingfield *et al.*, 2005). More generally, these findings are consistent with the idea that age-related

differences on a range of cognitive measures are reduced when aspects of the environment can be used to support performance on a demanding task (e.g., Craik, 1983, 1986). Convergent evidence of age-related compensation during perceptual and cognitive tasks has also been found in cognitive neuroscience research showing that there is more bilateral activation of the brain for older adults when they achieve the same performance as younger adults on a variety of tasks (e.g., Grady, 2000; Cabeza, 2002; Reuter-Lorenz, 2002).

An alternative explanation for the differences related to the presentation method used in the two experiments is that there was feedback provided in Experiment 1, but not in Experiment 2. In particular, the feedback provided in Experiment 1 may account for the improvement from List 1 to List 4 that was observed for both age groups, in agreement with previous studies demonstrating younger adults' abilities to learn to identify noise-vocoded speech (Davis *et al.*, 2005). Without the benefit from feedback in Experiment 2, our older participants may have been more disadvantaged than their younger counterparts. The idea that older adults may be facilitated differentially by feedback has been used to explain findings in a study of the effect of age on auditory lexical decision (Stine-Morrow *et al.*, 1999).

D. Carry-over of learning

Neither younger adults nor older adults seemed to be able to carry-over learning from Experiment 1 to improve their performance on Experiment 2, as seen by the lack of a significant difference between the mean band thresholds of the experienced (Experiment 2) and inexperienced groups (Experiment 3). An obvious explanation for this null effect is that there was not enough training. Perceptual training studies typically involve training sessions that take many hours over the course of many days whereas the training session in the current study lasted, at most, 2 h (Kraus *et al.*, 1995; Tremblay *et al.*, 2001). Nevertheless, in the current study, there was perceptual learning or familiarization within the training set for both age groups, as evidenced by the significant improvement over list presentations in Experiment 1. Furthermore, both age groups seemed to reach their plateau performance insofar as there was no significant difference for either group between their performance on Lists 3 and 4. The improvement of the participants in Experiment 1 is consistent with studies citing short-term perceptual adaptation as the explanation for improved performance in identification of distorted speech stimuli (Clarke, 2002; Mehler *et al.*, 1993; Davis *et al.*, 2005; Peelle and Wingfield, 2005).

Another possible explanation for the lack of an effect of learning is that learning does not carry-over across the different presentation methods used in Experiments 1 and 2. In the present study, the gating and feedback used in Experiment 1 differed in presentation from the single, blocked presentation of stimuli in Experiment 2. Although both experiments used simple, monosyllabic words, the words were not identical. In addition, the extent of experience with the bands tested in Experiment 2 was not uniform in Experiment 1 because of the differences in the gating and blocked organization of the presentation of the words. One or more of these differences between the experiments could have prevented carry-over. The idea that perceptual learning is task-specific and that training effects may not generalize has been suggested in other auditory learning research. For example, Irvine *et al.* (2000) found that the effect of training did not transfer to other frequencies for a frequency discrimination task. Similarly, Burk *et al.* (2006) found that training on isolated words presented in noise was not sufficient to yield large improvements on untrained words presented in noise. In the visual domain, Fahle and Morgan (1996) found no transfer of perceptual learning or training between similar stimuli tested in two different tasks. Our finding of no difference between the word identification thresholds of experienced listeners in Experiment 2 and inexperienced listeners in Experiment 3 may be in keeping with the more general finding that perceptual learning can be highly task-specific and does necessarily generalize to other tasks given only prior exposure to similar stimuli in a different context/task (Cohen *et al.*, 2006).

In summary, we observed perceptual learning in Experiment 1, but no significant carry-over to Experiment 2 for either younger or older adults. Many differences between the experiments may have prevented carry-over of learning. In any case, the lack of carry-over of learning does not seem to explain why there were age-related differences in Experiment 2 since the same pattern of results was found for an inexperienced group who completed the same tests in Experiment 3.

VI. CONCLUSIONS

Without the benefit of summing information from repetitions of a word and/or benefit from feedback, older adults do not use envelope cues as well as younger adults to identify noise-vocoded words in an open-set task. This new evidence that older adults show a deficit in using the temporal amplitude envelope cues relevant to suprasegmental aspects of speech perception extends previous evidence that there are age-related declines in other aspects of auditory temporal processing relevant to other levels of speech processing. Importantly, we also found that when noise-vocoded speech is presented with the opportunity to sum information in the speech signal over repetitions and/or feedback is provided after each presentation, as in Experiment 1, then age-related differences are eliminated. Lastly, we found that perceptual learning with noise-vocoded stimuli occurred similarly for both age groups, but did not generalize across experiments for either age group, suggesting that age-related differences in carry-over of learning did not explain the age-related differences that we observed in ability to use envelope cues to identify noise-vocoded words.

ACKNOWLEDGMENTS

The authors are grateful to Ewen MacDonald for preparing the noise-vocoding program. This research was funded by the Canadian Institutes of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

We can test whether or not the logistic function provides a good fit to the data using the normalized Pearson's χ^2 statistic where

$$\text{norm } \chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \frac{1}{1+e^{-\sigma[(\log_{10} x_i)-\mu]}}\right)^2}{\frac{1}{1+e^{-\sigma[(\log_{10} x_i)-\mu]}}} + \sum_{i=1}^{n} \frac{\left(\frac{1}{1+e^{-\sigma[(\log_{10} x_i)-\mu]}} - y_i\right)^2}{1-\frac{1}{1+e^{-\sigma[(\log_{10} x_i)-\mu]}}}.$$

x_i is the number of bands, y_i is the proportion of correctly identified words at x_i, μ, and σ are the threshold and slope parameters, respectively, of the logistic function, and n_i is the number of stimuli in the experiment. Specifically, the values of μ and σ are systematically varied to minimize the normalized χ^2 statistic. When the normalized χ^2 statistic is multiplied by the number of times, N, each stimulus is presented in the experiment, it becomes a Pearson's χ^2 statistic with $n-2$ degrees of freedom. In Experiment 2, $N=50$, and $n=4$. Hence the degrees of freedom are $4-2=2$.

Because the test statistic is distributed according to χ^2 with 2 degrees of freedom, we can test the null hypothesis to determine whether or not the logistic function provides a good fit to the data from each individual. All tests were conducted using a Bonferroni correction for the number of tests conducted (in this case 24 tests were conducted to see if the logistic function could describe individual data). Hence, to correct for the number of tests we used $\alpha=0.05/24=0.0021$ for each test. Using the Bonferroni correction, we failed to reject the null hypothesis for all of the participants, and therefore concluded that the logistic function provided a good fit to the individual data.

Abel, S. M., Krever, E. M., and Alberti, P. W. (**1990**). "Auditory detection, discrimination and speech processing in aging, noise-sensitive and hearing impaired listeners," Scand. Audiol. **19**, 43–54.

Alain, C., McDonald, K. L., Ostroff, J. M., and Schneider, B. (**2001**). "Age-related changes in detecting a mistuned harmonic," J. Acoust. Soc. Am. **109**, 2211–2216.

Bergerson, T. R., Schneider, B. A., and Hamstra, S. J. (**2001**). "Duration discrimination in younger and older adults," Can. Acoust. **29**, 3–9.

Burk, M. H., Humes, L. E., Amos, N. E., and Strauser, L. E. (**2006**). "Effect of training in word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners," Ear Hear. **27**, 263–278.

Cabeza, R. (**2002**). "Hemispheric asymmetry reduction in older adults: The HAROLD model," Psychol. Aging **17**, 85–100.

Callan, D. E., Taijma, K., Callan, A. M., Kubo, R., Masaki, S., and Akahane-Yamada, R. (**2003**). "Learning-induced neural plasticity associated with improved identification performance after training difficult second-language phonetic contrasts," Neuroimage **19**, 113–124.

CHABA (Committee on Hearing, Bioacoustics, and Biomechanics) Working Group on Speech Understanding and Aging, National Research Council (**1988**). "Speech Understanding and Aging," J. Acoust. Soc. Am. **83**, 850–805.

Clarke, C. M. (**2002**). "Perceptual adjustment to foreign-accented English with short-term exposure," Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, Co., Vol. **1**, pp. 253–256.

Cohen, Y. E., Russ, B. E., and Lee, Y. S. (**2006**). "Neural and behavioural correlates of auditory categorization," Programme and Abstracts for the International Conference on the Auditory Cortex, MRC Institute of Hearing Research, Nottingham, UK, pp. 53–54.

Craik, F. I. M. (**1983**). "On the transfer of information from temporary to permanent memory," Philos. Trans. R. Soc. London, Ser. B **302**, 341–359.

Craik, F. I. M. (**1986**). "A functional account of age differences in memory," in *Human Memory and Cognitive Capabilities, Mechanisms, and Performances*, edited by F. Klix and H. Hagendorf (Elsevier Science Publishers, North-Holland, Amsterdam), pp. 499–522.

Culter, A., Dahan, S., and van Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang. Speech **40**, 141–201.

de Boer, E., and Dreschler, W. A. (**1987**). "Auditory psychophysics: Spectrotemporal representation of signals," **38**, 181–202.

de Cheveigné, A. (**2003**). "Time domain auditory processing of speech," J. Phonetics **31**, 547–461.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. **134**, 222–241.

Divenyi, P., and Simon, H. (**1999**). "Hearing in aging: Issues old and young," Current Opinion in Otolaryngology and Head & Neck Surgery **7**, 282–289.

Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.

Drew, P. (**1997**). " 'Open' class repair initiators in response to sequential sources of troubles in conversation," J. Pragmatics **28**, 69–101.

Dupoux, E., and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psychol. **33**, 914–927.

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**, 2704–2710.

Fahle, M., and Morgan, M. (**1996**). "No transfer of perceptual learning between similar stimuli in the same retinal position," Curr. Biol. **6**, 292–297.

Fishman, K., Shannon, R. V., and Slattery, W. H. (**1997**). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," J. Speech Hear. Res. **40**, 1201–1215.

Fitzgibbons, P., and Gordon-Salant, S. (**1994**). "Age effects on measures of auditory duration discrimination," J. Speech Hear. Res. **37**, 662–670.

Fitzgibbons, P. J., and Gordon-Salant, S. (**1995**). "Age effects on duration discrimination with simple and complex stimuli," J. Acoust. Soc. Am. **98**, 3140–3145.

Fitzgibbons, P. J., and Gordon-Salant, S. (**1996**). "Auditory temporal processing in elderly listeners," J. Am. Acad. Audiol **7**, 183–189.

Friesen, L., Shannon, R., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Frisina, R. D. (**2001**). "Possible neurochemical and neuroanatomical bases of age-related hearing loss-presbycusis," Semin. Hear. **22**, 213–225.

Fu, Q. J., and Nogaki, G. (**2005**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," **6**, 19–27.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1993**). "Temporal factors and speech recognition performance in young and elderly listeners," J. Speech Hear. Res. **36**, 1276–1285.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1997**). "Selected cognitive factors and speech recognition performance among young and elderly listeners," J. Speech Hear. Res. **40**, 423–431.

Gordon-Salant, S., Yeni-Komishian, G. H., Fitzgibbons, P. J., and Barrett, J. (**2006**). "Age-related differences in identification and discrimination of temporal cues in speech segments," J. Acoust. Soc. Am. **119**, 2455–2466.

Grady, C. L. (**2000**). "Functional brain imaging and age-related changes in cognition," Biol. Psychol. **54**, 259–281.

Greenberg, S. (**1996**). "Auditory processing of speech," in *Principles of Experimental Phonetics*, edited by N. J. Lass (Mobsy, St. Louis), pp. 362–407.

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. **17**, 152–162.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Grosjean, F. (**1996**). "Gating," Lang. Cognit. Processes **11**, 597–604.
Hafter, E. R., and Dye, R. H., Jr. (**1983**). "Detection of interaural differences of time in trains of high-frequency clicks as a function of interclick interval and number," J. Acoust. Soc. Am. **73**, 644–651.
Hafter, E. R., Dye, R. H., Jr., and Wenzel, E. (**1983**). "Detection of interaural differences of intensity in trains of high-frequency clicks as a function of interclick interval and number," J. Acoust. Soc. Am. **73**, 1708–1713.
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.
Konkle, D. F., Beasley, D. S., and Bess, F. H. (**1977**). "Intelligibility of time-altered speech in relation to chronological aging," J. Speech Hear. Res. **20**, 108–115.
Kraus, N., McGee, T., Carrell, T., King, C., and Tremblay, K. (**1995**). "Central auditory system plasticity associated with speech discrimination training," J. Cogn Neurosci. **7**, 27–34.
Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.
Mehler, J., Sebastian, N., Altmann, G., Dupoux, E., Christophe, A., and Pallier, C. (**1993**). "Understanding compressed sentences—The role of rhythm and meaning," Ann. N.Y. Acad. Sci. **682**, 272–282.
Peelle, J. E., and Wingfield, A. (**2005**). "Dissociable components of perceptual learning revealed by adult age differences in adaptation to time-compressed speech," J. Exp. Psychol. Hum. Percept. Perform. **31**, 1315–1330.
Penrod, J. P. (**1985**). "Speech discrimination testing," in *Handbook of Clinical Audiology*, 4th ed., edited by J. Katz (Williams and Wilkins, Baltimore, MD), pp. 235–255.
Phillips, D. P. (**1995**). "Central auditory processing: A view from auditory neuroscience," Am. J. Otol. **16**, 338–352.
Pichora-Fuller, M. K., and Schneider, B. A. (**1992**). "The effect of interaural delay of the masker on masking-level difference in young and old adults," J. Acoust. Soc. Am. **91**, 2129–2135.
Pichora-Fuller, M. K., Schneider, B., Benson, N., Hamstra, S. J., and Storzer, E. (**2006**). "Effect of age on gap detection in speech and nonspeech stimuli varying in marker duration and spectral symmetry," J. Acoust. Soc. Am. **119**, 1143–1155.
Pichora-Fuller, M. K., Schneider, B., MacDonald, E., Brown, S., and Pass, H. (**2007**). "Temporal jitter disrupts speech intelligibility: A simulation of auditory aging," Hear. Res. **223**, 114–121.
Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (**1995**). "How young and old adults listen to and remember speech in noise," J. Acoust. Soc. Am. **97**, 593–608.
Pichora-Fuller, M. K., and Souza, P. (**2003**). "Effects of aging on auditory processing of speech," Int. J. Audiol. **42**, S11–S16.
Raven, J. C. (**1965**). *The Mill Hill Vocabulary Scale* (Lewis, London).
Reuter-Lorenz, P. A. (**2002**). "New visions of the aging mind and brain," Trends Cogn. Sci. **6**, 394–400.
Rosen, S. (**1992**). "Temporal information in speech: acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.
Schneider, B. A., Daneman, M., and Murphy, D. R. (**2005**). "Speech comprehension difficulties in older adults: Cognitive slowing or age-related changes in hearing?," Psychol. Aging **20**, 261–271.
Schneider, B. A., and Pichora-Fuller, M. K. (**2001**). "Age-related changes in temporal processing: Implications for listening comprehension," Semin. Hear. **22**, 227–239.
Schneider, B. A., Pichora-Fuller, M. K., Kowalochuk, D., and Lamb, M. (**1994**). "Gap detection and the precedence effect in younger and older adults," J. Acoust. Soc. Am. **95**, 980–991.
Shannon, R. V. (**2002**). "The relative importance of amplitude, temporal and spectral cues for cochlear implant processor design," Amer. J. Audiology **11**, 124–127.
Shannon, R. V., Zeng, F., Kamath, V., and Wygonski, J. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Snell, K. B. (**1996**). "Age-related changes in temporal gap detection," J. Acoust. Soc. Am. **101**, 2214–2220.
Snell, K. B., and Frisina, R. D. (**2000**). "Relationships among age-related differences in gap detection and word recognition," J. Acoust. Soc. Am. **107**, 1615–1626.
Sommers, M. S. (**1997**). "Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment," J. Acoust. Soc. Am. **101**, 2278–2288.
Souza, P. E., and Boike, K. T. (**2006**). "Combining temporal-envelope cues across channels: Effects of age and hearing loss," J. Speech Lang. Hear. Res. **49**, 138–149.
Sticht, T. G., and Gray, B. B. (**1969**). "The intelligibility of time compressed words as a function of age and hearing loss," J. Speech Hear. Res. **12**, 443–448.
Stine, E. A., and Wingfield, A. (**1987**). "Process and strategy in memory for speech among younger and older adults," Psychol. Aging **2**, 272–279.
Stine-Morrow, E. A., Soederberg-Miller, L. M., and Nevin, J. A. (**1999**). "The effects of context and feedback on age differences in spoken word recognition," J. Gerontol. B Psychol. Sci. Soc. Sci. **54**, 125–134.
Stuart, A., and Phillips, D. P. (**1996**). "Word recognition in continuous and broadband noise by young normal-hearing, older normal-hearing, and presbycusic listeners." Ear Hear. **17**, 478–489.
Summers, V., and Leek, M. R. (**1998**). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res. **41**, 1294–1306.
Tremblay, K., Kraus, N., Carrell, T. D., and McGee, T. (**1997**). "Central auditory system plasticity: Generalization to novel stimuli following listening training," J. Acoust. Soc. Am. **102**, 3762–3773.
Tremblay, K., Kraus, N., McGee, T. J., Ponton, C. W., and Otis, B. (**2001**). "Central auditory plasticity: Changes in the N1-P2 complex following speech-sound training," Ear Hear. **22**, 79–90.
Trout, J. D. (**2005**). "Lexical boosting of noise-band speech in open- and closed-set formats," Speech Commun. **47**, 424–435.
Turner, C. W., Souza, P. E., and Forget, L. N. (**1995**). "Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners," J. Acoust. Soc. Am. **97**, 2568–2576.
van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152–1161.
van Tasell, D. J., Greenfield, D. G., Logemann, J. J., and Nelson, D. A. (**1992**). "Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants," J. Acoust. Soc. Am. **92**, 1247–1257.
Vaughan, N., and Letowski, T. (**1997**). "Effects of age, speech rate, and type of test on temporal auditory processing," J. Speech Lang. Hear. Res. **40**, 1192–1200.
Versfeld, N. J., and Dreschler, W. A. (**2002**). "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," J. Acoust. Soc. Am. **11**, 401–408.
Villaume, W. A., Brown, M. H., and Darling, R. (**1994**). "Presbycusis communication and older adults," in *Interpersonal Communication in Older Adulthood*, edited by M. L. Hummert, J. M. Weiman, and J. F. Nussbaum (Sage, Thousand Oaks, CA), pp. 83–106.
Vongpaisal, T., and Pichora-Fuller, M. K. (**2007**). "Effect of age on F_0 difference limen and concurrent vowel identification," J. Speech Lang. Hear. Res. **50**, 1139–1156.
Wingfield, A. (**1996**). "Cognitive factors in auditory performance: Context, speed of processing, and constraints of memory," J. Am. Acad. Audiol **7**, 175–182.
Wingfield, A., Lindfield, K. C., and Goodglass, H. (**2000**). "Effects of age and hearing sensitivity on the use of prosodic information in spoken word recognition," J. Speech Lang. Hear. Res. **43**, 915–925.
Wingfield, A., Tun, P. A., Koh, C. K., and Rosen, M. J. (**1999**). "Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech," Psychol. Aging **14**, 380–389.
Wingfield, A., Tun, P. A., and McCoy, S. L. (**2005**). "Hearing loss in older adulthood: What it is and how it interacts with cognitive performance," Curr. Dir. Psychol. Sci. **14**, 144–148.
Wingfield, A., Wayland, S. C., and Stine, E. A. (**1992**). "Adult age differences in the use of prosody for syntactic parsing and recall of spoken sentences," J. Gerontol. **47**, 350–356.

Priming and sentence context support listening to noise-vocoded speech by younger and older adults

Signy Sheldon, M. Kathleen Pichora-Fuller,[a] and Bruce A. Schneider
Department of Psychology, University of Toronto, 3359 Mississauga Road N., Mississauga, Ontario L5L 1C6, Canada

(Received 17 January 2007; revised 30 July 2007; accepted 22 August 2007)

Older adults are known to benefit from supportive context in order to compensate for age-related reductions in perceptual and cognitive processing, including when comprehending spoken language in adverse listening conditions. In the present study, we examine how younger and older adults benefit from two types of contextual support, predictability from sentence context and priming, when identifying target words in noise-vocoded sentences. In the first part of the experiment, benefit from context based on primarily semantic knowledge was evaluated by comparing the accuracy of identification of sentence-final target words that were either highly predictable or not predictable from the sentence context. In the second part of the experiment, benefit from priming was evaluated by comparing the accuracy of identification of target words when noise-vocoded sentences were either primed or not by the presentation of the sentence context without noise vocoding and with the target word replaced with white noise. Younger and older adults benefited from each type of supportive context, with the most benefit realized when both types were combined. Supportive context reduced the number of noise-vocoded bands needed for 50% word identification more for older adults than their younger counterparts.
 [DOI: 10.1121/1.2783762]

PACS number(s): 43.71.Lz, 43.71.Es, 43.71.Sy, 43.66.Sr [PEI] Pages: 489–499

I. INTRODUCTION

Older adults, even those with clinically normal audiometric thresholds in the speech range, have more difficulty than younger adults comprehending spoken language, especially in adverse listening conditions (for a review see CHABA, 1988; Pichora-Fuller and Souza, 2003). Even when older adults are able to correctly identify words heard in challenging listening conditions, there is evidence that they engage in more effortful processing than younger adults (for reviews see Pichora-Fuller, 2003; Wingfield and Tun, 2007). It has been suggested that older adults depend on supportive context to compensate for age-related reductions in perceptual and cognitive processing (e.g., Craik, 1983, 1986), including those that contribute to difficulties in comprehending spoken language (for reviews see Pichora-Fuller, 2003; Wingfield, 1996). Although both younger and older adults use various types of cues, including acoustic and lexical cues, to support listening to speech in challenging auditory scenes (Gallacher, 2005), older adults may even benefit more than younger adults from these various types of supportive context in challenging listening conditions (for reviews see Schneider and Pichora-Fuller, 2000; Wingfield and Tun, 2007). What is not yet fully understood is how supportive context from different sources, that is, acoustic cues from the speech signal and lexical cues from the syntactic or semantic context, interact during spoken language comprehension for younger and older listeners. Therefore, in the present study, we focus on the possible interactive use of two types of supportive context that are available when younger and older adults try to understand noise-vocoded speech.

A. Noise vocoding

Noise vocoding is a form of speech distortion that involves dividing a speech signal into specific frequency bands and, within each band, extracting the amplitude envelope and using it to modulate noise of the same bandwidth. This results in the fine structure of the signal being replaced with noise. This type of distortion minimizes the contribution of fine structure cues, and preserves the use of temporal amplitude-envelope cues. As the number of bands is increased, more band-specific envelope information becomes available. The intelligibility of noise-vocoded speech improves as the number of bands is increased (e.g., Loizou *et al.*, 1999; Shannon *et al.*, 1995).

There is compelling evidence that the comprehension of spoken language relies heavily on information that is carried by envelope cues (van Tasell *et al.*, 1992; Shannon *et al.*, 1995; Dorman and Loizou, 1998; Shannon, 2002). Specifically, envelope cues are one type of cue that carries supra-segmental information involved in lexical and syntactic processing (Schneider and Pichora-Fuller, 2001; Shannon, 2002; Greenberg, 1996; Rosen, 1992), with these cues contributing to speech prosody (Cutler *et al.*, 1997). Noise vocoding permits the investigation of the use of temporal amplitude-envelope cues relevant to supra-segmental speech processing while minimizing the contributions of fine structure cues that carry pitch-related supra-segmental information.

It is important to examine noise-vocoded speech comprehension for a number of reasons. First, given that we

[a]Electronic mail: k.pichora.fuller@utoronto.ca

know there are age-related declines in auditory temporal processing for cues relevant to segmental (Schneider *et al.*, 1994; Snell and Frisina, 2000; Pichora-Fuller *et al.*, 2006) and sub-segmental aspects of speech processing (Abel *et al.*, 1990; Summers and Leek, 1998; Alain *et al.*, 2001; Vongpaisal and Pichora-Fuller, 2007), it is important to discover if age-related differences also exist with regard to the envelope cues involved in supra-segmental speech processing. Second, noise vocoding provides an opportunity to examine how different types of contextual support may be used by younger and older listeners to compensate for auditory processing problems induced by challenging listening conditions (i.e., noise-vocoded speech).

In a recent companion study (Sheldon *et al.*, 2008), we used noise-vocoded speech to investigate possible age-related differences in the auditory processing of envelope cues. We established that noise-vocoding undermines word identification more for older than for younger listeners when target words are presented only once and with no feedback. Interestingly, these age-related differences in word identification were eliminated when older adults were able to benefit from feedback and/or from the summing of information over sequential presentations of the to-be-identified word, beginning with the word presented with one frequency band and incrementing the number of bands by one until it was correctly identified. Thus, noise vocoding provides an interesting type of signal degradation insofar as age-related perceptual differences in word identification seem to be effectively offset when feedback and/or summing of information is possible. One limitation of our earlier study is that there was no opportunity for the listeners to use different types of supportive context because the target words were spoken in the fixed carrier phrase "*Say the word*" (Sheldon *et al.*, 2008). Therefore, in the present experiment, we expand on the repertoire of types of supportive context to examine the resulting effect on lexical access.

B. Lexical access

Identifying a word involves establishing a mapping between the speech signal and the corresponding semantic meaning. This process is complex and intuitively integrative, involving information from both low-level and high-level processes. Accordingly, lexical access is affected by processes at these multiple levels. For instance, single-word lexical access can be supported via semantic priming (e.g., Meyer and Schvaneveldt, 1971) and phonological priming (e.g., Goldinger *et al.*, 1992). At the word level, semantic and phonological contexts have been found to each influence lexical selection; for example, Ferreira and Griffin (2003) showed that semantic priming and homophone priming could each improve lexical selection. Specific to speech, different types of context are known to mediate lexical selection during quiet listening conditions (e.g., Radeau *et al.*, 1998).

Less is known about the way in which different types of context combine to facilitate lexical access in challenging listening conditions and how such facilitation may change with age. It has been well established that in challenging listening conditions, speech can often be understood if there is sufficient semantic or linguistic contextual support available to provide information about the degraded signal (e.g., Kalikow *et al.*, 1977; Bilger *et al.*, 1984). Further, older adults are particularly adept at using context to compensate for difficulties hearing a degraded acoustic signal, presumably having developed expertise because typical everyday listening conditions are often more perceptually challenging for them than they are for younger adults (Perry and Wingfield, 1994; Pichora-Fuller *et al.*, 1995; Gordon-Salant and Fitzgibbons, 1997; Sommers and Danielson, 1999; Wingfield *et al.*, 2005).

The acoustic signal itself may also support spoken language comprehension by supplementing or augmenting the use of context based on semantic knowledge. Various studies have demonstrated that listeners can use phonological or prosodic information to direct attentional or top-down resources during spoken word recognition (Gow and Gordon, 1995; Marslen-Wilson and Tyler, 1980; Pitt and Samuel, 1990). Moreover, other situational cues, such as priming with a semantically related sentence (e.g., Gagné *et al.*, 2002), presenting visual speech for speech reading (e.g., Sumby and Pollack, 1954), presenting written text or clear speech as feedback (e.g., Davis *et al.*, 2005), spatially separating concurrent sounds (e.g., Freyman *et al.*, 1999, 2001; Li *et al.*, 2004), and increasing the pitch differences among simultaneous talkers (e.g., Mackersie and Prida, 2001) can all enhance speech intelligibility. These cues presumably enhance listening to the relevant structural patterns in the acoustic signal and may, in turn, enhance the effectiveness of semantic cues. In other words, information at the perceptual level can influence the deployment of semantic information for understanding degraded speech.

The interaction between acoustic and semantic level cues during lexical access has been examined by Connine *et al.* (1997). In their study, young, healthy participants performed a series of experiments that showed that phoneme detection reaction times are influenced by similarity to a carrier stimulus both in terms of form and meaning. Similarly, Andruski *et al.* (1994) demonstrated that the perceptual clarity of a word-initial phoneme is one determinant for the amount of semantic priming. Thus, these experiments demonstrate that the activation of the meaning of words is intimately tied to both the activation of the phonetic form of the word and to its lexical meaning.

The purpose of the present study is to investigate how different types of context interact to support word identification in younger and older adults. Specifically, we examine how the predictability of a word from sentence context, and priming with sentence content, combine to affect word identification in an adverse listening condition (noise-vocoded speech).

II. METHOD

A. Participants

Sixteen younger adults (mean age=20.6 years, s.d. =2.7, range=17–25 years) and sixteen older adults (mean age=67.6 years, s.d.=3.0, range=65–73 years) participated in this experiment. Both age groups had audiometric pure-

TABLE I. Mean (s.d.) of audiometric air-conducted pure-tone thresholds (dB HL) for test ears of younger and older participants.

	Frequency (kHz)							
	0.25	0.5	1	2	3	4	6	8
Younger participants (N=16)								
Mean	7.50	3.13	0.31	−2.18	−0.94	−0.63	3.13	1. 25
(s.d.)	(6.32)	(5.44)	(5.00)	(5.15)	(5.54)	(6.02)	(7.50)	(8.66)
Older participants (N=16)								
Mean	6.25	5.93	4.37	10.93	11.87	20.93	30.62	45.00
s.d.	(8.47)	(7.35)	(7.71)	(8.76)	(8.80)	(8.13)	(11.58)	(15.90)

tone air-conduction thresholds in the test ear that were considered to be clinically normal in the speech range (less than or equal to 25 dB HL from 0.25 to 3 kHz; see Table I). All of the participants were highly educated, with 15.1 (s.d.=2.7) and 13.7 (s.d.=2.3) mean years of education, respectively, for the younger and older groups; there was no significant difference in education between the age groups [$t(30)=1.48$, $p>0.10$]. All participants had learned English before the age of 5 years. To measure verbal knowledge, each participant completed the Mill-Hill Vocabulary Scale (Raven, 1965); the scores for both age groups indicated that all participants had good knowledge of the English language. The mean scores out of 20 for the younger and older groups were 12.5 (s.d. =2.2) and 14.8 (s.d.=2.2), respectively, with the older group significantly outperforming the younger group [$t(30)=0.30$, $p<0.05$]. All of the participants were paid volunteers recruited from the local community and none had previously heard noise-vocoded speech or the sentence materials that were used in the study. The participants completed Part A of the experiment in one session and returned within a week to complete Part B of the experiment.

B. Experiment: Part A

1. Materials

A digitized version of the sentences of the Speech Perception in Noise Test (SPIN-R; Bilger *et al.*, 1984) was used in the study. The SPIN-R materials consist of eight lists of 50 sentences per list. Each of the eight lists consists of 25 sentences in which the sentence-final word is predictable from the sentence context (e.g., "*Stir your coffee with a spoon*.") and 25 sentences in which the sentence-final word cannot be predicted from the sentence context (e.g., "*He would think about the rag*."). The high-context and low-context sentences are presented in a fixed pseudorandom order within each list.

For each list, the sentences were noise-vocoded into 16-band, 8-band, 4-band, and 2-band versions. To do so, we followed the procedure described in detail by Eisenberg *et al.* (2000). First, using the Goldwave digit audio editor, the stimuli were converted into binary files with a sampling rate of 20 kHz. Using MATLAB software, stimuli were then processed through a pre-emphasis filter [a high-pass first-order Butterworth infinite impulse response (IIR)] with a cut-off frequency of 1.2 kHz and an attenuation rate of −6 dB per octave. The signal was split into a varying number of frequency bands ($n=2,4,8,16$) using fourth-order elliptical IIR bandpass filters with a maximum peak-to-peak ripple of 0.5 dB in the passband and a minimum attenuation of 40 dB in the stop band. The passband used to split the signal into frequency bands spanned a frequency range from 0.3 to 6 kHz for all conditions. The frequency spacing of the filter banks was based on the work of Greenwood (1990). The boundary frequencies for the band-processed conditions are shown in Table II. To extract the envelopes, the magnitude of the Hilbert transform was computed and passed through a low-pass filter (second-order Butterworth IIR with cut-off frequency of 160 Hz). One minor difference between the present and earlier procedures was that whereas Eisenberg *et al.* (2000) rectified and then low-pass filtered the filter bank outputs, we extracted the envelope using the magnitude of the Hilbert transform followed by a low-pass filter similar to that used by Eisenberg *et al.* (2000). Narrow-band noise was generated by passing a Gaussian white noise signal through the same Butterworth and elliptical filters. The envelopes extracted in the previous step were then used to modulate the corresponding band of noise. The bands of modulated noise were then summed together. Finally, the stimuli were converted to .wav format with a sampling rate of 24 kHz using a digital audio editor.

2. Procedure

During testing, the participant sat comfortably inside an International Acoustics Company (IAC) double-walled

TABLE II. Boundary frequencies (Hz) for the 2-, 4-, 8-, and 16-band noise-vocoded conditions.

2 band	4 band	8 band	16 band
300	300	300	300
1528	722	477	382
6000	1528	722	477
	3066	1061	590
	6000	1528	722
		2174	878
		3066	1061
		4298	1276
		6000	1528
			1825
			2174
			2584
			3066
			3632
			4298
			5080
			6000

sound-attenuating booth. The sound files were played by a Tucker Davis Technologies (TDT) System III monaurally to the participant's better ear over Sennheiser (model HD 265) headphones. The signal level was set at 70 dB SPL and was constant across conditions for both age groups.

To familiarize participants with noise-vocoded speech, each participant was first exposed to four lists of 50 noise-vocoded W-22 words (Martin and Pennington, 1971; Martin and Forbis, 1978; Penrod, 1994). The presentation of stimulus conditions was blocked, with the order of conditions progressing from easiest to hardest. First, participants heard a word list vocoded with 16 bands, followed by a word list in the 8-band condition, then a word list in the 4-band condition, and finally a word list in the 2-band condition. Words within a list were presented in random order and list order was counterbalanced across participants so that each list was presented four times in each order position for each group. For each word in each list, the participant was asked to identify the word. No feedback was provided either during the familiarization or during the experimental trials.

Following the familiarization with noise-vocoded words, participants were told that they would hear four lists of sentences and that they were to identify the last word of each sentence. One full SPIN-R list in the 16-band condition was presented, followed by another list in the 8-band condition, and then another list in the 4-band condition, and finally the last list in the 2-band noise-vocoded condition. Breaks were given as needed between lists. The list order was counterbalanced across participants so that each list was presented twice in each order position for each group. Guessing was strongly encouraged.

The experimenter used Sennheiser headphones (model HD 265) to listen to the participant's responses, which were scored immediately. If the experimenter was uncertain about any response, the participant was asked to repeat and spell the word aloud. Any phonemic difference between the response and the target word was marked as an error. No feedback was given after a response. All sessions were audio-taped to enable later verification of the responses recorded on the score sheet. The testing session lasted approximately 1–1.5 h.

C. Experiment: Part B

1. Materials

The same SPIN-R sentences were used in Part A and Part B of the experiment. Each participant heard all eight noise-vocoded SPIN-R lists, four lists in Part A and the other four lists in Part B of the experiment. No list was heard more than once by any participant. Part B differed from Part A because a priming utterance was presented prior to each sentence (following Experiment 2 of Freyman *et al.*, 2004). The priming utterance for each noise-vocoded sentence was constructed using the intact version of the sentence that had been used to produce the noise-vocoded sentence. For the prime, the initial portion of the intact sentence was presented, but the sentence-final word was replaced by a segment of white noise. The segment of noise used to replace the word was a randomly generated 700-ms white noise token that was scaled to have an average rms of 10 dB below the average rms of the noise-vocoded SPIN-R sentences. The wave form was edited to extract the sentence-final word from each sentence and to replace it with a noise segment such that the word and any coarticulatory cues were minimized in the priming utterance.

For presentation, the files were converted to single-channel .wav files and played monaurally to the participant's better ear over headphones. The signal level for the noise-vocoded sentences was set at 70 dB SPL and was constant across conditions for both age groups.

2. Procedure

During the testing session, the participant sat comfortably inside the double-walled sound-attenuating booth. Participants were told that they would hear four more lists of sentences and they were asked to listen to the prime (clear sentence context with the sentence-final word replaced with white noise) followed by the noise-vocoded sentence and then to identify the final word of the distorted sentence. First, they were presented one full SPIN-R list in the 16 band-processed condition, and then lists vocoded with 8, 4, and 2 bands. The order of the band conditions was fixed from easiest to hardest. Breaks were given as needed between lists. The list order was counterbalanced across participants so that each list was represented twice in each order position for each group. Guessing was strongly encouraged.

Participants' responses were scored immediately as in Part A. If the experimenter was uncertain about any response, the participant was asked to repeat and spell the word aloud. Any difference between the response and the target word was marked as an error. No feedback was given after a response. All sessions were audio-taped so that scoring could be confirmed as needed after the session. The testing session lasted approximately 1.5 h.

III. RESULTS

The percentage of sentence-final words correctly identified as a function of the number of frequency bands used in noise vocoding was calculated for each participant for each sentence type (low context or high context) in each experimental condition (with or without the priming utterance). We will refer to Part A of the experiment as the "without prime" experimental condition and Part B of the experiment as the "with prime" experimental condition.

Exponential functions of the form to follow were calculated for each of the two age groups to describe their performance in each of the four conditions; that is, with two levels of sentence context (low and high) and two levels of priming context (with and without prime):

$$y = c - e^{(a-bx)}.$$

Exponential functions were fit to the individual data, and to the average data for each age group. The functions describe the probability of correctly identifying sentence-final words (y) as a function of number of bands (x), subject to the restrictions that $a \geq 0$, and $c \leq 1$. Figure 1 plots the fitted functions for all of the participants in the low-context condi-

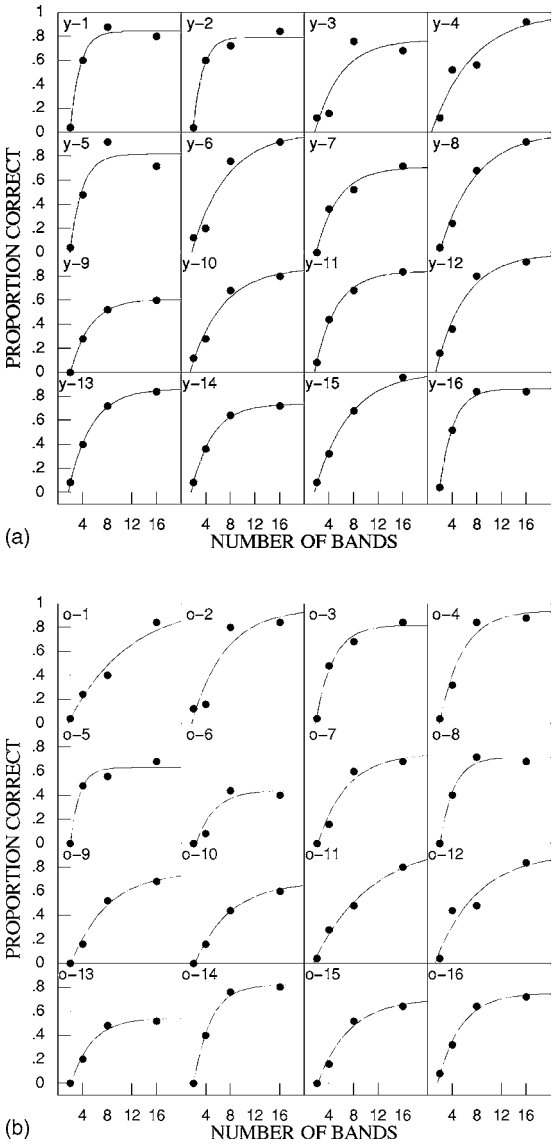


FIG. 1. (a) The proportion of words in low-context sentences that each younger participant (y) correctly identified, as a function of the number of bands used in noise vocoding. (b) The proportion of words in low-context sentences that each older participant (o) correctly identified, as a function of the number of bands used in noise vocoding.

tion with no prime. Figure 1 illustrates that the exponential function provides a good fit to the individual data.

The individual functions were used to estimate the band number that resulted in 50% of the target words being correctly identified in each of the four conditions (low context without prime; low context with prime; high context without prime; and high context with prime). For one older adult the 50% point could not be determined in the low-context condition without prime because for that participant the exponential function reached an asymptote that was less than 50%; for subsequent analyses, the data of this particular older participant were discarded and replaced with the corresponding mean of the data of the other older participants.

Figure 2 shows the raw mean band threshold for each of the four conditions for younger and older adults. As seen in Fig. 2, for both age groups, the mean band threshold is lowered by the presence of a prime, by context, and by the

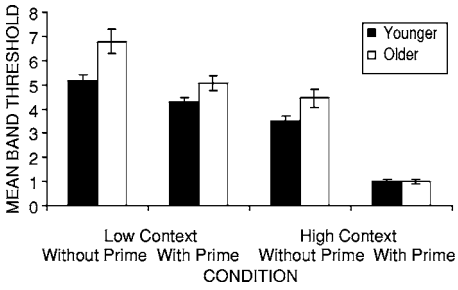


FIG. 2. The mean threshold values (the number of bands required to achieve 50% correct word identification) for older and younger participants in the four experimental conditions: Without prime, low context; with prime, low context; without prime, high context; with prime, high context. Standard error bars are shown.

presence of both combined. Figure 2 also illustrates that in the low-context condition without prime older adults have a mean threshold value that is nearly two bands greater than the mean threshold value of younger adults. Figure 2 also indicates that the age-related difference in mean threshold values is greatly reduced when a prime is present (low context, with prime condition) and when contextual cues are present (high context, without prime). Furthermore, when both prime and high-context cues are available, there is no age-related difference in the mean threshold value. Thus, Fig. 2 suggests that the mean threshold values of older adults are reduced to a greater extent than those of younger adults by the presence of the prime, and also by the presence of high context, but to the same extent when both the prime and high context are available in combination.

The description of Fig. 2 was verified by an analysis of variance (ANOVA) with age (younger or older) as a between-subjects factor and prime condition (with or without prime) and sentence context (low context or high context) as within-subjects factors. There were significant main effects of age $[F(1,30)=7.88, p<0.01]$, prime $[F(1,30)=146.47, p<0.0001]$, and context $[F(1,30)=273.40, p<0.0001]$ on the number of noise-vocoded bands needed to achieve 50% word identification. There were also significant two-way interaction effects of prime and context $[F(1,30)=36.30, p<0.0001]$, age and prime $[F(1,30)=6.30, p<0.02]$, and age and context $[F(1,30)=4.46, p<0.05]$. The three-way interaction effect of age, prime, and context $[F(1,30)<1]$ was not significant.

The two-way interaction between prime and context indicates that the combined effects of priming and context are greater than the sum of their individual effects. To help interpret the pattern of two-way interaction effects between context and priming, age and priming, and age and context, we next examined the benefit that younger and older participants received from context in the conditions without and with prime. Figure 3 plots the mean proportion of instances in which the sentence-final word was correctly identified in low-context and high-context sentences as a function of the number of bands in two conditions (with prime and without prime) for the younger group and for the older group. The separations between the 50% points on the low-context and high-context functions are indicated by horizontal lines located at 0.5 on the abscissa.[1] Figure 3 suggests that there is a

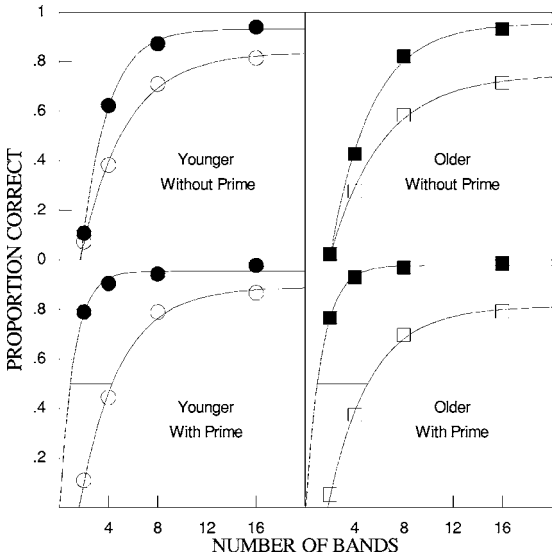


FIG. 3. The proportion of words correctly identified, averaged across participants, as a function of the number of bands for younger (circles) and older (squares) participants in the conditions without prime (Part A) and with prime (Part B). Closed symbols represent the conditions in which the sentence-final word was predictable from the sentence context; unfilled symbols represent the conditions in which the sentence-final word was not predictable from the sentence context.

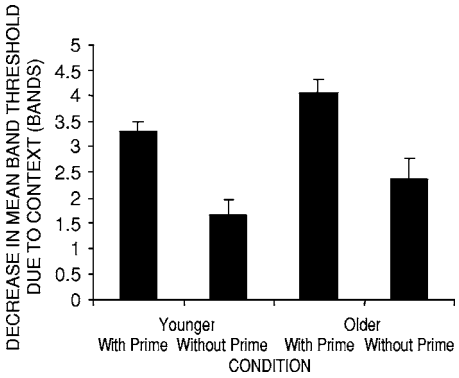


FIG. 4. The mean reduction in threshold (the number of bands required to achieve 50% correct word identification) due to sentence context for the four groups: younger adults, without prime; younger adults, with prime; older adults, without prime; and older adults, with prime. Standard error bars are shown.

larger difference in the number of bands required to achieve a threshold of 50% correct word identification between the low-context and high-context conditions in the condition with prime compared to the condition without prime. Moreover, the effect of priming on these thresholds appears to be about the same for both younger and older adults.

Although the extent of the age-related difference appears to be the same in the condition with prime as it is in the condition without prime, the threshold separation between the low-context and high-context functions appears to be larger in older adults than it is in younger adults in both the conditions with and without prime. To test this statistically, for the conditions with and without prime, we subtracted each individual's threshold for high-context sentences from his or her threshold for low-context sentences to obtain an estimate of the extent of the reduction in threshold due to context (the reduction in the number of bands needed to reach 50% correct).

Figure 4 plots the average reduction in threshold due to the addition of context for the two age groups in the conditions with and without prime (corresponding to the horizontal lines at 0.5 on the abscissa for each panel of Fig. 3). An ANOVA with age (younger or older) as a between-subjects factor and priming condition (with prime versus without prime) as a within-subjects factor confirmed that there was a significant main effect of age [$F(1,30)=4.53$, $p<0.05$], and a significant main effect of priming condition [$F(1,30)=36.27$, $p<0.0001$] on the benefit due to context, but there was no significant interaction between age and priming condition [$F(1,30)<1$)].

To complement Fig. 3, Fig. 5 plots the mean proportion of instances in which the sentence-final word was correctly identified in sentence contexts with and without prime as a function of the number of bands in two conditions (low context and high context) for the younger group and for the older group. In each panel of Fig. 5, the separation between the 50% points on the functions for the conditions with and without prime is indicated by horizontal lines located at 0.5 on the abscissa.[1] As seen in Fig. 5, the effects of priming are larger for high-context sentences than they are for low-context sentences by about the same amount in both younger and older adults. Figure 5 also shows that the effects of priming appear to be larger for older than they are for younger adults by about the same amount for both low-context and high-context sentences.

To complement Fig. 4, Fig. 6 plots the average reduction in threshold due to priming for the two age groups in the high-context and low-context conditions (corresponding to

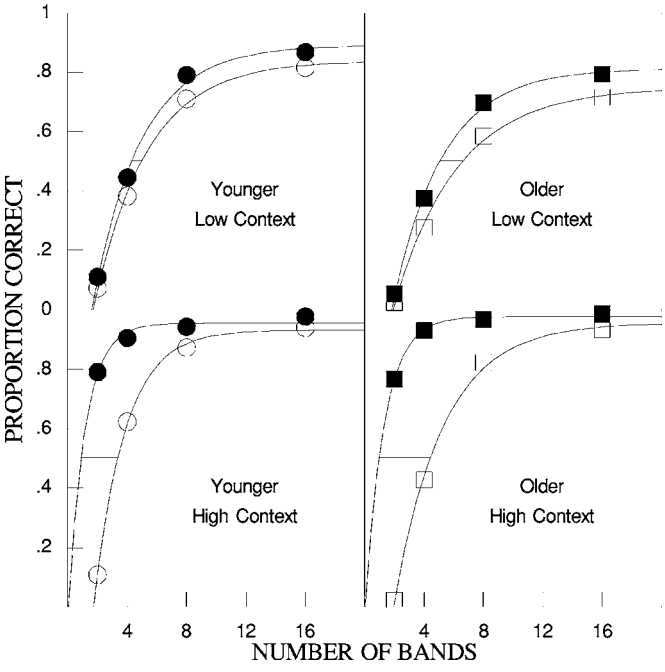


FIG. 5. The proportion of words correctly identified, averaged across participants, as a function of the number of bands for the conditions with prime (closed symbols) and without prime (open symbols) for younger adults (circles) and older adults (squares) when the sentence-final word was predictable from sentence context (high context) or when the sentence-final word was not predictable from sentence context (low context).

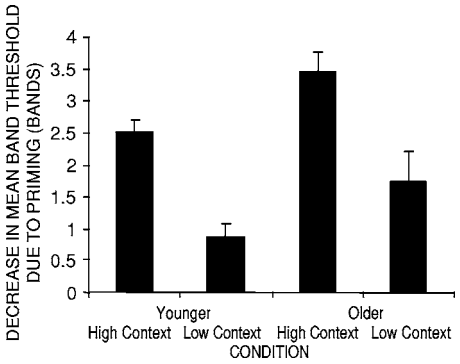


FIG. 6. The mean reduction in threshold (the number of bands required to achieve 50% correct word identification) due to the addition of the prime for the four groups: younger adults with low-context sentences; younger adults with high-context sentences, older adults with low-context sentences, and older adults with high context sentences. Standard error bars are shown.

the horizontal lines at 0.5 on the abscissa for each panel of Fig. 5). An ANOVA of the differences with age (younger and older) as a between-subjects factor and sentence context (low context or high context) as a within-subjects factor confirmed that there was a significant main effect of age $[F(1,30)=6.44,\ p>0.02]$ and context $[F(1,30)=36.27,\ p<0.0001]$, but no significant age by context interaction $[F(1,30)<1]$.

IV. DISCUSSION

The primary goal of the current study was to explore age-related differences in the contributions of two types of supportive context: Support provided by priming with the undistorted presentation of all but the final word of the sentence, and support provided by the predictability of a sentence-final word from the sentence context in an adverse listening condition. We used a form of speech distortion, noise vocoding, in which the fine structure cues in the speech signal are systematically reduced to create different degrees of adversity. In general, the band threshold (number of bands corresponding to 50% correct target word identification) was higher for older than for younger adults, lower when the prime was present, and lower for high as opposed to low-context conditions. In addition, the band threshold in the condition combining high context and priming was 5 bands lower than in the condition with low context and without priming. This 5-band benefit from the combination of supports was bigger than the sum (3.3 bands) of the individual benefits of adding a prime alone (1.3 bands), or changing the context from low to high alone (2.0 bands).

The effects of context and priming were larger for older than for younger adults. Specifically, for younger adults, adding context reduced the threshold from 4.7 to 2.2 bands (2.5 bands); for older adults, it reduced the threshold from 5.9 to 2.7 bands (3.2 bands). Furthermore, for younger adults, priming reduced the threshold from 4.3 to 2.6 bands (1.7 bands); for older adults, it reduced the threshold from 5.6 to 3.0 bands (2.6 bands). However, these age-related differences disappear if the effects of context and priming are considered in terms of the percentage decrease in threshold.[2] Specifically, the percentage decrease when going from low to high context was 48% for younger and 46% for older adults. Furthermore, going from conditions without prime to conditions with prime reduced thresholds by 61% and 54% for younger and older adults, respectively. Hence, depending on whether or not the threshold changes are expressed as differences or percentages, we would conclude either that older adults benefit more from priming and context that do younger adults (thresholds changes expressed as differences), or that younger and older adults benefit equally from these two factors (threshold changes expressed as percentages). In either case, we can conclude that older adults benefit at least as much as do younger adults from these two types of support.

It seems unlikely that the age differences found in the present study can be attributed to age-related differences in audiometric thresholds. All younger and older adults had clinically normal audiometric thresholds for frequencies from 0.25 to 3 kHz, although the pure-tone average (0.5, 1, and 2 kHz) of the older group was about 6 dB higher than that of the younger group, and there were larger differences at higher frequencies. Nevertheless, it is not obvious how small differences in hearing threshold could account for the pattern of results related to the number of bands used in noise vocoding the sentences. In particular, increasing the number of bands used to vocode the sentences provided the listeners with more frequency-specific envelope information, but the frequency range and the audibility of the noise-vocoded sentences remained constant across the different band conditions. Furthermore, all stimuli were presented in quiet so age-related differences related to signal-to-noise ratio cannot explain the pattern of findings. Importantly, our finding of an age effect in the present study is consistent with our finding of age-related differences in Experiment 2 in our previous study (Sheldon *et al.*, 2008), and it is also in line with the results of Souza and Boike (2006), who found that age, but not degree of hearing loss, was a significant predictor of the ability of listeners to identify noise-vocoded /aCa/ nonsense bisyllables in a 16-alternative closed-choice task.

Considering other possible factors associated with age, the participants did not differ in education level. However, older adults outperformed younger adults on the Mill-Hill vocabulary score (Raven, 1965). Hence, the possibility that the greater benefit from support realized by the older adults might be related to their superior lexical knowledge cannot be ruled out.

The results of the study will be discussed with an emphasis on age-related differences in the auditory processing of temporal amplitude-envelope cues relevant to suprasegmental speech processing and the use of the two types of support and how they may combine during spoken language comprehension.

A. Age-related differences in auditory temporal processing

The word identification scores in the low-context condition without prime (Part A) indicate that older adults are poorer than younger adults at processing envelope cues when little contextual support is available. The present results for sentence-final word identification in the low-context condi-

tion without prime are consistent with the results of our earlier study in which age-related deficits were found in the use of envelope cues to identify noise-vocoded monosyllable words presented with a fixed carrier phrase, at least when there was no feedback or opportunity to sum information from sequential presentations of the same word (Sheldon *et al.*, 2008). Nevertheless, for the participants in the present experiment, we found no significant correlation between their word identification performance in the familiarization task and their performance in any of the four experimental conditions (low context with or without prime, high context with or without prime).

Together with prior research, the present findings provide converging evidence that aging is associated with deficits at multiple levels of auditory temporal processing that contribute to different levels of speech processing: the subsegmental level (e.g., Abel *et al.*, 1990; Summers and Leek, 1998; Alain *et al.*, 2001; Vongpaisal and Pichora-Fuller, 2007), the segmental level (Schneider *et al.*, 1994; Snell and Frisina, 2000; Pichora-Fuller *et al.*, 2006), and the suprasegmental level (Sheldon *et al.*, 2008).

B. Sentence context supports spoken language comprehension

When younger and older participants listened to SPIN-R sentences in the four band-processed conditions, whether they heard a prime (Part B) or did not (Part A), word identification performance was better in high-context sentences than in low-context sentences for both age groups. However, the difference due to the addition of this type of context was greater for older than for younger listeners. In other words, older adults derived more benefit from sentence context than did younger adults when benefit was measured in terms of the difference in the number of bands required for 50%-correct identification. The present results are consistent with previous studies that have shown that older adults are the same (Dubno *et al.*, 2000) or even better than younger adults at using semantic context than younger adults to boost word identification in adverse situations where listening (or reading) is effortful (for reviews see Schneider, 2001; Pichora-Fuller, 2003). A general finding has been that older adults require a higher signal-to-noise ratio to achieve the same word identification score as younger adults when no supportive context is available. Because the signal-to-noise conditions typical in everyday life would often be challenging for older listeners even though these conditions would not be particularly challenging for younger listeners, one reasonable possibility is that older adults have much more practice at using context in a compensatory fashion (Pichora-Fuller *et al.*, 1995). It is important to note that the older participants in the present study scored higher than did the younger participants on the vocabulary test. The superior vocabulary of the older participants in the present study is consistent with the possibility that the larger lexicons and greater lexical familiarity among older adults could explain an age-related benefit from sentence context associated with better semantic knowledge (Wingfield *et al.*, 2005).

C. Priming directs auditory processing

The presence of the prime also decreased the mean band threshold for both age groups. Even when semantic and linguistic cues were minimal, as in the low-context sentences, there was still an increase due to priming in the word identification scores for both age groups. Because the comprehension of noise-vocoded speech is primarily based on information carried by the envelope, it is possible that listeners attend to the envelope cues of the intact prime to facilitate comprehension of the vocoded sentence. That is, the prime may direct a listener's attention to relevant aspects of structure of the speech signal, thereby contributing to improved sentence-final word identification. Support for this explanation comes from studies that stress the importance to comprehension of the supra-segmental speech information provided by the envelope (Martin, 1979; for a review, see Cutler *et al.*, 1997). Specifically, envelope cues preserve amplitude modulations that may provide cues to the beginnings and endings of words that are needed for segmenting running speech (Sanders *et al.*, 2002; Sanders and Neville, 2003a, b). It also seems that the prime may offer an opportunity for perceptual learning that could result in improved understanding of noise-vocoded speech. This is suggested by the finding that when clear speech or written text was provided following the presentation of a noise-vocoded utterance, word identification improved for subsequently presented novel utterances (Davis *et al.*, 2005).

The presence of the prime was of greater benefit to older than younger adults in the low-context conditions when the advantage is expressed in terms of differences in number of bands required for 50%-correct identification. To our knowledge, an age-related advantage from this sort of priming has not been reported previously. We offer two possible explanations for our observations.

1. Compensation. One possibility is that the age-related difference in ability to benefit from supportive context reflects how older adults compensate by engaging top-down processing to achieve the same performance as younger adults on various perceptual and cognitive tasks (for a review see Pichora-Fuller and Singh, 2006). The possibility that older adults engage in different and possibly more effortful types of information processing to compensate for deficits in auditory processing is consistent with the more general observation that older adults use more brain regions and show a reduction in hemisphericity of activation on a wide range of perceptual and cognitive tasks. This age-related difference in activation is thought to reflect compensatory adaptations (e.g., Grady, 2000; Cabeza, 2002; Reuter-Lorenz, 2002). The age-related difference in benefit from the prime in the low-context sentence conditions may reflect the greater compensatory use of supportive context by older adults in challenging listening conditions such as those that they frequently encounter in everyday life. We speculate that older adults are simply more skilled at using information provided by the prime to direct attention to the envelope cues in the vocoded speech that facilitate word segmentation and phonological processing, particularly in low-context sentences where other types of contextual support are not readily available.

2. Summing information. Another explanation is that there is an age-related difference in the processing strategies used when the prime is available. In our earlier study, older listeners were able to match the word identification accuracy of younger listeners by summing information across a sequence of versions of the same target word that were incremented progressively in the number of bands used to vocode the speech (Sheldon *et al.*, 2008). In the current study, the listener may be able to sum information by comparing the clear sentence context of the prime and the vocoded sentence context. It could be that older adults are particularly adept at summing knowledge to help decipher a target word.

D. Priming facilitates benefit from sentence context

Not surprisingly, when both types of support, priming and high sentence context, are available, listeners reach ceiling performance and no significant age-related differences are observed. The benefit derived when these two types of support are combined is greater than the sum of their independent contributions. In the high-context condition with prime, both younger and older listeners correctly identified over half of the words even in the 2-band condition, the lowest band condition tested. This high level of performance indicates that both age groups successfully use auditory verbal closure such that contextual information facilitates identification of the degraded target word (for a review see Elliott, 1995). The prime facilitates the use of sentence context because it provides an undistorted presentation of the sentence context. However, even when the sentence context is low, the prime provides information that may help the listener to narrow in on what to listen for and when to listen in the noise-vocoded sentence carrying the target word (Freyman *et al.*, 2004). It seems that listeners benefit from knowledge of the acoustical structure as well as semantic knowledge of the sentence context.

V. CONCLUSIONS

Envelope cues that provide supra-segmental speech information are integral to spoken language comprehension. When such cues are degraded, as they are when speech is noise vocoded, younger and older listeners compensate by relying to a greater extent on the available supportive context. Five conclusions can be drawn from this study regarding the use of such support when speech is noise vocoded:

(1) Words that are highly predictable from sentence context are better comprehended than words that are not predictable from sentence context.
(2) A prime, in the form of undistorted presentation of the sentence context, can facilitate comprehension, even in the absence of high-context semantic information, by directing a listener to relevant acoustic properties of the speech signal.
(3) When the prime is presented for a high-context sentence, the prime also facilitates benefit from the semantic context, thus demonstrating the interactivity of the two types of cues.
(4) Older adults do not identify words as well as younger adults when envelope cues are reduced by noise vocoding. Conversely, older adults need more bands than younger adults to achieve the same level of accuracy in identifying the final words of noise-vocoded sentences.
(5) Older adults benefit more than younger adults from sentence context and the prime; however, both age groups receive equal benefit when these two types of context interact.

ACKNOWLEDGMENTS

The authors are grateful to Ewen MacDonald for creating the noise-vocoding software. This research was funded by the Canadian Institutes of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada (NSERC). These experiments were conducted as part of the Master's Thesis of S.S.

[1]Note that to find the 50% correct points on the functions for the high-context sentences with prime shown in Figs. 3 and 5, it was necessary to extrapolate from the data. For both age groups, the mean percent correct score exceeded 70% when high-context sentences were presented with a prime in the 2-band noise-vocoding conditions. To find the 50% correct point, we extrapolated using the conservative assumption that both groups would score 0% correct if no stimulus were presented (0 band condition). Given that the sentence-final words in the high-context sentences of the SPIN test were designed to be highly predictable from preceding sentence context, a more realistic and less conservative assumption would have been a minimum score of 70% based on the paper and pencil test of auditory verbal closure that was conducted during the development of the original SPIN test (Kalikow *et al.*, 1977; see also Bilger *et al.*, 1984; Elliott, 1995). Had we used a less conservative method to estimate the 50% point, or had we used a percentage correct higher than 50% to compare conditions, then the difference due to context (Fig. 3) and the difference due to priming (Fig. 5) would have been larger for both age groups, but the differences for the older group would have remained larger than those for the younger group.

[2]To confirm the lack of interaction when changes in the 50% correct-identification thresholds are expressed as percentages, we conducted an ANOVA on the log band thresholds with age (younger or older) as a between-subjects factor and priming condition (with or without prime) and sentence context (low context or high context) as within-subjects factors. Note that when the 50% band threshold values are converted to logarithms, equal ratios between any two pairs of threshold values on the original measures correspond to equal intervals on the log-transformed values. All of the main effects remained significant: age $[F(1,30)=4.66, p<0.05]$, prime $[F(1,30)=516.75, p<0.0001]$, and context $[F(1,30)=437.43, p<0.0001]$ on the number of noise-vocoded log bands needed to achieve 50% word identification. The significant two-way interaction of prime and context also remained significant, $[F(1,30)=311.00, p<0.0001]$; however, the two-way interactions with age were not significant: age and prime $[F(1,30)=3.25, p>0.81]$, and age and context $[F(1,30)<1]$. The three-way interaction effect of age, prime, and context $[F(1,30)<1]$ was again not significant.

Abel, S. M., Krever, E. M., and Alberti, P. W. (**1990**). "Auditory detection, discrimination and speech processing in aging, noise-sensitive and hearing impaired listeners," Scand. Audiol. **19**, 43–54.

Alain, C., McDonald, K. L., Ostroff, J. M., and Schneider, B. (**2001**). "Age-related changes in detecting a mistuned harmonic," J. Acoust. Soc. Am. **109**, 2211–2216.

Andruski, J. E., Blumstein, S. E., and Burton, M. (**1994**). "The effect of subphonetic differences in lexical access," Cognition **52**, 163–187.

Bilger, R. C., Nuetzel, M. J., Rabinowitz, W. M., and Rzeczkowski, C. (**1984**). "Standardization of a test of speech perception in noise," J. Speech Hear. Res. **27**, 32–48.

Cabeza, R. (**2002**). "Hemispheric asymmetry reduction in older adults: The HAROLD model," Psychol. Aging **17**, 85–100.

CHABA (Committee on Hearing, Bioacoustics, and Biomechanics) Working Group on Speech Understanding and Aging, National Research Council.

(**1988**). "Speech understanding and aging," J. Acoust. Soc. Am. **83**, 850–805.
Connine, C. M., Titone, D., Deelman, T., and Blasko, D. (**1997**). "Similarity mapping in spoken word recognition," J. Mem. Lang. **37**, 463–480.
Craik, F. I. M. (**1983**). "On the transfer of information from temporary to permanent memory," Philos. Trans. R. Soc. London, Ser. B **302**, 341–359.
Craik, F. I. M. (**1986**). "A functional account of age differences in memory," in *Human Memory and Cognitive Capabilities, Mechanisms, and Performances*, edited by F. Klix, and H. Hagendorf (Elsevier Science, Amsterdam), pp. 499–522.
Cutler, A., Dahan, S., and van Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang Speech **40**, 141–201.
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. **134**, 222–241.
Dorman, M. F., and Loizou, P. C. (**1998**). "Identification of consonant and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal hearing subjects using simulations processors with two to nine channels," Ear Hear. **19**, 162–166.
Dubno, J. R., Ahlstrom, J. B., and Horwitz, A. R. (**2000**). "Use of context by young and aged adults with normal hearing," J. Acoust. Soc. Am. **107**, 538–546.
Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**, 2704–2710.
Elliott, L. L. (**1995**). "Verbal auditory closure and the Speech Perception in Noise (SPIN) Test," J. Speech Hear. Res. **38**, 1363–1376.
Ferreira, V. S., and Griffin, Z. M. (**2003**). "Phonological influences on lexical (mis)selection," Psychol. Sci. **14**, 86–90.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.
Gagné, J.-P., Rochette, A.-J., and Charest, M. (**2002**). "Auditory, visual, and audiovisual clear speech," Speech Commun. **37**, 213–230.
Gallacher, J. (**2004**). "Hearing, cognitive impairment and aging: A critical review," Reviews in Clinical Gerontology **14**, 199–209.
Goldinger, S. D., Luce, P. A., Pisoni, D. B., and Marcario, J. K. (**1992**). "Form-based priming in spoken word recognition; the roles of competition and bias," J. Exp. Psychol. Learn. Mem. Cogn. **18**, 1211–1238.
Gordon-Salant, S., and Fitzgibbons, P. J. (**1997**). "Selected cognitive factors and speech recognition performance among young and elderly listeners," J. Speech Hear. Res. **40**, 423–431.
Gow, D. W., and Gordon, P. C. (**1995**). "Lexical and prelexical influences on word segmentation: Evidence from priming," J. Exp. Psychol. Hum. Percept. Perform. **21**, 344–359.
Grady, C. L. (**2000**). "Functional brain imaging and age-related changes in cognition," Biol. Psychol. **54**, 259–281.
Greenberg, S. (**1996**). "Auditory processing of speech," in *Principles of Experimental Phonetics*, edited by N. J. Lass (Mobsy, St. Louis), pp. 362–407.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species - 29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence material with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.
Li, L., Daneman, M., Qi, J., and Schneider, B. A. (**2004**). "Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults?," J. Exp. Psychol. Hum. Percept. Perform. **30**, 1077–1091.
Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.
Mackersie, C. L., and Prida, T. L. (**2001**). "The role of sequential stream segregation and frequency selectivity in the perception of simultaneous sentences by listeners with sensorineural hearing loss," J. Speech Lang. Hear. Res. **44**, 19–28.
Marslen-Wilson, W. D., and Tyler, L. K. (**1980**). "The temporal structure of spoken language understanding," Cognition **8**, 1–71.
Martin, F. N., and Forbis, N. R. (**1978**). "The present status of audiometric practice: A follow-up study," ASHA **20**, 531–541.
Martin, F. N., and Pennington, C. D. (**1971**). "Current trends in audiometric practices," ASHA **13**, 671–677.
Martin, J. G. (**1979**). "Rhythmic and segmental perception are not independent," J. Acoust. Soc. Am. **65**, 1286–1297.
Meyer, D. E., and Schvaneveldt, R. W. (**1971**). "Facilitation in recognizing pairs of words; evidence of a dependence between retrieval operations," J. Exp. Psychol. **90**, 227–234.
Penrod, J. P. (**1994**). "Speech threshold and word recognition/discrimination testing," in *Handbook of Clinical Audiology*, 4th ed., edited by J. Katz (Williams and Wilkins, Baltimore, MD), pp. 147–164.
Perry, A. R., and Wingfield, A. (**1994**). "Contextual encoding by young and elderly adults as revealed by cued and free recall," Aging, Neuro., and Cogn. **1**, 120–139.
Pichora-Fuller, M. K. (**2003**). "Cognitive aging and auditory information processing." Int. J. Audiol., Suppl. 2, **42**, S26–S32.
Pichora-Fuller, M. K., Schneider, B. A., Benson, N., Hamstra, S., and Storzer, E. (**2006**). "Effect of age on gap detection in speech and nonspeech stimuli varying in marker duration and spectral symmetry," J. Acoust. Soc. Am. **119**, 1143–1155.
Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (**1995**). "How young and old adults listen to and remember speech in noise," J. Acoust. Soc. Am. **97**, 593–608.
Pichora-Fuller, M. K., and Singh, G. (**2006**). "Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiologic rehabilitation," Trends Amplif. **10**, 29–59.
Pichora-Fuller, M. K., and Souza, P. (**2003**). "Effects of aging on auditory processing of speech," Int. J. Audiol., Suppl. @, Suppl 2 **42**, S11–S16.
Pitt, M. A., and Samuel, A. G. (**1990**). "The use of rhythm in attending to speech," J. Exp. Psychol. Hum. Percept. Perform. **16**, 564–573.
Radeau, M., Besson, M., Fonteneau, E., and Castro, S. L. (**1998**). "Semantic, repetition and rime priming between spoken words: Behavioral and electrophysiological evidence," Biol. Psychol. **48**, 183–204.
Raven, J. C. (**1965**). *The Mill Hill Vocabulary Scale* (Lewis, London).
Reuter-Lorenz, P. A. (**2002**). "New visions of the aging mind and brain," Trends Cogn. Sci. **6**, 394–400.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.
Sanders, L. D., and Neville, H. J. (**2003a**). "An ERP study of continuous speech processing. II. Segmentation, semantics, syntax in non-native speakers," Brain Res. Cognit. Brain Res. **15**, 214–227.
Sanders, L. D., and Neville, H. J. (**2003b**). "An ERP study of continuous speech processing. I. Segmentation, semantics, syntax in native speakers," Brain Res. Cognit. Brain Res. **15**, 228–240.
Sanders, L. D., Newport, E. L., and Neville, H. J. (**2002**). "Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech," Nature (London) **5**, 700–703.
Schneider, B. A. (**2001**). "Sensation, cognition, and levels of processing in aging," in *Perspectives on Human Memory and Cognitive Aging: Essays in Honour of Fergus Craik*, edited by M. Naveh-Benjamin, M. Moscovitch, and H. L. Roediger III (Psychology Press, New York), pp. 298–314.
Schneider, B. A., and Pichora-Fuller, M. K. (**2000**). "Implications of perceptual deterioration for cognitive aging research." in *Handbook of Aging and Cognition*, 2nd, ed., edited by F. I. M. Craik and T. A. Salthouse (Erlbaum, Mahwah, NJ), pp. 155–220.
Schneider, B. A., and Pichora-Fuller, M. K. (**2001**). "Age-related changes in temporal processing: Implications for listening comprehension," Semin. Hear. **22**, 227–239.
Schneider, B. A., Pichora-Fuller, M. K., Kowalchuk, D., and Lamb, M. (**1994**). "Gap detection and the precedence effect in younger and older adults," J. Acoust. Soc. Am. **95**, 980–991.
Shannon, R. V. (**2002**). "The relative importance of amplitude, temporal and spectral cues for cochlear implant processor design," Am. J. of Audiol. **11**, 124–127.
Shannon, R. V., Zeng, F., Kamath, V., and Wygonski, J. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Sheldon, S., Pichora-Fuller, M. K., Schneider, B. A. (**2008**). "Effect of age, presentation method, and training on identification of noise-vocoded words," J. Acoust. Soc. Am. **123**, ■.
Snell, K. B., and Frisina, R. D. (**2000**). "Relationships among age-related differences in gap detection and word recognition," J. Acoust. Soc. Am. **107**, 1615–1626.

Sommers, M. S., and Danielson, S. M. (**1999**). "Inhibitory processes and spoken word recognition in young and old adults: The interaction of lexical competition and semantic context," Psychol. Aging **14**, 458–472.
Souza, P. E., and Boike, K. T. (**2006**). "Combining temporal-envelope cues across channels: Effects of age and hearing loss," J. Speech Lang. Hear. Res. **49**, 138–149.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Summers, V., and Leek, M. R. (**1998**). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res. **41**, 1294–1306.
van Tasell, D. J., Greenfield, D. G., Logemann, J. J., and Nelson, D. A. (**1992**). "Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants," J. Acoust. Soc. Am. **92**, 1247–1257.
Vongpaisal, T., and Pichora-Fuller, M. K. (**2007**). "Effect of age on F0 difference limen and concurrent vowel identification," J. Speech Lang. Hear. Res. **50**, 1139–1156.
Wingfield, A. (**1996**). "Cognitive factors in auditory performance: Context, speed of processing, and constraints of memory," J. Am. Acad. Audiol **7**, 175–182.
Wingfield, A., and Tun, P. A. (**2007**). "Cognitive supports and cognitive constraints on comprehension of spoken language," J. Am. Acad. Audiol **18**.
Wingfield, A., Tun, P. A., and McCoy, S. L. (**2005**). "Hearing loss in older adulthood: What it is and how it interacts with cognitive performance," Curr. Dir. Psychol. Sci. **14**, 144–148.

Spectral envelope sensitivity of musical instrument sounds

David Gunawan[a] and D. Sen
School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia

(Received 1 August 2006; revised 25 October 2007; accepted 2 November 2007)

It is well known that the spectral envelope is a perceptually salient attribute in musical instrument timbre perception. While a number of studies have explored discrimination thresholds for changes to the spectral envelope, the question of how sensitivity varies as a function of center frequency and bandwidth for musical instruments has yet to be addressed. In this paper a two-alternative forced-choice experiment was conducted to observe perceptual sensitivity to modifications made on trumpet, clarinet and viola sounds. The experiment involved attenuating 14 frequency bands for each instrument in order to determine discrimination thresholds as a function of center frequency and bandwidth. The results indicate that perceptual sensitivity is governed by the first few harmonics and sensitivity does not improve when extending the bandwidth any higher. However, sensitivity was found to decrease if changes were made only to the higher frequencies and continued to decrease as the distorted bandwidth was widened. The results are analyzed and discussed with respect to two other spectral envelope discrimination studies in the literature as well as what is predicted from a psychoacoustic model.
 [DOI: 10.1121/1.2817339]

PACS number(s): 43.75.Cd, 43.66.Jh [DD] Pages: 500–506

I. INTRODUCTION

Timbre research has found the spectral envelope to be a salient attribute.[1–4] In musical acoustics, the spectral envelope can be described in the frequency domain as an interpolation between the amplitudes of the sinusoidal components of a signal.[1,2] Sufficient modification of the spectral envelope of an instrument produces a change in perception of that instrument's timbre, and in some cases significant modification can lead to the instrument sounding similar to a different instrument. Grey's[3] work in developing perceptual spaces of timbre using multidimensional scaling led to the identification of the spectral energy distribution being one of the important dimensions of timbre. More recently, McAdams *et al.*[1] have identified the spectral envelope shape as being the most salient parameter in timbre discrimination when performing various simplifications to instrument spectrotemporal parameters. Caclin *et al.*[4] also verified the spectrum's importance in their confirmatory study using synthetic tones.

A thorough understanding of timbre therefore requires knowledge of how much spectral deviation is required before there is a perceptible change in timbre. The primary objectives of this paper are to analyze the discrimination thresholds of spectral change for various instruments and observe the sensitivity to change as a function of center frequency and bandwidth. We have chosen to study three instruments (trumpet, clarinet and viola) which represent the brass, woodwind and string families. While previous studies have analyzed sensitivity to musical instrument spectral envelopes,[5–7] none of them has investigated the sensitivity as a function of center frequency and bandwidth. Other studies have studied sensitivity as a function of frequency but not in the context of musical instruments. Due to the complex nature of musical instrument signals, the results of such studies are very difficult to translate into a musical instrument context.

Early studies by Plomp[5] investigated perceptual sensitivity to spectral change for static musical instrument and vowel spectra and found that spectral differences were good predictors of differences in timbre. Horner *et al.*[6] extended this work by observing instrument discrimination for random alterations to time-varying instrument spectra. The spectra of instruments were modified by various error levels (8%, 16%, 24%, 32% and 48%) by randomly altering the amplitudes of individual sinusoids. They observed that discrimination was very good for 32% and 48% error levels, moderate for the 16% and 24% error levels and poor for the 8% error levels. However, the spectral modifications were performed randomly over time and frequency and did not account for the varying sensitivities that may be apparent as a function of frequency.

Similar work has been done in the field of speech processing particularly for the purposes of speech coding. Paliwal[7] divided speech signals into frames of approximately 20 ms and observed that the average spectral distortion difference limen for perceptual indistinguishability is 1 dB, ensuring that no frames have average spectral distortions greater than 4 dB and less than 2% of the frames have average spectral distortions between 2 and 4 dB. These results have been used extensively in the design of vector quantizers for speech coders. However, once again, these observations are based on the entire spectrum and do not reveal sensitivity as a function of frequency.

[a]Author to whom correspondence should be addressed. Electronic mail: d.gunawan@student.unsw.edu.au

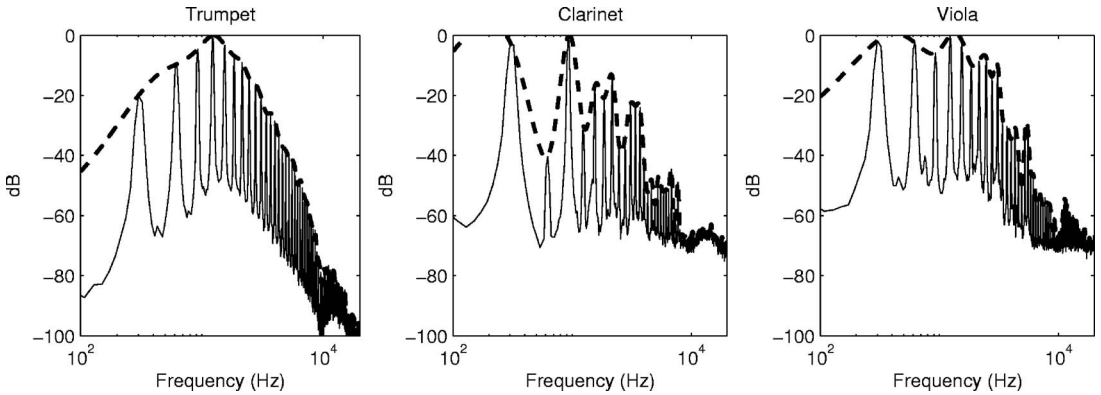


FIG. 1. Average musical instrument spectra (solid lines). Dashed lines illustrate the spectral envelope calculated using the SEEVOC method. Averages were taken over 32 frames, each of 2048 samples.

Auditory profile analysis is a field concerned with observations on the discrimination thresholds of spectrally modified sounds. Green[8] performed an analysis of discrimination thresholds for 21 component complexes; however, like most auditory profile analysis experiments, the stimuli considered were sums of sinusoids that were spectrally flat and with log-spaced frequencies. Thus, the stimuli were very different from realistic musical instrument spectra which are harmonically spaced and nonuniform. The results are therefore difficult to extrapolate to a musical instrument context.

In the present study, we aim to investigate the discrimination thresholds for changes to musical instrument spectral envelopes. Previous studies have often assumed that spectral envelope sensitivity is unchanged as a function of frequency,[3,6] however we hypothesize that there will be variations in the discrimination thresholds for modifications made as a function of center frequency and bandwidth. The experimental results are compared to a number of spectral distortion measures and then are discussed with reference to other experimental findings as well as predictions from a psychoacoustic model.

II. EXPERIMENTAL METHOD

In order to investigate the sensitivity to the spectral envelope, we endeavored to keep all other physical parameters constant. These included fundamental frequency, level and duration—the details of which are described in the following section. With the intent of understanding how sensitivity varies as a function of center frequency and bandwidth, each stimulus was modified by attenuating a band of frequencies by various amounts. Subjective tests were conducted to determine discrimination thresholds for different instruments.

A. Stimuli

Three musical instrument sounds were selected for analysis. Samples of trumpet, clarinet, and viola taken from a University of Iowa website[9] were used. The samples were chosen for their representation of three different instrument families—brass, woodwind and string. The sounds were recorded using 16 bits, and a 44 100 Hz sampling rate, and each sound was played at a pitch of E^b4, corresponding to a fundamental frequency of approximately 311.1 Hz—a frequency within the normal playing range of these instruments and commonly used in timbre experiments for this reason.[1,6] Average spectra of the three sounds are illustrated in Fig. 1. The duration of each sound was standardized to 1.5 s using a 100 ms half-Hanning window to taper the offsets. The onsets of each sample were left unmodified. The level of each sound was adjusted by a gain factor such that five independent subjects perceived them to be of equal loudness.

The three sounds were then each modified such that various bands across the frequency spectrum were attenuated by various amounts. The stimuli presentation was controlled using MATLAB on an Intel PC with an RME Multiface sound card. Each of the stimuli was presented monaurally at an average level of approximately 65 dB sound pressure level through Beyerdynamic DT770pro headphones in a sound-insulated (Acoustic Systems) anechoic chamber.

B. Stimuli modification

The system illustrated in Fig. 2 was employed to make the relevant modifications. As the stimuli are time varying in nature, time-invariant filters were employed to preserve the time resolution. Each stimulus was passed through a zero-phase bandpass filter and the output of the filter was then attenuated and subtracted from the original stimulus. Using 14 zero-phase filters of differing center frequencies and bandwidths, we compiled a set of stimuli where the output was the original signal with a certain frequency band attenuated. Note that the modified stimuli were not equalized for loudness as this would produce more audible changes than not equalizing the loudness.

The zero-phase filters were designed by taking 256-tap linear-phase bandpass filters (designed by the window method based on a Hamming window) and advancing the output signal by the group delays of the filters. Since the human auditory system has an nonlinear frequency resolution[10] which can be approximated by a logarithmic-

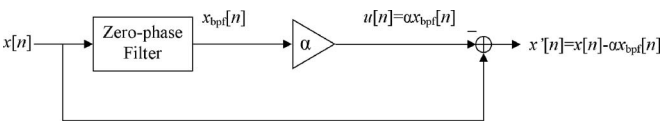


FIG. 2. System used for stimuli modification.

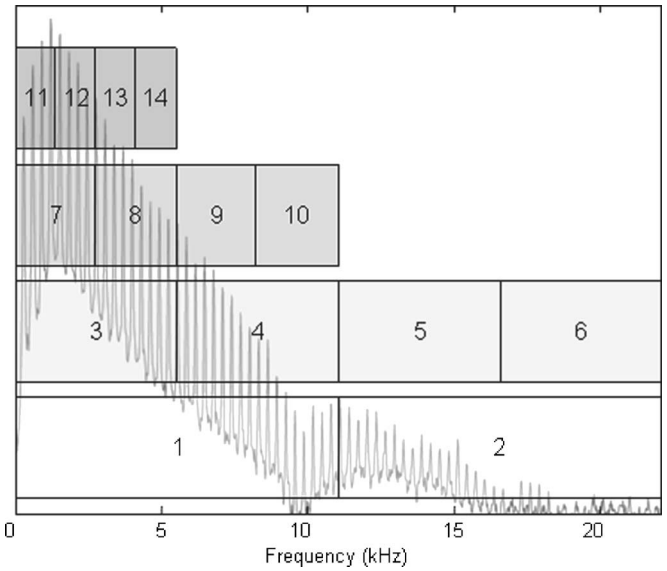


FIG. 3. Bandwidths of the 14 zero-phase filters with trumpet spectrum overlaid. (The rectangular boxes only indicate bandwidth and should not be associated with the *y* axis which indicates the spectral magnitude of the trumpet.)

like function, 14 logarithmically spaced filters were used as illustrated in Fig. 3. More low frequency filters with narrower bandwidths were selected to analyze the lower frequencies with higher resolution in similar fashion to the auditory system. The filters are labeled 1–14. As an example, the magnitude response of filter 4, which has a bandwidth of 5512.5 Hz and a center frequency of 8268.75 Hz, is plotted in Fig. 4. An attenuated viola spectrum using this filter is illustrated in Fig. 5. Preliminary tests using equivalent rectangular bandwidth (ERB) gammatone filters similar to those in Ref. 11 resulted in measurements being dependent on harmonic content rather than the spectral envelope. The ERB gammatone filters had bandwidths that were too fine for spectral envelope analysis and thus wider logarithmically spaced bandwidth filters were used to observe the effects of spectral envelope modification.

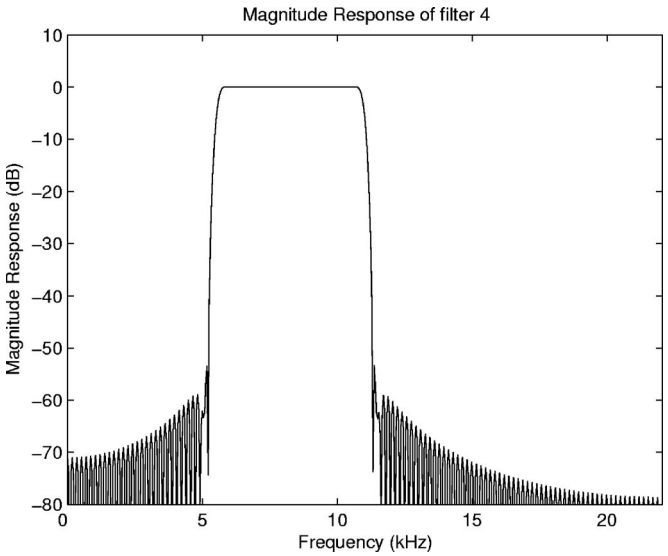


FIG. 4. Magnitude response filter 4.

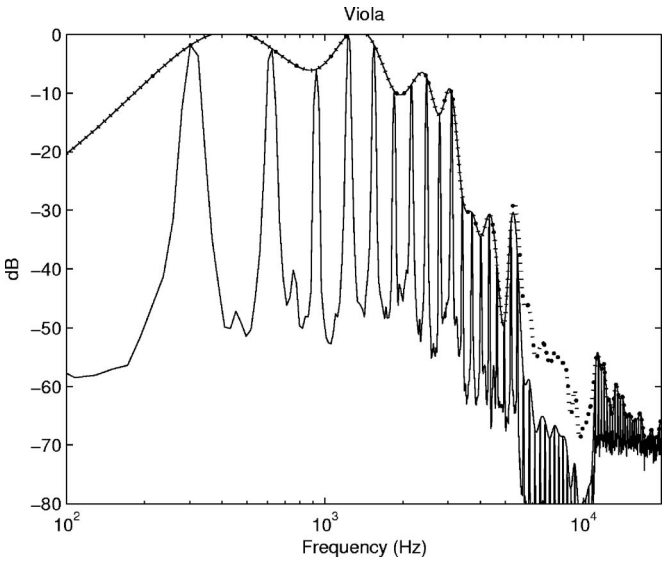


FIG. 5. Attentuation of a band of frequencies by filter 4 on the viola. The spectrum and spectral envelope of the attenuated signal are illustrated by the solid line while the dashed line illustrates the original unattentuated viola spectrum.

C. Participants

Five listeners aged between 20 and 26 years participated in the experiment. Four participants were male and one was female and all were tested and found to have normal hearing. Three of the participants had musical training with experience ranging between 5 and 10 years.

D. Procedure

A two-alternative forced-choice Reference AB, 1-up 2-down paradigm[12] was used for all our experimentation. For each trial, the participant heard three sounds: the reference sound (original, unfiltered) followed by two other sounds—one of which was filtered and the other which was the same as the reference. The order of presentation of the two latter sounds was independently randomized for each trial and 300 ms silence periods separated the presentation of each sound. For each trial, the participant was prompted with "Which sound has a different timbre to the reference?" and had to respond by clicking buttons marked A and B on the screen. Once a response was submitted, feedback was provided to the participant in the form of "Correct" or "Incorrect."

The first trial presented for each center frequency was always with the most attenuation and the attenuation was incrementally decreased to include more of the contents of the band. The attenuation step sizes changed from 4 to 2 dB and finally to 0.5 dB. The last three reversals were averaged to estimate the discrimination threshold. Listeners were trained for 15 min to familiarize themselves with the task prior to the experiment. Thresholds for the 14 filtered bands were recorded in a single 50 min block per instrument.

III. RESULTS

The results from the experiment were analyzed in four different ways. The first was a measurement of sensitivity

which analyzed the individual band attenuations (BA). Following that, we computed two different distortion measures as employed in Refs. 6 and 7 to compare the data to previous studies. Finally in Sec. IV, the results are discussed with what is predicted by a psychoacoustic difference limen model.

A. Band attenuation (BA)

If a listener is more sensitive to a change in a signal parameter, then a smaller change of that parameter is needed to hear the effect of the change. We define $x[n]$ to be the original discrete pressure stimulus (where n is the discrete sample index), $x'[n]$ to be the modified stimulus (as illustrated in Fig. 2), $x'^{*}[n]$ to be the just-noticeable modified stimulus and $(1-\alpha^{*})$ to be the just-noticeable attenuation that produces $x'^{*}[n]$. If only a small change in the energy of a band is required before it is detected, sensitivity is considered to be high. This is reflected by a small α, which in turn means a high attenuation $(1-\alpha)$. If we define the band attenuation (BA) to be $(1-\alpha)$ (expressed in decibels), then sensitivity is simply proportional to the BA and can be used as a measure of sensitivity (Eq. (1)).

$$BA = 20\log_{10}(1-\alpha^{*})(dB), \tag{1}$$

where $(1-\alpha^{*})$ is the just-noticeable attenuation of a particular band (in linear units).

The BA results are shown in Fig. 6, clearly indicating that there are obvious differences in the sensitivities for different bandwidths and center frequencies. Qualitatively, it can be observed that smaller changes at lower frequencies consistently trigger a perceptual change in timbre compared to changes at higher frequencies. The lower frequencies are therefore more sensitive than higher frequencies.

Another important observation is that the bands that include the first few harmonics also tend to set the upper bound for sensitivity (filters 1, 3, 7, 11), for all other bands have lower sensitivity than these. This implies that the maximum sensitivity can be estimated from the sensitivities of the lower frequencies and no other region of the spectral envelope will have higher sensitivity.

B. Distortion measures

The results can also be expressed in terms of the amount of modification required to perceive a change. Here we compare our results to two other studies from Refs. 6 and 7, which employ two different distortion measures.

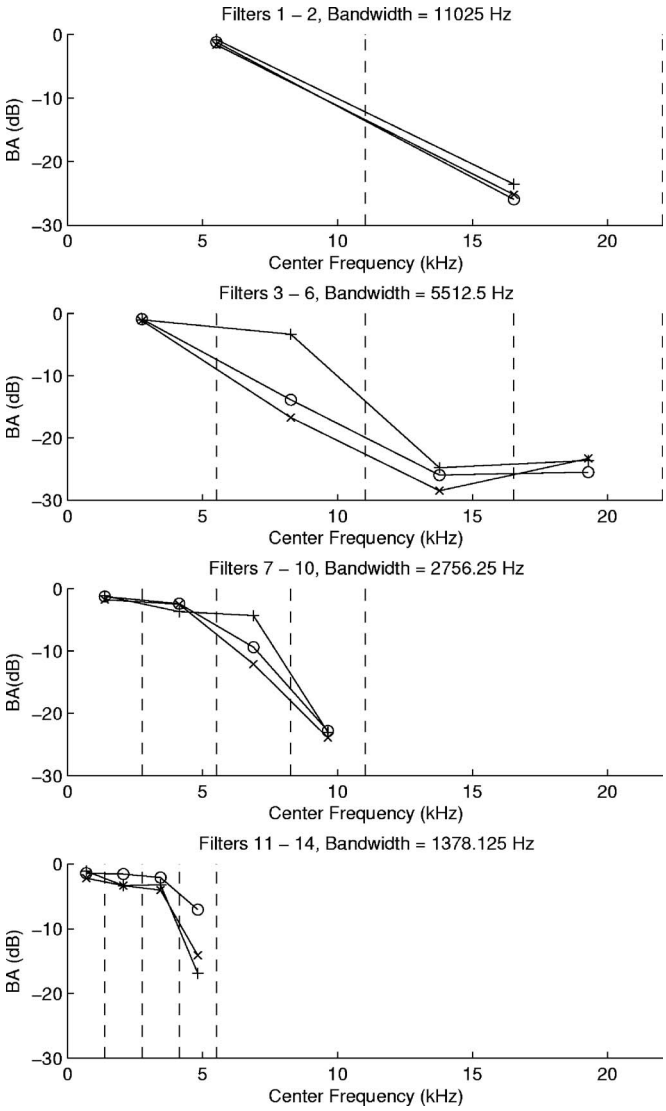


FIG. 6. Band attenuation plots for the trumpet (○), clarinet (+), viola (×) positioned at the center frequencies of filters 1–14. Dashed lines indicate the filter bandwidths.

1. Error level

The error level (EL) distortion metric is defined as

$$EL = \sqrt{\frac{\sum_{n=1}^{N} u[n]^2}{\sum_{n=1}^{N} x[n]^2}} \times 100\% \tag{2}$$

$$= \alpha^{*} \sqrt{\frac{\sum_{n=1}^{N} x_{bpf}[n]^2}{\sum_{n=1}^{N} x[n]^2}} \times 100\%, \tag{3}$$

where $x[n]$ is the original stimulus, $u[n]=x[n]-x'^{*}[n]$ $=\alpha^{*}x_{bpf}[n]$ is the minimum difference required to observe a change in a band, n is the sample index and N is the total number of samples.

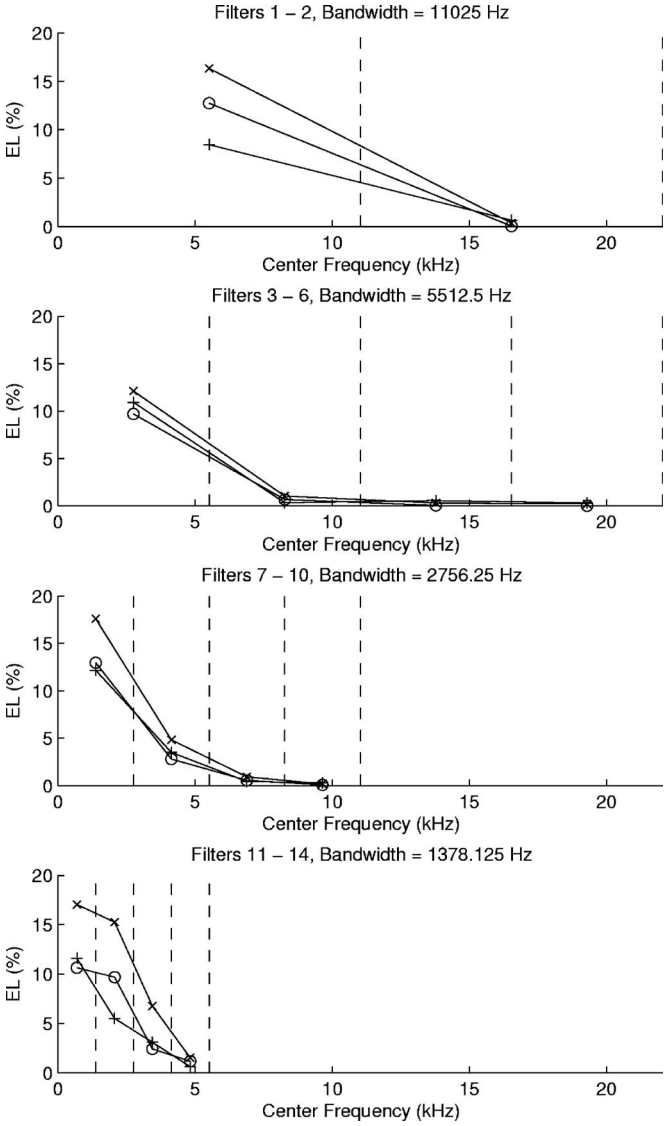


FIG. 7. Error level plots for the trumpet (○), clarinet (+), viola (×) positioned at the center frequencies of filters 1–14. Dashed lines indicate the filter bandwidths.

EL is a measure of the just-noticeable change as a percentage of the entire signal, and for fixed bandwidth and center frequency, EL varies linearly with α^*.

In a related study by Horner *et al.*,[6] the spectra of musical instrument samples were altered randomly and the spectral deviation was measured by observing average error levels as a percentage of the deviation from the original. Alteration of the harmonic spectra was performed by multiplying each amplitude of the kth harmonic at time m, $A_k[m]$, with a randomly selected scalar r_k:

$$A_k'[m] = r_k A_k[m]. \tag{4}$$

The scalars $\{r_k\}$ were selected uniformly in the range $[1-2\epsilon,\ 1+2\epsilon]$, where ϵ denotes the error level. While the methods in the study of Horner *et al.* differ from the methods employed in this study, it is useful to compare the results as they examine the sensitivity to amplitude modifications to the frequencies of harmonic musical instrument samples.

ELs from this study are shown in Fig. 7, and correspond to 70.7% discrimination on the psychometric curve.[12] These results indicate that the discrimination for the bands with low center frequencies (containing most of the signal) is around 13%. This agrees with the results in Ref. 6 where it was found that discrimination was approximately 16% at the 75% discrimination level. While the analyses for the bands with low center frequencies concur with Ref. 6, the additional analysis for various bandwidths in this study reveals that ELs vary for different bandwidths and center frequencies. Bands with higher center frequencies and with wider bandwidths can only undergo smaller changes relative to the entire signal before discrimination.

2. Spectral distortion

The spectral envelope analysis by Paliwal[7] employed a spectral distortion error metric to define the maximum error before the altered spectrum could be distinguished from the original. The spectral envelopes for each frame were calculated by the spectral envelope estimation vocode (SEEVOC) method[2] which proceeds as follows: A periodic signal is divided into frames and the discrete fourier transform (DFT) of each frame is calculated. Using the $F0$ of the signal, the harmonic peaks are located and a smooth curve is fitted through them. In this analysis, 1024 frequency points for the DFT were chosen and cubic interpolation was chosen to join the harmonic peaks.

The spectral distortion (defined for a given frame as the root-mean-square difference between the original log-power spectral envelope and the modified log-power spectral envelope), is averaged over a large number of frames to give the average spectral distortion

$$SD = \frac{1}{N_k}\sum_{k=1}^{N_k}\sqrt{\frac{1}{M}\sum_{\omega=0}^{M-1}(s(\omega)-s'^{*}(\omega))^2}, \tag{5}$$

where N_k is the number of frames, k is the frame number, M is the number of frequency points, $s(\omega)$ is the original log-power spectral envelope, $s'^{*}(\omega)$ is the just noticeable modified log-power spectral envelope and ω is the DFT frequency number.

Figure 8 illustrates the results with respect to spectral distortion. The spectral distortion for these envelopes was then calculated by Eq. (5) yielding a spectral distortion measure that was averaged over a number of frames. Interestingly, the results for the bands with lower center frequencies concur with the 1 dB value of distortion found for the spectral transparency of speech.[7] The present analysis sheds further insight into the spectral distortions allowable for various bandwidth modifications. A significantly larger amount of spectral distortion of up to 17 dB is allowable before discrimination occurs for bands with higher center frequency.

C. Difference limen model comparison

It would be of interest to compare the subjective experimental data to that predicted by a psychoacoustic difference limen model. A simultaneous auditory masking model, which approximates threshold levels as a function of signal level

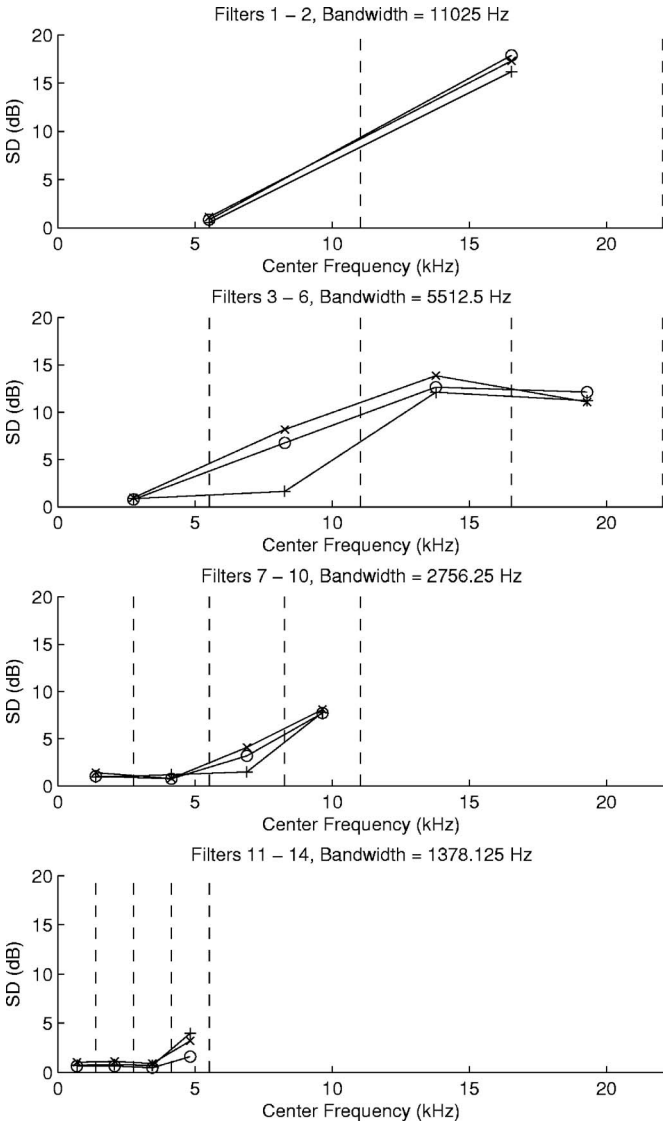


FIG. 8. Spectral distortion (SD) plots for the trumpet (○), clarinet (+), viola (×) positioned at the center frequencies of filters 1–14. Dashed lines indicate the filter bandwidths.

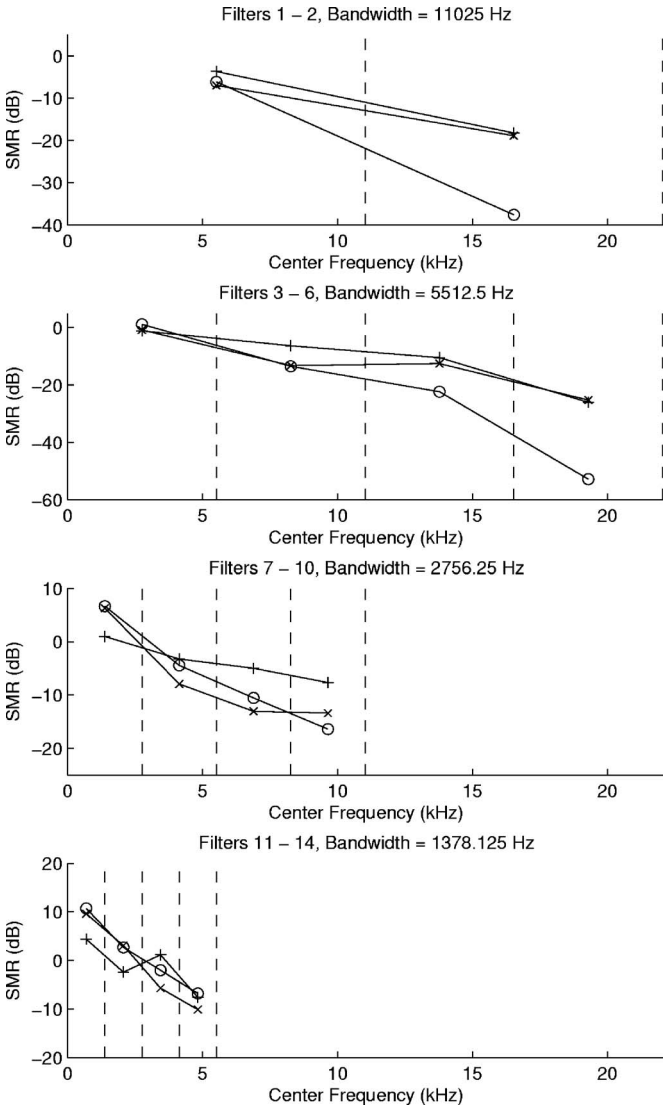


FIG. 9. Signal-to-masking ratio (SMR) plots for the trumpet (○), clarinet (+), viola (×) positioned at the center frequencies of filters 1–14. Dashed lines indicate the filter bandwidths.

and frequency content, was used in our study. Given the complex nature of the stimuli used in this study, simple frequency or level difference limen models are not appropriate. Fundamentals of our masking model are entrenched in the work by Zwicker[13] and Terhardt[14] (for analyzing partial loudness and virtual pitch, respectively). This work was later formalized by Moore[15] and essentially strives to emulate functions of the peripheral auditory organs. The model begins with critical band decomposition of the signal,[16] followed by a triangular spreading function[17] and finally accounts for the absolute threshold of hearing.[18] Further refinement is achieved by accounting for the asymmetry of masking between tonal and noise-like stimuli.[19,20]

Masking thresholds were calculated for each of the three original stimuli using overlapping frames of 512 samples. For each stimulus, the masking thresholds were then averaged over all the frames, and then for each band (see Fig. 3) the average signal-to-masking ratio (SMR) was calculated to represent the band's SMR. The SMR describes the relationship between the stimuli and the minimum distortion that is perceivable. A high SMR indicates that only a small deviation from the original signal can be tolerated, while a low SMR suggests the opposite. SMR can thus be viewed as an indication of sensitivity.

Figure 9 illustrates the average SMR for each of the instruments. The results clearly show that the bands with lower center frequency are more sensitive to change than the bands with higher center frequency and therefore agree with the BA results found in Fig. 6. The results also show that the lower bands indeed dominate the sensitivity and the higher bands become increasingly more sensitive as the bandwidth narrows. However, the SMR model is not an extremely accurate predictor of sensitivity in the lower bands, for while the experimental findings suggest a more consistent sensitivity as the bandwidth narrows, the SMR model clearly suggests an increase in sensitivity as the bandwidth narrows.

IV. DISCUSSION AND CONCLUSION

The results from the experiment highlight a number of important attributes about perceptual sensitivity to the spec-

tral envelope. The BA plot (Fig. 6) clearly shows that any assumption of sensitivity being equal over center frequency and bandwidth is inaccurate. The spectral envelope's sensitivity to change varies considerably over center frequency and bandwidth, and further studies that manipulate the spectral envelope of an instrument ought to consider such effects.

The experiment in this paper highlights that there are clear discrepancies between the amount of distortion tolerable over frequency, but also accentuates the importance of clarifying the reference for the measure of distortion. This can be seen by comparing the results from Figs. 7 and 8. What initially seems contradictory in fact proves to be complementary. Figure 7 shows that the error required to discriminate changes for higher band decompositions is much smaller than lower band decompositions. However, this is relative to the entire signal energy. Figure 8, on the other hand, gives the spectral distortion in dB. A greater level of distortion (in dB) is required for discrimination of the higher bands because the lower frequency components are extremely effective in masking higher frequency content. Thus the auditory system is able to tolerate noise at higher frequencies when there is significant lower frequency content. The converse is not true for two reasons. First there is not enough high frequency content to mask lower frequencies and second higher frequency components are not effective maskers of lower frequency content.

The studies in Refs. 6 and 7 sought to quantify how much spectral envelope modification could be made before a change in timbre was observed. The error level for 75% discrimination in Ref. 6 was approximately 16% and this result is similar to the 13% error level at 70.7% discrimination for bands with low center frequencies found in this experiment. The spectral distortion threshold result of 1 dB found in Ref. 7 is a criterion that is frequently employed in the design of speech vector quantizers. Interestingly in the context of musical instruments, this 1 dB result also aligns well with the 1 dB spectral distortion threshold for bands containing the lower harmonics as calculated in this paper. Thus, the results in this paper agree with previous results for bands with low center frequency, but shed further light into the nature of discriminability when considering change to only a certain bandwidth.

The comparison with a masking analysis model illustrates that our sensitivity measurements generally agreed with psychoacoustic masking theory. Despite some differences, particularly in the bands with lower center frequency, Figs. 6 and 9 seem to have the same fundamental appearance and would therefore suggest that sensitivity to the spectral envelope can be crudely approximated using the average SMR value for the band in question. However, the experimental results suggest a more consistent sensitivity of the lower bands than the masking model infers.

In summary, distortion of different portions of musical instrument spectral envelopes using band attenuation with different bandwidths and center frequencies results in different discrimination levels. This implies that sensitivity varies as a function of frequency and bandwidth. Sensitivity is maximum for the lower frequencies and decreases as the center frequency moves higher. For bands with lower center frequency, the sensitivity remains approximately the same while the bands with higher center frequency consistently decrease in sensitivity. Thus, from a perceptual standpoint, sensitivity has an upper bound governed by the first few harmonics and our sensitivity does not improve when extending the bandwidth any higher. However, if changes are made only to the higher harmonics, then our sensitivity is decreased and reduces further as the bandwidth distorted is widened.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable contributions and comments.

[1]S. McAdams, J. W. Beauchamp, and S. Meneguzzi, "Discrimination of musical instrument sounds resynthezied with simplified spectrotemporal parameters," J. Acoust. Soc. Am. **105**, 882–897 (1999).
[2]D. B. Paul, "The spectral envelope estimation vocoder," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-29**, 786–794 (1981).
[3]J. M. Grey, "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**, 1270–1277 (1977).
[4]A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," J. Acoust. Soc. Am. **118**, 471–482 (2005).
[5]R. Plomp, "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden, 1970).
[6]A. Horner, J. Beauchamp, and R. So, "Detection of random alterations to time-varying musical instrument spectra," J. Acoust. Soc. Am. **116**, 1800–1810 (2004).
[7]K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. Speech Audio Process. **1**, 3–14 (1993).
[8]D. M. Green and C. R. Mason, "Auditory profile analysis: Frequency, phase, and weber's law," J. Acoust. Soc. Am. **77**, 1155–1161 (1985).
[9]"The University of Iowa musical instrument samples," http://theremin.music.uiowa.edu/ (last viewed 3/21/2007).
[10]B. C. J. Moore, *Introduction to the Psychology of Hearing* (Macmillan, London, 1977).
[11]M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Technical Report No. 35, Apple Computer (1993).
[12]H. Levitt, "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477 (1971).
[13]E. Zwicker and B. Scharf, "A model of loudness summation," Psychol. Rev. **72**, 3–26 (1965).
[14]E. Terhardt, "Calculating virtual pitch," Hear. Res. **1**, 155–182 (1979).
[15]B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," Acustica **82**, 335–345 (1996).
[16]E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer-Verlag, Berlin, 1990).
[17]B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio Eng. Soc. **45**, 224–240 (1997).
[18]H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," J. Acoust. Soc. Am. **5**, 82–108 (1933).
[19]R. P. Hellman, "Asymmetry of masking between noise and tone," Percept. Psychophys. **11**, 241–246 (1972).
[20]J. L. Hall, "Asymmetry of masking revisited: Generalization of masker and probe bandwidth," J. Acoust. Soc. Am. **101**, 1023–1033 (1997).

Measurements and predictions of hooded crow (*Corvus corone cornix*) call propagation over open field habitats

Kenneth Kragh Jensen[a)]
School of Medicine, Jordan Hall, Indiana University, 1001 East Third Street, Bloomington, Indiana 47405

Ole Næsbye Larsen[b)]
University of Southern Denmark, Institute of Biology, Campusvej 55, DK-5230 Odense M, Denmark

Keith Attenborough[c)]
Department of Engineering, University of Hull, Kingston Upon Hull, HU6 7RX, United Kingdom

(Received 29 May 2007; revised 29 October 2007; accepted 5 November 2007)

In a study of hooded crow communication over open fields an excellent correspondence is found between the attenuation spectra predicted by a "turbulence-modified ground effect plus atmospheric absorption" model, and crow call attenuation data. Sound propagation predictions and background noise measurements are used to predict an optimal frequency range for communication ("sound communication window") from an average of crow call spectra predicted for every possible combination of the sender/receiver separations 300, 600, 900, and 1200 m and heights 3,6,9 m thereby creating a matrix assumed relevant to crow interterritorial communication. These predictions indicate an optimal frequency range for sound communication between 500 Hz and 2 kHz. Since this corresponds to the frequency range in which crow calls have their main energy and crow hearing in noise is particularly sensitive, it suggests a specific adaptation to the ground effect. Sound propagation predictions, together with background noise measurements and hearing data, are used to estimate the radius of the hooded crow active space. This is found to be roughly 1 km in moderately windy conditions. It is concluded that the propagation modeling of the sort introduced here could be used for assessing the impact of human noise on animal communication. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2817363]

PACS number(s): 43.80.Lb, 43.66.Gf, 43.28.Gq [JAS] Pages: 507–518

I. INTRODUCTION

Many studies of the propagation of animal vocalizations and the implications of propagation factors for communication and behavioral ecology have been based more or less on ad hoc methods for quantifying sound distortion and attenuation during propagation appropriate in the given study (Gish and Morton, 1981; Dabelsteen *et al.*, 1993; Fotheringham and Ratcliffe, 1995; Kime *et al.*, 2000). Consequently they offer limited opportunities for generalization. On the other hand, modeling of outdoor sound propagation has been of considerable interest for predicting noise levels from man-made sources close to the ground (Daigle *et al.*, 1983; Attenborough *et al.*, 1995; Embleton, 1996; Li *et al.*, 1998; Taherzadeh *et al.*, 1998; Attenborough *et al.*, 2000; Attenborough, 2002; Zaporozhets *et al.*, 2003). Many of the models have been shown to predict the resulting overall sound level and spectral profile of propagated sounds rather well when properties of ground impedance are known or can be assumed (Attenborough, 2002). Since the propagation models can be used for any sound source they could be of significant importance in bioacoustics (Michelsen and Larsen, 1983). Relevant tasks could include evaluation of field measurements, predicting communication conditions over multiple transmission transects too time consuming to perform experimentally, or estimating communication conditions when measurements in the field are impossible.

Here the usefulness of an established sound propagation model as a tool in bioacoustics is tested and demonstrated by performing a study on hooded crow communication. Crows live in open field habitats and have large territories of the order of 500–700 m in diameter between which they communicate (Coombs, 1978; Cramp and Perrins, 1994; Madge and Burn, 1994). Synthetic sounds and natural crow calls have been transmitted over open field habitats up to 320 m. The frequency dependent attenuation data have been compared with predictions. There is generally excellent correspondence between measurements and model fits. On this basis propagation modeling is taken further as a tool in estimating the maximum communication distance (active space) of crows, which is found to be roughly 1 km in diameter at their normal calling height. Predictions and data show that acoustic communication is greatly impaired when crows are on the ground. Finally, the best averaged transmitted frequency range is predicted by considering multiple sender/receiver geometries relevant to crow territorial communication. The resulting optimal communication range of frequencies significantly overlaps the bandwidth corresponding to low critical ratios (i.e., optimal hearing in noise) for hooded crows. This suggests that crow hearing has evolved

[a)]Author to whom correspondence should be addressed. Electronic mail: kkj@jensenkk.net
[b)]Electronic mail: onl@biology.sdu.dk
[c)]Electronic mail: k.attenborough@hull.ac.uk

to match the ground effect and typical conditions of sound propagation.

II. MATERIAL AND METHODS

A. Locations and meteorological conditions

The experiments were conducted during September 2004 on farmland fields in the vicinity of the University of Southern Denmark, Odense, Denmark, where hooded crows (*Corvus corone cornix*) are very common.

1. Location 1

Location 1 for the sound transmission measurements was a stubble field with dry soil. The 10 to 15-cm long stubble was lying flat on a rough ground consisting of 10 to 20-cm-high humps with 2–4 cm dispersed lumps of dry soil. The nearest reflecting object besides the ground was a small forest of tall deciduous trees 50–70 m away from the loudspeaker and orientated approximately at 90° in relation to the transmission line. The experiments started at 16:00 hours and ended at 23:30 hours. The wind direction was stable at approximate 90° angle relative to the transmission transects. The conditions were sunny. Wind speed, temperature, and humidity were measured approximately 3 m above ground with a hot-wire-based multiprobe electronic device (Testo term, type 452). The average wind speed during the day was 2 to 3 m/s with gusts of up to 5 to 6 m/s. At 23:30 hours the wind speed was 1.5–2.5 m/s. The temperature rose to between 21.5 and 23.5 °C during the day and steadily decreased after sunset to 15.5 °C at 23:30 hours. The relative humidity was between 45% and 75% during the measurements.

2. Location 2

Location 2 for the sound transmission measurements was in the middle of a ploughed field. Sound transmission measurements were made along a 15-m-wide stripe of mown fallow grassland consisting of 10 to 20-cm-long stems of mostly soft green grass, somewhat compressed from the mowing. The soil was moist and the ground was rough (10 to 20-cm-high humps). The nearest reflecting objects besides the ground was a line of approximately 10 m high, foliaged trees 20–30 m behind the loudspeaker. The experiments started at 13:00 hours and ended at 18:00 hours. It was a sunny day and the wind direction was stable and at an approximate 90° angle relative to the transmission transects. The wind speed was between 1 and 2 m/s with gusts of up to 4 m/s. The temperature varied from 16 to 19 °C. The relative humidity was between 70% and 75% throughout the day.

3. Background noise measurements

Natural background noise levels were measured at three different locations minimally affected by human activities. The measurements were performed at typical crow locations during summertime (August) in open field crop covered habitats with wind speed <1 m/s and a temperature around 20 °C. Some distant small songbird and dove vocalizations could be heard as well as very distant traffic noise. Since the noise measurements at the two locations on the same day were very similar they have been averaged.

The third location was also at a typical crow location but the measurements were performed during winter (February) on an open snow covered field. No traffic noise could be heard. The ground was frozen and covered by between 5 and 10 cm snow (with a 1–1.5 cm hard crust on top). The wind speed was between 5 and 6 m/s and the temperature just below 0 °C.

B. Experimental setup

1. Propagation measurements

A Marantz Professional recorder (PMD670/W1B) connected to a Nagra Kudelski DSM-Monitor was used to play the test sounds. Transmitted test sounds were re-recorded onto two channels simultaneously by two wind shielded Brüel & Kjær $\frac{1}{2}$ in. microphones (type 4190; Pre-amplifier type 2669), a Brüel & Kjær microphone amplifier (type 5935), and a second Marantz Professional recorder (PMD670/W1B). The recording system was calibrated with a sound level calibrator (Brüel & Kjær, type 4230). At both locations, the sounds were played across transects of 40, 80, 160, and 320 m. The centers of the monitor and of the receiving microphone were placed at 2.8 m height on two 5-cm-diam aluminum telescope poles secured by three stabilizing ropes. In addition, at Location 2 the loudspeaker and receiving microphone were placed at 0.28 m height corresponding approximately to the position of the head of a crow walking on the ground.

2. Measurements of background noise

To determine the maximum communication distance (see Sec. II F) an estimate of the natural background noise level was made from measurements at the three different locations mentioned in Sec. II A 3. In the two August locations the noise was measured in third octave filters using a Brüel & Kjær sound level meter (type 2250) with linear weighting, and the corresponding pressure spectrum densities were calculated. The sound level meter was positioned as for the August measurements on a tripod 1.4 m above the ground and the microphone, which had a windscreen attached, was pointed horizontally. At the February location, the noise was recorded using a Brüel & Kjær sound level meter (type 2235) with "random incidence" and "linear weighting" connected to a Marantz Professional recorder (PMD670/W1B). The system was calibrated by a sound level calibrator (Brüel & Kjær, type 4230) at 94 dB SPL (rms re 20 μPa). The sound level meter was positioned on a tripod 1.4 m above the ground and the microphone, which had a windscreen attached, was pointed horizontally. The spectrum level of the background noise was calculated in SASLAB PRO (version 4.39) over a sequence of 40 s where no nearby identifiable or locatable noise sources of any kind could be heard (including rustling leaves, bird calls, etc.). The wind unavoidably generated some noise at the microphone. Consequently the true background noise level is somewhat overestimated in the lower frequencies.

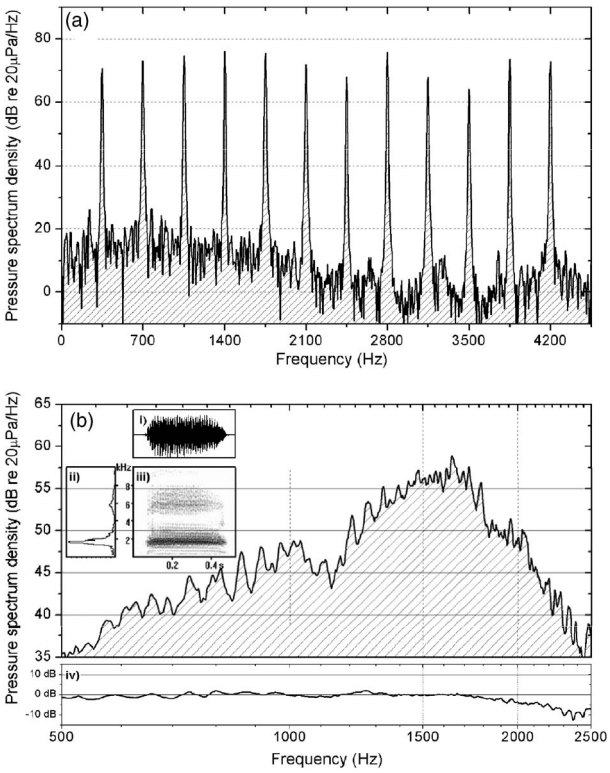


FIG. 1. Sound source characteristics. (a) Pressure spectrum density (dB re 20 μPa/Hz) recorded at 1 m distance of the *harmonic complex* used for measuring attenuation spectra. The relative amplitudes of the harmonics indicate the frequency response of the playback system. (b) The average pressure spectrum density (dB re 20 μPa/Hz) of the 16 different, individual *crow calls* used for transmission recorded at 1 m distance. The insets show (i) the oscillogram of a single crow call together with (ii) its normalized mean power spectrum and (iii) spectrogram. The bottom graph (iv) shows the detailed frequency response of the playback system within the crow call range recorded at 1 m distance.

C. Test sounds and playback level

The test sounds consisted of: (1) a 15-s-long harmonic complex with a natural crow call fundamental frequency of 350 Hz and all harmonics up to 4.2 kHz synthesized by a Tucker-Davis System 3 Real time Processor (TDT, RP2.1) [Fig. 1(a)] and (2) a succession of calls from 16 different crow individuals normalized digitally to the same rms value [Fig. 1(b)]. The harmonic complex was played back at an overall sound pressure level of 85 dB (re 20 μPa) at 1 m at Location 1 and 87 dB (re 20 μPa) at 1 m at Location 2. The crow calls were played back in both locations at an overall sound pressure level of 83±1 dB (re 20 μPa) at 1 m. The average pressure spectrum density at 1 m from the loudspeaker of played back natural crow calls recorded from 16 different individuals is shown in Fig. 1(b). The average pressure spectrum density peaks broadly around approximately 1.6 kHz and has a 10 dB bandwidth ($Q_{10\ \mathrm{dB}}$) of 1.2 kHz. Insets (i)–(iii) in Fig. 1(b) show characteristics of single crow calls. Inset (i) shows the harsh, amplitude modulated nature of crow calls. From insets (ii) and (iii) it can be seen that the power spectra of single calls are generally more peaked than the average of the 16 individuals. Figure 1(b), inset (iv) shows the frequency response of the loudspeaker which is rather flat with a 10 dB deviation from approximately 2 to 2.5 kHz. Thus the artificially reproduced crow calls were representative of natural calls.

D. Sound analysis

The sound attenuation of the individual harmonic components after propagation across the different transects was measured in SASLAB PRO (version 4.39) by the "power spectrum (logarithmic)" feature as the level difference between the source level at 1 m and the level at the end of the transmission transect of interest. The pressure spectrum density was calculated for background noise levels and crow calls in SASLAB by the "Power Spectrum (spectrum level units)" feature. The pressure spectrum density given for crow calls is an average of 16 calls from 16 different individuals [Fig. 1(b)]. The high background traffic noise and noise from roosting jackdaws (*Corvus monedula*) and rooks (*Corvus frugilegus*) at location one meant that analysis of crow call propagation at Location 1 was not possible. However the approximately 10 dB higher sound levels of most of the individual tones in the harmonic complex signals enabled analysis of their attenuation at this site [compare levels in Figs. 1(a) and 1(b)].

E. Propagation model and predictions

1. Propagation factors included in the model

Propagation of sound outdoors involves geometric spreading, air absorption, refraction associated with wind and temperature gradients and effects of atmospheric turbulence. There may be interaction with the ground, barriers and buildings, topography and vegetation. Distance alone will result in wave front spreading. In the simplest case of a sound source radiating equally in all directions (an omnidirectional source) far from the ground wave front spreading means that the sound pressure level decreases at a rate of 6 dB per doubling of distance at all frequencies. In addition a proportion of sound energy is converted to heat as it travels through the air. There are heat conduction losses, shear viscosity losses, and molecular relaxation losses. The resulting air absorption becomes significant at high frequencies and at long range so air acts as a low pass filter at long range. Here attenuation due to atmospheric absorption has been calculated according to ISO 9613-1 (ISO, 1993).

When source and receiver are elevated but close to the ground there is interference between sound traveling directly from source to receiver and sound reflected from the ground. This is ground effect. Sometimes the phenomenon is called ground absorption but, since the interaction of outdoor sound with the ground involves interference, there can be enhancement associated with constructive interference as well as attenuation, in excess of that due to wave front spreading and atmospheric absorption, resulting from destructive interference. Atmospheric turbulence reduces the coherence between the direct and ground reflected sound paths and hence the extent of the destructive and constructive interference.

The propagation model used in this study allows for frequency-dependent ground effect for a point source, air absorption, and turbulence. The model predicts excess attenuation, i.e., attenuation in excess of that expected from simple geometric spreading due to the turbulence-modified ground effect and atmospheric absorption, as a function of frequency. Excess attenuation is independent of the source spectrum and depends only on the geometry (sender and re-

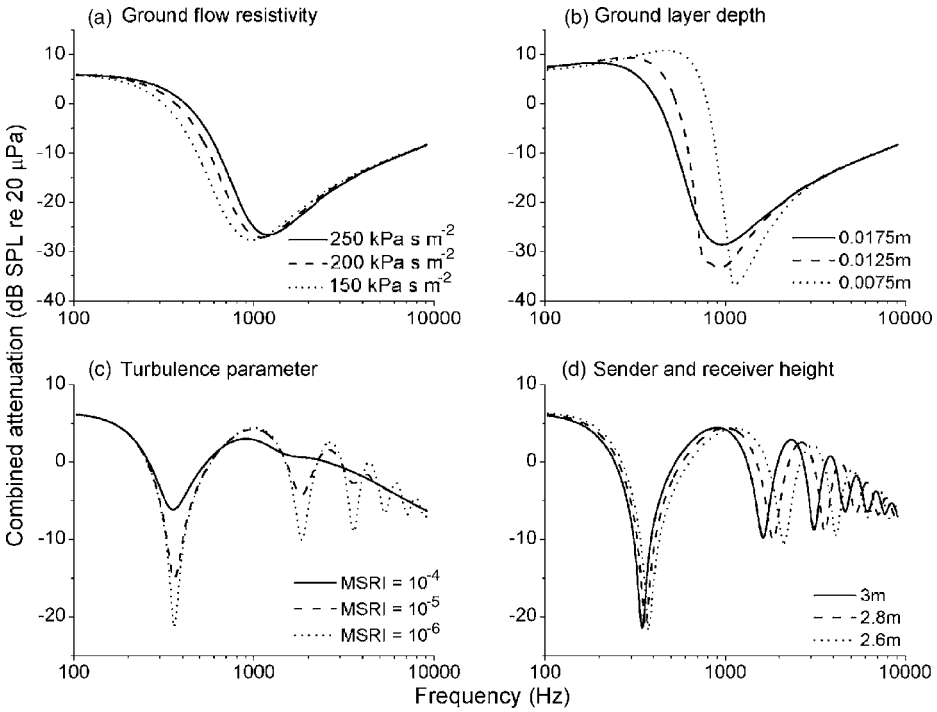


FIG. 2. Examples of the effect of model parameter adjustment on predicted combined attenuation within the ranges used in the current study. (a) Effects of different effective flow resistivities at a separation of 40 m with a sender/receiver height of 0.28 m, a MSRI of 10^{-6}, and an infinitely thick ground. (b) Effects of different layer depths at a separation of 40 m with sender/receiver height of 0.28 m, a MSRI of 10^{-6}, and a fixed effective flow resistivity of 200 kPa s m^{-2}. (c) Effects of different mean squared refraction indices (MSRI) at a separation of 80 m with at sender/receiver height of 2.8 m, flow resistivity of 200 kPa s m^{-2}, and layer depth of 0.0125 m. (d) Effects of different sender/receiver heights with 80 m separation and MSRI of 10^{-6}, flow resistivity of 200 kPa s m^{-2}, and layer depth of 0.0125 m.

ceiver heights and their horizontal separation) and the nature of the intervening ground. It can be either positive or negative due to either constructive or destructive interference between the direct and ground reflected waves. In this study, positive values of excess attenuation signify sound levels *above* those expected from simple geometric spreading as a result of *constructive* interference and negative values indicate *destructive* interference. Although the model assumes propagation from a point source, i.e., an omnidirectional source, it is applicable as long as the sound radiation from the source of interest does not deviate too much from this assumption. Any deviation becomes less important at distances that are long compared with the source dimensions.

Prediction of the ground effect requires knowledge of the acoustical properties of the ground, specifically its surface impedance, i.e., the ratio of sound pressure to acoustically induced normal velocity at the ground surface. Here the ground impedance has been modeled as that of a hard-backed rigid-porous layer by using the semiempirical Delany and Bazley formulas (Attenborough, 2002). This introduces two parameters: an effective flow resistivity and a layer depth. Other ground impedance models might give better predictions for the rough ground at the locations of interest but these have not been explored here (Attenborough, 1992; Attenborough *et al.*, 2000; 2006). The predictions have used the temperature and humidity measured at the test locations to calculate the sound speed. The model assumes a Gaussian turbulence spectrum with the outer scale length of turbulence fixed at 1 m (a value of the same order as the source heights of interest) and a value of the mean squared refractive index (MSRI) chosen between 10^{-6} and 10^{-4} indicating the strength of the turbulence. A more detailed account of the sound propagation model and explanations of the parameters can be found elsewhere (Attenborough *et al.*, 2006). The sum of the excess attenuation from the turbulence-modified ground effect and atmospheric absorption has been termed "combined attenuation" (see Figs. 2–6).

2. Fitting procedure

Figures 2(a)–2(d) show predictions, for the nominal sender/receiver geometries used in the field experiments. The effects of varying effective flow resistivity, layer depth, sender and receiver heights, and MSRI are shown. Changes in the ground impedance parameters change the depth and frequency of the lowest frequency dip in particular [Figs. 2(a) and 2(b)]. In the subsequent graphs the ground impedance is calculated from an effective flow resistivity of 200 kPa s m^{-2} and a layer depth of 0.0125 m. An effective flow resistivity of 200 kPa s m^{-2} is fairly typical of grassland (Attenborough *et al.*, 2006). Adjustment of the MSRI changes the depth and height of the dips and peaks without much affecting their frequencies [Fig. 2(c)]. Slight adjustment of the sender/receiver heights (±20 cm) causes slight changes in the frequencies of the peaks and dips [Fig. 2(d)]. The horizontal sender/receiver separation was not adjusted since changes of ±20 cm comparable to those considered for the sender/receiver heights have a negligible effect on the predictions.

For each transmission geometry, the best fit ground impedance parameters were deduced from the shortest range data first. Approximate fits of the frequencies and depths of the dips are obtained with the nominal sender/receiver heights, a MSRI value of 10^{-6}, and ground impedance parameter values of 200 kPa s m^{-2} and 0.0125 m. The fitting of the data was improved next by varying the sender/receiver heights by ±20 cm from their nominal values [see Fig. 2(d)]. This is justified since the ground at the field sites is rough, with roughness heights of this order, and the effective sender

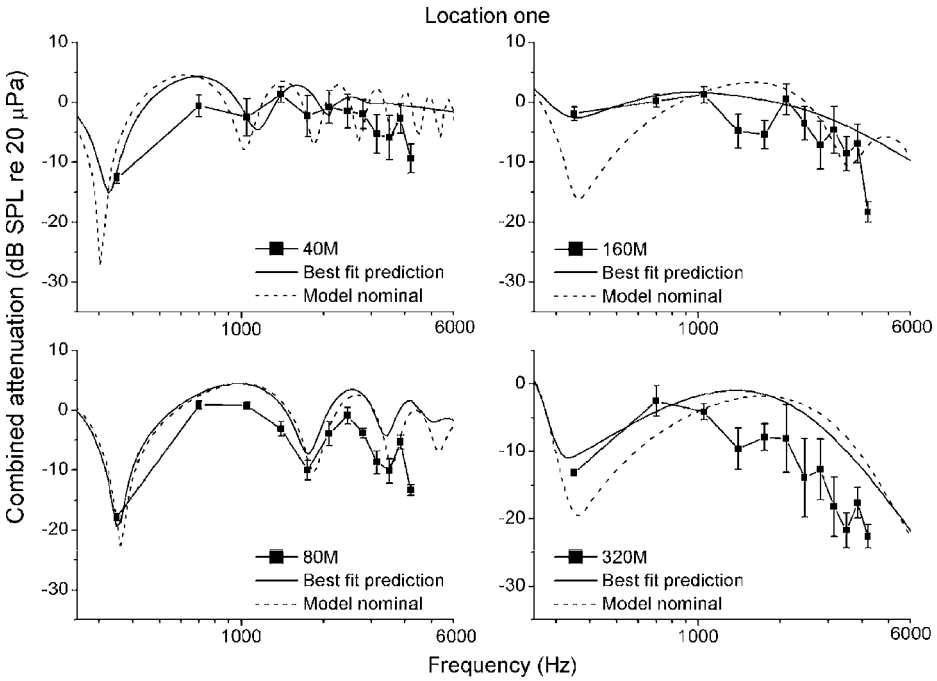


FIG. 3. Measured combined attenuation (ground effect and atmospheric absorption) during air-to-air transmission of harmonic complexes at 40, 80, 160, and 320 m at Location 1 with nominal sender and receiver heights of 2.8 m (joined squares). The solid lines show the best model fits (see the text for parameter values) and the dashed lines show the predicted model attenuation with nominal sender/receiver heights measured at the sender/receiver locations and an assumed MSRI of 10^{-6} indicating low turbulence.

and receiver heights for sound transmission are likely to differ from their nominal values. Also this adjustment is an approximate way of allowing for the effects of atmospheric refraction which are not otherwise included in the model. A final improvement of the predictions follows from adjusting the MSRI value. Since the turbulence parameters were not measured, the MSRI has been varied between 10^{-6} (low turbulence) and 10^{-4} (strong turbulence) to obtain the best fit to data. After fixing the ground impedance, the low frequency ground effect dip between 300 and 400 Hz is particularly sensitive to the assumed value of MSRI [Fig. 2(c)] and it is straightforward to determine the best fit MSRI value. Rather than apply computerized fitting the best visual fits have been obtained. In Figs. 3–6, as well as the best fit predictions, are shown predictions using the nominal sender/receiver heights and a low turbulence MSRI value of 10^{-6}. This indicates the validity of the adjustments used to obtain best fit. These adjustments are discussed in detail in the following.

F. Method of estimating maximum communication distance

Frequently the term "active space" has been used to signify the approximate distance at which a territorial bird can signal its presence acoustically to a conspecific receiver (Brenowitz, 1982; Dooling *et al.*, 2000; Lohr *et al.*, 2003). The active space of hooded crows has been calculated relative to their masked threshold (Jensen and Klokker, 2006).

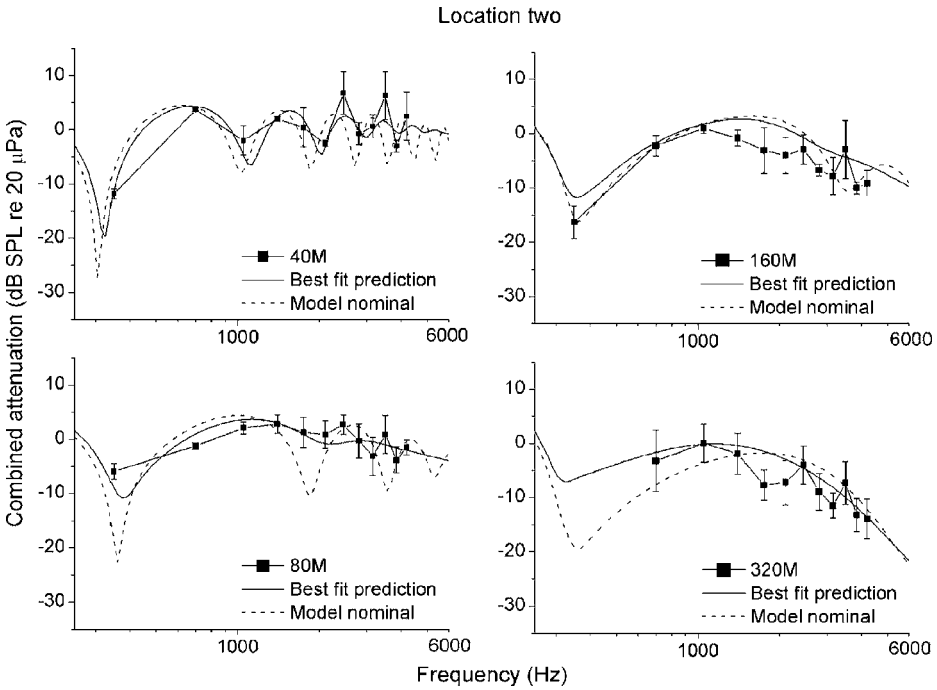


FIG. 4. Measured combined attenuation (ground effect and atmospheric absorption) during air-to-air transmission of harmonic complexes at 40, 80, 160, and 320 m at Location 2 with sender and receiver heights of 2.8 m (joined squares). The solid lines show the best model fits (see the text for parameters) and the dashed lines show predicted attenuations with the nominal sender and receiver heights of 2.8 m measured at the sender/receiver locations and an assumed MSRI of 10^{-6} indicating low turbulence.

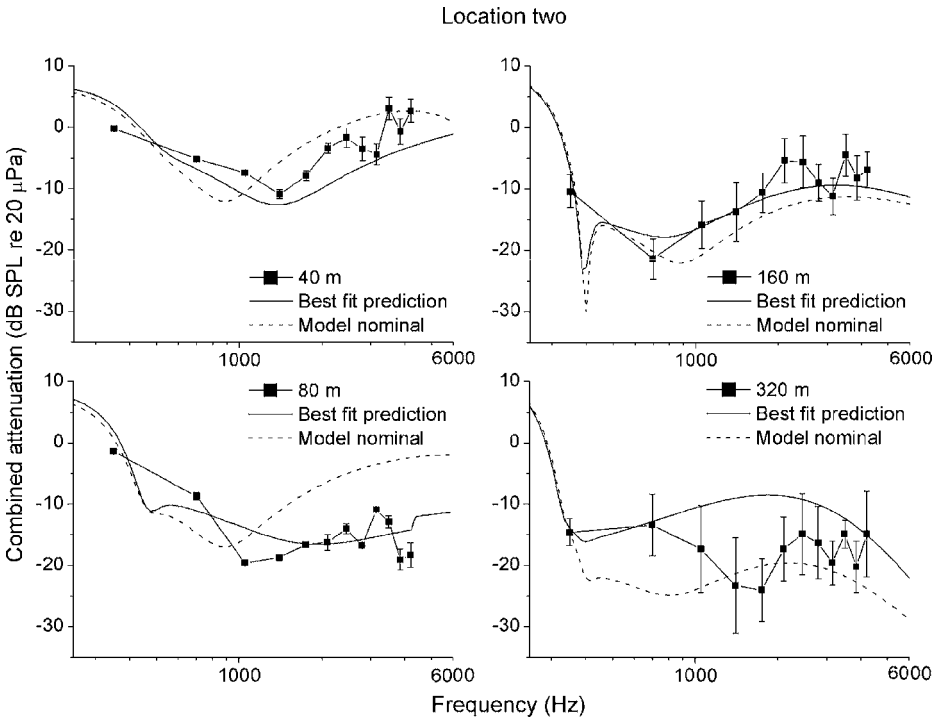


FIG. 5. Measured combined attenuation (ground effect and atmospheric absorption) during air-to-ground transmission of harmonic complexes at 40, 80, 160, and 320 m at location two with sender height of 2.8 m and receiver height of 0.28 m (joined squares). The solid lines show the best model fits (see the text) and the dashed lines show the predicted attenuations with the nominal sender and receiver heights of 2.8 and 0.28 m, respectively, measured at the sender/receiver locations and an assumed MSRI of 10^{-6} indicating low turbulence.

Two exemplary masked thresholds have been calculated here: one based on the average pressure spectrum density of background noise recorded during *low wind speed* conditions in August and one based on *moderately windy* (or moderate wind speed) conditions recorded in February (see earlier text). The masked threshold for each condition was obtained by adding the crow critical ratio (Jensen and Klokker, 2006) to the pressure spectrum density of the noise (Klump *et al.*, 1996). To determine if call energy exceeds the masked threshold for a broadband signal, the energy of the call must be summed within the auditory filters, and if the energy level exceeds or equals the masked threshold at any given frequency the crow calls can presumably be detected (Gässler, 1954; Spiegel, 1981). The width of the auditory filter, or its "effective bandwidth" the critical bandwidth (CB), can be approximated by the critical ratio (CR). The CR for humans underestimates the CB by a factor of 0.4 (Moore, 2003). Similar differences apply to budgerigars (*Melopsittacus undulatus*) (Dooling and Searcy, 1979), but in starlings (*Sturnus vulgaris*) the CR corresponds well to the CB (Langemann *et al.*, 1995). Because the crow is a passerine bird like the starling it will be assumed that the uncorrected CR corresponds to the CB. This will result in a conservative measure of the active space. Again, as a conservative measure,

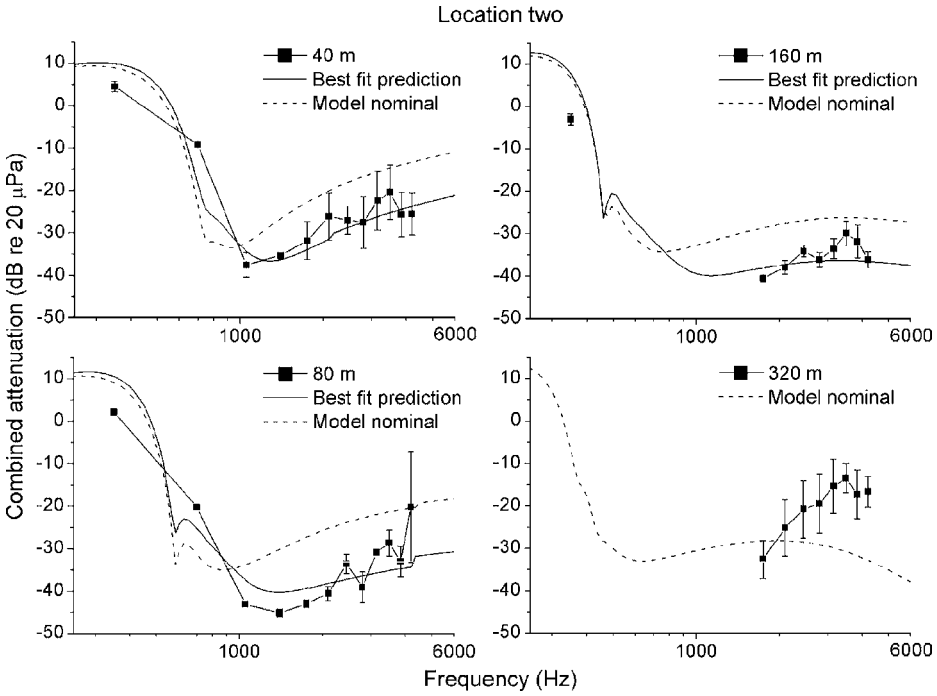


FIG. 6. Measured combined attenuation (ground effect and atmospheric absorption) during ground-to-ground transmission of harmonic complexes at 40, 80, 160, and 320 m at Location 2 with sender and receiver heights of 0.28 m (joined squares). The solid lines show the best model fit (see the text) and the dashed lines show the predicted attenuations with the nominal sender and receiver heights of 0.28 m measured at the sender/receiver location and a MSRI of 10^{-6} indicating low turbulence.

the smallest auditory filter width was used for energy summation measured in the low CR region of crow hearing, which corresponds to 132 Hz (Jensen and Klokker, 2006).

III. RESULTS

A. Propagation of the harmonic complex signals

Combined attenuation spectra for three appropriate sender–receiver geometries have been considered: (a) "air-to-air" corresponding to both calling and listening crows perched above ground, (b) "air-to-ground" corresponding to the caller perched above ground but the listener walking on the ground, and (c) "ground-to-ground" corresponding to both caller and listener walking on the ground, respectively.

1. Air to air transmission

Figure 3 shows the combined attenuation spectra at Location 1 for the propagation distances of 40–320 m with the loudspeaker and microphone elevated to 2.8 m. The model predicts combined attenuation spectra with several peaks and dips at shorter ranges and only a single peak at 160 and 320 m. At 40 m it can be seen that predictions lie within or very close to the standard deviations in the data when a receiver (RH) and sender (SH) heights of 2.6 m and MSRI of $10^{-4.5}$ are chosen. At 80 m the overall trend of the data is predicted by using RH=2.9 m, SH=2.8 m, and MSRI $=10^{-5.5}$. Nevertheless the prediction is generally between 3 and 6 dB above the data. At 160 m the data are best fitted with the nominal RH and SH of 2.8 m and an assumed MSRI of 10^{-5}. Between 1.4 and 1.75 kHz there seems to be a dip in the data of approximately 6 dB not predicted. The measurements at 320 m, like those at 160 m, are best fitted with RH=SH=2.8 m and MSRI=10^{-5}. Again the apparent dip in the data between 1 and 2 kHz is not predicted.

Figure 4 shows measured sound propagation values of combined attenuation at Location 2 together with best model fits for propagation distances between 40 and 320 m with the loudspeaker and microphone both placed at a height of 2.8 m. At 40 m, the nominal heights of RH=SH=2.8 m and an assumed MSRI of 10^{-5} results in the best fits to the data such that the predictions lie within or very close to the standard deviations. The best fits to data from the 80 m propagation transect lie within or close to the standard deviations of the data using RH=SH=2.5 m and MSRI=$10^{-4.5}$. At 160 m a good fit is obtained using RH=SH=2.8 m and MSRI=10^{-5}. Nevertheless there is a slight indication of a dip in the data around 2 kHz, which is not predicted. A good fit of the data at 320 m is obtained with a RH=SH=2.5 m and MSRI=$10^{-4.5}$ m, but again the predictions do not account for the dip in the data around 2 kHz.

2. Air to ground transmission

Figure 5 shows the measured combined attenuation at Location 2 at transects from 40 to 320 m with the loudspeaker placed at 2.8 m and the receiving microphone at 0.28 m (this geometry was used at Location 2 only). When the microphone is placed close to the ground the predicted combined attenuation spectra have a single, broad dip between approximately 500 Hz and 2 kHz with a magnitude of between −10 and −20 dB. At 40 m the measured combined attenuation has a maximum negative value at 1.4 kHz, which can be predicted by using RH=0.12 m, SH=2.5 m, and MSRI=10^{-4}. Nevertheless, above 300 Hz, the resulting predictions are between 2 and 6 dB below the data. The data at 80 m are fitted well by using RH=0.05 m, SH=2.5 m, and MSRI=10^{-4}. The slight dip in the data around 1 kHz is not predicted. At 160 m a good fit is achieved with the nominal values of RH=0.28 m, SH=2.8 m, and an assumed MSRI $=10^{-5}$. The combined attenuation data at 320 m show great variance but predictions that fit within most of the standard deviations are obtained with RH=0.28 m, SH=2.8 m, and MSRI=$10^{-4.5}$. Again the dip in the data around approximately 2 kHz and higher frequency dips are not predicted.

3. Ground to ground transmission

In Fig. 6 the combined attenuation spectra measured at Location 2 with loudspeaker and microphone heights of 0.28 m are shown for transects between 40 and 320 m. In general, the predicted spectra have single, broad dips between approximately 500 Hz and 2 kHz and combined attenuation values as low as approximately −40 dB. The data for the 40 m transect shows at maximum negative combined attenuation close to −40 dB around 1 kHz, which is predicted by using source and receiver heights of 0.15 m and a MSRI value of 10^{-6}. The data at 80 m show the same overall pattern. The maximum negative combined attenuation is approximately −45 dB around 1.4 kHz and this is well predicted using RH=SH=0.15 m and MSRI=10^{-6}. Below 1.75 kHz, at ranges of 160 and 320 m, except for the 350 Hz tone at 160 m, the absolute levels of most of the harmonic components were below the background noise level and were not measurable. At 160 m the best fit prediction with RH=SH=0.10 m and MSRI=10^{-6} lies close to the existing data, but at 320 m the usable data are too few to fit so only the predictions using the nominal SH and RH of 0.28 m and an assumed MSRI of 10^{-6} are shown.

B. Propagation of crow calls

In Figs. 7 and 8 the measured pressure spectrum densities of the transmitted crow calls at all distances are compared to those predicted by the model using the parameters obtained by fitting the harmonic complex data. The predicted spectra are the result of subtracting the expected losses due to geometric spreading, atmospheric absorption, and then adding the modeled ground effect to the pressure spectrum density recorded at 1 m from the source.

Figure 7 shows the measured average pressure spectrum densities (thick lines) of the 16 crow calls recorded at ranges between 40 and 320 m with both the loudspeaker and the microphone placed at 2.8 m together with the predicted pressure spectrum density (thin lines) and the pressure spectrum density of the background noise at the experimental location (dashed lines). The predicted pressure spectrum densities are close to the measured ones. The biggest deviations occur beyond 40 m range and at low frequencies as the pressure spectrum densities of the crow calls fall to within approximately 10 dB of the background noise. This is to be expected

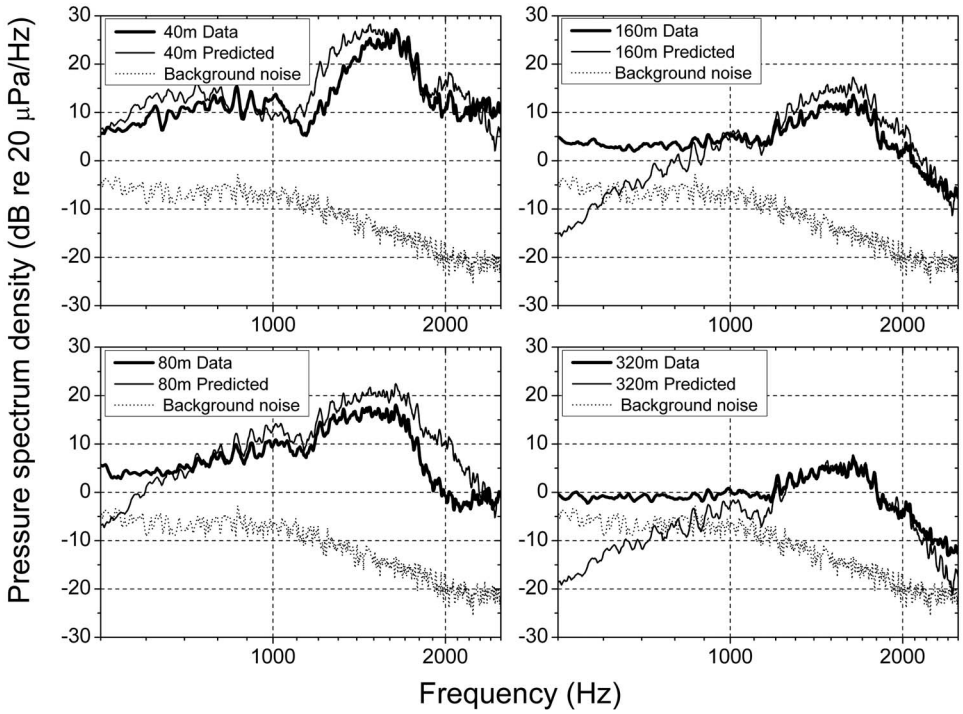


FIG. 7. Comparison of measured air-to-air transmission of crow calls (thick lines) at Location 2 with predicted pressure spectrum densities (thin lines). The data represent the pressure spectrum densities (dB re 20 μPa/Hz) of crow calls ($N=16$) transmitted over distances of 40, 80, 160, and 320 m distance with sender and receiver heights of 2.8 m. Predictions use the parameters (sender/receiver heights, ground parameters, and MSRI) obtained from the harmonic complex transmissions for each respective distance. Also shown are the pressure spectrum densities of the background noise at Location 2 (thin dashed line).

since the result of adding two sound levels is affected significantly by the lower level when it is within 10 dB of the higher level. This suggests that the deviations at lower frequencies are more likely to be the result of poor signal to noise ratio than poor predictions. Notice how the peak in the pressure spectrum densities at 40 m becomes less pronounced with distance.

When the receiver height is 0.28 m (Fig. 8), the predicted pressure spectrum density is in very close agreement with the measured one, but due to the larger attenuation by the ground effect, the measurements are greatly affected by the background noise at frequencies below approximately 1.2 kHz. The measured and predicted values are approximately 10 dB less than when both receiver and sender were at 2.8 m. Although the crow call transmission measurements formed part of the same experiments as the harmonic complex measurements, the good agreement between data and predictions supports the applicability of the best fit model parameters. The discrepancy between predictions and crow call data at 320 m in Fig. 8 corresponds to the discrepancy between predictions and harmonic complex data at 320 m in

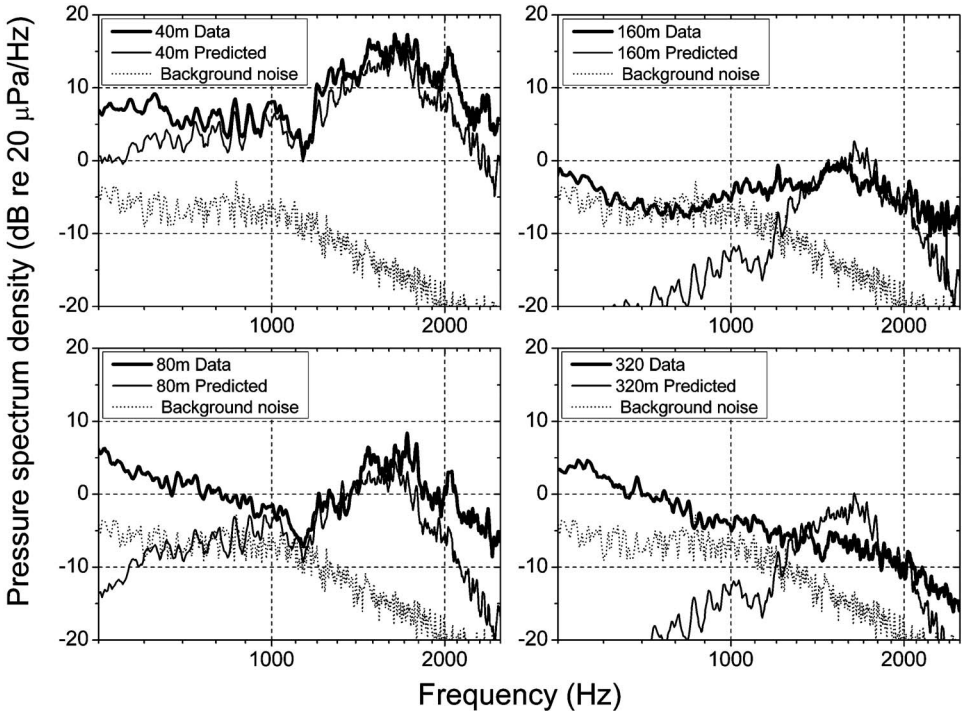


FIG. 8. Air-to-ground transmission of crow calls: Pressure spectrum density (dB re 20 μPa/Hz) of crow calls ($N=16$) transmitted over open field at 40, 80, 160, and 320 m distance with sender height of 2.8 m and receiver height of 0.28 m (thick lines) together with model predicted pressure spectrum densities (thin lines) using the best fitted values (sender/receiver heights, ground parameters, and MSRI) evaluated from the harmonic complex transmissions for each respective distance. Also shown is the pressure spectrum density of the background noise at Location 2 (thin dashed line).

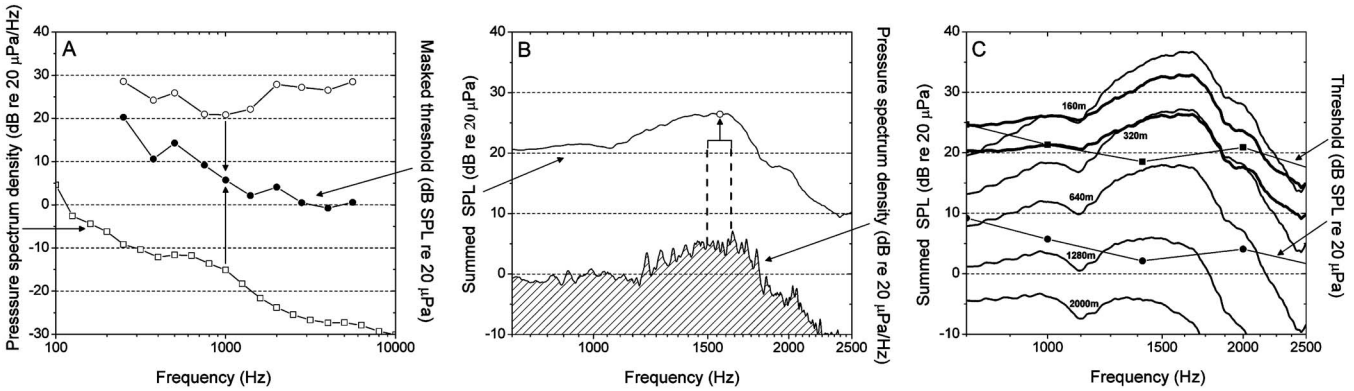


FIG. 9. Deduction of maximal hooded crow communication distances under two different environmental conditions. (a) Open squares denote the pressure spectrum density of the background noise recorded at a quiet location at low wind speeds (1 m/s); closed circles represent the resulting masked threshold calculated by adding the pressure spectrum density of the noise at low wind speed to the critical ratios (open circles). (b) Pressure spectrum density (dB re 20 Pa/Hz) of the air-to-air transmission of calls at 320 m distance at Location 2 (hatched pattern cf. Fig. 7). The summed sound pressure level was calculated by a running summation of the pressure spectrum density within filters of 132 Hz width (see Sec. II F). Open dot represent an example. (c) Thick line traces labeled "160 m" and "320 m" show the summed pressure levels based on data and thin lines represent predictions at the indicated distances. Predictions at 640, 1280, and 2000 m are also shown. If any given summed energy level at any given frequency is above a given masked threshold, the crow is assumed to be able to hear the calls at that condition.

Fig. 5. The particularly good fits between call transmission data and predictions between 1 and 2 kHz in Fig. 7 support the conclusions in Sec. III A.

C. Prediction of active space

First the energy summations (cf. Sec. II F) were carried out using the experimentally determined pressure spectrum density of the crow calls at 160 and 320 m distance (Fig. 7) and then using the corresponding predictions for 160 and 320 m. This was done for each frequency by summing the pressure spectrum densities over a 132 Hz wide band centered at that frequency. The best fit to the harmonic complex transmission at Location 2 at 160 m was obtained by assuming sender and receiver heights of 2.8 m and an MSRI value of 10^{-5} and at 320 m was obtained by assuming sender and receiver heights of 2.5 m and an MSRI value of $10^{-4.5}$ (see Sec. III A 1 and Fig. 4). These values were used for the predictions of the crow's active space. To assess the detectability of crow calls at longer transmission distances not covered by the experiment, predictions of the summed energy level in crow auditory filters have been made at 640, 1280, and 2000 m using the values of ground impedance parameters, sender/receiver heights, and MSRI as those given earlier for 320 m.

Figure 9(c) shows summed energy of crow calls within auditory filters of 132 Hz width in relation to two masked thresholds: one for lower and one for higher wind speeds (see Sec. II F) and Figs. 9(a) and 9(b) illustrate how the summed energies were computed. At 160 and 320 m [Fig. 9(c)], the summed energies based both on experimental data (thick lines) and model predictions (thin lines) are shown. In general there is good agreement between the data and predictions. At the lower frequencies the summed energy level based on data is between 2 and 6 dB above the predicted energy level since, as discussed earlier and shown in Fig. 7, the measurement is affected by the background noise. On this basis, the energy summations based on model predictions fit those based on recordings in the field very well. Therefore, there is good reason to believe that the energy levels predicted for 640, 1280, and 2000 m are close to what would be expected in the field.

1. Results at ranges >320 m

At 640 m the predicted energy level is slightly below the masked threshold in the noisier condition with moderate wind speeds, whereas it is approximately 20 dB above the masked threshold in the low wind speed and relatively quiet condition. The predicted energy level at 1280 m is approximately 15 dB below the windy threshold condition, whereas it is approximately 5 dB above the calm threshold condition. Thus, given a source sound pressure level of 83 dB (re 20 μPa) at 1 m, a crow is likely to detect a conspecific call up to approximately 640 m in moderate wind conditions and up to 1280 m or more in calm conditions. However, calling crows recorded in an anechoic room can produce a sound pressure level of up to at least 95 dB (re 20 μPa) at 1 m (125 ms averaging time; Jensen, *et al.*, private communication) so, probably, at least 10 dB could be added to the assumed source level. This means that the crow calls would be detectable at approximately 1280 m during moderate wind speeds, and as far as 2000 m under very calm conditions.

It has to be noted that the masked thresholds are based on a psychoacoustic task where thresholds were defined corresponding roughly to 50% likelihood of detection. The same is thus true for the assessed maximum detection distances. Also the sender/receiver geometry is important. For example, under low turbulence conditions (MSRI=10^{-6}) with sender/receiver heights of 8 m the energy level at 1280 m would be predicted to be 6 dB higher compared to a sender/receiver height of 2.5 m If, on the contrary, the sender/receiver height is lowered to 1.5 m the energy level is predicted to be 4 dB lower relative to that at 2.8 m.

To summarize, although the detection conditions in a changing natural environment with shifting sender/receiver geometries are extremely variable, based on open field background noise measurements, hooded crow critical ratios,

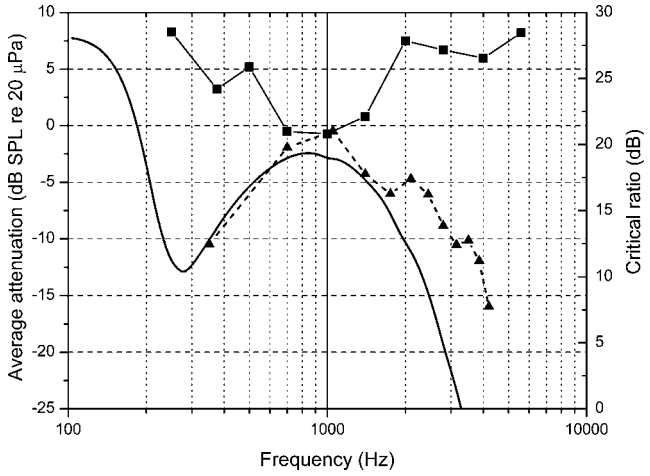


FIG. 10. The sound communication window of hooded crows is here defined as the average combined attenuation of all sender/receiver combinations within a spatial matrix representative of hooded crow territorial communication (heights of 3, 6, and 9 m and distances of 300, 600, 900, and 1200 m, respectively) (solid line). The measured attenuation is expressed as an average of measurements at 160 m and 320 m (triangles). Also shown is the critical ratios of hooded crows (Jensen and Klokker, 2006) (squares). The special hearing sensitivity in noise between 500 Hz and 2 kHz strongly overlaps the sound communication window, and experimental results, and suggests that crow hearing is adapted to the ground effect.

hooded crow source levels, and propagation predictions assuming sender/receiver heights of 2.5 m, it is estimated that crows are able to detect a conspecific call at roughly 1 km distance during moderately windy conditions. Theoretically, the detection range might be as far as 2 km under quiet and calm conditions.

D. "Sound communication window"

To assess the overall communication conditions within the crow's active space, the combined attenuation has been averaged within a spatial matrix that defines the space within which crows communicate territorially. Twenty-four model calculations were made of all possible sender/receiver geometries restricted to sender and receiver heights of 3, 6, and 9 m and sender/receiver separations of 300, 600, 900, and 1200 m over flat ground with impedance parameters corresponding to those used earlier. The calculations were performed assuming a temperature of 25 °C, relative humidity of 70%, and a MSRI value of 10^{-5} corresponding to moderate turbulence conditions. The average combined attenuation in that sound communication window is shown as the black trace in Fig. 10. An optimum range for communication is predicted between approximately 500 Hz and 1.6 kHz ($BW_{3\ dB}$). Since, to a great extent, this range overlaps the frequency range where crow critical ratios are lowest and crow hearing in noise is consequently best, it is suggested that the special critical ratio pattern of crows is an adaptation to communication within the sound communication window mainly created by the ground effect.

IV. DISCUSSION

A. Propagation modeling as a bioacoustic tool

Comparisons between predictions and data using synthetic harmonic complexes have demonstrated that, even at wind speeds up to 4 m/s, sound propagation modeling including only ground effect and turbulence, to a large degree, can account for the overall spectral profile of a propagated sound at different sender/receiver geometries. Despite the fact that the best fit predictions overestimated the observed attenuation between approximately 1 and 2 kHz at both locations with ground elevated sender and receivers, the contrast between the air-to-air transmissions and air-to-ground or ground-to-ground transmissions clearly demonstrates the different consequences that ground effect can have on sound propagation. It should be noted that above 1 kHz, sound has wavelengths comparable to the roughness of the ground (10–20 cm) at both locations. Although the effects of ploughing and ground roughness have been studied (Aylor, 1972; Peng and Lines, 1995; Attenborough *et al.*, 2000; Boulanger *et al.*, 2005; Attenborough *et al.*, 2006), there is a case for further study of ground roughness effects where the roughness scale is comparable to the wavelengths of interest. Another important omission from the model is refraction due to wind and temperature gradients. To allow for these would require more meteorological information than was available in the experiments described here. In principle, additional parameters and a heuristic ray trace extension of the model (Li, 1993) could be used but this will be the subject of future work. In spite of its limitations, the "turbulence-modified ground effect plus atmospheric absorption" model was able to predict the spectral profile of hooded crow calls remarkably well. This supports the wider use of modeling in predicting communication conditions for any given sender/receiver geometry.

B. Implications for hooded crow communication ecology

Calculation of the theoretical sound communication window within a spatial matrix relevant for interterritorial communication has predicted a sound communication window between 500 Hz and 1.6 kHz ($BW_{3\ dB}$). This is a frequency range in which hooded crows, uniquely among songbirds, have a particularly sensitive hearing ability in noise, i.e., low critical ratios (CR) (Jensen and Klokker, 2006) (Fig. 10). This special hearing range does not correlate with close range crow calls. However, it correlates surprisingly well with the predicted sound window, suggesting that crow hearing is specifically adapted to the ground effect.

The data and predictions have shown that, when a crow is walking on the ground, a place where crows spend a lot of their time foraging, it will experience approximately 10 dB greater attenuation relative to a receiver height of 2.8 m and thus have some difficulties in hearing signaling conspecifics. If both sender and receiver are placed on the ground acoustic communication is greatly impaired even at 40 m distance. Thus crows will have great difficulties in hearing conspecific calls when walking on the ground. If they wish to listen optimally to signaling crows (territorial neighbors or incoming predatory strangers) they need to rise from the ground. Crows can be observed to take off from the foraging grounds every now and then and fly up in trees or other high points where they appear to be looking out for conspecifics and

dangers. It is very possible that part of this behavior is to improve listening and signaling conditions. A similar behavior observed in the European blackbird (*Turdus merula*) has been explained with the same arguments (Dabelsteen *et al.*, 1993).

We estimated the active space of hooded crows at 2.5 m height to be roughly 1 km with wind speeds between 5 and 6 m/s. Since the average wind speed in Denmark during the most important "acoustic season" is 4.5–5.6 m/s for spring time (March to May) and 3.9–5.0 m/s for summer time (Danish Meteorological Institute, DMI), the calculated active space of hooded crows is probably representative. Note though, that the active space at 2.5 m height is based on what was practically measurable at open field and is probably in the lowermost range of natural crow calling heights. The active space will vary with height as discussed in Sec. III C 1. With territories of up to between 500 and 700 m in diameter (Coombs, 1978) and an active space of 1 km crows are probably able to communicate well between territories. This is known to be a rather important issue in songbird communication and breeding success (Kroodsma *et al.*, 1982; Catchpole and Slater, 1995; Kroodsma *et al.*, 1996).

It is important to note that the ground effect not only attenuates the sounds but also provides amplification depending on the sender/receiver geometry (cf. Michelsen and Larsen, 1983). This is for instance evident in Figs. 3 and 4 where the combined attenuation is generally above zero at 40 and 80 m around 500 Hz to 1 kHz, and even slightly above or below zero at 160 and 320 m around 1 to 2 kHz despite the atmospheric attenuation. At even longer distances, say 1200 m, with a sender/receiver height of 9 m, the sound propagation model predicts an amplification effect of 5 dB around 1.2 kHz (MSRI=10^{-5}). Thus, the ground effect provides a sound communication window that not only provides a range where attenuation is minimal, but can also provide significant amplification.

C. Turbulence

In the bioacoustic literature turbulence is assumed to have a large effect on *attenuation* and to depend strongly on frequency by a quadrate frequency dependence (Wiley and Richards, 1978; Wiley *et al.*, 1982). This assumption has been based on some early work (Lighthill, 1952; Ingard, 1953; Lighthill, 1953). However, more recent work does not support this notion (Brown and Clifford, 1976; Piercy *et al.*, 1977; Pan, 2003) and points to errors in Lighthill's work (Brown and Clifford, 1976). The more recent work points to a very small effect of turbulence and a weak frequency dependence of approximately $1.63(f/1000)^{1/3}$ in dB/km, where f is the frequency (Brown and Clifford, 1976; Pan, 2003). For the highest frequencies used in the present study, for instance, this would amount to approximately 1 dB reduction in sound pressure level over the longest transmission path of 320 m. Thus the turbulence in windy open field habitats is not likely to impair hooded crow communication to any noticeable degree.

On the other hand, turbulence has a significant influence on the ground effect. It varies the phase relationship between direct and reflected wave, thereby reducing their coherence, and reducing or eliminating the excess attenuation due to the ground effect (Peng and Lines, 1995; Attenborough *et al.*, 2000). Specifically the first maximum in the excess attenuation spectrum is reduced and the peaks and dips at the higher frequencies destroyed [cf. Fig. 2(c)] (Attenborough *et al.*, 2000). Since there is more prominent reduction in attenuation dips relative to peaks, on average turbulence will *increase* the likelihood that call sound energy is transmitted to the receiver. This possibility should be investigated further.

D. Perspectives and limitations

Ground effect has rarely been considered in a quantitative approach in previous songbird or other communication studies (however see e.g. Roberts *et al.*, 1979; Michelsen and Larsen, 1983; Garstang *et al.*, 1995; Nelson, 2003). Our study shows that it is very important to consider the ground effect and that it plays a major role in shaping sound propagation and animal sound communication. However, although sound propagation models based solely on the ground effect, turbulence effects, and atmospheric absorption have been shown useful in the current study, more complete propagation models, for example taking into account refraction and vegetation effects, are likely to be needed in further work.

There is currently a large interest in human noise impact on animal communication and welfare (Slabbekoorn and Peet, 2003; Brumm and Slabbekoorn, 2005; Brumm, 2006; Slabbekoorn *et al.*, 2007; Swaddle and Page, 2007). The method introduced for assessing the active space in crows could prove useful for quantitative predictions of the impact of noise on animal communication. Hearing data are available for a great number of birds and other vertebrate species (Fay, 1988; Dooling *et al.*, 2000) making a general prediction scheme possible. However, experimental validation of the method is crucial. Data for this are available from the classic field study of the active space of the red-winged blackbird (*Agelaius phoeniceus*) (Brenowitz, 1982). Here it was demonstrated that the birds were only able to detect and respond to conspecific song played back at or above a signal-to-noise (S/N) ratio of 3 dB defined as the song energy relative to the noise energy in a 4 kHz octave filter. Psychoacoustic detection depends on the S/N ratio within the auditory filters (Moore, 2003) and is indicated to be 0 dB in passerines (Langemann *et al.*, 1995). Based on a critical ratio of 29.5 dB for the red-winged blackbird, its auditory filter is approximately 890 Hz at 4 kHz (Hienz and Sachs, 1987). With this filter width a reassessment of Brenowitz (1982) reveals that the signal-to-noise ratio at detection threshold is indeed approximately 0 dB between 2 and 4 kHz where the main song energy lies. This supports a close correlation between psychoacoustic data and hearing abilities in natural settings (however see Klump *et al.*, 1996) and validates their use for assessing the impact of human noise in the biotope.

ACKNOWLEDGMENTS

The authors are grateful to Gerardo Obando Calderón for help in conducting some of the field measurements and for recording some of the crow calls used in the experiments, to Paola Laiolo for providing some of the hooded crow calls,

and to Simon Boel Pedersen for advice on signal analysis. This study was supported by a grant from the Danish Science Research Council (SNF 23155-4) to O.N.L.

Attenborough, K. (**1992**). "Ground parameter information for propagation modeling," J. Acoust. Soc. Am. **92**, 418–427.

Attenborough, K. (**2002**). "Sound propagation close to the ground," Annu. Rev. Fluid Mech. **34**, 51–82.

Attenborough, K., Li, K. M., and Horoshenkov, K. (**2006**). *Predicting Outdoor Sound* (Spon Press: an imprint of Francis and Taylor, London).

Attenborough, K., Taherzadeh, S., Bass, H. E., Di, X., Raspet, R., Becker, G. R., Gudesen, A., Chrestman, A., Daigle, G. A., Lesperance, A., Gabillet, Y., Gilbert, K. E., Li, Y. L., White, M. J., Naz, P., Noble, J. M., and Vanhoof, H. A. J. M. (**1995**). "Benchmark cases for outdoor sound-propagation models," J. Acoust. Soc. Am. **97**, 173–191.

Attenborough, K., Waters-Fuller, T., Li, K. M., and Lines, J. A. (**2000**). "Acoustical properties of farmland," J. Agric. Eng. Res. **76**, 183–195.

Aylor, D. (**1972**). "Noise-reduction by vegetation and ground," J. Acoust. Soc. Am. **51**, 197–205.

Boulanger, P., Attenborough, K., and Qin, Q. (**2005**). "Effective impedance of surfaces with porous roughness: Models and data," J. Acoust. Soc. Am. **117**, 1146–1156.

Brenowitz, E. A. (**1982**). "The active space of red-winged blackbird song," J. Comp. Physiol. **147**, 511–522.

Brown, E. H., and Clifford, S. F. (**1976**). "Attenuation of sound by turbulence," J. Acoust. Soc. Am. **60**, 788–794.

Brumm, H. (**2006**). "Animal communication: City birds have changed their tune," Curr. Biol. **16**, R1003–R1004.

Brumm, H., and Slabbekoorn, H. (**2005**). "Acoustic communication in noise," Adv. Stud. Behav. **35**, 151–209.

Catchpole, C. K., and Slater, P. J. B. (**1995**). *Bird Song-Biological Themes and Variations* (Cambridge University Press, Cambridge).

Coombs, C. J. F. (**1978**). *The Crows: A Study of the Corvids of Europe* (Batsford, London).

Cramp, S., and Perrins, C. M. (**1994**). *Crows to Finches*, Handbook of the Birds of Europe, the Middle East, and North Africa: The Birds of the Western Palaearctic, Vol. **VIII** (Oxford University Press, Oxford).

Dabelsteen, T., Larsen, O. N., and Pedersen, S. B. (**1993**). "Habitat-induced degradation of sound signals: Quantifying the effects of communication sounds and bird location on blur ratio, excess attenuation, and signal-to-noise ratio in blackbird song," J. Acoust. Soc. Am. **93**, 2206–2220.

Daigle, G. A., Piercy, J. E., and Embleton, T. F. W. (**1983**). "Line-of-sight propagation through atmospheric turbulence near the ground," J. Acoust. Soc. Am. **74**, 1505–1513.

Dooling, R. J., Lohr, B., Dent, M., Dooling, R. J., Fay, R. R., and Popper, A. N. (**2000**). "Hearing in birds and reptiles," in *Comparative Hearing: Birds and Reptiles*, edited by S. Greenberg, W. Ainsworth, A. N. Popper, and R. R. Fay (Springer, New York), pp. 308–359.

Dooling, R. J., and Searcy, M. H. (**1979**). "Relation among critical ratios, critical bands, and intensity difference limens in the parakeet (*Melopsittacus undulatus*)," Bull. Psychon. Soc. **13**, 300–302.

Embleton, T. F. W. (**1996**). "Tutorial on sound propagation outdoors," J. Acoust. Soc. Am. **100**, 31–48.

Fay, R. R. (**1988**). *Hearing in Vertebrates: A Psychophysics Databook* (Hill-Fay Assoiciates, Winnetika, IL).

Fotheringham, J. R., and Ratcliffe, L. (**1995**). "Song degradation and estimation of acoustic distance in black-capped chickadees (*Parus atricapillus*)," Can. J. Psychol. **73**, 858–868.

Garstang, M., Larom, D., Raspet, R., and Lindeque, M. (**1995**). "Atmospheric controls on elephant communication," J. Exp. Biol. **198**, 939–951.

Gässler, G. (**1954**). "Über die Hörschwelle für Schallereignisse mit verschieden breitem Frequenzspektrum," (On the threshold of sounds with different frequency bandwidths), Acustica **4**, 408–414.

Gish, S. L., and Morton, E. S. (**1981**). "Structural adaptation to local habitat acoustics in Carolina wren songs," Z. Tierpsychol **56**, 74–84.

Hienz, R. D., and Sachs, M. B. (**1987**). "Effects of noise on pure-tone thresholds in blackbirds (*Agelaius phoeniceus and Molothrus ater*) and pigeons (*Columba livia*)," J. Comp. Psychol. **101**, 16–24.

Ingard, U. (**1953**). "A review of the influence of meteorological conditions on sound propagation," J. Acoust. Soc. Am. **25**, 405–411.

ISO (**1993**). "Acoustics—Attenuation of sound during propagation outdoors. 1. Calculation of the absorption of sound by the atmosphere," ISO 9613-1 (ISO, New York).

Jensen, K. K., and Klokker, S. (**2006**). "Hearing sensitivity and critical ratios of hooded crows (*Corvus corone cornix*)," J. Acoust. Soc. Am. **119**, 1269–1276.

Kime, N. M., Turner, W. R., and Ryan, M. J. (**2000**). "The transmission of advertisement calls in Central American frogs," J. Creat. Behav. **11**, 71–83.

Klump, G. M., Kroodsma, D. E., and Miller, E. H. (**1996**). "Bird communication in the noisy world," in *Ecology and Evolution of Acoustic Communication in Birds* (Cornell University Press, Ithaca, NY), pp. 321–338.

Kroodsma, D. E., Miller, E. H., Kroodsma, D. E., and Miller, E. H. (**1982**). *Acoustic Communication in Birds* (Academic, New York), vols. **I** and **II**.

Kroodsma, D. E., Miller, E. H., Kroodsma, D. E., and Miller, E. H. (**1996**). *Ecology and Evolution of Acoustic Communication in Birds* (Comstock, Ithaca, NY).

Langemann, U., Klump, G. M., and Dooling, R. J. (**1995**). "Critical bands and critical-ratio bandwidth in the European starling," Hear. Res. **84**, 167–176.

Li, K. M. (**1993**). "On the validity of the heuristic ray-trace-based modification to the Weyl-Van der Pol formula," J. Acoust. Soc. Am. **93**, 1727–1735.

Li, K. M., Taherzadeh, S., and Attenborough, K. (**1998**). "An improved ray-tracing algorithm for predicting sound propagation outdoors," J. Acoust. Soc. Am. **104**, 2077–2083.

Lighthill, M. J. (**1952**). "On sound generated aerodynamically," Proc. R. Soc. London, Ser. A **211**, 564–587.

Lighthill, M. J. (**1953**). "On the energy scattered from the interaction of turbulence with sound or shock waves," Proc. Cambridge Philos. Soc. **49**, 531–555.

Lohr, B., Wright, T. F., and Dooling, R. J. (**2003**). "Detection and discrimination of natural calls in masking noise by birds; estimating the active space of a signal," Anim. Behav. **65**, 763–777.

Madge, S., and Burn, H. (**1994**). *Crows and Jays: A Guide to the Crows, Jays and Magpies of the World* (Black, London).

Michelsen, A., and Larsen, O. N. (**1983**). "Strategies for acoustic communication in complex environments," in *Neuroethology and Behavioral Physiology*, edited by F. Huber and H. Markl (Springer, Berlin), pp. 321–331.

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing* (Academic, San Diego).

Nelson, B. S. (**2003**). "Reliability of sound attenuation in Florida scrub habitat and behavioral implications," J. Acoust. Soc. Am. **113**, 2901–2911.

Pan, N. X. (**2003**). "Excess attenuation of an acoustic beam by turbulence," J. Acoust. Soc. Am. **114**, 3102–3111.

Peng, C., and Lines, J. A. (**1995**). "Noise-propagation in the agricultural environment," J. Agric. Eng. Res. **60**, 155–165.

Piercy, J. E., Embleton, T. F. W., and Sutherland, L. C. (**1977**). "Review of noise propagation in the atmosphere," J. Acoust. Soc. Am. **61**, 1403–1418.

Roberts, J., Kacelnik, A., and Hunter, M. L. (**1979**). "Model of sound interference in relation to acoustic communication," Anim. Behav. **27**, 1271–1273.

Slabbekoorn, H., and Peet, M. (**2003**). "Ecology: Birds sing at a higher pitch in urban noise—Great tits hit the high notes to ensure that their mating calls are heard above the city's din," Nature (London) **424**, 267–267.

Slabbekoorn, H., Yeh, P., and Hunt, K. (**2007**). "Sound transmission and song divergence: A comparison of urban and forest acoustics," Condor **109**, 67–78.

Spiegel, M. F. (**1981**). "Thresholds for tones in maskers of various bandwidths and for signals of various bandwidths as a function of signal frequency," J. Acoust. Soc. Am. **69**, 791–795.

Swaddle, J. P., and Page, L. C. (**2007**). "High levels of environmental noise erode pair preferences in zebra finches: Implications for noise pollution," Anim. Behav. **74**, 363–368.

Taherzadeh, S., Li, K. M., and Attenborough, K. (**1998**). "Some practical considerations for predicting outdoor sound propagation in the presence of wind and temperature gradients," Appl. Acoust. **54**, 27–44.

Wiley, R. H., and Richards, D. G. (**1978**). "Physical constraints on acoustic communication in atmosphere—Implications for evolution of animal vocalizations," Behav. Ecol. Sociobiol. **3**, 69–94.

Wiley, R. H., Richards, D. G., and Kroodsma, D. E. (**1982**). "Adaptations for acoustic communication in birds: Sound transmission and signal detection," in *Acoustic communication in Birds: Communication and Behavior* (Academic, New York), pp. 131–181.

Zaporozhets, O., Tokarev, V., and Attenborough, K. (**2003**). "Predicting noise from aircraft operated on the ground," Appl. Acoust. **64**, 941–953.

Harmonic pulsed excitation and motion detection of a vibrating reflective target[a)]

Matthew W. Urban[b)] and James F. Greenleaf
Department of Physiology and Biomedical Engineering, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, Minnesota 55905

(Received 22 June 2007; revised 12 October 2007; accepted 13 October 2007)

Elasticity imaging is an emerging medical imaging modality. Methods involving acoustic radiation force excitation and pulse-echo ultrasound motion detection have been investigated to assess the mechanical response of tissue. In this work new methods for dynamic radiation force excitation and motion detection are presented. The theory and model for harmonic motion detection of a vibrating reflective target are presented. The model incorporates processing of radio frequency data acquired using pulse-echo ultrasound to measure harmonic motion with amplitudes ranging from 100 to 10,000 nm. A numerical study was performed to assess the effects of different parameters on the accuracy and precision of displacement amplitude and phase estimation and showed how estimation errors could be minimized. Harmonic pulsed excitation is introduced as a multifrequency radiation force excitation method that utilizes ultrasound tonebursts repeated at a rate f_r. The radiation force, consisting of frequency components at multiples of f_r, is generated using 3.0 MHz ultrasound, and motion detection is performed simultaneously with 9.0 MHz pulse-echo ultrasound. A parameterized experimental analysis showed that displacement can be measured with small errors for motion with amplitudes as low as 100 nm. The parameterized numerical and experimental analyses provide insight into how to optimize acquisition parameters to minimize measurement errors. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2805666]

PACS number(s): 43.80.Vj, 43.25.Qp, 43.35.Yb [FD] Pages: 519–533

I. INTRODUCTION

A. Elasticity imaging

Elasticity imaging is a new medical imaging modality that seeks to noninvasively produce high-resolution images of the spatial distribution of the stiffness or elasticity of human tissue. For centuries, the practice of medicine has relied on the use of palpation to detect abnormalities in tissue stiffness because studies have shown that stiffness is related to pathology in different types of soft tissues.[1–3]

Many different methods have been created to perform elasticity imaging each with its own advantages and drawbacks. The two common components to all elasticity imaging methods are an induced stress and a measurement of the resulting deformation. Elasticity imaging methods have been reviewed by Greenleaf *et al.* and Parker *et al.*.[4,5]

The use of ultrasound radiation force has emerged as a way to noninvasively induce a localized stress in tissue. Ultrasound imaging methods have then been used to measure the resulting motion. Sugimoto *et al.* reported using impulsive radiation force as a way to excite tissue to investigate their stiffness.[6]

Walker *et al.*, reported using repeated tonebursts of ultrasound with duration of 2 μs to create radiation force in gel phantoms for mechanical characterization of the phantoms.[7] Nightingale *et al.*, demonstrated that short tonebursts (0.01–1 ms) of ultrasound could be used to generate ultrasound radiation force, and then the tissue motion could be tracked using pulse-echo ultrasound and correlation methods.[8] In this process, called acoustic radiation force impulse (ARFI) imaging method, the radiation force "push" pulse is transmitted, and then the scanner switches to B-mode imaging using pulse-echo ultrasound.[8] The raw radio frequency (rf) data are then used to perform cross correlation to estimate the displacement of the tissue. The push pulse is then repeated at other positions and the motion tracking ensues from the corresponding radiation force application. The pushing pulse never occurs while motion tracking is being performed.

Supersonic shear imaging is a method that uses a radiation force pulse moving at a supersonic speed, with respect to the shear wave speed, to create shear waves in tissue.[9] An ultrafast imaging system measures the propagation of the shear waves and the shear wave speed can be used to evaluate the shear modulus of the tissue.

The aforementioned methods use what is termed a static or quasi-static acoustic radiation force. Static radiation force can be generated with continuous wave ultrasound. In ARFI, a quasi-static radiation force can be generated using toneburst of ultrasound where the tissue is pushed with a fixed force, released, and then the tissue relaxes.[8] Dynamic ultrasound radiation force has been explored as another method to induce motion in tissue. Amplitude modulated ultrasound was used to generate radiation force to create shear waves in tissue and the shear wave speed was measured by another ultrasound transducer and used to find the elastic properties

[a)]Portions of this work were presented in "Motion detection for vibroacoustography," 151st Meeting of the Acoustical Society of America.
[b)]Author to whom correspondance should be addressed. Electronic mail: urban.matthew@mayo.edu

of the tissue.[10] Fatemi and Greenleaf introduced a method to create dynamic radiation force that uses ultrasound beams at slightly different frequencies, f_0 and $f_0+\Delta f$, where f_0 is typically in the megahertz range and Δf usually ranges from a few hundred Hertz to the kilohertz range.[11,12] When these two beams interfere in the focal region of the transducer a dynamic radiation force is created at frequency Δf.[13] This force locally vibrates the object in the focal region, and the resulting motion can be measured with a method called vibrometry. Vibro-acoustography uses this same force to create vibration which creates an acoustic field called acoustic emission that can be detected with a hydrophone.

Localized harmonic motion (LHM) imaging, proposed by Konofagou *et al.*, uses either amplitude modulated ultrasound or the interaction of two ultrasound beams to produce dynamic radiation force, and a separate transducer confocal with the transducer(s) producing the radiation force is used to perform pulse-echo ultrasound for motion detection.[14,15] Since motion detection is performed simultaneously with radiation force excitation, the radiation force is produced using ultrasound at 2.27 MHz, and the pulse-echo motion detection is performed at 1.1 MHz so that signals from the radiation force could be separated from those used for motion tracking.

Michishita *et al.*, proposed using gated tonebursts of amplitude modulated ultrasound to produce dynamic radiation force. In between these excitation tonebursts, pulse-echo ultrasound was performed to detect the induced motion. In their study, separate 5.0 MHz transducers were used for radiation force excitation and motion detection.[16]

B. Vibrometry

Vibrometry provides complementary information to the acoustic emission information obtained during vibro-acoustic imaging. Information about the velocity or displacement could be used in conjunction with models to extract information about material properties of the object or tissue under inspection.

Vibration phase information has been shown to be useful in differentiating between materials of different mass density.[17] Another area in which the vibration phase can be used is in shear wave speed measurement. Chen *et al.*[18] demonstrated that shear wave speed, c_s, can be measured by measuring the phase difference of the propagating harmonic shear wave of frequency ω_s at two different locations

$$c_s = \frac{\omega_s \Delta r}{\Delta \phi}, \tag{1}$$

where $\Delta\phi=\phi(r_1)-\phi(r_2)$ is the difference in phase at two different locations, r_1 and r_2, and $\Delta r=r_1-r_2$ is the distance between measurement points. The shear wave speed measured at different frequencies can be used to find the material properties of the tissue using[19]

$$c_s = \sqrt{\frac{2(\mu_1^2+\omega_s\mu_2^2)}{\rho(\mu_1+\sqrt{\mu_1^2+\omega_s\mu_2^2})}}, \tag{2}$$

where μ_1 and μ_2 are the shear elasticity and viscosity and ω_s is the angular frequency.

There are some limitations to current practice of vibrometry. In studies examining the mechanical response of spheres and tubes to radiation force, a Doppler laser vibrometer was used to measure the resulting motion.[17,20–23] The laser vibrometer is only appropriate for measuring the motion of an entire object or the surface of an object, and is not suitable for the study of tissue because the laser is not able to penetrate deep into the tissue.

In studies involving the measurement of shear waves or propagating waves along an artery, separate transducers were used for radiation force excitation and pulse-echo ultrasound motion detection.[18,24] In the LHM method and the method proposed by Michishita *et al.*, separate transducers are also used for radiation force and motion detection.[14–16] However, using two separate transducers may not be desirable in clinical situations.

Some methods of radiation force excitation and motion detection have been performed with a single transducer. In ARFI, the same transducer is used for radiation force excitation and motion detection. LHM has been performed with a phased array for thermal surgery monitoring using radiation force excitation with 1.1 MHz ultrasound and motion detection with pulse-echo ultrasound at the fifth harmonic (4.86 MHz).[25]

The goals of this paper are to describe new methods for radiation force excitation and motion detection. We will present a model for motion detection of a reflective target that is applicable to vibrometry calibration, nondestructive evaluation applications, and investigation of reflective targets in the human body such as vessel walls. The model uses numerically generated rf data to assess the effects of different parameters on the performance of amplitude and phase estimation of harmonic vibration. We will describe and characterize a radiation force excitation method which we call harmonic pulsed excitation (HPE) that utilizes repeated tonebursts of ultrasound to produce a multifrequency radiation force. Last, we will present an experimental method to exert the multifrequency force produced by HPE on a spherical target and measure the resulting motion with a single transducer. A parameterized experimental analysis was performed, and the results are compared with the model results.

II. METHODS

A. Harmonic motion detection

The motion of a harmonically vibrating scatterer causes a Doppler shift in interrogating ultrasonic waves.[26] Huang *et al.*, reported that if the scatterer is vibrating with a velocity much lower than the wave speed of the backscattered waves and the vibration frequency is much lower than the interrogating acoustic waves, then the spectrum of the detected motion can be modeled as a frequency modulated spectrum.[26]

Many groups have used this model to estimate motion with knowledge about the measured spectrum[26–30] using continuous wave (cw) ultrasound. However, it was noted that nonlinear propagation of cw ultrasound could produce effects that mimic motion in the Doppler spectrum and confuse the results.[28–30] Because the nonlinear propagation produces effects similar to harmonic motion, the frequency range for

which reliable motion detection is possible is limited. This problem can largely be alleviated by using pulse-echo ultrasound and comparing consecutive echoes to detect the motion of harmonically vibrating scatterers.

Ultrasonic waves at frequency ω_f are used to interrogate a scatterer or a collection of scatterers vibrating with frequency ω_s, where $\omega_s \ll \omega_f$. For the notation in this paper the subscript "*f*" will refer to fast time corresponding to the ultrasonic time scale on the order of microseconds and the subscript "*s*" will refer to slow time corresponding to the vibration time scale on the order of milliseconds. The displacement and the velocity of the vibrating scatterer is modeled as

$$D(t_s) = D_0 \sin(\omega_s t_s + \phi_s), \tag{3}$$

$$v(t_s) = v_0 \cos(\omega_s t_s + \phi_s), \tag{4}$$

where D_0 is the displacement amplitude, t_s is slow time, ϕ_s is the vibration phase, and v_0 is the velocity amplitude ($v_0 = D_0\omega_s$). If an ultrasonic pulse at frequency ω_f is used, the echo from the vibrating scatterer can be modeled[31]

$$r(t_f,t_s) = A(t_f,t_s)\cos(\omega_f t_f + \phi_f + \beta \sin(\omega_s t_s + \phi_s)), \tag{5}$$

where t_f is fast time, A is the echo amplitude, ϕ_f is the initial phase of the ultrasound signal, and β is defined as

$$\beta = \frac{2D_0\omega_f \cos(\theta)}{c}, \tag{6}$$

where θ is the Doppler angle, which will always be assumed to be 0°, and c is the longitudinal sound speed of the medium.

The quantities to be determined in measurements are D_0 and ϕ_s, but both variables are embedded in the phase of the ultrasound echo in Eq. (5). Many groups have approached this problem from the standpoint that a phase shift produces a time shift between consecutive echoes. Therefore, if the phase or time shift can be estimated, the displacement amplitude and vibration phase could also be estimated.

Two early methods used a two-dimensional autocorrelation approach.[32,33] The one striking difference between the algorithm proposed by Loupas *et al.* and Kasai *et al.* is that the Loupas method corrects for the mean ultrasound echo center frequency whereas the Kasai method assumes the transmitted ultrasound center frequency. For applications where displacement occurs as a result of impulsive radiation force excitation or static mechanical compression, cross-correlation combined with interpolation and spline estimation techniques have been used.[34–40] The interpolation and especially the spline based methods can be computationally expensive.

In the method proposed by Zheng *et al.*[31] quadrature demodulation is used to obtain the signal $y(t_s) = \beta \sin(\omega_s t_s + \phi_s)$. However, for short echo signals the low-pass filtering step in the quadrature demodulation can introduce artifacts because of transient effects of the filters. A different method was used in these studies to obtain the $y(t_s)$ signal.

Hasegawa and Kanai proposed a cross-spectrum method that corrects for the center frequency of the ultrasonic echo to evaluate the displacement tracked using pulse-echo ultrasound.[41] If we denote the nth and $(n+1)$-th received echoes as $r(n)$ and $r(n+1)$ and their corresponding frequency spectrums as $R_n(f)$ and $R_{n+1}(f)$ then the cross spectrum is calculated as

$$R_n^*(f)R_{n+1}(f) = |R_n(f)||R_{n+1}(f)|e^{j\Delta\theta_n(f)}, \tag{7}$$

where $\Delta\theta_n(f)$ is the phase shift between the two echoes and * represents complex conjugation. The motion of the vibrating scatterers can then be extracted by performing this cross-spectral analysis for all echoes after each echo was windowed using a Hann window. The velocity can be estimated by

$$v_n = \frac{c \cdot \Delta\theta_n(f_0)}{2\omega_f T_{\text{prf}}}, \tag{8}$$

where the phase shift is evaluated at the center frequency, f_0, of the cross-spectrum, and T_{prf} is the pulse repetition period of the pulse-echo interrogation. The center frequency of the echo is estimated by finding the frequency at which the maximum occurs in the magnitude of the cross spectrum. The correction for the center frequency is performed because absorption mechanisms can downshift the center frequency of the received echo from that of the transmitted pulse. Hasegawa and Kanai showed that if no correction is performed for the center frequency the results can be biased.[41] Because the vibration is harmonic at ω_s, the displacement signal can be obtained by $D_n = v_n/\omega_s$.

Once D_n has been found, a Kalman filter is used to extract amplitude and phase of the displacement.[31] The Kalman filter only requires the vibration frequency, ω_s, as an input. The Kalman filter is a state space based filter that recursively estimates the state variables using a least mean squared error criteria.[31] The Kalman filter is implemented digitally and is computationally efficient.

B. Parameterized model of harmonic motion detection

Many parameters enter into modeling the motion detection of a vibrating reflective target. First, it is assumed that there will be only one scatterer that acts similar to a point reflector. The parameters that are expected to most affect performance of this method are the displacement amplitude, D_0, signal-to-noise ratio (SNR) of the ultrasound echoes, the number of cycles of vibration used, N_c, and the number of points sampled per vibration cycle, N_p.

Displacement amplitude is determined by the radiation force amplitude. Since radiation force is proportional to the ultrasound intensity, the force that can be used is limited in practice because of bioeffect concerns. The intensities that lie within the limits of the Food and Drug Administration (FDA) produce small displacements, <10 μm, in tissue. Therefore, the lower limit of displacement amplitude necessary to obtain reliable results needs to be determined.

The SNR of the ultrasound echoes will primarily be determined by scatterer backscatter strength and the electronic noise introduced in the pulse-echo system. For a reflective target the SNR is expected to be 30 dB or higher. In a scattering medium, it could be lower.

TABLE I. Parameter study default parameters.

D_0	Displacement amplitude	1000 nm
ϕ_s	Vibration phase	0°
N_c	Cycles of vibration	5
N_p	Sampled points per vibration cycle	20
f_v	Vibration frequency	200 Hz
f_f	Ultrasound frequency	9.0 MHz
F_s	Sampling frequency	100 MHz
BW	Transducer bandwidth	6.5%
l_g	Gate length	1.0 mm
c	Speed of sound	1480 m/s
N	Iterations	1000

The number of vibration cycles in each measurement directly affects acquisition time for a measurement, $T=N_c/f_v$, where T is the acquisition time and f_v is the vibration frequency. For static point measurements, this may not have much impact, but if this method is employed for imaging, the acquisition time for the image will be governed by N_c. Also, increasing N_c will increase the processing time for the displacement estimate.

The number of points sampled per vibration cycle, N_p, affects the pulse repetition frequency, f_{prf}, used in the experiment as $N_p=f_{\text{prf}}/f_v$. To satisfy the Shannon sampling theorem, $N_p \geqslant 2$. The value of f_{prf} is limited by the distance of the vibrating scatterer from the transducer, where the maximum pulse repetition frequency $f_{\text{prf},m}=c/2z$ where c is the sound speed of the medium and z is the axial depth of the vibrating scatterer from the transducer. Increasing the value of N_p will also increase processing time for the displacement estimate. Since N_c and N_p are dimensionless, the results of this model can be extended for any value of f_v and f_{prf} within the limits described above.

For this study, D_0 will vary from 100 to 10,000 nm, SNR will vary from 0 to 60 dB, N_c will vary from 3 to 20, and N_p will vary from 5 to 30. Default parameters for the study are given in Table I. A default value of $D_0=1000$ nm was chosen because this is a typical value seen in experimentation. In vibrometry experiments, low frequency vibration was on the order of a few hundred Hertz so the default value of f_v was chosen as 200 Hz. The values of $N_c=5$ and $N_p=20$ give values of $T=25$ ms and $f_{\text{prf}}=4.0$ kHz. The interrogating pulses have been windowed with a Gaussian window to reflect the bandwidth of the transducer, BW. The gate length, l_g, is the spatial extent used for comparison of consecutive echoes and phase shift calculation. The sampling frequency, F_s, is the fast time sampling frequency for the ultrasound echoes.

To evaluate the performance of the method, we will introduce two metrics, bias and jitter. These metrics originate from the estimation of time delays. The bias, x_B, is the mean of the error, and the jitter, σ_J, is the standard deviation of the error and they are calculated as

$$x_B = \frac{1}{N}\sum_{n=1}^{N}(x_n - x_T), \tag{9}$$

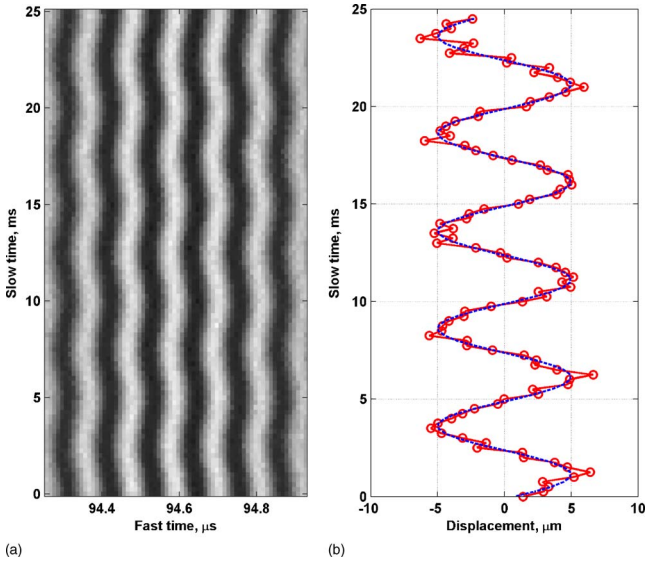


FIG. 1. (Color online) (a) Sample simulated echo data for vibration of a scatterer with $D_0=5000$ nm, $\phi_s=0°$, $N_c=5$, $N_p=20$, and SNR=20 dB, (b) displacement signal for data in (a). The solid curve with the data points marked by the open circles is the displacement signal estimated from the data, and the dashed curve is the true displacement signal. The estimated vibration amplitude and phase are $D_0=4990.2$ nm and $\phi_s=0.054°$ while the true values are $D_0=5000$ nm and $\phi_s=0°$.

$$\sigma_J = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - x_T - \bar{x})^2}, \tag{10}$$

where N is the number of data samples, x_T is the true value, and $\bar{x}$ is the mean of the data samples.[35] Minimal bias and jitter is desired. The bias reflects the accuracy of the motion detection, and the jitter reflects the precision of the motion detection. Bias and jitter measures will be evaluated on both displacement amplitude and phase for 1000 iterations with different initial conditions for the noise added to adjust the SNR. The added noise, $n(t_f,t_s)$, was normally distributed and added to the simulated ultrasound echo data

$$r(t_f,t_s) = A(t_f,t_s)\cos(\omega_f t_f + \phi_f + \beta\sin(\omega_s t_s + \phi_s)) + n(t_f,t_s). \tag{11}$$

Figure 1(a) shows sample data for $D_0=5000$ nm, $\phi_s=0°$, $N_c=5$, $N_p=20$, SNR=20 dB, $f_f=9.0$ MHz, BW=6.5%, $l_g=1.0$ mm, and $F_s=100$ MHz. The displacement signal after the phase estimation is shown in Fig. 1(b). The dashed line represents the true displacement signal. The estimates for the vibration amplitude and phase for this case were $D_0=4990.2$ nm and $\phi_s=0.054°$ which are very close to the true values.

C. Parameterized model results

Figures 2(a) and 2(b) show the displacement amplitude and phase results, respectively, for the default conditions in Table I. Each data point represents the mean of the 1000 iterations and the error bars represent one standard deviation. The dashed lines in these figures show the target values. The distance of the mean values from the target value is larger at

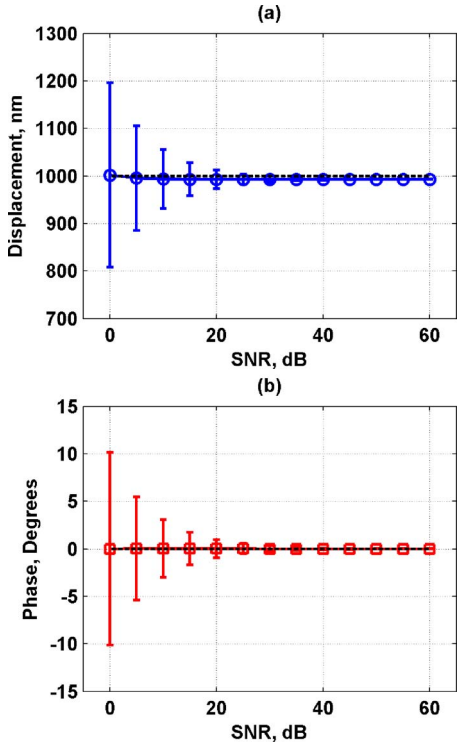


FIG. 2. (Color online) (a) Displacement amplitude results for default parameters in Table I. Each data point represents the mean of the 1000 iterations and the error bars represent one standard deviation. The dashed line represents the target value of D_0=1000 nm, (b) displacement phase results for default parameters in Table I. Each data point represents the mean of the 1000 iterations and the error bars represent one standard deviation. The dashed line represents the target value of ϕ_s=0°.

lower SNR indicating bias is dependent on SNR. The error bars which indicate the measurement jitter get smaller at higher values of SNR.

Figure 3 shows the displacement amplitude and phase bias and jitter while varying D_0. As SNR of the ultrasonic echoes increases the bias and jitter stabilize to a certain value. For the amplitude and phase jitter, increasing SNR yields lower jitter values. The magnitude of the amplitude bias increases as D_0 increases, but as a percent of D_0 it remains relatively low. The amplitude jitter does not change significantly with increasing values of D_0. As D_0 increases, the phase bias and jitter decrease. The phase bias is very low for all cases except D_0=100 nm. The phase jitter requires higher values of D_0 to decrease to levels comparable to the bias.

Figure 4 shows the results of the amplitude and phase bias and jitter for varying N_c. The amplitude bias does improve with increasing N_c but not to a significant degree. Increasing N_c does serve to decrease the amplitude bias at low SNR values but does not have a large effect at higher SNR. The phase bias was found to be consistently small for all values of N_c. The amplitude and phase jitter progressively decreases with increasing N_c.

Figure 5 shows the results for amplitude and phase bias and jitter for varying N_p. Increasing the value of N_p has a significant effect on decreasing the amplitude bias and has a less dramatic effect on decreasing the amplitude jitter. The phase bias is consistently low, but increasing N_p does serve to decrease the phase jitter.

D. Parameterized model discussion

The absolute value of the amplitude bias increased as D_0 increased, but if the bias was considered as a fraction of D_0, then the error can be considered small. The amplitude bias and jitter at D_0=100 nm are large enough to cause consider-

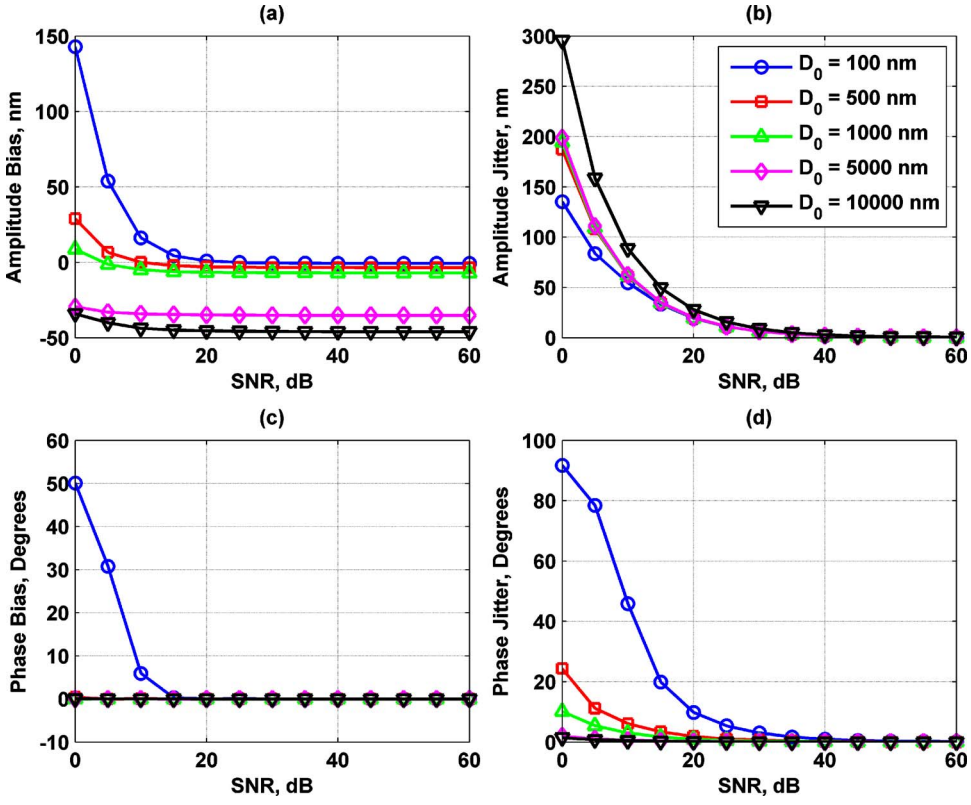


FIG. 3. (Color online) Displacement amplitude and phase bias and jitter for variation of D_0=100 (○), 500 (□), 1000 (△), 5000 (◇), and 10,000 (▽) nm. (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (b) applies to each panel.

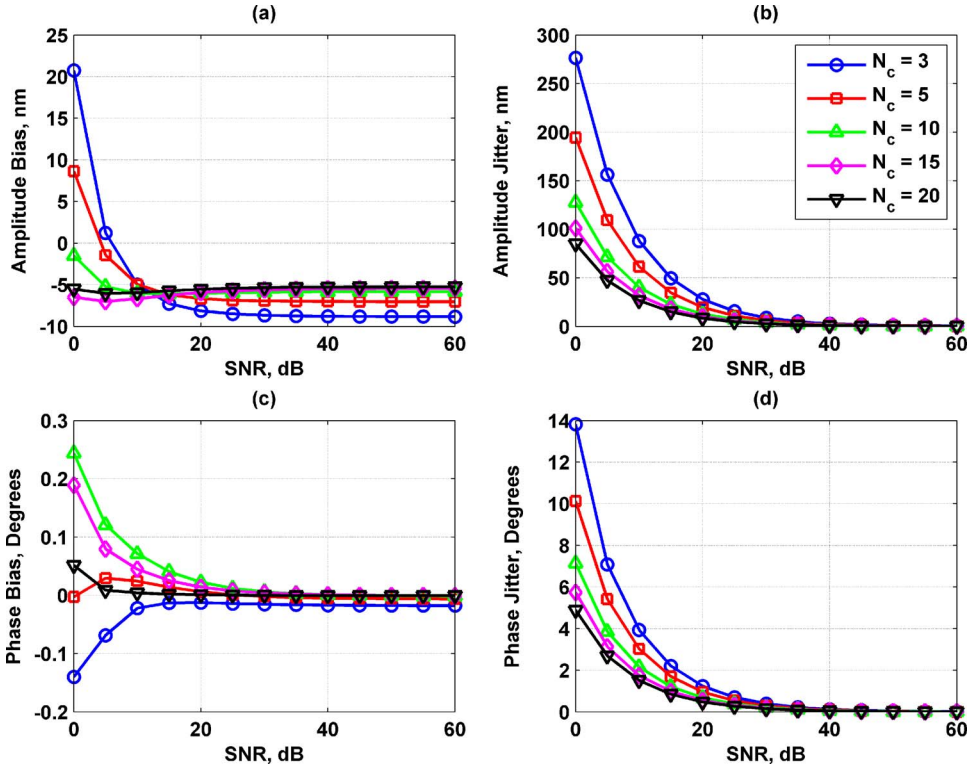


FIG. 4. (Color online) Displacement amplitude and phase bias and jitter for variation of N_c=3 (○), 5 (□), 10 (△), 15 (◇), and 20 (▽). (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (b) applies to each panel.

able error for that measurement to be trusted. The phase bias and jitter are only considerable for SNR $<$ 20 dB and D_0 $<$ 500 nm.

Increasing the value of N_c minimizes the amplitude and phase bias and jitter almost universally. Zheng *et al.*, showed that using more cycles of vibration reduced errors in shear wave speed measurements which are related to phase measurement errors.[31] One explanation for these results is that the Kalman filter acts as an averaging filter as estimation of the correct values improves as the filter is given more data to work with because of its recursive nature.

By increasing N_p, the results show a decrease in errors; however, for a given increase in N_p the bias or jitter does not decrease as much as in the case of increasing N_c. By sampling more points per vibration cycle, the harmonic function

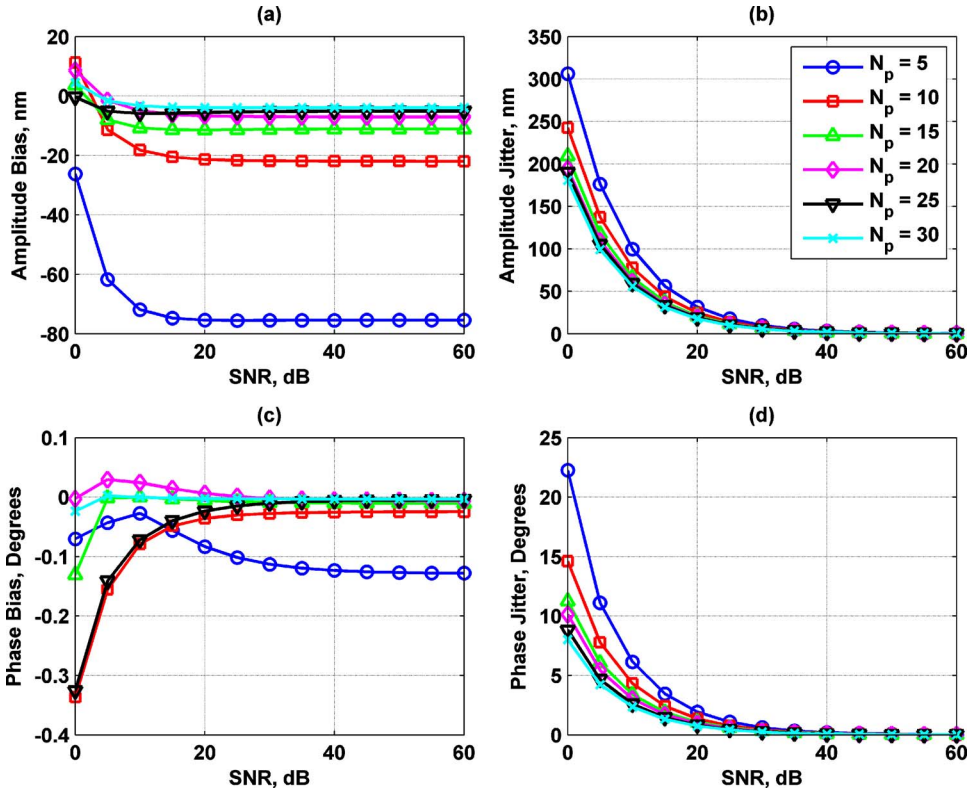


FIG. 5. (Color online) Displacement amplitude and phase bias and jitter for variation of N_p=5 (○), 10 (□), 15 (◇), 20 (◇), 25 (▽), and 30 (*). (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (b) applies to each panel.

becomes more sinusoidal in shape and the Kalman filter can operate on the improved data set to extract the amplitude and phase with better accuracy and precision.

The amplitude bias and the amplitude and phase jitter exhibited an exponential decay to a constant value as SNR increased. In almost all results, for SNR <30 dB, and varying any parameter the amplitude or phase jitter had larger values than the corresponding amplitude or phase bias. Above this threshold, the amplitude and phase jitter decreased to values that were comparable or lower than the associated bias.

The results from this parametric model study are useful for serving to optimize the implementation of this method in an experimental setting. The results show that to decrease displacement bias, D_0 should decrease and all other parameters should increase. To minimize displacement jitter and the phase bias and jitter, all parameters should be increased.

However, in an experimental setting, maximizing all of the parameters will have to be weighed against trade-offs associated with each parameter. To maximize D_0, the maximum safe ultrasonic intensity should be used to achieve maximal radiation force while preventing tissue or transducer heating. The increase of N_c will need to be weighed against time constraints of the experiment. If point measurements are being performed, then it may worthwhile to use 10–20 cycles of vibration, but in an imaging situation where many points are required to make an image, five cycles may be sufficient. Increasing the value of N_p will not affect acquisition time but will increase data size and processing time. However, this study showed that increasing N_p above 15 does not change the results significantly.

The dimensionless nature of N_c and N_p allow this analysis to be extended to different values of f_v. Two perspectives can be taken. For a given application and experiment, the values of D_0, N_c, N_p, and SNR used in the experiment can be used in the context of the model to find the level of error that can be expected. The other perspective is given a set of specifications for allowable bias and jitter, proper values of D_0, N_c, and N_p can be chosen for the experiment.

The method proposed by Hasegawa and Kanai for phase shift estimation with cross-spectral analysis is fast and does not suffer from finding false peaks as might be encountered with cross-correlation techniques. This method also avoids the errors introduced by low-pass filter transient effects in quadrature demodulation. The method also corrects for the center frequency of the ultrasound echo that may change due to frequency dependent attenuation during wave propagation. This correction reduces potential bias errors.

The Kalman filter based on harmonic vibration at a known frequency, ω_s, is very powerful and provides a robust estimate even in the face of noise as shown in Fig. 1. Another advantage of the Kalman filter is that it can be used to process multifrequency vibration data by processing the same data with different values of ω_s used as an input.

These simulations on a reflective target provide a model for motion detection in applications such as detecting wave motion in vessels,[24] for phase aberration correction methods based on tissue vibration,[42] nondestructive evaluation, and model-based analysis for objects or tissue regions that are ultrasonically reflective.

This model also provides a basis to analyze the vibration phase and the associated error in measurements when the phase measurements are used to estimate shear wave speed.[8,31]

E. Harmonic pulsed excitation

Harmonic pulsed excitation (HPE) is a new method that combines attributes of ARFI and the method proposed by Michishita and his colleagues. In this method, gated tonebursts of ultrasound are applied in a repetitive manner to produce a dynamic radiation force. In between the tonebursts, pulse-echo ultrasound is used to obtain rf data for motion detection. This method does not require amplitude modulation of the ultrasound used for radiation force generation, which eliminates the need for a modulating signal. Another attribute is that radiation force excitation and motion detection can be performed with the same transducer as in the ARFI method.

A timing diagram of HPE is shown in Fig. 6. The transmission of the ultrasound tonebursts for radiation force excitation is shown in Fig. 6(a) and the radiation force wave form is shown in Fig. 6(b). The timing gates for the transmitted and received pulses for motion tracking are shown in Figs. 6(c) and 6(d), respectively. The excitation tonebursts have length T_b and are repeated at a rate of f_r, where $f_r = 1/T_r$ and T_r is the repetition period. The motion detection pulses are transmitted with a pulse repetition frequency of f_{prf} where $f_{\text{prf}} = 1/T_{\text{prf}}$ and T_{prf} is the pulse repetition period. There may also be a delay, t_d, for the onset of the transmission of the motion detection pulses.

The radiation force function for this type of excitation is proportional to the short-term time average of the energy density[12] which is proportional to a low-pass filtered version of the square of the ultrasound pressure, so the function becomes a rectified rectangular pulse train as shown in Fig. 6(b).

Using Bracewell's conventions, we can describe this radiation force function, $f(t)$, as the convolution of an impulse train, $\text{III}(t)$, with a time-offset rectangle function, $\text{II}(t)$,[43]

$$f(t) = f_r \text{III}(f_r t) \otimes a\text{II}\left(\frac{t - T_b/2}{T_b}\right), \tag{12}$$

where a is the radiation force amplitude, which for this derivation will be unity and $\otimes$ indicates convolution. The Fourier transform of the radiation force function is

$$F(f) = af_r T_b \sum_{n=-\infty}^{\infty} e^{-i\pi T_b n f_r} \text{sinc}(T_b n f_r)\delta(f - nf_r), \tag{13}$$

where $\delta(\cdot)$ is the impulse function.

A few things should be noted about the spectrum described in Eq. (13). First, there are components at all multiples of f_r. Therefore, this method is a multifrequency excitation method. The velocity or displacement at vibration frequencies $f_v = nf_r$ can be analyzed with the method detailed previously by processing the data with the Kalman filter and

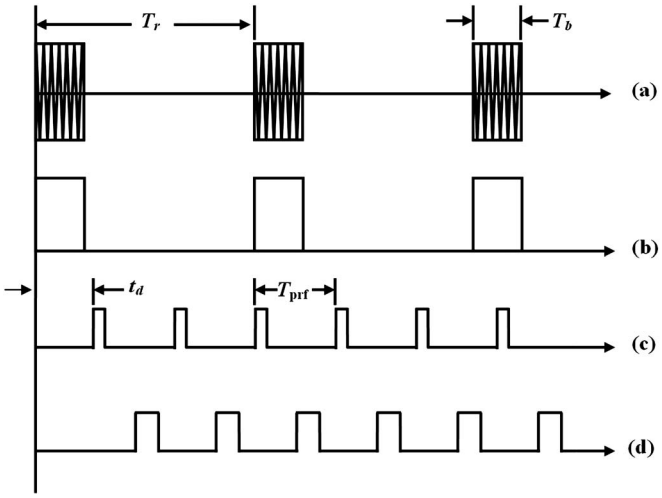


FIG. 6. Timing diagram for harmonic pulsed excitation and pulses used for motion tracking. (a) Ultrasound tonebursts with length T_b and repetition period of T_r, (b) radiation force produced by ultrasound tonebursts, (c) transmission gate for ultrasound tracking pulses with an onset delay of t_d and repetition period of T_{prf}, (d) reception gate for echoes of transmitted tracking pulses.

choosing an appropriate vibration frequency input. Second, the magnitude of those components is modulated by a sinc shaped envelope with zeros at frequencies that are multiples of $1/T_b$. The magnitude spectrum for the radiation force function, $|F(f)|$, with $T_b=200$ μs and $f_r=500$ Hz is shown in Fig. 7(a).

III. EXPERIMENTAL METHODS

A. HPE characterization experiment

To verify the shape of the radiation force function and its spectrum, an experiment was performed. The ultrasonic

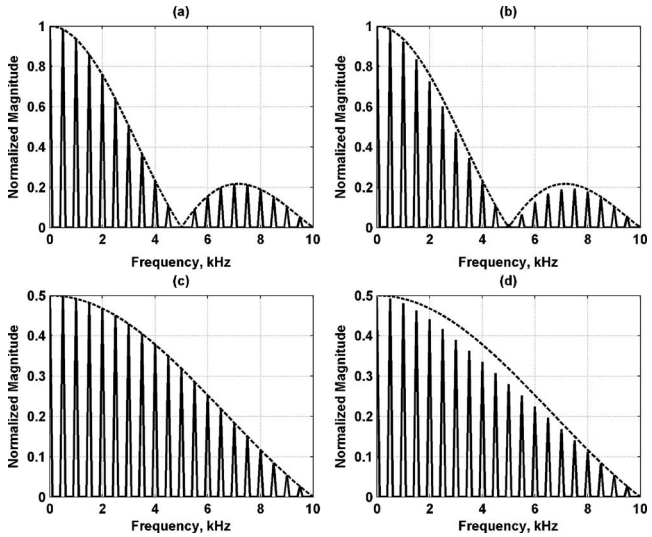


FIG. 7. Measured and calculated normalized magnitude spectrum of radiation force function for HPE with $f_r=500$ Hz μs. The dashed curves are the theoretical sinc envelope, (a) calculated magnitude spectrum for $T_b=200$ μs normalized to maximum value, (b) measured magnitude spectrum for $T_b=200$ μs normalized to maximum value, (c) calculated magnitude spectrum for $T_b=100$ μs normalized to maximum value in calculations for $T_b=200$ μs, (d) measured magnitude spectrum for $T_b=100$ μs normalized to maximum value in measurements for $T_b=200$ μs.

pressure from the transducer was measured from a two-element confocal 3.0 MHz transducer with outer diameter of 45 mm and a focal length of 70 mm. The two elements were driven with the same signal. The pressure was measured using a needle hydrophone (NTR Systems, Seattle, WA) in a large water tank. The needle hydrophone was placed at the focus of the transducer for these pressure measurements.

A HPE sequence was used with $f_r=500$ Hz and $T_b=200$ μs. To find the low-frequency radiation force, the squared pressure wave form was low-pass filtered and its frequency spectrum was calculated. The spectrum of the measured radiation force function was compared with a theoretical calculation using the given parameters, f_r and T_b. Figure 7 shows the differences in the spectrum of using tonebursts with $T_b=100$ μs. The two spectra are normalized with respect to the maximum value in the spectrum for the $T_b=200$ μs. A dashed line representing the envelope of the sinc function is also shown.

The measured results in Fig. 7 match very well with the theoretical calculations of the radiation force function. At a few frequencies in the measured radiation force function, the radiation force is underestimated. When the toneburst length is decreased by a factor of 2, the magnitude of the radiation force function is reduced by a factor of 2 as Eq. (12) predicts. Also, the zero positions of the sinc function occur at different frequency positions. For $T_b=200$ μs, the zeros occur at multiples of 5.0 kHz, and in the case of $T_b=100$ μs, the zeros occur at multiples of 10.0 kHz.

B. HPE and motion detection experiment

To assess the performance of the HPE and motion detection methods, an experiment was performed using a 440-C stainless steel sphere of diameter 1.59 mm embedded in a gelatin phantom made using 300 Bloom gelatin powder (Sigma-Aldrich, St. Louis, MO) with a concentration of 10% by volume. A preservative of potassium sorbate (Sigma-Aldrich, St. Louis, MO) was also added with a concentration of 10 g/L.

The sphere was placed in the focal region of the transducer. The resulting motion was detected with a Doppler laser vibrometer (Polytec, Waldbronn, Germany) which is used as the gold standard. A block diagram of this experimental setup is shown in Fig. 8(a).

The HPE method was initiated using a trigger signal from a signal generator producing one cycle of a transistor-transistor logic (TTL) pulse. This master trigger signal initiated a signal generator (33120A, Agilent, Palo Alto, CA) to produce a 20 cycle rectangular pulse train with a frequency of $f_r=100$ Hz. This pulse train triggered a signal generator (33120A, Agilent, Palo Alto, CA) that produced a toneburst of length $T_b=50$ or 100 μs. The voltage amplitude of this toneburst was varied to change the radiation force amplitude. The master trigger also triggered another signal generator (33250A, Agilent, Palo Alto, CA) to produce a rectangular pulse train with a frequency of f_{prf}. The value of the f_{prf} was varied to values of $f_{\text{prf}}=2.0$, 2.5, 3.0, and 4.0 kHz during the experiment to assess its effects on the results. This pulse train was used as a trigger input to the analog-to-digital converter

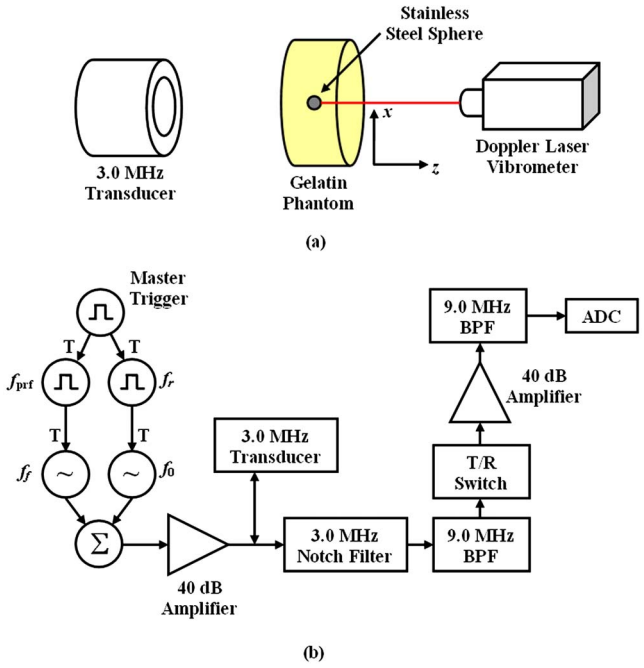


FIG. 8. (Color online) Experimental setup for harmonic pulsed excitation and motion detection. (a) Experimental setup for excitation and measurement of motion of a stainless steel sphere embedded in a gelatin phantom. The 3.0 MHz transducer creates the radiation force and in a later experiment will also be used to track the motion ultrasonically. The Doppler laser vibrometer provides a calibrated measurement of the sphere's motion. (b) For the excitation, a pulse train with frequency f_r=100 Hz is initiated, and each positive pulse triggers a toneburst of ultrasound at f_0=3.0 MHz. (The T indicates a trigger input.) For the tracking a pulse train at f_{prf} is initiated with specified time delay, t_d, and each positive pulse triggers a three cycle pulse at f_f=9.0 MHz to be transmitted. The excitation and tracking signals are summed together and amplified before being sent to the 3.0 MHz transducer. For tracking, the gated echoes are filtered with a notch filter centered at 3.0 MHz and a bandpass filter centered at 9.0 MHz before passing through a transmit/receive (T/R) switch. The signal is amplified and filtered again with a bandpass filter centered at 9.0 MHz before being sent to the digitizer (ADC).

(ADC) board in a personal computer. The ADC produced a trigger signal for every pulse in the pulse train and this triggered a signal generator (33250A, Agilent, Palo Alto, CA) to generate a three cycle pulse at 9.0 MHz.

The radiation force toneburst and tracking pulse were added together using a hybrid junction (M/A-COM, Inc., Lowell, MA) and this signal was amplified with a 40 dB amplifier. This signal passed through a diode bridge to eliminate low-level noise from the amplifier and through a matching transformer to the transducer. The echoes were received and filtered with a 3.0 MHz notch filter with a 50% bandwidth and a 9.0 MHz bandpass filter with 33% bandwidth. The echoes then passed through a transmit/receive (T/R) switch, amplified by a broadband amplifier and finally filtered by another 9.0 MHz bandpass filter before being digitized at a sampling frequency of 100 MHz. A block diagram of the experimental setup of HPE and motion detection is shown in Fig. 8(b).

The pulse-echo motion detection is performed at 9.0 MHz to separate the motion detection signals from the excitation in frequency. This is possible because the transducer is air-backed and responds well at its third harmonic. The frequency response of the transducer was measured us-

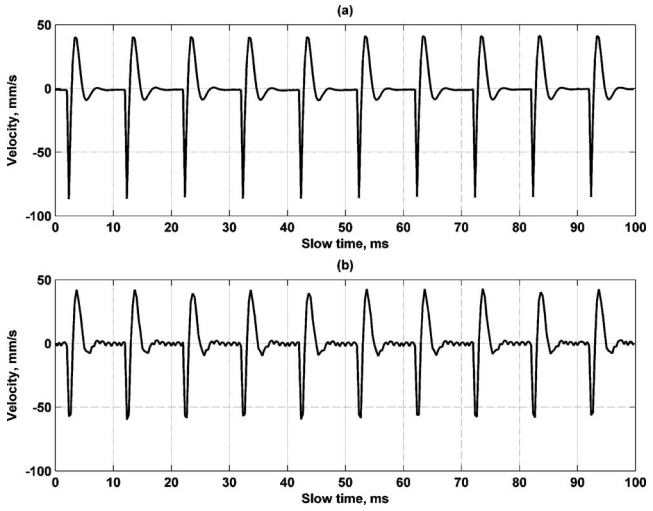


FIG. 9. Velocity of sphere measured by (a) Doppler laser vibrometer, (b) ultrasound based motion detection.

ing the needle hydrophone after excitation from a broadband pulser (5050PR, Panametrics, Waltham, MA). With reference to the peak at 3.0 MHz, the 9.0 MHz frequency component is only attenuated by 16 dB. This provides enough sensitivity to perform the pulse-echo measurements. Also, it is advantageous to perform the motion detection using a higher ultrasound frequency because small displacements can be detected without interpolation or other computationally expensive signal processing. Another group has reported performing a similar technique using a phased array and performed radiation force excitation at 1.1 MHz and tracked the motion with ultrasound near the fifth harmonic (4.86 MHz).[25]

A representative sample of the vibration data is shown in Fig. 9. The velocity is measured along the z direction as shown in Fig. 8. It is observed that the velocity response has a very short rise time, the sphere oscillates and the amplitude decays to rest until another excitation pulse is applied.

It was found through experimentation that a bandpass filter implemented before the Kalman filter was beneficial in obtaining better results. To process the data a windowed version of the displacement signal is used as the input to the Kalman filter. However, when a rectangular window is used, the starting sample of the window can cause large amounts of variation in the result. Therefore, a Hann window has been employed to window the data. The use of the Hann window decreases the strength of the signal by a factor of 2 as given by the equation for a Hann window,[44]

$$w[n] = \begin{cases} \frac{1}{2}\left[1 + \cos\left(\frac{2\pi n}{2M+1}\right)\right], & -M \leq n \leq M \\ 0, & \text{otherwise} \end{cases}. \quad (14)$$

As a result the final displacement result acquired after the Kalman filter is multiplied by two.

C. Experimental parameter analysis

To assess the effects of different parameters on the results, different variations of experimental parameters were performed. Results for two different values of T_b will be

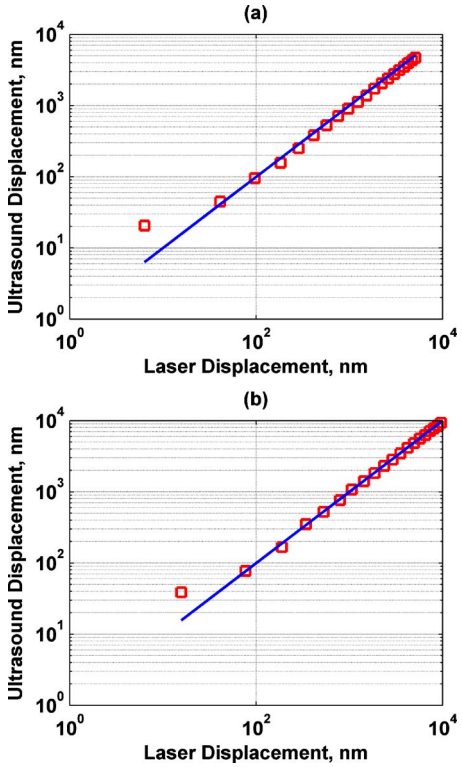


FIG. 10. (Color online) Comparison of displacement amplitude measured by laser vibrometer and ultrasound based detection for f_v=200 Hz using regression line $y=x$. Each data point represents the average of five measurements. (a) T_b=50 μs, R^2=0.9809, (b) T_b=100 μs, R^2=0.9970.

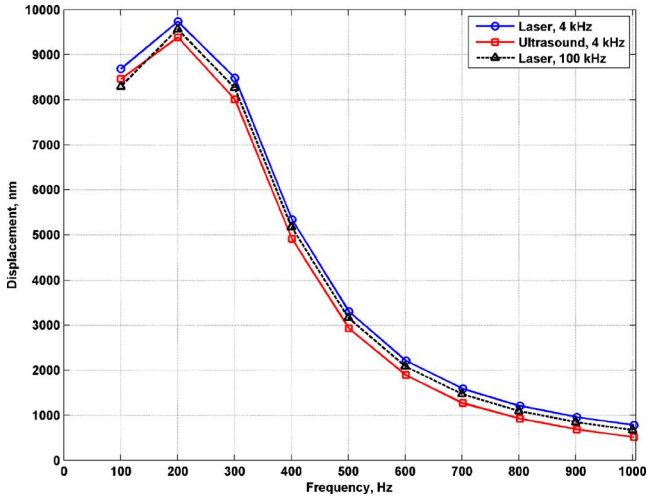


FIG. 11. (Color online) Multifrequency displacement measurements. Displacement motion was acquired with the laser vibrometer and ultrasound method with $F=F_0$, f_{prf}=4.0 kHz. The motion was also measured with the same radiation force excitation but the laser vibrometer signal was sampled at 100 kHz to provide an unaliased measurement.

shown and compared. Since different values of applied voltage for the excitation toneburst were used, the results can be compared versus the radiation force applied. For the purposes of reporting the results, the radiation force was normalized based on the maximum force produced by the highest voltage setting used and is denoted as F_0. The value of f_{prf} was varied to explore the differences in the results. Also, the value of T_s, the length of the temporal window used for analysis in slow time, was varied, such that the product N_cN_p was constant where N_c is the number of cycles analyzed and N_p is the number of samples per vibration cycle. For example, if T_s=50 ms, f_v=100 Hz, f_{prf}=4.0 kHz, then N_c=5 and N_p=40 and N_cN_p=200. If f_v is increased to 200 Hz, and all other values are not changed, N_p=20 and N_c=10 because twice as many cycles of vibration will occur in the same slow time window and N_cN_p=200. With this understanding of the parameter analysis, we can compare the results from the experiment with the simulation results. For the parameter analysis, the default values for analysis are $F=F_0$, f_v =200 Hz, f_{prf}=4.0 kHz, and T_s=50 ms.

D. Motion detection experimental results

Figure 10 compares the ultrasound based displacement versus the measured laser displacement for T_b=50, 100 μs. In this and following figures, each data point is the average of five measurements. The phase estimates had to be corrected for a frequency dependent phase shift because of the constant time delay associated with wave propagation between the transducer and sphere. The results were compared to the line $y=x$, which is the ideal case where the results from the laser and ultrasound measurements would match exactly. The R^2 values of the data fitted to this ideal regression line were calculated for each case. For T_b=50 μs, R^2 =0.9809 and for T_b=100 μs, R^2=0.9970. It was observed that the ultrasound measurements slightly underestimate the true displacement measured by the laser, but the regression fits are quite good. At around 100 nm, the data points start to deviate from the regression line. This represents a lower threshold for accurate measurement.

Figure 11 shows the results of multifrequency measurements made by the laser vibrometer and the ultrasound method. The motion was measured with the laser vibrometer and ultrasound methods with $F=F_0$ and f_{prf}=4.0 kHz. To ensure that no aliasing was occurring, the same radiation force was used, and the laser signal was sampled at 100 kHz. The results in Fig. 11 show the results from these three sets of measurements and good agreement was observed between the different measurements. Another notable feature was that the amplitude of vibration decreased with higher vibration frequencies.

Figure 12 shows the displacement amplitude and phase bias and jitter for T_b=100 μs. The results are plotted versus measured laser displacement for f_v=100, 200, 300, 400 Hz. The results show that amplitude bias and jitter decreases in measurements with large displacement amplitudes. The amplitude jitter is on the same order of the amplitude bias across all values of measured displacement amplitude. The phase bias is nearly constant for varying displacement amplitude for a given vibration frequency. The phase jitter decreases to very small values as displacement amplitude increases.

Figure 13 shows the displacement amplitude and phase results while varying the normalized radiation force to values of $F=F_0$, $F_0/2$, $F_0/4$, and $F_0/8$. The amplitude bias is larger for larger radiation force, probably because the displacement is larger. The amplitude jitter decreases with decreased force

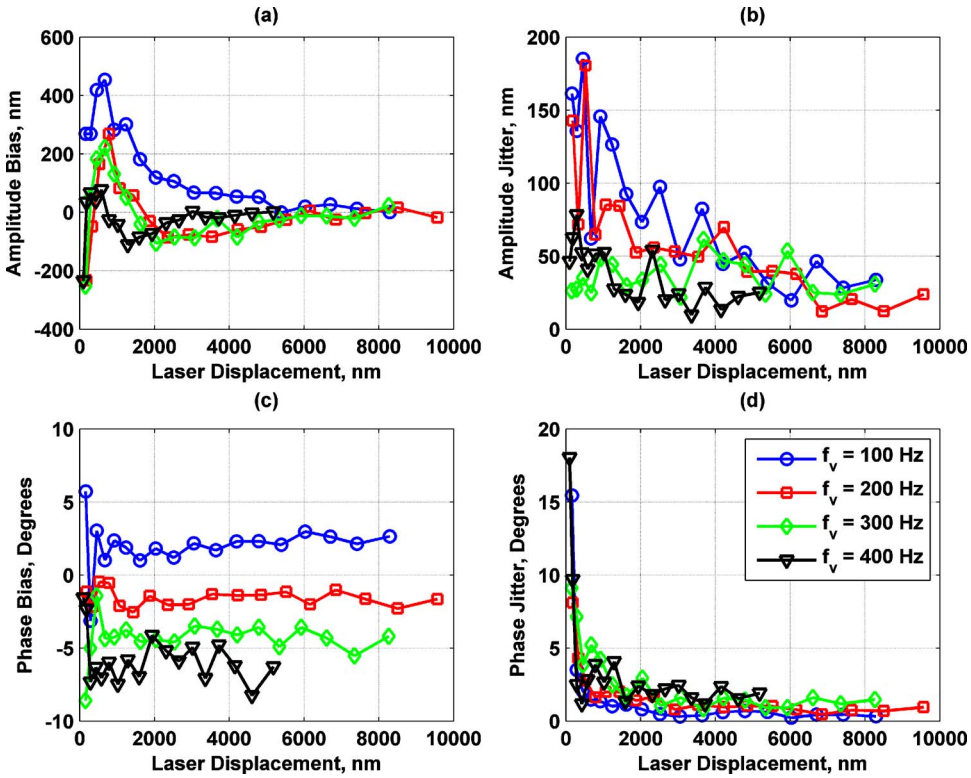


FIG. 12. (Color online) Displacement amplitude and phase bias and jitter for T_b=100 μs, F=F_0, f_{prf}=4.0 kHz, T_s=50 ms, and f_v=100 (○), 200 (□), 300 (△Δ), and 400 (▽) Hz. (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (d) applies to each panel.

while the phase bias and jitter do not change significantly for different levels of applied force. The phase bias exhibits a negative linear trend with increasing frequency. This increase may be a result of higher errors associated with decreased motion amplitude at higher frequencies as well as an unaccounted constant time delay. The phase jitter also increases with increasing frequency, probably because of errors associated with small motion amplitudes.

The value of f_{prf} was varied to values of f_{prf}=2.0, 2.5, 3.0, and 4.0 kHz and the results are shown in Fig. 14. As a function of frequency, the use of f_{prf}=2.0 kHz yielded results with larger amplitude and phase bias than other values of f_{prf}. The increase in f_{prf} provided graded differences in the displacement and phase bias. However, the use of f_{prf}=4.0 kHz gave much higher amplitude and phase jitter val-

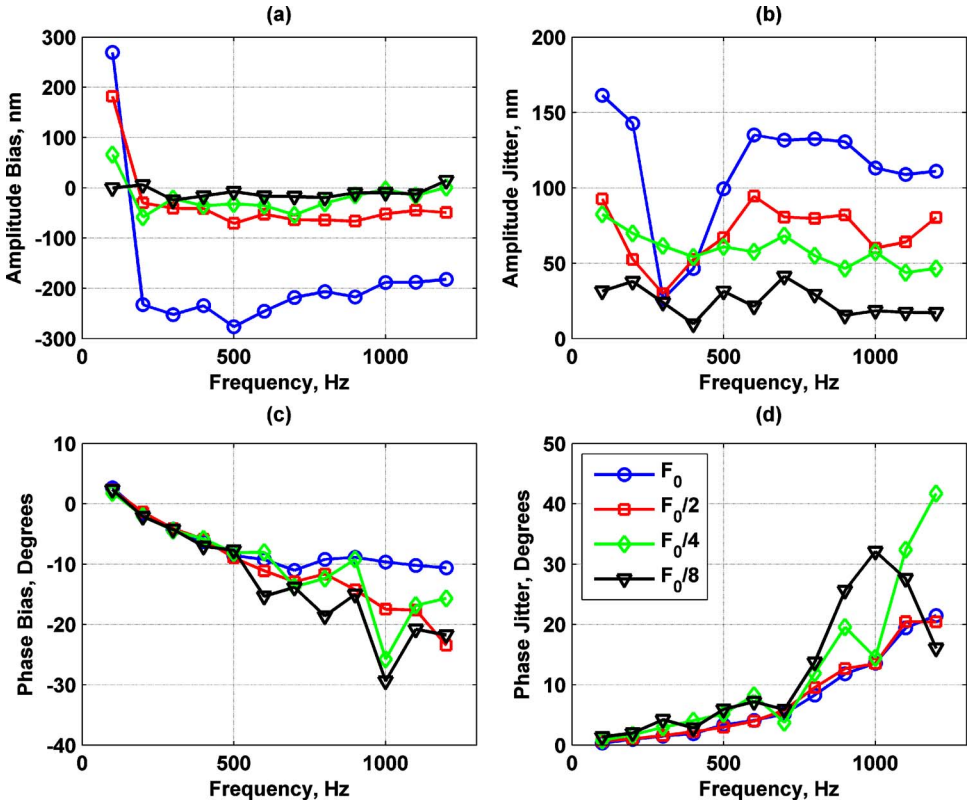


FIG. 13. (Color online) Displacement amplitude and phase bias and jitter for T_b=100 μs, f_{prf}=4.0 kHz, T_s=50 ms, and F=F_0 (○), $F_0/2$ (□), $F_0/4$ (△), and $F_0/8$ (▽). (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (d) applies to each panel.

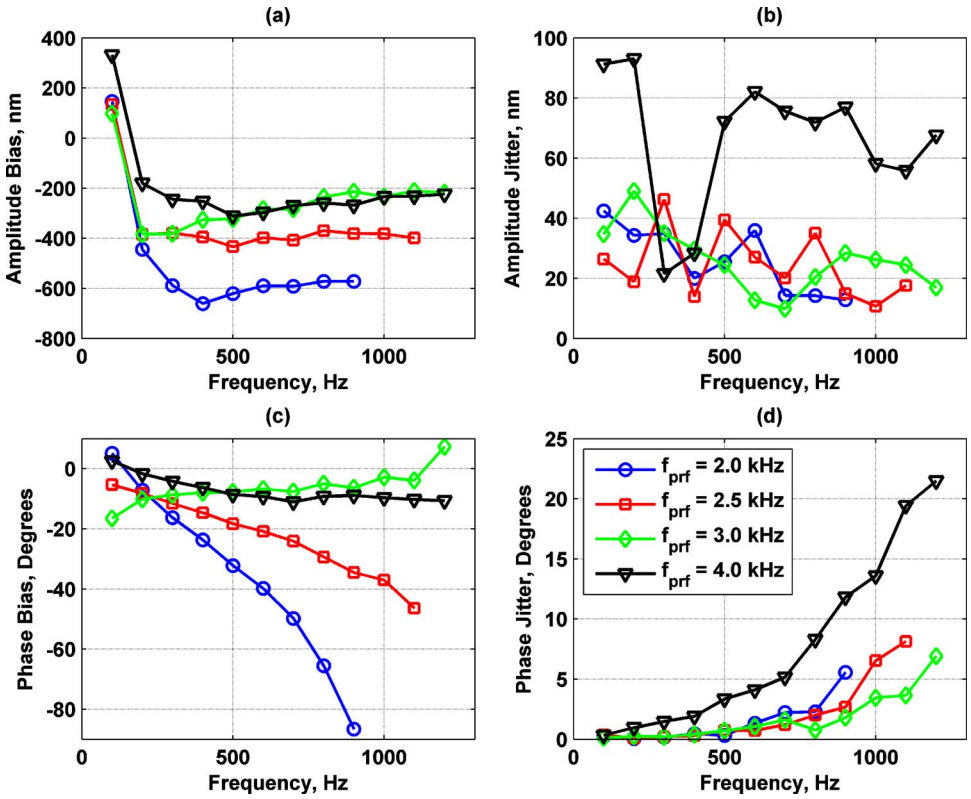


FIG. 14. (Color online) Displacement amplitude and phase bias and jitter for T_b=100 μs, $F=F_0$, f_v=200 Hz, T_s=50 ms, and f_{prf}=2.0 (○), 2.5 (□), 3.0 (△), 4.0 (▽) kHz. (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (d) applies to each panel.

ues. This may be due to interference between the excitation and tracking pulses. A value of f_{prf}=3.0 kHz was observed to give the best results.

Figure 15 shows the results of varying the length of the slow time processing window. The length of the window was varied to be T_s=30, 50, 100, 150 ms. A graded decrease was observed in the amplitude and phase jitter as the value of T_s increases. Phase bias seems unaffected by varying this parameter at low frequencies but shows graded differences at higher vibration frequencies.

E. Experimental discussion

Harmonic pulsed excitation is an effective method to provide multifrequency radiation force excitation. With an

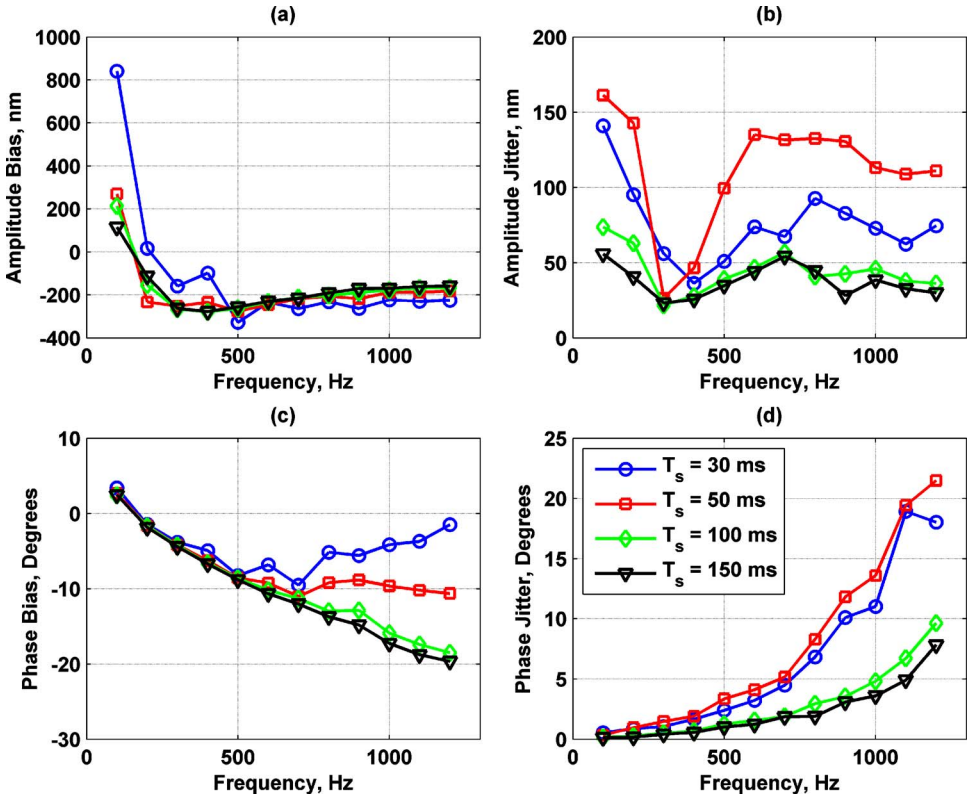


FIG. 15. (Color online) Displacement amplitude and phase bias and jitter for T_b=100 μs, $F=F_0$, f_v=200 Hz, f_{prf}=4.0 kHz, and T_s=30 (○), 50 (□), 100 (△), and 150 (▽) ms. (a) Amplitude bias, (b) amplitude jitter, (c) phase bias, (d) phase jitter. The legend in (d) applies to each panel.

analytic description, the method can be tailored to experimental situations and the radiation force function can be accurately estimated as shown by the results in Fig. 7.

The HPE method can also be used to obtain the same information acquired using the ARFI method by evaluating the response for the first pulse of the pulse sequence. However, this assumes that f_r is low enough to allow relaxation of the object or tissue so that analysis of the relaxation may be performed.

Along with the newly characterized excitation method, motion detection has been performed at multiple frequencies with the same data. This provides the prospect for dispersive measurements for applications that require such measurements. Figures 10 and 11 show that the results from the ultrasound based method provide results close to those measured by a laser vibrometer. This comparison provides confidence in the measurements made in the rest of the experiment.

A parameterized analysis of the method with experimental data was performed. A few general conclusions can be made from this analysis. The amplitude bias and phase bias and jitter decreased with increased value of displacement amplitude. These results agree with the model in that the phase jitter is reduced with increases in displacement amplitude. The model predicts higher amplitude bias and jitter with increased D_0 whereas the experimental results show the opposite. However, as a percentage of D_0, the bias and jitter are small for both the model and experimental results. Increased values of f_{prf} led to decreases in the amplitude and phase bias and jitter. This same trend was observed in the simulation results as N_p increased. The use of longer slow time processing windows lowered the amplitude and phase jitter values in a graded fashion, but overall, the bias was not sensitive to this parameter. As N_c was increased, the model yielded similar results. The model was mostly validated with the experimental results except for the behavior as D_0 was increased. In the future, a complete analytic model would be desired. The parameters used in this model would serve as inputs, and error propagation analysis would be performed, taking into account error associated with the time/phase estimation algorithm and the Kalman filter.

There are a few sources of error that may be encountered using HPE and ultrasound based motion detection on a reflective target. First, there is the risk of aliasing. If there is significant motion at high frequencies and the value of f_{prf} is low, then motion information from frequencies above the Nyquist sampling limit, $f_{\text{prf}}/2$, could alias down to corrupt the information at frequencies of interest. This can be prevented by increasing f_{prf} and/or reducing the radiation force used. When varying the value of f_{prf}, it was found that $f_{\text{prf}}=2.0$ kHz yielded results much different from those with higher values of f_{prf}. Aliasing errors may have been a factor since the Nyquist frequency in this case would only be 1.0 kHz. For most cases in biomedical applications, there is a viscous component to the tissue that acts as a low-pass filter, which essentially makes the tissue a physical anti-aliasing filter.

McAleavey *et al.*, have reported that in ARFI imaging measurement bias of the displacement can occur because of beam shapes used for excitation and motion detection.[45] They modeled the beam shapes with Gaussian functions and describe how only an average displacement estimate will be gained based on how much of the tracking beams intercept the excitation beam. If the tracking beam is wide with respect to the excitation beam, then more decorrelation or bias will result in the measurements. Ideally, the tracking beam should be much thinner than the excitation beam so that the tracking beam only intercepts the peak displacement of the scatterer(s). In this experiment, the full width at half maximum (FWHM) of the excitation beam is 0.80 mm and the FWHM of the tracking beam is 0.37 mm. These differences in the FWHM are accomplished with the same aperture, but the excitation beam is created using 3.0 MHz ultrasound while the tracking beam uses 9.0 MHz ultrasound, resulting in the decrease in the size of the beam.

Using the results from McAleavey *et al.*, with a ratio of excitation to tracking width of $W_x=W_y=2.19$, we should expect that the best results attainable would be tracking 92% of the peak displacement or would have at the least an 8% error.[45] However, the results presented show that the motion detection method obtained errors less than this theoretical threshold. The reflective spherical target serves to scatter the ultrasound from the tracking pulses directly back to the transducer and effectively decreases the tracking FWHM such that the motion detected using the ultrasound method is closer to the peak displacement as measured by the laser vibrometer.

One remaining error is the large phase bias that occurs with increasing frequency. The fact that the bias increases with frequency, linearly in most cases, leads us to believe that some systematic error in processing or data acquisition may be present. The laser signal was filtered with a low-pass Bessel filter with a cutoff frequency of 20.0 kHz, which would produce a linear phase shift with a 5.7° bias at 1.0 kHz.[46] The linear nature of the remaining phase bias may indicate a constant time delay in the processing electronics that could not be identified and corrected.

To produce adequate motion in the tissue, the intensity of the ultrasound toneburst can approach the limits set by the FDA. Measurements of one toneburst used for radiation force were performed with the 3 MHz transducer focused on a calibrated polymer polyvinylidene fluoride membrane hydrophone (GEC-Marconi Research Centre, Chelmsford, Essex, U.K.). For $T_b=100$ μs, the mechanical index calculated after derating by 0.3 dB/cm/MHz was 1.46 which is less than the FDA limit of 1.9.[47] The other FDA limit that must be met for diagnostic systems is the derated spatial-peak, temporal average intensity, $I_{\text{SPTA.3}}$, which is set at 720 mW/cm^2. (see Ref. 47) For the pulse sequence used, $I_{\text{SPTA.3}}=1442$ mW/cm^2 which is higher than the FDA limit. To use this pulse sequence for *in vivo* applications, either a shorter toneburst could be used at the same intensity or the intensity could be reduced. For an intensity that is less than the FDA limit, the maximum displacement may still be on the order of 5000 nm. For nondestructive evaluation applications, reduction of the intensity is not necessary. If tissue heating is evaluated for this pulse sequence using the bioheat equation without perfusion or convection described by Palm-

eri and Nightingale, the rise in temperature is 0.005 °C at the focus.[48] Implementation of this method for *in vivo* applications should not yield any negative bioeffects.

IV. CONCLUSIONS

The theory and model for motion detection of a vibrating reflective target were presented. Results from a parametric study show that the method works well and these results can be used for optimization of the method. Amplitude bias and jitter at $D_0 = 100$ nm were high enough to establish a theoretical lower limit of accurate estimation. The phase bias and jitter were only found to be considerable for the case when SNR < 20 dB and $D_0 < 500$ nm. It was shown that by maximizing displacement amplitude, SNR, N_c, and N_p will provide results that have low bias and jitter in estimation of the vibration amplitude and phase.

Harmonic pulsed excitation was introduced as a multifrequency radiation force excitation method that allows motion detection tracking to be implemented between excitation tonebursts. The multifrequency radiation force was characterized analytically and experimentally. Motion detection could reliably be performed for displacement with amplitudes down to 100 nm. An experimental analysis showed that motion can measured with low displacement and phase bias and jitter for displacement with larger amplitudes and longer slow time processing windows. Numerical and experimental parameterized analyses provided insight into how to optimize acquisition of the displacement data for minimization of errors for harmonic motion detection. This study introduced ways to use a multifrequency radiation force and sensitive harmonic motion detection for analyzing the mechanical response of tissue and other objects.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Shigao Chen and Randall Kinnick for experimental assistance, Dr. Yi Zheng for helpful correspondence and MATLAB code for the Kalman filter, and Jennifer Milliken for administrative assistance. This study was supported in part by Grants Nos. EB002640 and EB002167 from the National Institute for Biomedical Imaging and Bioengineering.

[1] T. A. Krouskop, T. M. Wheeler, F. Kallel, B. S. Garra, and T. Hall, "Elastic moduli of breast and prostate tissues under compression," Ultrason. Imaging **20**, 260–274 (1998).
[2] W. C. Yeh, P. C. Li, Y. M. Jeng, H. C. Hsu, P. L. Kuo, M. L. Li, P. M. Yang, and P. H. Lee, "Elastic modulus measurements of human liver and correlation with pathology," Ultrasound Med. Biol. **28**, 467–474 (2002).
[3] J. A. Schaar, C. L. de Korte, F. Mastik, L. C. van Damme, R. Krams, P. W. Serruys, and A. F. van der Steen, "Three-dimensional palpography of human coronary arteries. Ex vivo validation and in-patient evaluation," Herz **30**, 125–133 (2005).
[4] J. F. Greenleaf, M. Fatemi, and M. Insana, "Selected methods for imaging elastic properties of biological tissues," Annu. Rev. Biomed. Eng. **5**, 57–78 (2003).
[5] K. J. Parker, L. S. Taylor, S. Gracewski, and D. J. Rubens, "A unified view of imaging the elastic properties of tissue," J. Acoust. Soc. Am. **117**, 2705–2712 (2005).
[6] T. Sugimoto, S. Ueha, and K. Itoh, "Tissue hardness measurement using the radiation force of focused ultrasound," *1990 IEEE International Ultrasonics Symposium*, **1377–1380**, Honolulu, HI (1990).
[7] W. F. Walker, F. J. Fernandez, and L. A. Negron, "A method of imaging viscoelastic parameters with acoustic radiation force," Phys. Med. Biol. **45**, 1437–1447 (2000).
[8] K. R. Nightingale, M. L. Palmeri, R. W. Nightingale, and G. E. Trahey, "On the feasibility of remote palpation using acoustic radiation force," J. Acoust. Soc. Am. **110**, 625–634 (2001).
[9] J. Bercoff, M. Tanter, and M. Fink, "Supersonic shear imaging: A new technique for soft tissue elasticity mapping," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 396–409 (2004).
[10] A. P. Sarvazyan, O. V. Rudenko, S. D. Swanson, J. B. Fowlkes, and S. Y. Emelianov, "Shear wave elasticity imaging: A new ultrasonic technology of medical diagnostics," Ultrasound Med. Biol. **24**, 1419–1435 (1998).
[11] M. Fatemi and J. F. Greenleaf, "Ultrasound-stimulated vibro-acoustic spectrography," Science **280**, 82–85 (1998).
[12] M. Fatemi and J. F. Greenleaf, "Vibro-acoustography: An imaging modality based on ultrasound-stimulated acoustic emission," Proc. Natl. Acad. Sci. U.S.A. **96**, 6603–6608 (1999).
[13] G. T. Silva, S. Chen, J. F. Greenleaf, and M. Fatemi, "Dynamic ultrasound radiation force in fluids," Phys. Rev. E **71**, 056617 (2005).
[14] E. E. Konofagou and K. Hynynen, "Localized harmonic motion imaging: Theory, simulations and experiments," Ultrasound Med. Biol. **29**, 1405–1413 (2003).
[15] E. E. Konofagou, M. Ottensmeyer, S. Agabian, S. L. Dawson, and K. Hynynen, "Estimating localized oscillatory tissue motion for assessment of the underlying mechanical modulus," Ultrasonics **42**, 951–956 (2004).
[16] K. Michishita, H. Hasegawa, and H. Kanai, "Ultrasonic measurement of minute displacement of object cyclically actuated by acoustic radiation force," Jpn. J. Appl. Phys., Part 1 **42**, 4608–4612 (2003).
[17] M. W. Urban, R. R. Kinnick, and J. F. Greenleaf, "Measuring the phase of vibration of spheres in a viscoelastic medium as an image contrast modality," J. Acoust. Soc. Am. **118**, 3465–3472 (2005).
[18] S. Chen, M. Fatemi, and J. F. Greenleaf, "Quantifying elasticity and viscosity from measurement of shear wave speed dispersion," J. Acoust. Soc. Am. **115**, 2781–2785 (2004).
[19] Y. Yamakoshi, J. Sato, and T. Sato, "Ultrasonic imaging of internal vibration of soft tissue under forced vibration," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **37**, 45–53 (1990).
[20] S. Chen, M. Fatemi, and J. F. Greenleaf, "Remote measurement of material properties from radiation force induced vibration of an embedded sphere," J. Acoust. Soc. Am. **112**, 884–889 (2002).
[21] S. Chen, G. T. Silva, R. R. Kinnick, J. F. Greenleaf, and M. Fatemi, "Measurement of dynamic and static radiation force on a sphere," Phys. Rev. E **71**, 056618 (2005).
[22] X. M. Zhang, M. Fatemi, R. R. Kinnick, and J. F. Greenleaf, "Noncontact ultrasound stimulated optical vibrometry study of coupled vibration of arterial tubes in fluids," J. Acoust. Soc. Am. **113**, 1249–1257 (2003).
[23] X. Zhang, M. Zeraati, R. R. Kinnick, J. F. Greenleaf, and M. Fatemi, "Vibration mode imaging," IEEE Trans. Med. Imaging **26**, 843–852 (2007).
[24] X. Zhang and J. F. Greenleaf, "Noninvasive generation and measurement of propagating waves in arterial walls," J. Acoust. Soc. Am. **119**, 1238–1243 (2006).
[25] A. Zaitsev, R. Raymond, J. Thierman, J. Juste, and K. Hynynen, "Focused ultrasound thermal surgery, imaging, and elastometry using the same phase array: Feasibility study," *IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint 50th Anniversary Conference*, 2231–2234, Montreal, QC, Canada (2004).
[26] S.-R. Huang, R. M. Lerner, and K. J. Parker, "On estimating the amplitude of harmonic vibration from the Doppler spectrum of reflected signals," J. Acoust. Soc. Am. **88**, 2702–2712 (1990).
[27] J. Holen, R. C. Waag, and R. Gramiak, "Representations of rapidly oscillating structures on the Doppler display," Ultrasound Med. Biol. **11**, 267–272 (1985).
[28] O. B. Matar, J. P. Remenieras, C. Bruneel, A. Roncin, and F. Patat, "Noncontact measurement of vibration using airborne ultrasound," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **45**, 626–633 (1998).
[29] N. Mujica, R. Wunenburger, and S. Fauve, "Scattering of a sound wave by a vibrating surface," Eur. Phys. J. B **33**, 209–213 (2003).
[30] R. Wunenburger, N. Mujica, and S. Fauve, "Experimental study of the Doppler shift generated by a vibrating scatterer," J. Acoust. Soc. Am. **115**, 507–514 (2004).
[31] Y. Zheng, S. Chen, W. Tan, R. Kinnick, and J. F. Greenleaf, "Detection of tissue harmonic motion induced by ultrasonic radiation force using pulse-echo ultrasound and Kalman filter," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **54**, 290–300 (2007).

[32]C. Kasai, K. Namekawa, A. Koyano, and R. Omoto, "Real-time two-dimensional blood flow imaging using an autocorrelation technique," IEEE Trans. Sonics Ultrason. **SU-32**, 458–464 (1985).
[33]T. Loupas, R. B. Peterson, and R. W. Gill, "Experimental evaluation of velocity and power estimation for ultrasound blood-flow imaging, by means of a 2-dimensional autocorrelation approach," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **42**, 689–699 (1995).
[34]M. O'Donnell, A. R. Skovoroda, B. M. Shapo, and S. Y. Emelianov, "Internal displacement and strain imaging using ultrasonic speckle tracking," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **41**, 314–325 (1994).
[35]W. F. Walker and G. E. Trahey, "A fundamental limit on delay estimation using partially correlated speckle signals," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **42**, 301–308 (1995).
[36]M. L. Palmeri, S. A. McAleavey, G. E. Trahey, and K. R. Nightingale, "Ultrasonic tracking of acoustic radiation force-induced displacements in homogeneous media," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 1300–1313 (2006).
[37]F. Viola and W. F. Walker, "A comparison of the performance of time-delay estimators in medical ultrasound," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 392–401 (2003).
[38]F. Viola and W. F. Walker, "A spline-based algorithm for continuous time-delay estimation using sampled data," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**, 80–93 (2005).
[39]G. F. Pinton, J. J. Dahl, and G. E. Trahey, "Rapid tracking of small displacements with ultrasound," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 1103–1117 (2006).
[40]G. F. Pinton and G. E. Trahey, "Continuous delay estimation with polynomial splines," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 2026–2035 (2006).
[41]H. Hasegawa and H. Kanai, "Improving accuracy in estimation of artery-wall displacement by referring to center frequency of RF echo," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 52–63 (2006).
[42]M. W. Urban, M. Bernal, and J. F. Greenleaf, "Phase aberration correction using ultrasound radiation force and vibrometry optimization," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **54**, 1142–1153 (2007).
[43]R. N. Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. (McGraw Hill, Boston, 2000).
[44]S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, 2nd ed. (McGraw-Hill Irwin, Boston, 2001).
[45]S. A. McAleavey, K. R. Nightingale, and G. E. Trahey, "Estimates of echo correlation and measurement bias in acoustic radiation force impulse imaging," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 631–641 (2003).
[46]Polytec, *Laser Doppler Vibrometer* (Polytec, GmbH, Waldbronn, Germany, 2000).
[47]B. A. Herman and G. R. Harris, "Models and regulatory considerations for transient temperature rise during diagnostic ultrasound pulses," Ultrasound Med. Biol. **28**, 1217–1224 (2002).
[48]M. L. Palmeri and K. R. Nightingale, "On the thermal effects associated with radiation force imaging of soft tissue," IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 551–565 (2004).

Hearing sensitivity during target presence and absence while a whale echolocates

Alexander Ya. Supin[a)]
Institute of Ecology and Evolution of the Russian Academy of Sciences, 33 Leninsky Prospect, 119071 Moscow, Russia

Paul E. Nachtigall[b)] and Marlee Breese
Marine Mammal Research Program, Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, Hawaii 96744-1106, USA

(Received 14 June 2007; revised 10 October 2007; accepted 22 October 2007)

Hearing sensitivity was measured in a false killer whale during echolocation. Sensitivity was measured using probe stimuli as sinusoidally amplitude modulated signals with a 22.5-kHz carrier frequency and recording auditory evoked potentials as envelope-following responses. The probes were presented and responses were recorded during short 2-s periods when the animal echolocated to detect the presence or absence of a target in a go/no-go paradigm. In the target-absent trials, a hearing threshold of 90.4 dB re 1 μPa was found; in the target-present trials, the threshold was 109.8 dB. Thus, a 19.4-dB difference was found between thresholds in the target-present and target-absent trials. To check the possibility that this difference was the result of different masking degree of the probe by the emitted sonar clicks, click statistics were investigated in similar trials. No indication was found that the energy of the emitted clicks was higher in the target-present than in target-absent trials; on the contrary, mean click level, mean number of clicks per train, and overall train energy was slightly higher in the target-absent trials. Thus the data indicate that the hearing sensitivity of the whale varied depending on target presence or absence.
 [DOI: 10.1121/1.2812593]

PACS number(s): 43.80.Lb [WWA] Pages: 534–541

I. INTRODUCTION

The biosonar of odontocetes (toothed whales, dolphins, and porpoises) features a fascinating adaptation of the auditory system to specific conditions when the "sound image" of an object may be controlled by the subject itself. Although the odontocete biosonar has been a subject of investigation over the past few decades (Nachtigall and Moore, 1988; Au, 1993; Thomas *et al.*, 2003), many of the basic mechanisms underlying its functioning remain unexplored. In particular, the problem of automatic gain control in the odontocete's biosonar has attracted attention recently. Indeed, the target strength and distance to the target investigated by the animal's sonar may vary widely. Therefore, the echo level may vary by many tens of decibels. Successful performance of the biosonar shows that its receiving part (the auditory system) is capable of analyzing echoes under these widely varying conditions. This ability may require a kind of gain control to keep the response of the auditory system to the echo within the dynamic range where successful analysis of the echo signal is possible.

A few mechanisms of the gain control have been described previously. One of them is the decrease of variation of the echo level by compensating variation of the emitted pulse level. Experiments in the wild have shown that dolphins and whales vary the level of their echolocation pulses according to the distance to the target, roughly at a rate of 20 dB per distance decade (Rasmussen *et al.*, 2002; Au and Benoit-Bird, 2003; Au and Herzing, 2003; Au and Würsig, 2004; Au *et al.*, 2004). To some extent, this may compensate for the echo attenuation with distance, although it does not influence the ratio of the echo level to transmitted pulse level (for brevity, specified in the following as the echo-to-transmission ratio). Another mechanism involves gain control in the auditory system based on forward-masking interaction between the emitted sonar pulse and the echo: With increasing the distance to a target, the echo becomes fainter, but at the same time, the echo delay increases and therefore the echo releases from masking by the emitted pulse; these two processes compensate one another (Supin *et al.*, 2004, 2005, 2007).

A remaining question is: Does the gain-control system in the odontocete's biosonar include an active variation of the sensitivity of the auditory system? There are some observations indicating this possibility. Measurement of evoked responses to emitted sonar pulses has shown that the emitted pulses are perceived as faint as around −35 dB relative to the pulse source level in target-present trials but markedly higher (−20 dB) in target-absent trials (Supin *et al.*, 2006). This finding is difficult to explain in any other way than to suppose that the subject was capable of actively varying the hearing sensitivity according to the current situation (the presence or absence of a target producing an echo loud enough).

[a)]Electronic mail: alex-supin@mail.ru
[b)]Electronic mail: nachtiga@hawaii.edu

 0001-4966/2008/123(1)/534/8/$23.00

In order to clarify the possibility of variation of hearing sensitivity during echolocation, it was reasonable to perform direct measurements of hearing sensitivity during echolocation in odontocetes. In the present study, we made those sorts of measurements using the evoked-potential method which has been previously demonstrated to be comparatively efficient for producing audiometric measurements in odontocetes (Supin *et al.*, 2001; Yuen *et al.*, 2005; Popov *et al.*, 2007; Nachtigall *et al.*, 2007).

II. MATERIALS AND METHODS

A. Subject and experimental conditions

The experiments were carried out at the Hawaii Institute of Marine Biology, Marine Mammal Research Program. The subject was a false killer whale *Pseudorca crassidens*, an approximately 30-year-old female maintained in a wire-net enclosure in Kaneohe Bay, HI. The animal had a significant hearing loss at frequencies above 30 kHz (Yuen *et al.*, 2005) however its echolocation pulses had the spectrum peak below 30 kHz (Supin *et al.*, 2006) fitting the best-sensitivity frequency range. The animal was trained to accept soft latex suction cups containing EEG electrodes to pick up the evoked potentials, to ensonify and recognize targets by echolocation, and to report the target presence or absence using a go/no-go reporting paradigm.

The experimental facilities were laid out as follows (Fig. 1). The experimental enclosure was constructed of a floating pen frame (1), 8×10 m in size, supported by floats and bearing an enclosing wire net. This enclosure (the animal section) linked to a target section—another floating frame (2), 6×8 m in size that served to mount targets and hydrophones and did not bear net. In the net divider separating these two sections, there was an opening bounded by a hoop (3), 55 cm in diameter that served as a hoop station for the animal. In front of the hoop, a hydrophone (4) was positioned 1 m from the level of the animal's blowhole to record the echolocation pulses. At a distance 2 m in front of the animal's blowhole, a transducer (5) to play test sound signals was positioned. A target (6) was hung from a thin monofilament line at a distance of 3 m from the animal's head and could be pulled up out of water and lowered down into the water. The targets were hollow aluminum cylinders with an outer diameter of 38 mm (1.5 in.) and 25.4 mm (1 in.) inner diameter, 180 or 45 mm long, axis vertical. The target strengths were −22 and −34 dB, respectively, as measured by clicks imitating those of the subject. The hoopstation (3), the hydrophone (4), the transducer (5), and the lowered target (6) were in a longitudinal straight line; altogether at a depth of 80 cm. In front of the animal, there was a movable baffle (7). When pulled up, this baffle screened the target area from the animal positioned in the hoop station; when it was lowered down, it opened the space in front of the animal thus allowing inspection for target presence or absence by echolocation. Behind the baffle, there was a screen (8) made of thin black polyvinylchloride film that was sound transparent but not light transparent. This screen served to prevent visual detection of the target. Near the hoop station, a response ball (9) was mounted above the water surface serving as a target-present response indicator. The trainer kept a position (10) to give instructions to the animal and to reward it with fish for correct responses. The animal's position in the stationing hoop was monitored through an underwater video camera (11). The electronic equipment and the operator were housed in a shack (12).

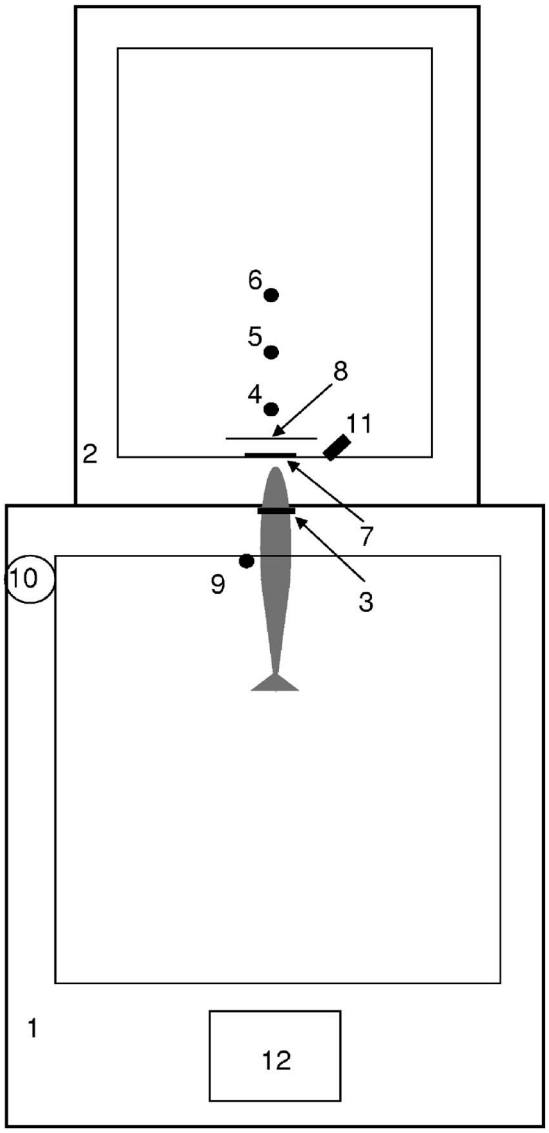


FIG. 1. Experimental conditions. Items 1–12 are designated in the text.

B. Experimental procedure

Two types of experimental sessions were performed—brain-response recording sessions and sonar-click recording sessions.

1. Brain-response recording sessions

Each session consisted of 30 trials, 15 target-present and 15 target-absent, randomly alternated. The experimental procedure was as follows.

(i) Each session began with the trainer attaching the suction-cup electrodes for evoked-potential recording (see the following for details). (ii) The animal was given a signal to go to the hoop station (3 in Fig. 1). During the animal's

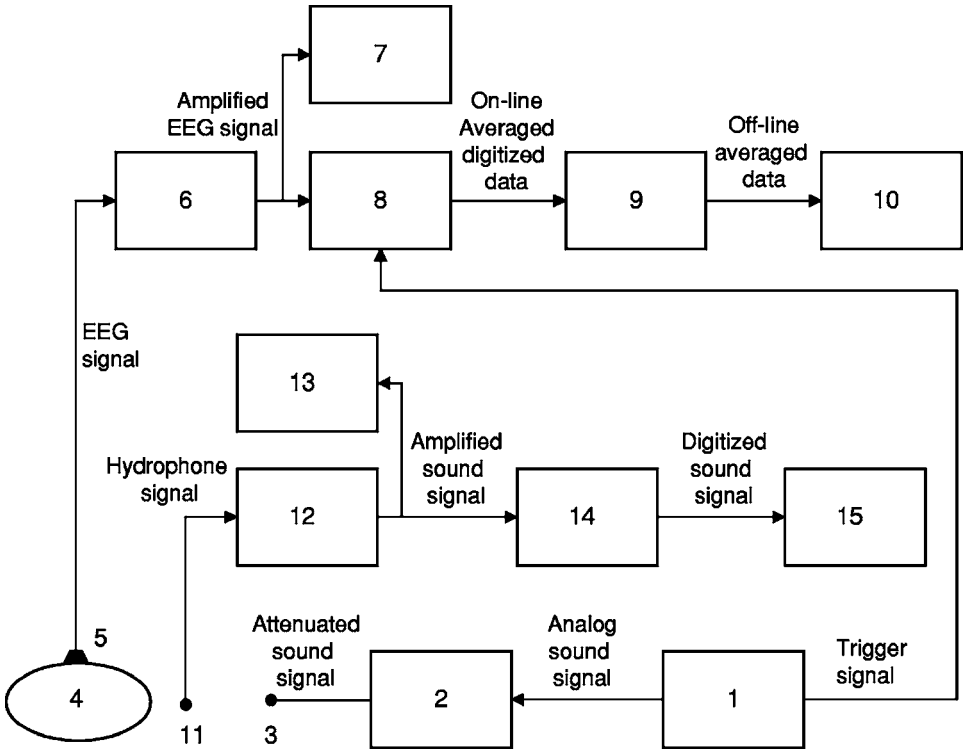


FIG. 2. Instrumentation. Items 1–15 are designated in the text.

positioning, the baffle (7 in Fig. 1) screened the target from the animal. The target (6 in Fig. 1) was either lowered down into water (a target-present trial) or pulled up out of water (a target-absent trial) in advance. During each experimental session, only one target size (either of −22 or of −34 dB target strength) was used. (iii) As soon as the animal took the position in the hoop station, the baffle was lowered down, thus opening the space in front of the animal. Immediately after that, the animal emitted a train of echolocation clicks, as a rule, 10–40 clicks in a train lasting from 1.5 to 5 s. Simultaneously with the baffle opening, a 2-s long series of test sound stimuli (R-features of the stimuli see the following) was played through the transducer (5 in Fig. 1) nonsynchronously with the echolocation clicks, and brain responses to the stimuli were recorded. (iv). This 2-s stimulus was either entirely overlapped by a longer click train or almost entirely overlapped when the train was slightly shorter than 2 s. If the target was present, the animal was required to signal its detection by leaving the hoop and touching the response ball (9 in Fig. 1), then coming to the trainer for the fish reward. During the no-target trails, the animal was required to wait until it was signaled to leave the hoop and come for the fish reward. Errors were not reinforced.

2. Echolocation-click recording sessions

The click recording sessions were organized in the very same manner as the brain-response recording sessions, except the test sound stimuli were not played and transmitted sonar clicks were recorded through the hydrophone (4) instead of brain-response recording.

C. Instrumentation, stimulation, and data collection

The stimulation and recording equipment was designed as shown in Fig. 2. The test stimuli were sinusoidally amplitude-modulated tones of 22.5-kHz carrier frequency [this frequency fitted the best-sensitivity frequency range of the animal as defined by Yuen *et al.* (2005)] and 875-Hz modulation rate, 100% modulation depth, 19.4-ms duration (17 whole modulation cycles), presented at a rate of 20/s. Thus, during the 2-s time of stimulation, 40 stimulating bursts were presented in each trial. The stimulus carrier frequency and modulation rate were chosen as optimal based on previous audiometric investigations of this subject (Yuen *et al.*, 2005). The signals were digitally generated at a sampling rate of 512 kHz and digital-to-analog converted by a DAQ-6062E (National Instruments) card installed in a standard laptop computer (1 in Fig. 2), amplified and attenuated by a custom-made power amplifier-attenuator (2) and played through an ITC-1032 (International Transducer Corporation) spherical transducer (3). The stimulus level varied by 5-dB steps within a range from 90 to 150 dB relative 1 μPa rms.

Brain potentials were picked up from the subject (4) by EEG electrodes (5) which were gold-plated disks 10 mm in diameter mounted within rubber suction cups 50 mm in diameter. The active electrode was attached with conductive gel at the dorsal head surface, at the midline, 5–7 cm behind the blowhole. The reference electrode was also attached, along with conductive gel, on the animal's dorsal fin. Brain potentials were led by shielded cables to a balanced amplifier (6) and amplified by 2.5×10^4 within a frequency range from 200 to 5000 Hz. The amplified signal was monitored by an oscilloscope Tektronix TDS1002 (7) and entered into a 12-bit analog-to-digital converter (8) of the same data acquisition card DAQ-6062E, sampling rate of 16 kHz. The com-

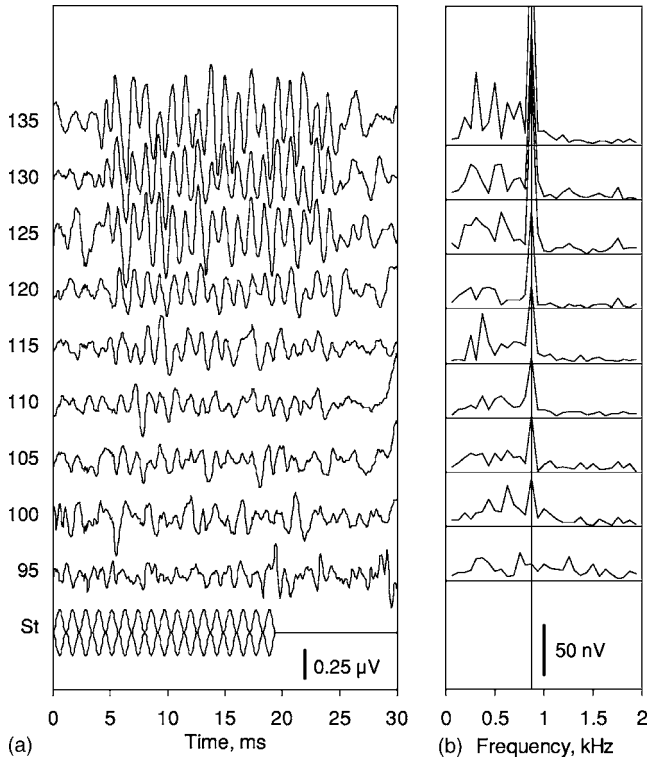


FIG. 3. Wave forms (a) and frequency spectra (b) of responses to probe stimuli of various intensities in the target-absence trials. The spectra (b) were obtained by Fourier analysis of 16-ms-long (from 6 to 22 ms) windows of records in (a). Stimulus intensities (dB re 1 μPa rms) are indicated next to the records; St—stimulus envelope. Vertical straight line in (b) marks the response peak at the modulation frequency of 875 Hz.

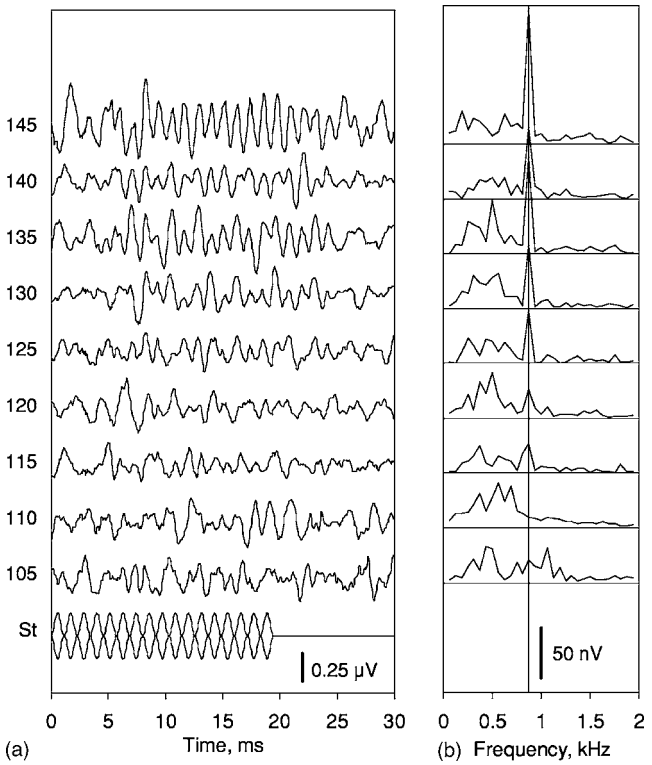


FIG. 4. The same as Fig. 3, for the target-presence trials.

puter performed on-line averaging of responses within a 30-ms time window; the average was coherent to the triggering signal from the generator (1).

In all cases, averaging 40 sweeps collected in each trial was not enough for good extraction of the response from noise. For the final extraction, intertrial off-line average of on-line averaged records was performed (9), and the final records were stored in computer memory (10).

The echolocation clicks were picked up by a B&K 8103 hydrophone (11), amplified by a custom-made 40-dB amplifier (12), monitored by a Tektronix TDS1002 oscilloscope (13), led to an analog-to-digital converter (14) of the same data acquisition card DAQ-6062E, and stored in computer memory (15). Sampling rate was 256 kHz. The clicks were collected in an analog internal-triggering mode using a 100-μs-long acquisition window including 10-μs pretrigger time.

Stimulus generation, brain-potential recording, and sonar-click recording were all controlled by a custom-made program (virtual instrument) designed on the basis of LabVIEW (National Instruments) software.

III. RESULTS

A. Auditory evoked potential wave form and thresholds

In each of the four experimental conditions (target presence and target absence of −22- and −34-dB target strength) and at every stimulus level from 95 to 145 dB, from 8 to 22 on-line averaged records (40 individual records each) were obtained. Thus, the final off-line averaged records were based on 320 to 880 individual records. In total, 34 final envelope-following response (EFR) wave forms were obtained, eight to nine wave forms at each of the four experimental conditions and at stimulus intensities from 110 to 145 dB (5-dB steps) at the target-present conditions and from 95 to 135 dB (also 5-dB steps) at the target-absent conditions.

As a result of successive on-line and off-line averaging, typical EFR wave forms as described in previous studies (review Supin *et al.*, 2001) were extracted [Figs. 3(a) and 4(a)]. The EFR was a quasisinusoidal wave form of the same frequency as the stimulus envelope (875 Hz) and the same duration as the stimulating burst (19.4 ms) but delayed relative to the stimulus by 4–4.5 ms (the response lag). The Fourier transform of the response records revealed a definite peak at the stimulus modulation rate of 875 Hz [Figs. 3(b) and 4(b)] showing that the recorded wave form was really the response to stimulus amplitude modulation.

The EFR amplitude was dependent on stimulus level: The higher stimulus level, the higher EFR amplitude. This dependence was qualitatively similar in all trial types: With the target presence or absence, with the target of higher (−22 dB) or lower (−34 dB) target strength. However, quantitatively the amplitude-versus-level dependence was different in different trial types. The most significant difference was between the target-present and target-absent conditions. Comparison of Figs. 3 and 4 shows that similar EFR amplitudes were obtained at much lower stimulus intensities in the target-absent conditions (Fig. 3) than in the target-present conditions (Fig. 4).

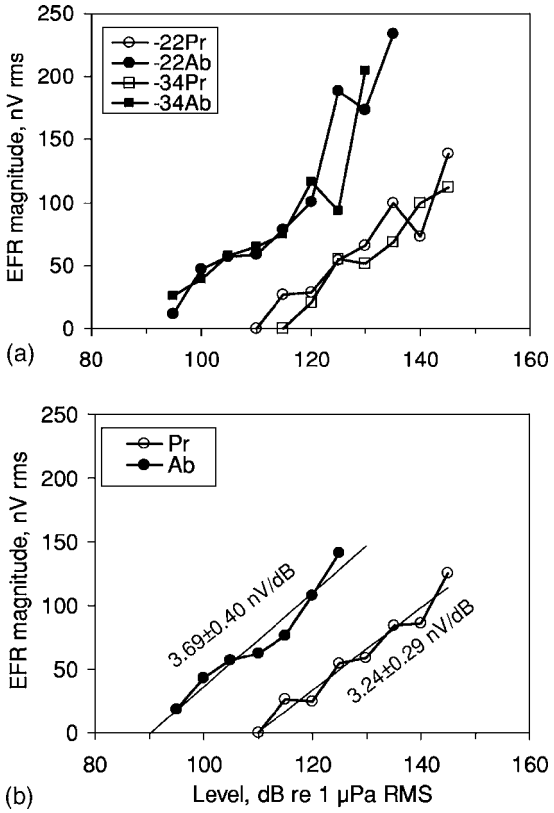


FIG. 5. (a) EFR magnitude dependence on stimulus level for four experimental conditions: −22-dB target presence (-22Pr), −22-dB target absence (-22Ab), −34-dB target presence (-34Pr), and −34-dB target absence (-34Ab). (b) Combined data for all target presence trial (Pr) and all target absence trials (Ab); straight lines—approximations by regression lines; the regression slopes (nV/dB) are indicated next to the plots.

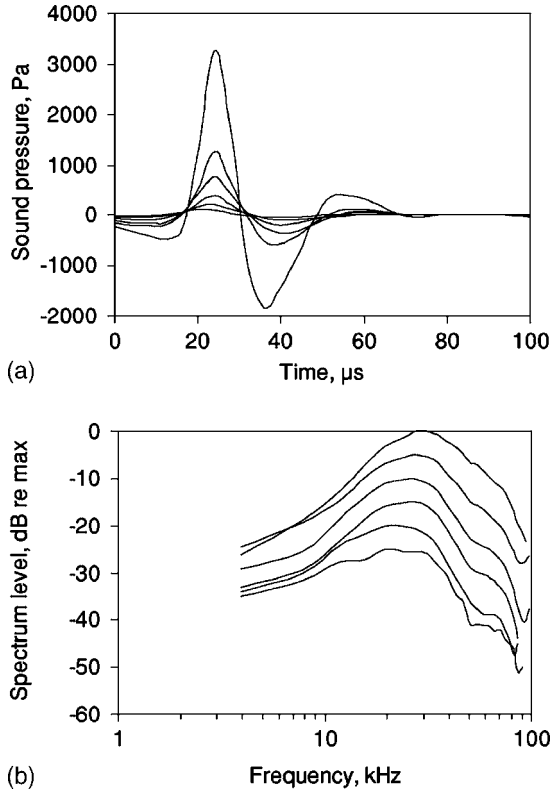


FIG. 6. (a) Sonar click wave forms from 165 to 190 dB re 1 μPa p/p (from the lowest to highest amplitude) with 5-dB steps. (b) Frequency spectra of the presented wave forms (from the lowest to highest, respectively).

All the data obtained in both target-present and target-absent conditions and with the use of both targets are summarized in Fig. 5(a) as EFR magnitude dependence on stimulus level. The magnitude was estimated as excess of the spectrum peak at the frequency of 875 Hz over adjacent (background) spectrum components. The excess was calculated by subtraction of the squared magnitude (which is a relative estimate of power) of the adjacent spectrum components from that of the 875-Hz component:

$$M = \sqrt{M_i^2 - (M_{i-1}^2 + M_{i+1}^2)/2}, \quad (1)$$

where M is the response magnitude, M_i is the spectrum component magnitude at the stimulus modulation frequency (875 Hz), and M_{i-1} and M_{i+1} are magnitudes of spectrum components one step below and one step above the modulation frequency.

The results show little difference between sessions using the targets of higher (−22 dB) and lower (−34 dB) strength. However, there was a significant difference between the target-present and target-absent conditions. Stimuli of one and the same level produced much higher EFR amplitude in target-absent as compared to target-present conditions. Respectively, one and the same EFR amplitude could be obtained at much lower stimulus intensities in target-absent rather than in target-present conditions.

Because of the slight difference between data obtained with different targets (−22- and −34-dB target strength), we considered it reasonable to average these data. The averaging was done for a near-threshold range where averaged response magnitude did not exceed 150 nV. The result is presented in Fig. 5(b). It shows a substantial (around 20 dB) shift between the target-present and target-absent magnitude-versus-level functions. Both the functions were approximated by regression straight lines within a range of 110–145 dB for the target-present trials and 95–125 dB for the target-absent trials where the response magnitude varied from 0 to 150 nV. The approximation was satisfactory (r^2 =0.96 for target-presence trials and r^2=0.94 for target-absence trials), and slopes of the regression lines were similar (3.24±0.29 and 3.69±0.40 nV/dB for target-present and target-absent trials, respectively). Estimations of EFR thresholds as crossing points of the regression lines with the zero-magnitude level gave values of 90.4 dB for the target-absent condition and 109.8 dB for the target-present condition, i.e., the threshold difference between the target-present and target-absent conditions was as large as 19.4 dB.

B. Sonar click features and statistics

The wave forms and spectra of echolocation clicks are presented in Figs. 6(a) and 6(b), respectively. Each wave form was obtained by averaging of 150–9500 clicks sorted according to their levels in 5-dB bins, from 160 to 195 dB re 1 μPa peak-to-peak sound pressure. The number of averaged clicks in each bin depended on probability of their appearance. Only wave forms from 165 to 190 dB are illustrated in Fig. 6(a) because the difference between the lowest (160 dB)

TABLE I. Statistics of clicks.

Trial type	No. of trains	No. of clicks total	No. of clicks per train	Mean ± s.d. level, dB re 1 μPa p/p	Relative energy per train, dB re single pulse 1 μPa p/p
−22-dB target present	373	7 719	20.7	170.7±6.9	190.1
−34-dB target present	366	7 952	21.7	171.6±7.9	191.8
−22-dB target absent	307	8 270	26.9	173.6±8.5	195.4
−34-dB target absent	315	8 478	26.9	173.1±8.1	195.4
Target present total	**739**	**15 671**	**21.2**	**171.1±7.4**	**191.0**
Target absent total	**622**	**16 748**	**26.9**	**173.3±8.6**	**195.4**

and highest (195 dB) levels (56 times of amplitude ratio) is too large to present them on the same amplitude scale. Figure 6 demonstrates a rather constant wave form, little dependent on the click level. Respectively, the spectra of these wave forms were also rather similar, except for a small shift of the peak frequency when the level increased, from 27.3 kHz at 165 dB to 31.2 kHz at 190 dB.

For statistical analysis of click number and level, a total 1361 click trains were collected, from 307 to 373 trains in each of the four experimental conditions (−22- and −34-dB strength target present and absent). These trains contained 32 419 clicks, from 7719 to 8270 clicks in total and from 20.7 to 26.9 clicks per train in each of the four experimental conditions (Table I). Distributions of click levels are presented in Fig. 7. In both target-present [Fig. 7(a)] and target-absent trials [Fig. 7(b)], there was little difference between the distributions for the trials with -22-dB- and -37-dB targets; the means differed as little as 0.9 dB for target-present and 0.5 dB for target-absent trials (Table I). However, there was a small but noticeable shift of the target-absent distribution to higher click levels as compared to the target-present trials. The average of the both target-present distributions and both target-absent distributions ([Fig. 7(c)] also revealed this difference with the target-absent mean being 2.2±0.06(SE) dB *higher* than the target-present mean (Table I).

IV. DISCUSSION

A. Estimates of hearing sensitivity

The most essential finding of the measurements presented herein is a remarkable difference between estimates of hearing sensitivity of the subject in the target-present and the target-absent conditions. When targets were absent, the sensitivity was almost 20 dB better than when targets were present. Figure 5 shows not only that the threshold was lower, but that all response magnitudes up to around 150 nV rms required around 20 dB less stimulus level in the target-absent than in the target-present conditions. So it is hardly possible that this difference is a result of measurement error or imprecision in the threshold computation. Does this found difference in sensitivity *estimates* reflect real difference in sensitivity?

To answer this question, at least four explanations must be considered and discussed:

(i) Since the behavior of the animal was different in the target-present and target-absent trials (go- and no-go responses, respectively), the stimulation conditions were somehow influenced, making the stimulus less intensive in target-present than in target-absent conditions.

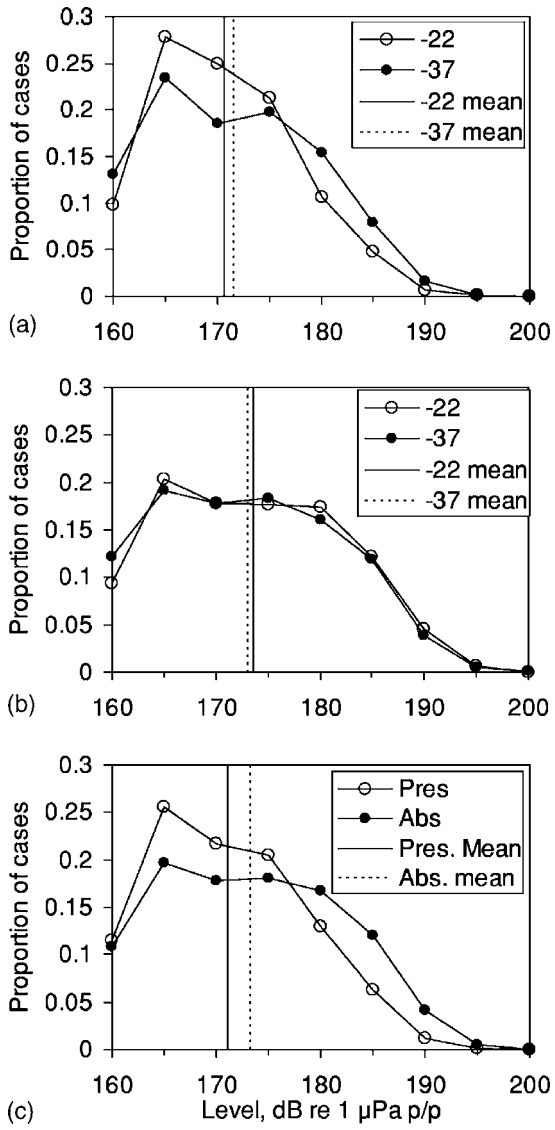


FIG. 7. Amplitude distributions of sonar clicks (proportion per 5-dB bin). (a) Target-presence trials with −22- and −34-dB targets, as indicated in the legend. (b) Target-absence trials with −22- and −34-dB targets. (c) Combined data for target-presence and target-absence trials; vertical lines indicate means of the distributions.

(ii) Since sensitivity was measured during active echolocation, the probe stimuli were partially masked by the outgoing sonar pulses, and this masking was more intensive in the target-present than in target-absent conditions.
(iii) The probe stimuli were masked by echoes in target-present trials and not masked in target-absent trials.
(iv) The hearing sensitivity of the whale really was better when targets were absent compared to when targets were present.

B. Difference in stimulation conditions between target-present and target-absent trials

In the target-present trials, the animal left the hoop station. When this rear movement began before the end of the 2-s probe stimulus, the distance between the transducer and the animal increased, whereas in target-absent trials the animal's position and distance to the transducer were kept constant during the whole probe duration. The increase of the distance to the transducer resulted in some decrease of the sound-pressure level near the animal's head. However, this decrease could not be the source of the observed difference between sensitivity estimates. With the initial distance of 2 m, the increase by 0.5 m or so results in sound-pressure level decrease of only about 2 dB. Actually, the effect must be even less than 2 dB, since not all the probe stimulus but only its final part was played at the decreased level. Thus, difference of around 20 dB cannot be explained by this effect. The decrease of sound pressure level by 20 dB might appear at ten-times a distance increase, i.e., up to 20 m, but that was obviously impossible in our experimental conditions. Thus, the idea based on inconstancy of the animal-to-transducer distance does not provide an explanation for the observed difference in sensitivity between target-present and target-absent trials.

C. Difference in echolocation click energy between target-present and target-absent trials

Probe stimuli might be differently masked by the outgoing sonar clicks if some parameters of these clicks were markedly different in target-present and target-absent trials. To produce the found 20-dB difference of sensitivity estimates, sonar clicks in the target-present trials should be much more intensive than those in target-absent trials. Bearing in mind this possibility, an investigation of sonar click parameters was undertaken under the very same conditions as those for sensitivity measurements. This investigation did not provide any evidence of more intensive sonar click emission in the target-present as compared to the target-absent conditions. On the contrary, slightly but noticeably larger clicks were found in the target-absent conditions: It manifested itself both in higher mean level of the clicks (173.3 dB at target absent as compared to 171.1 dB with target present) and in a larger mean number of clicks per train (26.9 at target absence contrary to 21.2 at target presence). To combine both these parameters, we computed a relative estimate of overall energy of clicks per train as

$$E = 10 \log \sum N_I \times 10^{I/10}, \quad (2)$$

where E is the train energy, I is a click-level class (160, 165, etc.) in dB re 1 μPa peak-to-peak, and N_I is the mean number of pulses of the level I per train; click duration was assumed constant. With this computation, the energy E is expressed in decibels re energy of a single click of 1 μPa p/p level. The result was 195.4 dB for the target-absent trains and 191.0 dB for the target-present trains, i.e., there is a 4.4-dB difference *in favor of the target-absent* trains. The difference is not very big; but at least there is obviously no indication that target-present trains are more intense than the target-absent trains. It is therefore obvious that the target-present trains did not produce masking of the probe stimuli 20-dB stronger than in the target-absent trials. Thus, the explanation based on different masking effects of the sonar pulses in the target-present and target-absent trials cannot be accepted either.

D. Masking of the probe stimuli by echoes

The presence or absence of the echo is a remarkable difference between the target-present and target-absent trials. If the echo substantially masked the probe stimuli, it could result in increased thresholds in target-present trials. To estimate the possibility of this effect, the echo level should be estimated and compared to the level of the probe stimuli.

In the target-present trials, the mean peak-to-peak source level of transmitted click was 171 dB (see Table I). So at the target strengths of −22 and −34 dB and at a distance of 3 m, the mean echo level was 139 and 127 dB, respectively. This was a peak-to-peak level of short (a few tens of microseconds each, see Fig. 6) and widely separated pulses. Assuming that all 21 pulses (see Table I) appeared during the 2-s acquisition time, the mean interpulse interval was 95 ms. For a wave form presented in Fig. 5, computation over the 95-ms interval resulted in a rms value of −44 dB relative to the peak-to-peak value. Thus, the overall rms level of the echo train may be roughly estimated as 95 and 83 dB for the two target strengths. This is a conservative estimate based on assuming that the train exactly coincides with the acquisition 2-s interval. If not all of the echolocation clicks fell into the 2-s interval, the overall rms was even lower; if some click trains were shorter than the 2-s interval, a part of probe stimuli might be not masked at all. Even the conservative level estimates of 95 and 83 dB are markedly lower than the threshold in target-present trials (110 dB) and extremely lower than probe levels still revealing the difference between target-present and target-absent trials (up to 135 dB, see Fig. 5). Obviously a masker *below* the probe down to 52 dB (83-echo rms level and 135-dB probe level) cannot produce any noticeable masking effect. Thus, the explanation based on a masking effect of the echo cannot be accepted either.

E. Hearing sensitivity variation depending on the target presence or absence

Denying the three previous explanations, there seems only one remaining. It agrees well with the data previously collected (Supin *et al.*, 2006) showing that at a certain source

level of transmitted echolocation clicks, their sensation level was also as much as 15 dB higher in the target-absent trials than in the target-present trials.

In principle, the functional regulation of hearing sensitivity is a commonly known physiological event. Hearing sensitivity may be regulated at both conductive (the stapedial reflex) and sensorineural levels (adaptation). These mechanisms are known as being provoked by acoustical stimuli themselves, reducing the hearing sensitivity to high-level sounds. We cannot exclude the possibility that in whales and dolphins similar regulations of sensitivity are controlled by the echolocation activity. If such a mechanism truly exists, then better sensitivity in the target-absent rather than in target-present trials makes obvious sense: in the absence of an echo loud enough, the increased sensitivity should help to pick up fainter echoes. It seems very reasonable that the role of this hearing gain control deserves further investigation to find out its role in odontocete sonar functioning.

ACKNOWLEDGMENTS

The authors graciously thank Bob Gisiner of the U.S. Office of Naval Research for his support through Grant No. N00014.05.1.07.38 to P.E.N. and further thank The Russian Ministry of Science and Education for Grant No. NSh-7117.2006.4 to A.Ya.S. Work was conducted under a Scientific Research Permit to Take Marine Mammals from NOAA NMFS Office of Protected Resources Permit No. 978-1567-02 and a Research Protocol approved by the University of Hawaii IACUC No. 93-005-13 to PEN. This is contribution No. 1296 of the Hawaii Institute of Marine Biology.

Au, W. W. L. (**1993**). *The Sonar of Dolphins* (Springer, NewYork).

Au, W. W. L., and Benoit-Bird, K. J. (**2003**). "Automatic gain control in the echolocation system of dolphins," Nature (London) **423**, 861–863.

Au, W. W. L., Ford, J. K. B., Horne, J. K., and Allman, K. A. N. (**2004**). "Echolocation signals of free-ranging killer whales (*Orcinus orca*) and modeling of foraging for chinook salmon (*Oncorhynchus tshawytscha*)," J. Acoust. Soc. Am. **115**, 901–909.

Au, W. W. L., and Herzing, D. L. (**2003**). "Echolocation signals of wild Atlantic spotted dolphin (Stenella frontalis)," J. Acoust. Soc. Am. **113**, 598–604.

Au, W. W. L., and Würsig, B. (**2004**). "Echolocation signals of dusky dolphins (Lagenorhynchus obscurus) in Kaikoura, New Zealand," J. Acoust. Soc. Am. **115**, 2307–2313.

Nachtigall, P. E., Mooney, T. A., Taylor, K. A., and Yuen, M. M. L. (**2007**). "Hearing and auditory evoked potential methods applied to odontocete cetaceans," Aquat. Mamm. **33**, 6–13.

Nachtigall, P. E., and Moore, P. W. B., eds. (**1988**). *Animal Sonar: Processes and Performance* (Plenum, New York).

Popov, V. V., Supin, A. Ya., Pletenko, M. G., Tarakanov, M. B., Klishin, V. O., Bulgakova, T. N., and Rosanova, E. I. (**2007**). "Audiogram variability in normal bottlenose dolphins (*Tursiops truncatus*)," Aquat. Mamm. **33**, 24–33.

Rasmussen, M. H., Miller, L. A., and Au, W. W. L. (**2002**). "Source levels of clicks from free-ranging white beaked dolphins (*Lagenorhinchus albirostris* Gray 1846) recorded in Icelandic waters," J. Acoust. Soc. Am. **111**, 1122–1125.

Supin, A. Ya., Nachtigall, P. E., Au, W. W. L., and Breese, M. (**2004**). "The interaction of outgoing echolocation pulses and echoes in the false killer whale's auditory system: Evoked-potential study," J. Acoust. Soc. Am. **115**, 3218–3225.

Supin, A. Ya., Nachtigall, P. E., Au, W. W. L., and Breese, M. (**2005**). "Invariance of evoked-potential echo-responses to target strength and distance in an echolocating false killer whale," J. Acoust. Soc. Am. **117**, 3928–3935.

Supin, A. Ya., Nachtigall, P. E., and Breese, M. (**2006**). "Source-to-sensation level ratio of transmitted biosonar pulses in an echolocating false killer whale," J. Acoust. Soc. Am. **120**, 518–526.

Supin, A. Ya., Nachtigall, P. E., and Breese, M. (**2007**). "Evoked-potential recovery during double click stimulation in a whale: A possibility of biosonar automatic gain control," J. Acoust. Soc. Am. **121**, 618–625.

Supin, A. Ya., Popov, V. V., and Mass, A. M. (**2001**). *The Sensory Physiology of Aquatic Mammals* (Kluwer, Boston).

Thomas, J. A., Moss, C. F., and Vater, M., eds. (**2003**). *Echolocation in Bats and Dolphins* (University of Chicago Press, Chicago).

Yuen, M. M. L., Nachtigall, P. E., Breese, M., and Supin, A. Ya. (**2005**). "Behavioral and auditory evoked potential audiograms of a false killer whale (*Pseudorca crassidens*)," J. Acoust. Soc. Am. **118**, 2688–2695.

Estimating bottlenose dolphin (*Tursiops truncatus*) hearing thresholds from single and multiple simultaneous auditory evoked potentials

James J. Finneran[a)]
U.S. Navy Marine Mammal Program, Space and Naval Warfare Systems Center, San Diego, Code 71510, 53560 Hull Street, San Diego, California 92152

Dorian S. Houser
Biomimetica, 7951 Shantung Drive, Santee, California 92071

Dave Blasko, Christie Hicks, Jim Hudson, and Mike Osborn
The Mirage Dolphin Habitat, 3400 Las Vegas Boulevard South, Las Vegas, Nevada 89109

(Received 20 July 2007; revised 12 October 2007; accepted 22 October 2007)

Hearing thresholds were estimated in four bottlenose dolphins by measuring auditory evoked responses to single and multiple sinusoidal amplitude modulated tones. Subjects consisted of two males and two females with ages from 4 to 22 years. Testing was conducted in air using a "jawphone" transducer to couple sound into each subject's lower right jaw. Carrier frequencies ranged from 10 to 160 kHz in one-half octave steps. Amplitude modulated stimuli were presented individually and as the sum of four, five, and nine simultaneous tones with unique carrier and modulation frequencies. Evoked potentials were noninvasively recorded using surface electrodes embedded in silicon suction cups. The presence or absence of an evoked response at each modulation frequency was assessed by calculating the magnitude-squared coherence from the frequency spectra of the recorded sweeps. All subjects exhibited traditional "U-shaped" audiograms with upper cutoff frequencies above 113 kHz. The time required for threshold estimates ranged from 23 to 37 min for single stimuli to 5–9 min for nine simultaneous stimuli. Agreement between thresholds estimated from single stimuli and multiple, simultaneous stimuli was generally good, indicating that multiple stimuli may be used for quick hearing assessment when time is limited.
 [DOI: 10.1121/1.2812595]

PACS number(s): 43.80.Lb [WWA] Pages: 542–551

I. INTRODUCTION

There is substantial interest in understanding the auditory capabilities of marine mammals in order to predict and mitigate the effects of anthropogenic noise on wild animals (NRC, 1994; Richardson *et al.*, 1995; NRC, 2000, 2003, 2005) and to facilitate training and improve the health care of captive animals (e.g., Houser and Finneran, 2006b). Although hearing capabilities have traditionally been assessed using psychophysical techniques (see Nachtigall *et al.*, 2000); the required level of access, time, and training have prevented the application of these methods to large numbers of subjects and species. To overcome these limitations, electrophysiological techniques, which measure changes in the electroencephalogram (EEG) that are time-locked to a sound stimulus, have been increasingly applied (e.g., Bullock *et al.*, 1968; Ridgway *et al.*, 1981; Popov and Supin, 1985, 1990; Popov *et al.*, 1992; Dolphin *et al.*, 1995; Supin and Popov, 1995; Dolphin, 1996; Szymanski *et al.*, 1999; Popov *et al.*, 2005; Yuen *et al.*, 2005). In these methods, parameters of the sound are manipulated and the resulting voltages from synchronous neural discharges, called auditory evoked potentials (AEPs), are measured to determine the effects of the sound on the neural activity within the auditory pathway (Eggermont, 2007). AEP measurements are relatively fast and do not require active subject participation, which allows their use with individuals not specifically trained for a hearing test and under opportunistic circumstances (e.g., Nachtigall *et al.*, 2005; Cook *et al.*, 2006; Houser and Finneran, 2006b; Nachtigall *et al.*, 2007).

Although AEPs may be elicited with a variety of stimuli, frequency-specific measurements in marine mammals have often used sinusoidally modulated tones (Dolphin *et al.*, 1995; Supin and Popov, 1995; Dolphin, 1996; Supin and Popov, 2000; Nachtigall *et al.*, 2005; Yuen *et al.*, 2005). These stimuli produce a periodic evoked response, called the envelope following response or auditory steady-state response (ASSR), whose fundamental frequency is related to the stimulus modulation frequency (Campbell *et al.*, 1977; Hall, 1979; Stapells *et al.*, 1984; Picton *et al.*, 1987). We use the more general term ASSR here to remain consistent with the large volume of human literature and to reflect the fact that the ASSR may be generated by stimuli such as sinusoidal frequency modulated tones that do not possess a temporal envelope. An advantage of the ASSR is that multiple sinusoidally modulated tones, each with a unique modulation frequency, may be used to simultaneously test hearing at multiple frequencies (Picton *et al.*, 1987; Regan and Regan,

a)Electronic mail: james.finneran@navy.mil

1988; Lins *et al.*, 1995; Lins and Picton, 1995; Dolphin, 1996; Popov *et al.*, 1997, 1998; Finneran and Houser, 2007; Finneran *et al.*, 2007c). The evoked response to each tone occurs at the corresponding modulation rate and, if sufficient frequency separation exists, hearing thresholds estimated with multiple tones match those obtained with single stimuli (Lins and Picton, 1995; John *et al.*, 1998; John *et al.*, 2002). The multiple ASSR technique therefore offers the potential for substantial reductions in testing time, making this method particularly attractive for use with individuals for whom access time may be limited (e.g., stranded or rehabilitating animals). The multiple ASSR technique also offers the possibility of quickly performing periodic hearing assessments on captive marine mammals to track changes in hearing ability over time and/or before and after specific events that could potentially affect hearing (i.e., periods of increased environmental noise, antibiotic treatment).

There are limited data regarding multiple ASSR measurements in marine mammals. Dolphin (1996) demonstrated that the bottlenose dolphin (*Tursiops truncatus*) auditory system could simultaneously track multienvelope stimuli created from two to four pure tones (one, three, or six dominant envelope components). Popov *et al.* (1997, 1998) reported suppression of the ASSR to a 76 kHz tone when a second tone with lower amplitude was simultaneously presented at 85 kHz, a frequency separation of approximately 1/6 octave. Finneran and Houser (2007) found good agreement between bottlenose dolphin hearing thresholds and input-output functions (i.e., ASSR amplitude as a function of stimulus level) in response to one to four simultaneous amplitude modulated tones with frequency separations as small as 0.4 octave. Finneran *et al.* (2007c) measured ASSR hearing thresholds and temporary threshold shift in a dolphin using seven simultaneous frequency modulated tones. The multiple ASSR results closely matched behavioral data with respect to the audiogram shape and upper frequency limit.

The present paper describes a series of experiments to characterize the hearing ability of a small (n=4) group of captive bottlenose dolphins. The specific goal was to compare thresholds obtained with single and multiple (four, five, and nine) simultaneous stimuli to determine the feasibility of using a large number of stimuli to quickly assess hearing over a broad frequency range. The primary application of this technique was envisioned to be hearing characterization of individuals for whom access was very limited.

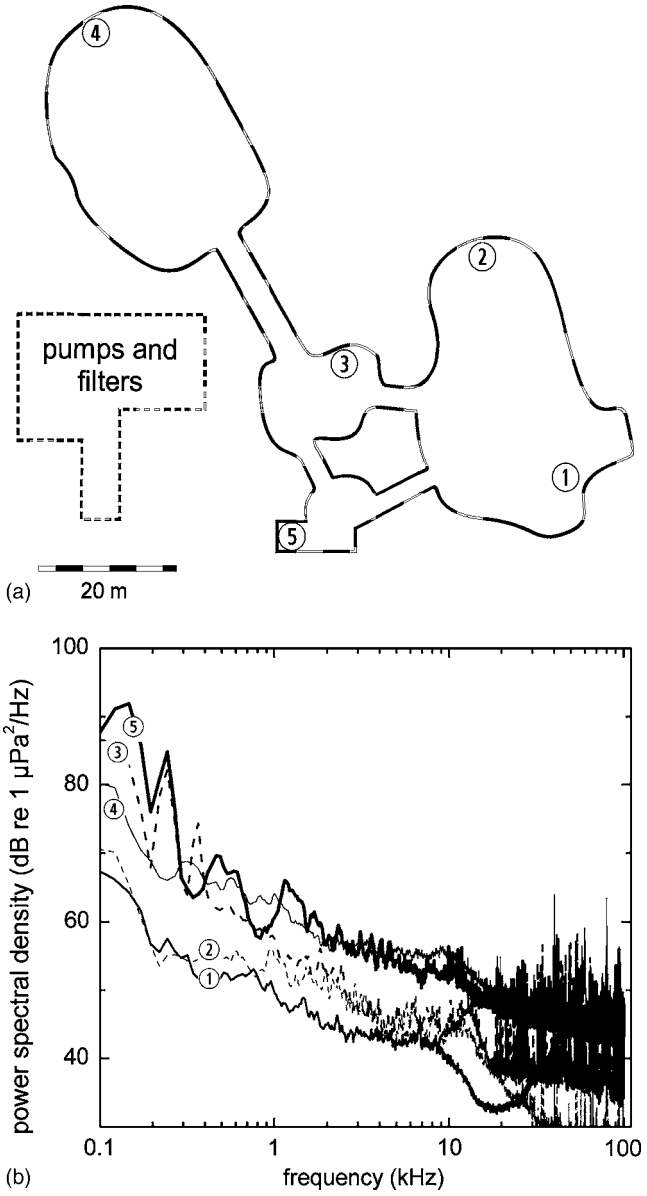


FIG. 1. (a) Geometry of the pools at the Mirage Dolphin Habitat showing the locations of the ambient noise measurements. (b) Noise power spectral densities at each location. Each trace represents the mean power spectral density (n=500) with a 24.4 Hz frequency resolution. Signals were high-pass filtered at 100 Hz and a Hanning window was applied before spectral analysis.

II. METHODS

A. Subjects

Experiments were conducted with four bottlenose dolphins: SA9701 (female, 9 years old), HU0001 (female, 7 years), MA0301 (male, 4 years), and LI0601 (male, 22 years). All tests were conducted in air with the subject resting on a foam mat. To reduce tension and anxiety, the dolphin LI0601 was given 50 mg of diazepam approximately 1 h prior to testing. Diazepam is a benzodiazepine which has been shown to have a mild effect on the latencies, but not the amplitudes, of short latency auditory evoked responses in humans (Adams *et al.*, 1985). Subjects were housed in a series of concrete pools with a total volume of 9460 m^3 [Fig. 1(a)] at the Mirage Dolphin Habitat in Las Vegas, NV. Ambient noise levels from 100 Hz to 100 kHz [Fig. 1(b)] were measured at various locations in the pools using a low-noise hydrophone (Reson TC4032) and preamplifier (Reson VP1000). Noise spectral densities between 1 and 10 kHz were generally between 45 and 70 dB re 1 $\mu Pa^2/Hz$. Spectra above 10 kHz were significantly influenced by echolocation pulses from nearby animals as well as some high frequency electrical noise.

B. Stimulus generation

Sound stimuli were presented using a "jawphone" consisting of a piezoelectric sound projector (ITC 1042 or Reson

TC 4013) embedded in a silicon rubber suction cup attached to the lower right jaw. The jawphone containing the TC 4013 was smaller than the other (2.5 cm diameter compared to 4 cm) and was used with subjects HU0001 and MA0301 because they were too small for the larger jawphone. Received sound pressure levels (SPLs) were estimated from underwater direct field measurements with a calibrated hydrophone placed 15 cm from the jawphone face (see Finneran and Houser, 2006 for details). There is no universally accepted technique for calibrating jawphones and the suitability of jawphone-measured thresholds as predictors of free field or direct field underwater thresholds is questionable; however, comparisons of AEP/jawphone and underwater behavioral thresholds in five dolphins showed that the jawphone measurements provide reasonable approximations to underwater behavioral thresholds (in dolphins), especially with respect to audible bandwidth (Finneran and Houser, 2006; Houser and Finneran, 2006a). Since the present study is concerned with differences between thresholds obtained with single and multiple stimuli delivered through the same jawphone, the specific jawphone calibration method is of secondary importance compared to the consistency in approach between measurements on the same individual.

Stimulus generation and evoked response recording were performed using custom software (the Evoked Response Study Tool, EVREST) running on a rugged notebook computer with a multifunction data acquisition board (National Instruments NI PCI-6251). Stimuli were digitally generated then converted to analog at a rate of 2 MHz with 16 bit resolution. Analog signals were then filtered (Krohn-Hite 3C series, bandpass 0.2–150 kHz), passed through a custom programmable attenuator (0–65 dB attenuation), and applied to the jawphone.

Stimuli consisted of the sum of one, four, five, or nine individual sinusoidal amplitude modulated tones with instantaneous voltage, $v(t)$, defined as

$$v(t) = \sum_{n=1}^{N} \frac{A_n}{2} \sin(2\pi F_n t)[1 - \cos(2\pi f_n t)], \quad (1)$$

where N is the number of waveform components (N=1, 4, 5, or 9), n indicates the nth component, A_n is the amplitude, F_n is the carrier frequency, and f_n is the modulation frequency. There were nine carrier frequencies of primary interest, chosen to cover the range from 10 to 160 kHz with a 1/2 octave frequency spacing: 10, 14.1, 20, 28.3, 40, 56.6, 80, 113.1, and 160 kHz (Table I). Some additional carrier frequencies (130, 140, or 150 kHz) were sometimes tested during single ASSR measurements to better define the steep increase in threshold occurring at high frequencies. All single ASSR stimuli used a 1 kHz modulation rate, had a duration of 22 ms (including a 1 ms rise and fall), and were presented intermittently at a rate of approximately 30 stimuli/s. Multiple ASSR stimuli consisted of the sum of either four, five, or nine amplitude modulated tones with durations of 62 ms (including a 1 ms rise and fall) and were presented intermittently at a rate of approximately 14 stimuli/s. Carrier frequencies were based on the nine 1/2 octave frequencies from 10 to 160 kHz identified earlier. Modulation frequencies were selected to be near those known to produce robust ASSRs in dolphins (Dolphin *et al.*, 1995; Supin and Popov, 1995; Finneran *et al.*, 2007b) and allow individual spectral components to be resolved with a 16.7 Hz resolution. Stimuli with nine components used carrier (and modulation) frequencies of 10 (0.90), 14.1 (0.95), 20 (1.00), 28.3 (1.05), 40 (1.10), 56.6 (1.15), 80 (1.20), 113.1 (1.25), and 160 (1.30) kHz. The frequency spectra of these components is shown in Fig. 2. Stimuli with five components used carrier (and modulation) frequencies of 10 (0.90), 20 (1.00), 40 (1.10), 80 (1.20), and 160 (1.30) kHz, while stimuli with four components used carrier (and modulation) frequencies of 14.1 (0.95), 28.3 (1.05), 56.6 (1.15), and 113.1 (1.25) kHz. The intent was to cover the frequency range from 10 to 160 kHz in a single measurement using a 1/2 octave carrier frequency separation (nine simultaneous components) and in two measurements, each with a full octave carrier frequency separation (four and five component stimuli).

TABLE I. Parameters for single and multiple amplitude modulated tones.

Number of components	Duration (ms)	Carrier frequencies (kHz)	Modulation frequencies (kHz)
1	22	10	1.00
		14.1	1.00
		20	1.00
		28.3	1.00
		40	1.00
		56.6	1.00
		80	1.00
		113.1	1.00
		160	1.00
4	62	14.1	0.95
		28.3	1.05
		56.6	1.15
		113.1	1.25
5	62	10	0.90
		20	1.00
		40	1.10
		80	1.20
		160[a,b]	1.30
9	62	10	0.90
		14.1	0.95
		20	1.00
		28.3	1.05
		40	1.10
		56.6	1.15
		80	1.20
		113.1	1.25
		160[a]	1.30

[a]145 kHz was substituted for 160 kHz with MA0301.
[b]The 160 kHz component was not included with HU0001.

C. Evoked potential recording and analysis

Evoked responses were measured using 10 mm gold cup surface electrodes embedded in silicon rubber suction cups and attached to the subject using conductive paste. Three

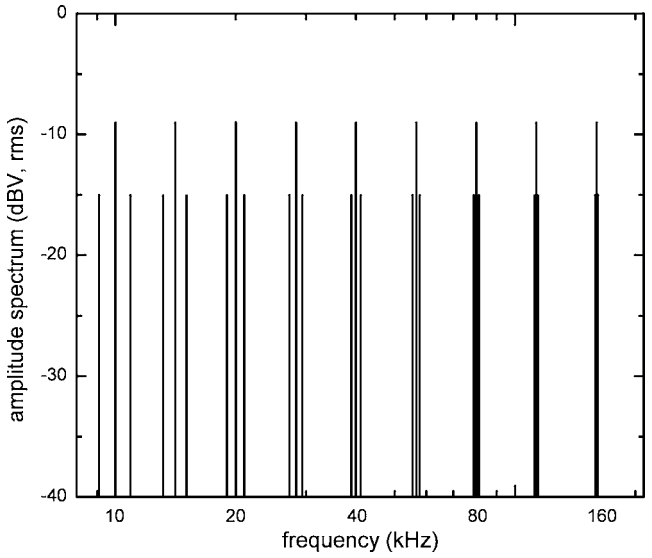


FIG. 2. Spectral amplitudes for amplitude modulated tone components with 1 V peak instantaneous voltage. The specific carrier (and modulation frequencies) shown are 10 (0.90), 14.1 (0.95), 20 (1.00), 28.3 (1.05), 40 (1.10), 56.6 (1.15), 80 (1.20), 113.1 (1.25), and 160 (1.30) kHz.

electrodes were used: A noninverting electrode placed just behind the blowhole, an inverting electrode placed just behind the external auditory meatus contralateral to the jawphone, and a common electrode placed near the dorsal fin. A biopotential amplifier (Grass ICP-511) was used to amplify ($\times 10^5$) and filter (0.3–3 kHz) the voltage between the noninverting and inverting electrodes before digitization at 10 kHz with a 16 bit resolution (NI PCI-6251). Any sweeps with peak voltage exceeding 12 μV were excluded from analysis. Frequency analysis was performed on the recorded time waveforms using analysis windows with durations of 20 ms (single ASSR) or 60 ms (multiple ASSR) centered on the evoked response, i.e., the response intervals corresponding to the stimulus rise and fall were excluded from the analysis. There was always an integral number of cycles of the modulating waveforms within the analysis window.

The presence or absence of an evoked response was determined after integral multiples of 250 sweeps were collected. If a response was detected at all of the modulation frequencies, the measurement was complete; if not, an additional 250 sweeps were collected and the process repeated (using all the available sweeps). The maximum number of sweeps collected was 1000 (single ASSR) or 500 (multiple ASSR). The longer stimuli for the multiple ASSR measurements resulted in better frequency resolution during spectral analysis and thus lower noise levels, therefore fewer averages were necessary to achieve comparable signal-to-noise ratios. At each integral multiple of 250 sweeps, coherent averaging in the frequency domain was used to obtain 20 unique "subaverages," each created from an equal number of consecutive sweeps (12, 25, 37, or 50 sweeps), and a single "grand average" created from all of the sweeps (240, 500, 740, or 1000 sweeps). Magnitude-squared coherence (MSC), which is a ratio of the power in the grand average to the average power of the subaverages (a ratio of signal power to signal-plus-noise power, Dobie and Wilson, 1989; 1996; Finneran *et al.*, 2007a) was then calculated. If the MSC was greater than the critical value (α=0.01, Brillinger, 1978), the response at a particular modulation frequency was considered to be detected.

D. Test sequence

Threshold testing began with each individual waveform component at a SPL of 96 or 110 dB re 1 μPa, depending on the jawphone. After each measurement, the component SPLs were independently adjusted using a modified up/down staircase approach. If a response was detected, the SPL was reduced by the step size ΔL; if a response was not detected, the SPL was increased by ΔL. The initial step size was 30 dB, except for 160 kHz (10 dB). After each reversal (a transition from detection to nondetection or vice versa), the step size was reduced according to the rules:

$$\Delta L_{k+1} = 0.4\Delta L_k \quad \text{(for reversals following detections)},$$

$$\Delta L_{k+1} = 0.45\Delta L_k \quad \text{(for reversals following nondetections)}, \qquad (2)$$

where ΔL_k is the step size for the kth measurement. The staircase was terminated when the step size for the next measurement was $<$3 dB. To preserve the number of waveform components within the stimulus, individual components were not turned off after reaching threshold, but instead the SPL was adjusted to fill in the largest gap within the ASSR input-output function. The threshold was defined as the mean of the stimulus SPLs corresponding to the lowest detection and the next highest nondetection.

Since the peak instantaneous voltage of the sum of several stimuli may be larger than that of any of the stimuli presented alone, some stimulus amplitudes that were tested during single ASSR measurements could not be tested using the multiple ASSR technique. This was particularly common at 160 kHz, where the required stimulus levels were relatively high, and with the smaller jawphone, which had a lower transmitting voltage response compared to the larger jawphone. To prevent clipping, the EVREST software calculated the instantaneous peak voltage of the combined stimulus before generation and, if necessary, proportionally reduced the amplitudes of all waveform components to allow the summed waveform to remain within the hardware output voltage limits. However, this led, at times, to situations where the measurements could not properly converge on the thresholds. For example, if the 160 kHz component could not attain the SPL necessary to elicit a response without the peak voltage exceeding the hardware limits, all components would be progressively reduced, which could affect another component's ability to reach threshold. For these reasons, 145 kHz was substituted for 160 kHz in the five and nine component stimuli for subject MA0301. For HU0001, tests with the nine component stimulus were repeated using different starting SPLs (a higher SPL at 160 kHz) and the five component stimulus did not include 160 kHz, making it actually a four component stimulus.

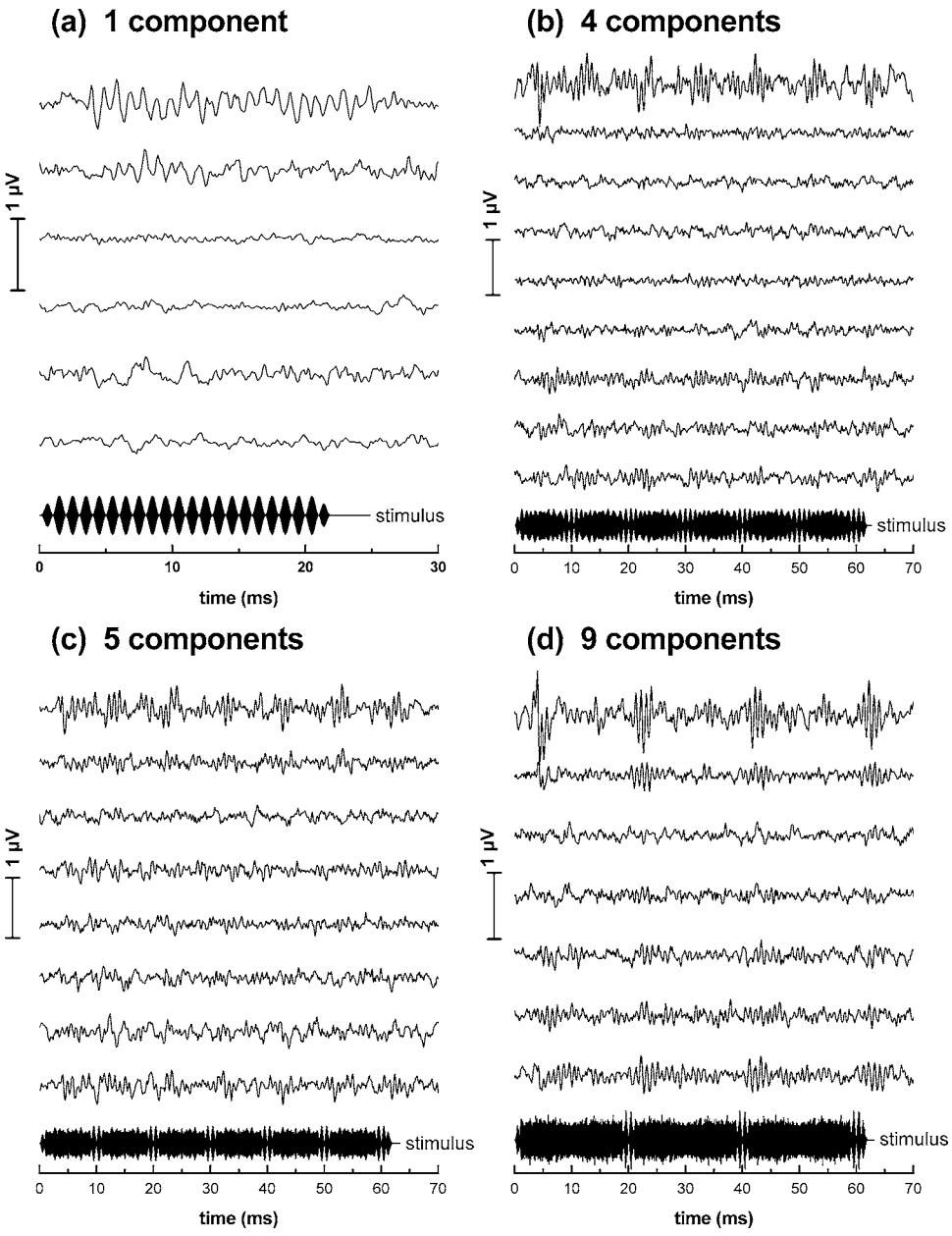


FIG. 3. Evoked response waveforms for the dolphin SA9701 obtained with (a) single (80 kHz), (b) four-component, (c) five-component, and (d) nine-component stimuli. The waveforms are arranged within each panel from top to bottom in the order of the measurements; see Fig. 4 for the individual SPLs of each waveform component. The stimulus waveforms shown in each panel correspond to the first evoked response.

III. RESULTS

Figures 3 and 4 show representative examples for the evoked response waveforms and amplitude spectra, respectively, for series of (a) single, (b) four-component, (c) five-component, and (d) nine-component stimuli. The stimulus levels were independently adjusted in staircase fashion, thus there is no consistent relationship between the order of the waveforms and frequency spectra and the individual component SPLs; however, Fig. 4 includes the SPL of each component beneath the spectral peak corresponding to its response. At relatively high levels the response to each waveform component was clear in the frequency domain, except at 160 kHz where it was difficult to generate sufficient SPLs for large amplitude responses. Background noise levels in the electrophysiological signals were relatively low and signals with spectral amplitude $\geqslant$12 nV were normally detected.

Figure 5 shows the hearing thresholds for each subject as functions of the (carrier) frequency (i.e., displayed as audiograms) for the single and multiple component stimuli. All of the audiograms exhibited the traditional "U-shape" common to mammals, with thresholds increasing steeply above 113 kHz. Upper cutoff frequencies, based on the frequency where the thresholds reached 120 dB re 1 μPa (as in Houser and Finneran, 2006b), were 150–160 kHz or higher, revealing full auditory bandwidth in these subjects and no significant high-frequency hearing loss. Audiograms produced from tests with single and multiple stimuli were similar in shape and consistent with respect to the upper cutoff frequencies.

Figure 6 shows the differences between single and multiple ASSR thresholds at each frequency. Note that since the nine-component test was performed twice for HU0001, each with different starting SPLs, two sets of nine-component re-

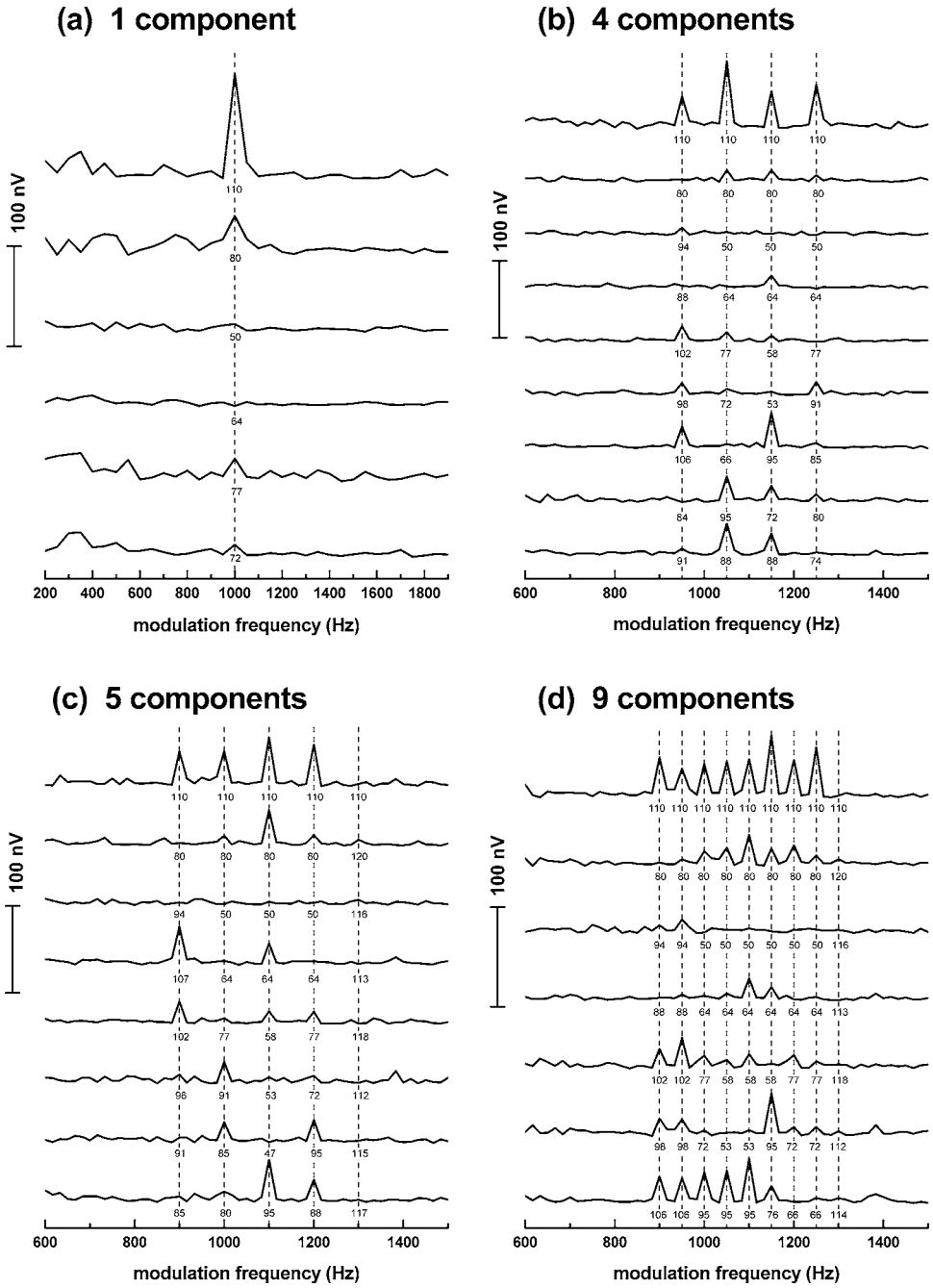


FIG. 4. Evoked response amplitude spectra for the dolphin SA9701 obtained with (a) single (80 kHz), (b) four-component, (c) five-component, and (d) nine-component stimuli. The dashed lines indicate the modulation frequencies for the individual waveform components. The spectra are arranged within each panel from top to bottom in the order of the measurements. The stimulus levels were independently adjusted in staircase fashion; the numerical values beneath each trace indicate the SPL of the component with the corresponding modulation rate (see Table I).

sults are presented. Fifty-four percent of the multiple ASSR thresholds were within ±5 dB of the single ASSR thresholds and 90% were within ±10 dB. The largest differences tended to occur in the areas of better sensitivity (lower thresholds); differences at the lower and higher frequencies, where thresholds were relatively high, were typically smaller. Multiple ASSR thresholds obtained with the full octave spacing were on average 1 to 2 dB closer to the single ASSR thresholds than the multiple ASSR thresholds obtained with the 1/2 octave spacing.

The total amount of time required to reach threshold at all nine 1/2 octave frequencies using the single and multiple ASSR methods is shown in Fig. 7. Mean times were 28.7 min when testing all nine frequencies separately, 12.1 min when performing two measurements, one with five stimuli and the other with four (frequency separation of one octave; only three values since HU0001 was not tested with five components), and 6.5 min when testing all nine components simultaneously (1/2 octave frequency separation). There were no significant differences between the total times for the sum of the four- and five-component stimuli compared to the nine-component stimuli, but both techniques were significantly faster than testing all nine frequencies independently (repeated measures analysis of variance, $p < 0.05$,GraphPad Software, 2003; data from HU0001 were excluded). Speed increases (the ratio of single ASSR time to multiple ASSR time) were 2.4 and 4.4 when using four and five stimuli and when using nine stimuli, respectively, and thus were not directly proportional to the number of components (i.e., using nine components was not nine times faster).

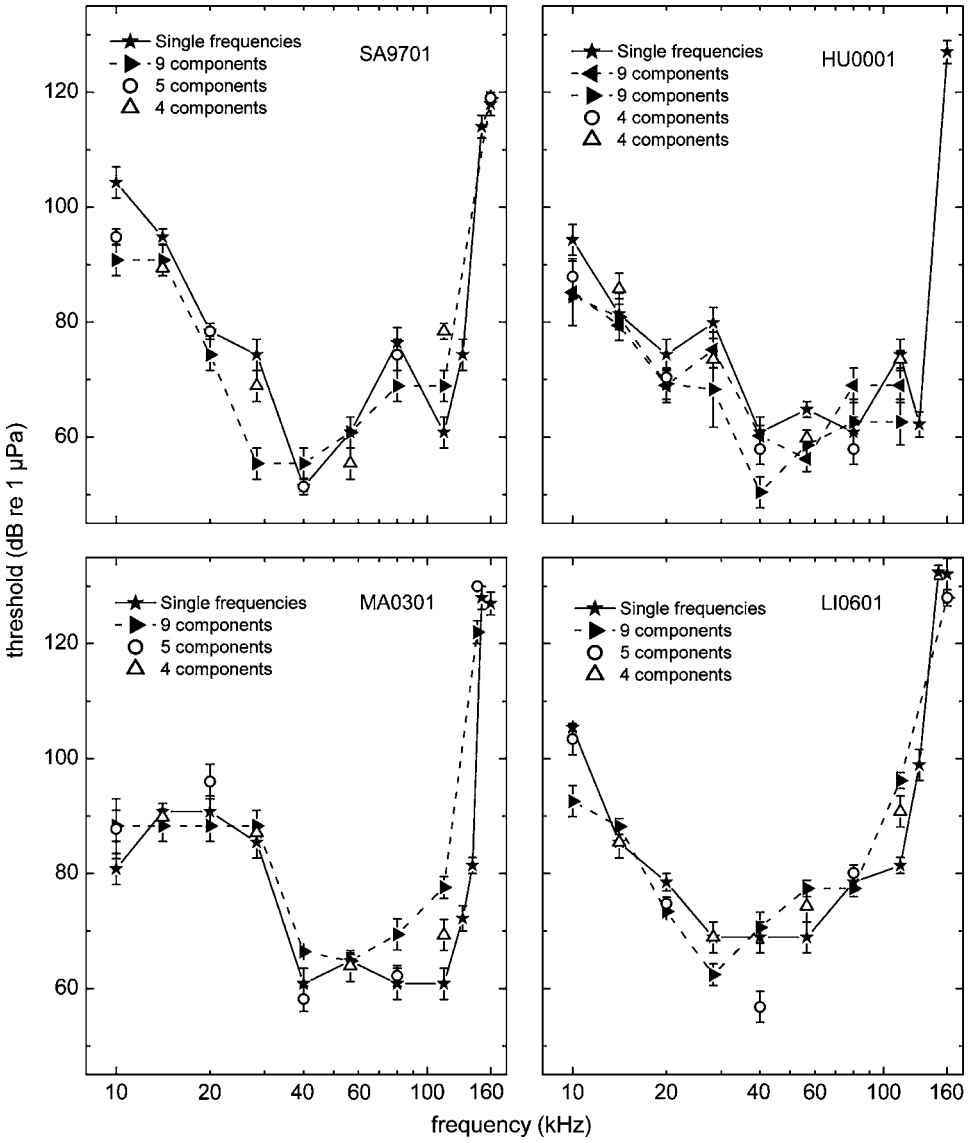


FIG. 5. Evoked response thresholds as functions of carrier frequency for the dolphins SA9701, HU0001, MA0301, and LI0601 obtained with single and multiple stimuli. Thresholds were defined as the mean of the SPLs corresponding to the lowest detection and the highest nondetection. For HU0001 160 kHz was not included in the five-component stimulus (making it a four-component stimulus) and there were no detections at 160 kHz with the nine-component stimulus, which was tested twice.

IV. DISCUSSION

Multiple ASSR measurements are relatively common for hearing assessment in people, where up to four components may be tested in each ear simultaneously (Lins and Picton, 1995; John *et al.*, 1998). Previous multiple ASSR measurements in dolphins have tested up to six (Dolphin, 1996) or seven (Finneran *et al.*, 2007c) simultaneous tones. The present study extended the number of simultaneous stimuli to nine, with a 1/2 octave frequency separation between adjacent components.

Subjects possessed upper cutoff frequencies (defined at 120 dB re 1 μPa) above 150 kHz, indicating no high-frequency hearing loss as has been observed in other captive populations of bottlenose dolphins (Houser and Finneran, 2006b; Houser *et al.*, 2007). The subjects were relatively young (4, 7, 9, and 22 years), with the oldest just reaching the age range where hearing loss began to appear in the largest group of captive dolphins tested to date (Houser and Finneran, 2006b). The ambient noise levels were reasonable for concrete pools operating with closed-circuit filtration and nearby pumps, with noise levels 15–20 dB lower than those measured in San Diego Bay (Finneran *et al.*, 2005).

Audiograms created from the single and multiple ASSR thresholds (Fig. 5) were similar, especially with respect to the general shape and upper cutoff frequency. Most thresholds for a particular subject and frequency were within ±10 dB of each other, regardless of the number of stimuli. The agreement between the single and nine-component multiple ASSR was particularly good considering the relatively close spacing between stimulus carrier frequencies (1/2 octave), less than the frequency separation typically considered necessary to avoid significant interactions (Lins and Picton, 1995; John *et al.*, 1998, 2002). In other words, thresholds were similar despite the likely occurrence of interactions between adjacent stimuli.

The required testing time for the multiple ASSR method was significantly faster than the single ASSR, with audiograms from 10 to 160 kHz obtained in as little as 5 min when using nine simultaneous stimuli. This makes the multiple ASSR technique ideal for situations where access to subjects is severely limited, e.g., with stranded or rehabilitated animals, juveniles, or pregnant females. In many circumstances (e.g., prior to release of a rehabilitated individual), potential errors introduced by interactions between

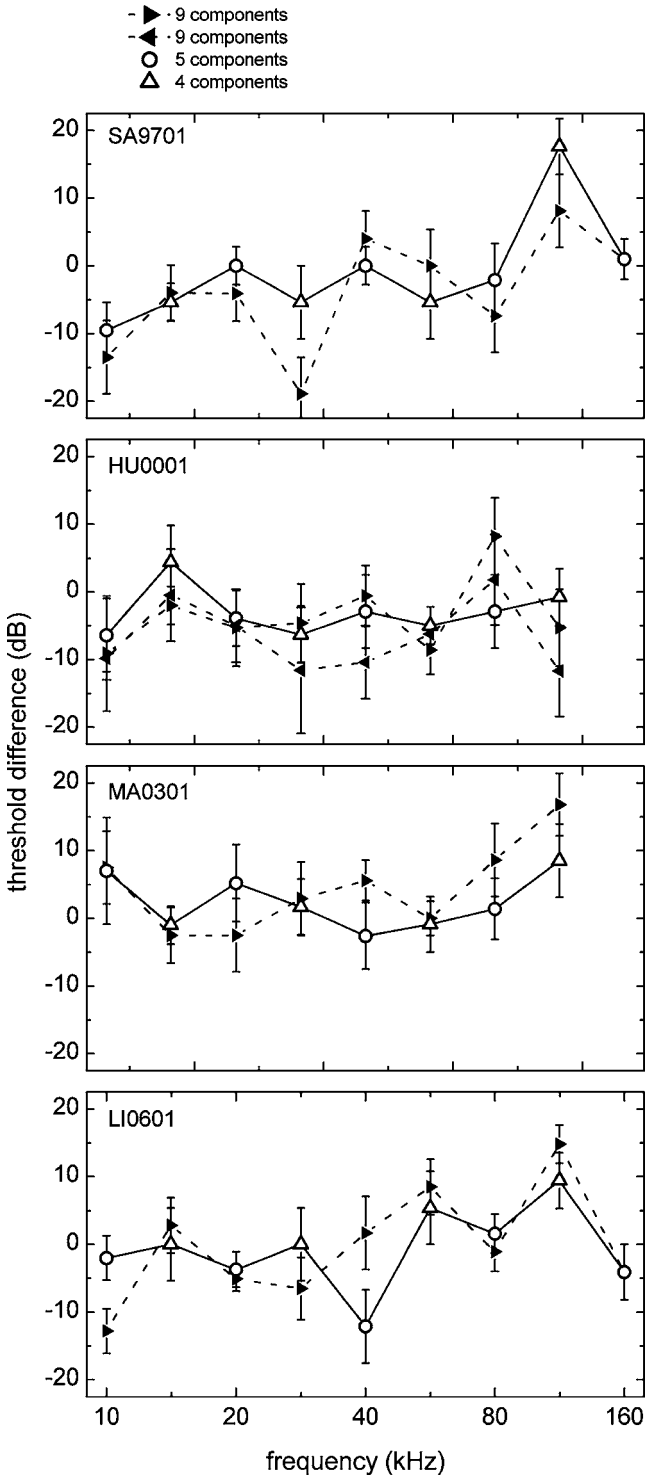


FIG. 6. Differences between single and multiple ASSR thresholds for each subject, defined as the threshold with a single stimulus subtracted from those from the multiple component stimuli.

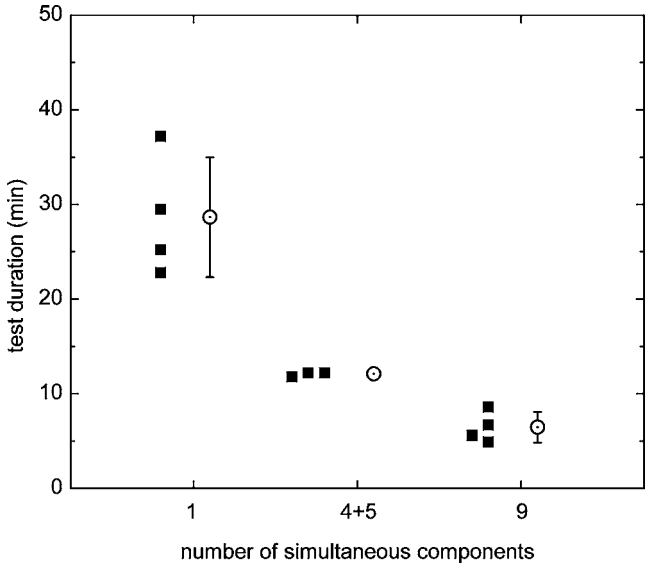


FIG. 7. Total amount of time required for threshold measurements using single and multiple component stimuli. Abscissa legend: 1=nine single ASSR thresholds; 4+5=total time for five-component multiple ASSR plus four-component multiple ASSR performed sequentially; 9 =nine-component multiple ASSR. Circles and error bars indicate the mean ±s.d. The squares show the individual values, with the *x* values offset for clarity when necessary. Test times for single ASSR thresholds are based only on the nine 1/2 octave frequencies from 10 to 160 kHz. The mean time from the two nine-component stimulus measurements was used for HU0001. There were only three data values for the 4+5 condition since 160 kHz was not included in what would have been the five-component stimulus for HU0001.

closely spaced stimuli are of secondary importance to the need to assess auditory system health in as little time as possible and thus the use of stimuli with 1/2 octave separation may be justified, despite the potential for interactions. In situations where more time is available, the multiple ASSR technique may still be used to provide a quick estimate of sensitivity, which could be followed with more detailed measurements using fewer, more widely spaced multiple components or single frequencies if desired.

One of the chief limitations of the multiple ASSR technique as applied in the current study was the difficulty of simultaneously generating suprathreshold stimuli over frequency regions where sensitivity differed greatly (sometimes referred to as areas of "sloping" hearing loss). In particular, it was difficult to generate SPLs at 160 kHz sufficient to produce detectable responses when multiple stimuli were presented. This issue was exacerbated by the EVREST software, which featured an output "limiter" to prevent clipping of the output waveform when the sum of the waveform components exceeded the hardware capabilities. Because the limiter proportionally reduced all waveform components, a single component with large amplitude could have a dramatic effect on other, smaller amplitude components and prevent the measurements from converging properly, in the sense that threshold levels would not be attained or would require an inordinate amount of time. To help avoid this problem in future measurements, the limiter function has been modified to only reduce waveform components if they are within 30 dB of the single largest amplitude component.

Another limitation of the present implementation of the multiple ASSR technique was the requirement that all measurements within a series used the same number of waveform components. This meant that once a particular component reached threshold it was not turned off but was instead adjusted to fill in the largest existing gaps in the ASSR input-output function (to improve the resolution). This could result in large amplitude components at frequencies close to small amplitude components, increasing the potential for masking or suppression of near-threshold responses (Popov *et al.*,

1997; John *et al.*, 1998; Popov *et al.*, 1998). Areas of sloping hearing loss also increase the potential for masking of near-threshold tones by nearby stimuli at much higher levels. Indeed, some of the larger differences between single and multiple ASSR thresholds occurred near the upper cutoff frequency and thus may have resulted from such an effect. Performance of the multiple ASSR technique would probably improve if waveform components remained at near-threshold levels or components were turned off after reaching threshold.

V. CONCLUSIONS

Multiple ASSR measurements with up to nine simultaneous stimuli may be used for quick hearing threshold estimates in bottlenose dolphins when time is limited. Multiple ASSR measurements faithfully capture the major features of the audiogram, especially the upper cutoff frequency; however, discrepancies between single and multiple ASSR results may exceed ±10 dB, especially when the threshold changes substantially with small changes in frequency.

ACKNOWLEDGMENTS

The authors thank the staff and volunteers at the Mirage Dolphin Habitat. Development of the EVREST software was supported by the US Office of Naval Research. Access to LI0601 was provided courtesy of Kevin Walsh (Gulf World Marine Park).

Adams, D. A., McClelland, R. J., Houston, H. G., and Gamble, W. G. (**1985**). "The effects of diazepam on the auditory brain stem responses," Br. J. Audiol. **19**, 277–280.

Brillinger, D. R. (**1978**). "A note on the estimation of evoked response," Biol. Cybern. **31**, 141–144.

Bullock, T. H., Grinnell, A. D., Ikezono, E., Kameda, K., Katsuki, K., Nomoto, M., Sato, O., Suga, N., and Yanagisawa, K. (**1968**). "Electrophysiological studies of central auditory mechanisms in cetaceans," Zeitschrift fur Vergleichende Physiologie **59**, 117–156.

Campbell, F. W., Atkinson, J., Francis, M. R., and Green, D. M. (**1977**). "Estimation of auditory thresholds using evoked potentials," in *Progress in Clinical Neurophysiology*, edited by J. E. Desmedt (Karger, Basel), pp. 68–78.

Cook, M. L. H., Varela, R. A., Goldstein, J. D., McCulloch, S. D., Bossart, G. D., Finneran, J. J., Houser, D., and Mann, D. A. (**2006**). "Beaked whale auditory evoked potential hearing measurements," J. Comp. Physiol., A **192**, 489–495.

Dobie, R. A., and Wilson, M. J. (**1989**). "Analysis of auditory evoked potentials by magnitude-squared coherence," Ear Hear. **10**, 2–13.

Dobie, R. A., and Wilson, M. J. (**1996**). "A comparison of *t* test, *F* test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids," J. Acoust. Soc. Am. **100**, 2236–2246.

Dolphin, W. F. (**1996**). "Auditory evoked responses to amplitude modulated stimuli consisting of multiple envelope components," J. Comp. Physiol., A **179**, 113–121.

Dolphin, W. F., Au, W. W., Nachtigall, P. E., and Pawloski, J. (**1995**). "Modulation rate transfer functions to low-frequency carriers in three species of cetaceans," J. Comp. Physiol., A **177**, 235–245.

Eggermont, J. J. (**2007**). "Electric and magnetic fields of synchronous neural activity: Peripheral and central origins of AEPs," in *Auditory Evoked Potentials: Basic Principles and Clinical Applications*, edited by R. F. Burkard, J. J. Eggermont, and M. Don (Lippencott Williams & Wilkins, Philadelphia, PA), pp. 2–21.

Finneran, J. J., Carder, D. A., Schlundt, C. E., and Ridgway, S. H. (**2005**). "Temporary threshold shift (TTS) in bottlenose dolphins (*Tursiops truncatus*) exposed to mid-frequency tones," J. Acoust. Soc. Am. **118**, 2696–2705.

Finneran, J. J., and Houser, D. S. (**2006**). "Comparison of in-air evoked potential and underwater behavioral hearing thresholds in four bottlenose dolphins (*Tursiops truncatus*)," J. Acoust. Soc. Am. **119**, 3181–3192.

Finneran, J. J., and Houser, D. S. (**2007**). "Bottlenose dolphin (*Tursiops truncatus*) steady-state evoked responses to multiple simultaneous sinusoidal amplitude modulated tones," J. Acoust. Soc. Am. **121**, 1775–1782.

Finneran, J. J., Houser, D. S., and Schlundt, C. E. (**2007a**). "Objective detection of bottlenose dolphin (*Tursiops truncatus*) steady-state auditory evoked potentials in response to AM/FM tones," Aquat. Mamm. **33**, 43–54.

Finneran, J. J., London, H. R., and Houser, D. S. (**2007b**). "Modulation rate transfer functions in bottlenose dolphins (*Tursiops truncatus*) with normal hearing and high-frequency hearing loss," J. Comp. Physiol., A **193**, 835–843.

Finneran, J. J., Schlundt, C. E., Branstetter, B., and Dear, R. L. (**2007c**). "Assessing temporary threshold shift in a bottlenose dolphin (*Tursiops truncatus*) using multiple simultaneous auditory evoked potentials," J. Acoust. Soc. Am. **122**, 1249—1264.

GraphPad Software (**2003**). "GraphPad Prism," GraphPad Software, San Diego, CA.

Hall, J. W. (**1979**). "Auditory brainstem frequency following responses to waveform envelope periodicity," Science **205**, 1297–1299.

Houser, D. S., and Finneran, J. J. (**2006a**). "A comparison of underwater hearing sensitivity in bottlenose dolphins (*Tursiops truncatus*) determined by electrophysiological and behavioral methods," J. Acoust. Soc. Am. **120**, 1713–1722.

Houser, D. S., and Finneran, J. J. (**2006b**). "Variation in the hearing sensitivity of a dolphin population obtained through the use of evoked potential audiometry," J. Acoust. Soc. Am. **120**, 4090–4099.

Houser, D. S., Gomez-Rubio, A., and Finneran, J. J. (**2008**). "Evoked potential audiometry of 13 Pacific bottlenose dolphins (*Tursiops truncatus gilli*)," Mar. Mammal Sci., **24** (in press).

John, M. S., Lins, O. G., Boucher, B. L., and Picton, T. W. (**1998**). "Multiple auditory steady-state responses (MASTER): Stimulus and recording parameters," Audiology **37**, 59–82.

John, M. S., Purcell, D. W., Dimitrijevic, A., and Picton, T. W. (**2002**). "Advantages and caveats when recording steady-state responses to multiple simultaneous stimuli," J. Am. Acad. Audiol **13**, 246–259.

Lins, O. G., Picton, P. E., Picton, T. W., Champagne, S. C., and Durieux-Smith, A. (**1995**). "Auditory steady-state responses to tones amplitude-modulated at 80–110 Hz," J. Acoust. Soc. Am. **97**, 3051–3063.

Lins, O. G., and Picton, T. W. (**1995**). "Auditory steady-state responses to multiple simultaneous stimuli," Electroencephalogr. Clin. Neurophysiol. **96**, 420–432.

Nachtigall, P. E., Lemonds, D. W., and Roitblat, H. L. (**2000**). "Psychoacoustic studies of dolphin and whale hearing," in *Hearing by Whales and Dolphins*, edited by W. W. L. Au, A. N. Popper, and R. R. Fay (Springer, New York), pp. 330–363.

Nachtigall, P. E., Supin, A. Y., Amundin, M., Roken, B., Møller, T., Mooney, T. A., Taylor, K. A., and Yuen, M. (**2007**). "Polar bear *Ursus maritimus* hearing measured with auditory evoked potentials," J. Exp. Biol. **210**, 1116–1122.

Nachtigall, P. E., Yuen, M. M. L., Mooney, T. A., and Taylor, K. A. (**2005**). "Hearing measurements from a stranded infant Risso's dolphin, *Grampus griseus*," J. Exp. Biol. **208**, 4181–4188.

National Research Council (NRC) (**1994**). *Low-Frequency Sound and Marine Mammals: Current Knowledge and Research Needs* (National Academy Press, Washington, DC).

National Research Council (NRC) (**2000**). *Marine Mammals and Low-Frequency Sound: Progress Since 1994* (National Academy Press, Washington, DC).

National Research Council (NRC) (**2003**). *Ocean Noise and Marine Mammals* (National Academies Press, Washington, DC).

National Research Council (NRC) (**2005**). *Marine Mammal Populations and Ocean Noise* (National Academies Press, Washington, DC).

Picton, T. W., Skinner, C. R., Champagne, S. C., Kellett, A. J., and Maiste, A. C. (**1987**). "Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone," J. Acoust. Soc. Am. **82**, 165–178.

Popov, V. V., and Supin, A. Y. (**1985**). "Determining hearing characteristics in dolphins using evoked potentials of brain stem," Dokl. Akad. Nauk SSSR **283**, 496–499.

Popov, V. V., and Supin, A. Y. (**1990**). "Auditory brainstem responses in characterization of dolphin hearing," J. Comp. Physiol., A **166**, 385–393.

Popov, V. V., Supin, A. Y., and Klishin, V. O. (**1992**). "Electrophysiological

study of sound conduction in dolphins," in *Marine Mammal Sensory Systems*, edited by J. A. Thomas, R. A. Kastelein, and A. Y. Supin (Plenum, New York), pp. 269–276.
Popov, V. V., Supin, A. Y., and Klishin, V. O. (**1997**). "Paradoxical lateral suppression in the dolphin's auditory system: Weak sounds suppress response to strong sounds," Neurosci. Lett. **234**, 51–54.
Popov, V. V., Supin, A. Y., and Klishin, V. O. (**1998**). "Lateral suppression of rhythmic evoked responses in the dolphin's auditory system," Hear. Res. **126**, 126–134.
Popov, V. V., Supin, A. Y., Wang, D., Wank, K., Xiao, J., and Li, S. (**2005**). "Evoked-potential audiogram of the Yangtze finless porpoise *Neophocaena phocaenoides asiaeorientalis* (L)," J. Acoust. Soc. Am. **117**, 2728–2731.
Regan, D., and Regan, M. P. (**1988**). "The transducer characteristic of hair cells in the human ear: A possible objective measure," Brain Res. **438**, 363–365.
Richardson, W. J., Greene, C. R., Jr., Malme, C. I., and Thomson, D. H. (**1995**). *Marine Mammals and Noise* (Academic, New York).
Ridgway, S. H., Bullock, T. H., Carder, D. A., Seeley, R. L., Woods, D., and Galambos, R. (**1981**). "Auditory brainstem response in dolphins," Neurobiology **78**, 1943–1947.
Stapells, D. R., Linden, D., Suffield, J. B., Hamel, G., and Picton, T. W. (**1984**). "Human auditory steady-state potentials," Ear Hear. **5**, 105–113.
Supin, A. Y., and Popov, V. V. (**1995**). "Envelope-following response and modulation transfer function in the dolphin's auditory system," Hear. Res. **92**, 38–46.
Supin, A. Y., and Popov, V. V. (**2000**). "Frequency-modulation sensitivity in bottlenose dolphins, *Tursiops truncatus*: Evoked-potential study," Aquat. Mamm. **26**, 83–94.
Szymanski, M. D., Bain, D. E., Kiehl, K., Pennington, S., Wong, S., and Henry, K. R. (**1999**). "Killer whale (*Orcinus orca*) hearing: Auditory brainstem response and behavioral audiograms," J. Acoust. Soc. Am. **106**, 1134–1141.
Yuen, M. M. L., Nachtigall, P. E., Breese, M., and Supin, A. Y. (**2005**). "Behavioral and auditory evoked potential audiograms of a false killer whale (*Pseudorca crassidens*)," J. Acoust. Soc. Am. **118**, 2688–2695.

Evidence for double acoustic windows in the dolphin, *Tursiops truncatus*

Vladimir V. Popov, Alexander Ya. Supin,[a] Vladimir O. Klishin, Mikhail B. Tarakanov, and Mikhail G. Pletenko
Institute of Ecology and Evolution of the Russian Academy of Sciences, 33 Leninsky Prospect, 119071 Moscow, Russia

(Received 18 June 2007; revised 22 October 2007; accepted 26 October 2007)

In a bottlenose dolphin positions of sound receiving areas on the head surface were determined by comparing the acoustic delays from different sound-source positions. For this investigation, auditory brainstem responses (ABRs) to short tone pips were recorded and their latencies were measured at different sound source positions. After correction for the latency dependence on response amplitude, the difference in ABR latencies was adopted as being the difference of the acoustic delays. These delay differences were used to calculate the position of the sound-receiving point. Measurements were conducted at sound frequencies from 16 to 128 kHz, in half-octave steps. At probe frequencies of 16 and 22.5 kHz, the receiving area was located 21.7–26 cm caudal of the melon tip, which is near the bulla and auditory meatus. At higher probe frequencies, from 32 to 128 kHz, the receiving area was located from 9.3 to 13.1 cm caudal of the melon tip, which corresponds to a proximal part of the lower jaw. Thus, at least two sound-receiving areas (acoustic windows) with different frequency sensitivity were identified. © *2008 Acoustical Society of America.*
[DOI: 10.1121/1.2816564]

PACS number(s): 43.80.Lb [WWA] Pages: 552–560

I. INTRODUCTION

Pathways of sound propagation to the ear of cetaceans, in particular odontocetes (toothed whales, dolphins, and porpoises) are still a matter of discussion (reviewed in detail by Ketten, 1990, 1992a, b, 1997, 2000). As a result of adaptation to underwater hearing conditions, all parts of the odontocete outer, middle, and inner ears are modified as compared to the ear of terrestrial mammals. Therefore, the sound-propagation pathway characteristic of terrestrial mammals (through the external auditory canal and tympanic membrane) does not function in odontocetes. Although some past investigators suggested a role of the external auditory canal for sound transmission in odontocetes (Fraser and Purves, 1954, 1959, 1960), it is currently accepted now that it is not a way of sound propagation to the middle ear. The canal is very narrow, filled with cells and cerumen, and does not connect with the tympanic membrane. Respectively, the tympanic membrane is thick and modified into an elongated, conical structure, the tympanic cone, which does not connect to the middle-ear ossicles. Thus, neither external auditory canal nor tympanic membrane plays the same role in sound reception as in terrestrial mammals.

The best known hypothesis of the sound conduction in cetaceans implicates the lower jaw as the primary pathway (Norris, 1968, 1969, 1980). The dolphin mandibular channel was found to contain a fatty body with an acoustic impedance close to that of sea water (Varanasi and Malins, 1971, 1972). The distal end of the fatty body contacts the outer surface of the lower jaw through a thin bony plate and its proximal end is close to the tympanic bulla which houses the middle ear. It was supposed that sounds enter the fat channel through the thin bony plate, and the channel functions as a specific pathway conducting sounds to the middle ear which is located just near the rear edge of the lower jaw. The region of the lower-jaw surface where sounds enter the fatty body was referred to as the *acoustic window*.

This "mandibular hypothesis" was supported by various data. Both intracranially recorded evoked responses (Bullock *et al.*, 1968) and cochlear action potentials (McCormick *et al.*, 1970, 1980) have revealed the lowest threshold when a sound source is placed on or near the lower jaw. A similar study with the use of a specially designed contact transducer in a suction cup and noninvasively recorded auditory brainstem responses (ABRs) also revealed the highest sensitivity at the middle of the lower jaw surface (Møhl *et al.*, 1999). An attempt to investigate the role of the lower jaw in echolocation performance was made by Brill (1988, 1991) and Brill *et al.* (1988) who placed a sound-shielding neoprene hood on the lower jaw of a dolphin. The hood impaired echolocation performance and this result was considered as evidence in favor of the mandibular hypothesis.

However, not all experimental data could be explained based on the mandibular hypothesis. In evoked-potential experiments, Popov and Supin (1990) determined the location of sound-receiving points by measuring acoustic delays from sound sources of varying positions. In three odontocete species, the sound-receiving point was found to be next to the tympanic bulla and far behind of the lower jaw. On the other hand, those results did not indicate that sounds were transmitted *only* through a point next to the bulla because the delay-vs-azimuth function was more complicated than should be for a single receiving point (Popov *et al.*, 1992).

[a]Author to whom correspondence should be addressed. Electronic mail: alex_supin@mail.ru

Further evidence in favor of multiple sound-receiving points in odontocetes was found when directivity diagrams (threshold-vs-azimuth functions) were measured in a dolphin at various sound frequencies (Popov *et al.*, 2006). The best-sensitivity azimuth depended on sound frequency: the lower the frequency, the more deviation of the best-sensitivity azimuth from the head midline. This feature cannot be explained based on a single receiving aperture since an aperture is always the most sensitive at its axis. Therefore, the data were explained by the presence of two (or more) sound-receiving apertures with different frequency sensitivities: a high-frequency receiving aperture with the axis close to the midline and a low-frequency aperture with the axis deviated from the midline.

Although the presence of more than one acoustic window in cetaceans is seemingly probable, it cannot yet be considered as indisputably confirmed. Little is known concerning the frequency selectivity of the supposed multiple windows. The goal of the present study was to fill in this gap.

To localize acoustic windows, we measured acoustic delays from differently positioned sound sources. The delay measurement was based on latency of the auditory evoked potentials (AEPs). In principle, this method was already used in our preceding studies (Popov and Supin, 1990; Popov *et al.*, 1992). However, the previous studies only used wideband clicks. In the present study, narrowband sound probes were used to investigate how the sound-receiving point position depends on sound frequency.

II. METHODS

A. Measurement paradigm

The principle of the method used to localize sound-receiving points is presented in Fig. 1. Suppose a sound receiving device (the receiver itself or an acoustic window channeled to the receiver) is located in an unknown point X [Fig. 1(a)]. The sound source S is moved around the receiver in a certain manner, for example, by a circle of a radius R with a center at a certain reference point O. The distance d from the sound source to the receiver varies depending on the source angular position: in the presented example, it is the shortest at the position S_1 (distance d_1) and the longest at the position S_2 (distance d_2). If $R \gg r$ (where r is the distance from the reference point O to the receiver X), the distance-vs-angle dependence is a cosine function [Fig. 1(b)] with an amplitude r equal to the distance r from the reference point to the receiver, and the phase shift α equal to the angular position α of the receiver.

The variation of the source-to-receiver distance results in a corresponding variation in the acoustic delay which adds to the latency of responses to the sound signal. The resulting latency variation can be measured experimentally. Assuming that the physiological latency of the response is constant, the variation of the response latency can be attributed to the acoustic delay variation. Conversion of the acoustic delay variation to the distance variation by multiplying by the sound velocity results in a cosine function like that in Fig. 1(b). Its amplitude and phase indicate, respectively, the distance and azimuth of the receiver X relative to the reference point O.

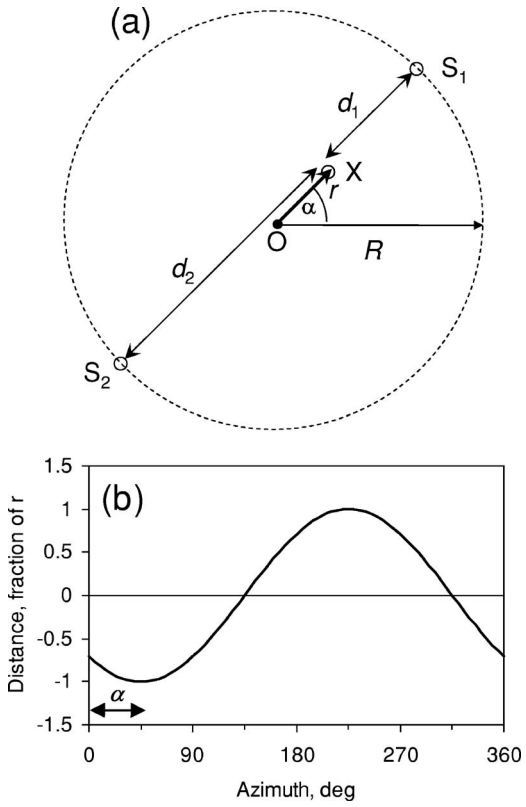


FIG. 1. (a) Experimental paradigm for finding a sound-receiving point by acoustic delays. O—reference point, X—searched-for sound receiving point, S_1 and S_2—opposite positionS of the sound source S, d_1 and d_2—distances from positions S_1 and S_2 to the point X, R—radius of the circle of sound-source movement, r—distance from the reference point O to the point X, α—azimuth of the point X. (b) Dependence of source-to-receiver distance on sound-source azimuth; the middle distance is taken as arbitrary zero; r and α—cosine function amplitude and phase.

If sounds reach a receiver trough a specific channel connecting the receiver and a certain receiving aperture (acoustic window), variation of the acoustic delays reveals the position of the window, because the transmission time through the channel is independent of the sound-source position, and variation of the acoustic delays depends on distance between the sound source and receiving window. Thus, the procedure is appropriate for finding acoustic window positions.

Based on this paradigm, the measurement procedure was adopted as follows:

(i) Auditory evoked responses (AEPs) were recorded to sound stimuli emitted from a number of sound-source positions, all positions at equal distance R from an arbitrarily chosen reference point of the head.

(ii) AEP latencies were measured as a function of the sound-source azimuth.

(iii) Measures were taken to compensate for possible azimuth-dependent variation in the physiological latency.

(iv) The resulting delay-vs-azimuth function was approximated by a cosine function.

(v) The cosine delay-vs-azimuth function was converted to a distance-vs-azimuth function.

(vi) The amplitude and phase of this cosine distance-vs-azimuth function was taken as the distance and azimuth of the receiving point relative to the reference point.

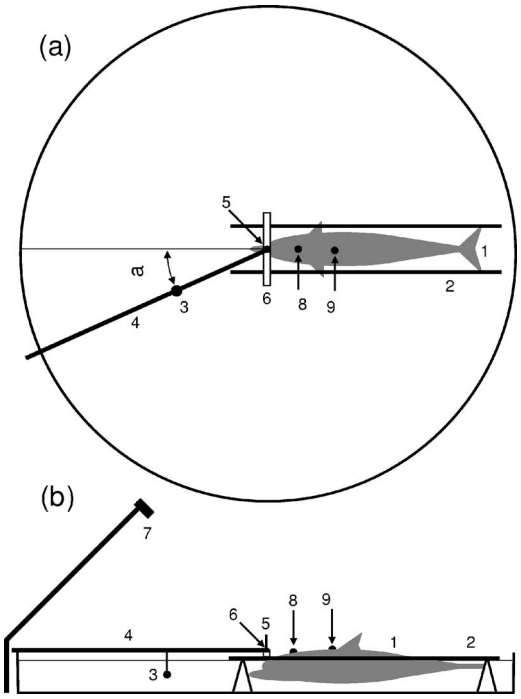


FIG. 2. Experimental design. (a) Dorsal view. (b) Lateral view. Explanation in text.

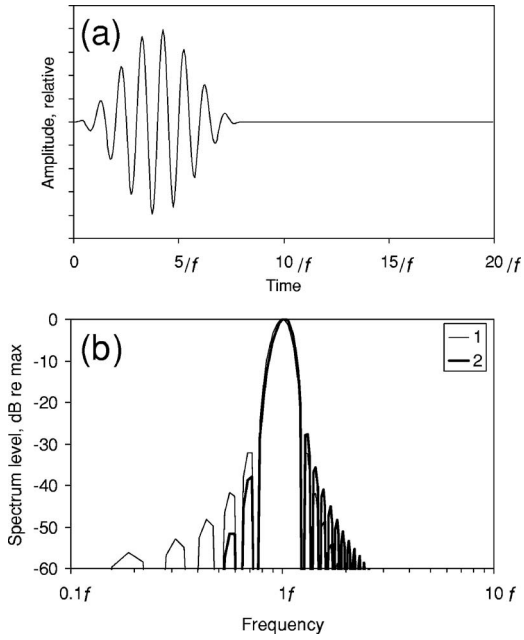


FIG. 3. (a) Stimulus waveform. (b) Stimulus spectrum, 1—electric signal, 2—sound signal played through a transducer frequency response of 12 dB/oct.

B. Subject

The experimental animal was an adult bottlenose dolphin, *Tursiops truncatus*, female, kept in the Utrish Marine Station of the Russian Academy of Sciences on the Black Sea Coast. The animal was housed in an on-land seawater pool $9\times4\times1.2$ m. The use and care of the animal adhered to the guidelines of "Ethical Principles of the Acoustical Society of America for Research Involving Human and Non-Human Animals in Research and Publishing and Presentations."

C. Measurement conditions

During the measurements, the animal was housed in a circular experimental tank filled with sea water, 6 m in diameter, 0.45 m deep (Fig. 2). The animal, (1) rested on a stretcher, (2) was positioned in such a way that the main part of its body was submerged but the blowhole and a part of the back were above water. A sound-emitting transducer (3) was mounted on a bar (4), which could be rotated around a center pin (5) mounted on a support (6). The support (6) was mounted on the stretcher so that the rotation center (5) took one of four positions: above the animal's melon tip, 12.5, 25, or 37.5 cm behind the melon tip, all at the head midline. The animal was positioned in the tank in such a manner that the rotation center coincided with the tank center. The transducer was located at a distance of 1.2 m from the rotation center. Rotation of the bar allowed the placement of the transducer at varying azimuth angles relative to the longitudinal head axis. Since the rotation center coincided with the tank center, the distance from transducer to the tank walls was 1.8 m. Therefore, the path for sounds reflected from the walls to the animal's head was at least 3.6 m longer (the delay at least 2.4 ms longer) than the direct way from the transducer. The position of the animal's head was monitored by a video camera (7).

D. Acoustic stimulation

The acoustic probes were short sound pips [Fig. 3(a)] designed using modulation of a carrier frequency by one cycle of a cosine envelope. Carrier frequencies varied from 16 to 128 kHz by 1/2-octave steps, i.e., 16, 22.5, 32, 45, 64, 90, and 128 kHz. The frequency of the cosine envelope was always 8 times as low as the carrier frequency, that is, from 2 kHz at the 16 kHz carrier to 16 kHz at the 128 kHz carrier. Thus, at all carrier frequencies, the pip contained a constant number (eight) of carrier cycles, and the pip spectrum was constant as expressed on a relative frequency scale [Fig. 3(b,1)]. The primary lobe width of the spectrum was ±0.25 of the carrier frequency, and sidelobes did not exceed −30 dB. Due to deformation of the spectrum by the hydrophone frequency response, the sidelobes were not higher than −37 dB below the main lobe and not higher than −28 dB above the main lobe [Fig. 3(b,2)].

The signals were digitally synthesized and digital-to-analog converted at an update rate of 512 kHz by an acquisition board E-6040 (National Instruments). The analog signals were amplified, attenuated and played through a B&K-8104 transducer. The transducer was placed at a distance of 1.2 m from the rotation center of the transducer-holding bar, at a depth of 22 cm (the mid-depth between the tank bottom and water surface). The stimulus sound pressure levels (peak-to-peak re 1 μPa) were 135 dB at 16 kHz, 140 dB at 22.5 kHz, 145 dB at 32 kHz, 150 dB at 45 kHz, and 160 dB at 64, 90, and 128 kHz. These levels were determined by maximal available voltage of the sound-power amplifier and

frequency response of the transducer. The azimuth position of the transducer varied within a range of ±165° from the longitudinal head axis, in steps of 15°.

E. Evoked-potential recording

Evoked potentials were recorded noninvasively using 1 cm stainless-steel disk electrodes secured at the body surface by rubber suction cups. The active electrode was placed at the vertex midline, 6–7 cm behind the blowhole [Fig. 2(8)]. The reference electrode was placed near the dorsal fin, above the water surface [Fig. 2(9)]. The recorded potentials were amplified, bandpass filtered between 200 and 5000 Hz, digitized at a sampling rate of 40 kHz using a 12 bit analog-to-digital converter and averaged by the data acquisition board E-6040. Each evoked response was collected by averaging 1000 poststimulus sweeps. The program for both stimulus generation and evoked response recording (a "virtual instrument") was designed using LabVIEW software (National Instruments).

F. Computation of the receiving point position

To evaluate the latency difference between AEPs recorded at different sound-source positions, the cross-correlation function (CCF) between these responses was calculated. The lag featuring the highest CCF value was taken as the latency difference.

To compensate for the possible dependence of AEP latency on stimulus efficiency (which may be unequal at different sound-source positions), AEPs were recorded at a constant (zero azimuth) sound-source position and at a variety of stimulus intensities, from the threshold to that producing maximum available (saturation) AEP amplitude. For this series of AEP records, the latency-vs-amplitude dependence was approximated by a straight regression line. The resulting latency-vs-amplitude slope was used to add/subtract a latency correction according to AEP amplitude.

The resulting and corrected latency estimates were averaged between the symmetrical right and left sound-source positions. The resulting latency-vs-azimuth dependence (within an azimuth range from 0° to 165°) was approximated by a 165°-long fragment of a function

$$l(\alpha) = C + d\cos[\pi(\alpha - \varphi)/180],$$

where l (ms) is the latency, C (ms) is a constant, d (ms) is the cosine amplitude, α (degrees) is the azimuth and φ (degrees) is the phase shift. For approximation, the parameters C, d, and φ were iteratively adjusted until reaching the best fit to the experimental data according to the least-mean-square criterion. The resulting value of φ was taken as the direction from the rotation center point to the searched-for sound receiving point and the amplitude d multiplied by the sound velocity was taken as the distance from the rotation center point to the receiving point. The sound velocity was adopted 1505 m/s as calculated by equation of Wilson (1960) at temperature of 22 °C, salinity of 15‰, and pressure of 0.1 MPa. The constant C was ignored as not bearing information of the receiving point position.

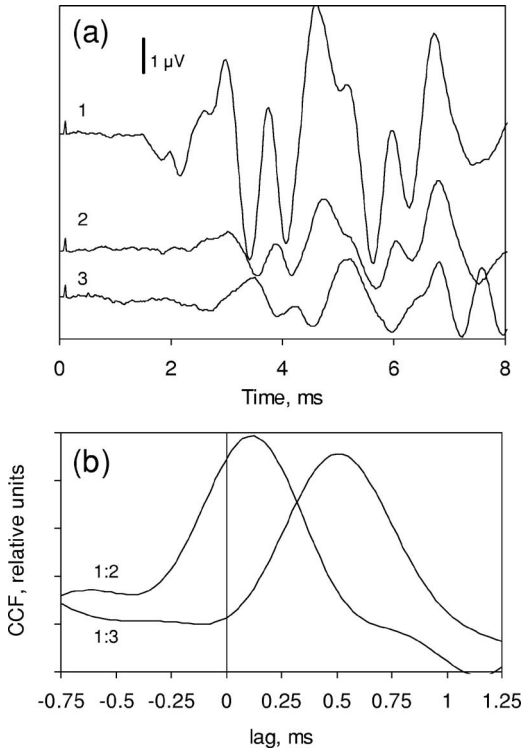


FIG. 4. (a) Examples of ABRs at different sound-source positions and sound intensities. Stimulus frequency 128 kHz. 1—azimuth 0°, intensity 160 dB; 2—0° 125 dB, 3—165° 160 dB. (b) Cross-correlation functions of waveforms 1 and 2 and waveforms 1 and 3 (functions 1:2 and 1:3, respectively).

III. RESULTS

A. AEP waveforms and latencies

Typical AEP waveforms recorded in the present study are exemplified in Fig. 4(a). The responses consisted of a few waves, each shorter than 1 ms, with an overall response duration of 3.5–4 ms and an onset latency of 1.5–2.5 ms. This response waveform was typical of the auditory brainstem evoked responses (ABRs) in odontocete whales, dolphins, and porpoises (rev. Supin *et al.*, 2001).

In many cases, the record contained two similar wave complexes with an interval of 2.4 to 4 ms. This feature is also exemplified in Fig. 4(a). The delay between the two responses corresponds to the delay between sounds directly spreading from the transducer to the animal's head and sound reflected from the nearest tank wall. Therefore, these two wave complexes were considered as responses to direct and reflected sounds, respectively. The delay between the two responses was long enough to measure the amplitude and latency of the first response (to the direct signal) while ignoring the second response (to the reflected sound).

The response latency was dependent on both stimulus intensity and sound-source position. Figure 4(a) exemplifies responses to probes of 128 kHz frequency, from one and the same sound-source position (0° azimuth) but of different intensities: 160 dB (1) and 125 dB re 1 μPa (2). Apart from lower amplitude, the response to lower probe intensity (2) featured longer latency as compared to the response to higher intensity (1). To evaluate the latency difference, a CCF between the waveforms (1) and (2) was calculated [Fig. 4(b)].

This CCF featured its peak at a lag of 0.125 ms; i.e., the waveform (2) was delayed by 0.125 ms relative to the waveform (1).

The influence of sound-source position on the response parameters is also exemplified in Fig. 4. The waveform (3) in Fig. 4(a) presents the response evoked by a sound source of 165° azimuth. The response featured both lower amplitude and longer latency than the response (1) evoked by the probe of 0° azimuth and the same 160 dB intensity. The lower response amplitude indicated lower hearing sensitivity at the 165° azimuth as compared to that at 0°.

The CCF between the waveforms (1) and (3) peaked at a lag of 0.5 ms; i.e., the waveform (3) was delayed by 0.5 ms relative to the waveform (1). This latency difference might be partially attributed to the amplitude-dependent variation. However, there was also a delay between responses (2) and (3) of almost equal amplitudes: the time shift between CCFs (1:2) and (1:3) was 0.375 ms. Assuming that this shift arose because of different acoustic delays, the distances from the two sound source positions to the sound receiver were estimated as differing by 0.375 ms × 150 cm/ms = 56 cm.

B. Acoustic delay dependence on sound-source azimuth

As Fig. 4 shows, AEP latency depended on sound-source azimuth both because of variation of the acoustic delay and because of azimuth-dependent variation of sensitivity which, in turn, influenced the response latency. In the present study, the acoustic delay was the matter of interest. Therefore, for making the necessary compensation possible, it had to be determined, which portion of the azimuth-dependent latency variation was a result of variation of hearing sensitivity.

For this purpose, each measurement session included AEP recordings at a constant (zero) sound source azimuth and at various probe intensities, from response threshold to maximum (saturation), in 5 dB steps. To evaluate both the amplitude variation and the latency shift, CCFs were calculated between the highest of the responses and each of other responses. Figure 5(a) exemplifies the magnitude and latency dependencies on stimulus intensity in a measurement session with a probe frequency of 128 kHz. When the response amplitude reached its maximum ("saturation" at 150 dB), the latency reached its minimum too and did not shorten with further intensity increase up to 160 dB. Thus, the latency variation was associated with the response magnitude rather than with the intensity itself. The latency dependence on response magnitude was nearly a straight line [Fig. 5(b)]. In this particular case, the regression line slope was −0.115 ms/μV.

The next step was AEP recordings at different azimuths, while keeping the probe intensity constant. Using these records, AEP delays relative to that at the zero azimuth were found using the same CCF calculation technique. At each of the sound frequencies, the measurements were repeated four times. The results of the four measurement sets were averaged. Figure 6(a) demonstrates the results of measurements at a probe frequency of 128 kHz and reference point position of 25 cm behind the melon tip, presented as delay-vs-azimuth functions. In this particular measurement session,

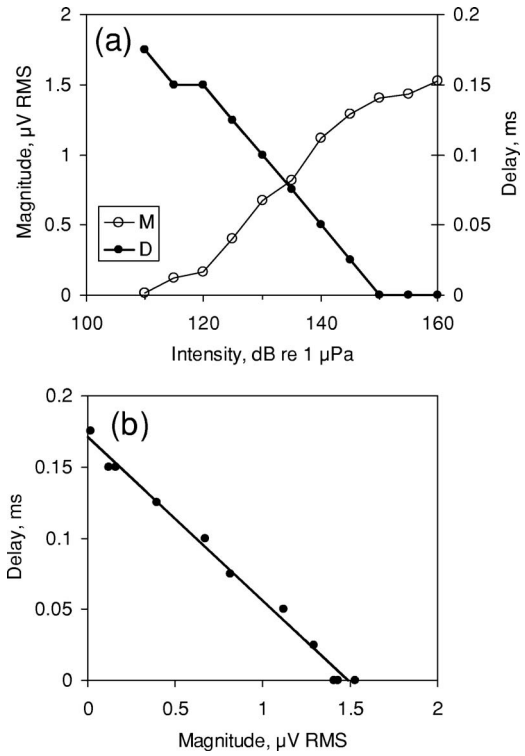


FIG. 5. (a) Dependence of ABR magnitude (M, left scale) and delay relative to the shortest latency (D, right scale) on stimulus intensity. (b) Dependence of ABR delay on magnitude. Stimulus frequency 128 kHz.

both the four original plots and their average featured a minimum latency at the zero azimuth; at all other azimuths, the responses latencies were longer. The delay reached 0.3–0.5 ms at the 165° azimuth.

The response magnitude was also maximal at the zero azimuth [Fig. 6(b)]; thus, a part of the delay-vs-azimuth dependence should be attributed to the latency dependence on response magnitude. Therefore, a correction was done as

$$d_c(\alpha) = d_\alpha + k(m_0 - m_\alpha),$$

where d_c is the corrected delay value at a certain azimuth α, d_α is the original delay value obtained at the azimuth α, m_0 is the response magnitude at the zero (reference) azimuth, m_α is response magnitude at the azimuth α, and k is the delay-vs-magnitude factor. In the measurement session exemplified in Fig. 6, correction using a factor of −0.115 ms/μV (see above, Fig. 5) resulted in a delay-vs-azimuth function presented in Fig. 6(c), 1. This function was adopted as the acoustic delay dependence on azimuth.

Assuming the right and left sound-conductive structures of the head were roughly symmetric, the right and left branches of the found function were symmetrically averaged. Thus, the final delay-vs-azimuth function was obtained by averaging the total of eight measurements within the azimuth range of 0 to 165°. Figure 6(c), 2 presents this average together with standard errors based on both the response delay [Fig. 6(a)] and magnitude [Fig. 6(b)] data scatters. Approximation of this final plot by a 165° fragment of a cosine function gave the best fit to experimental data at a cosine amplitude of 0.14 ms and a phase (minimum delay) of 32°.

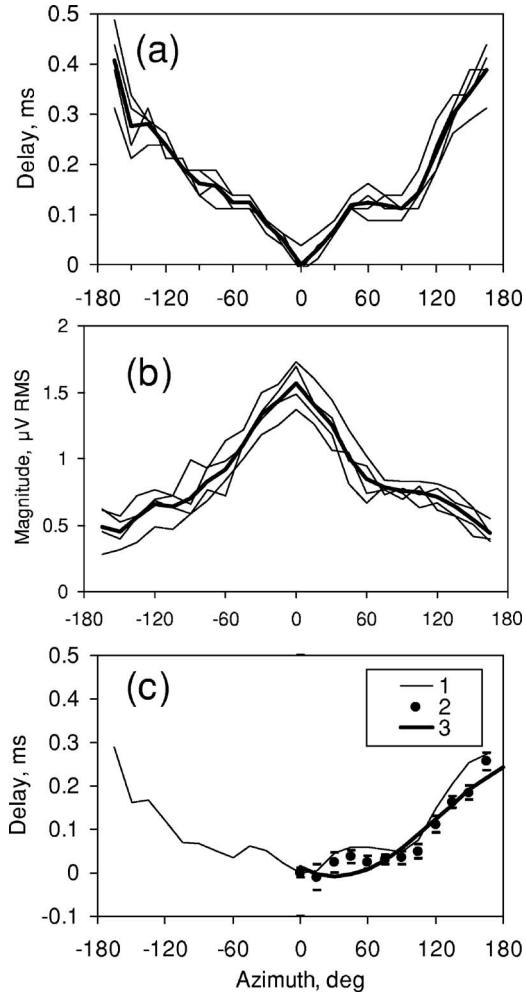


FIG. 6. (a) ABR delay (relative to the shortest latency) dependence on sound-source azimuth. Probe frequency 128 kHz. Thin lines—four sets of measurements, solid line—average of the four. (b) ABR magnitude dependence on sound-source azimuth [thin and solid lines—the same as in (a)]. (c) 1—delay dependence on azimuth corrected for delay-vs-magnitude dependence; 2—average (with standard errors) of the left and right branches of (1); 3—approximating cosine function.

C. Position of sound-receiving points

In the measurement session exemplified above, the obtained parameters of the delay-vs-azimuth function (0.14 ms amplitude, 32° phase) indicated a receiving point at an azimuth of 32° relative to the longitudinal axis and at a distance of $0.14\ \text{ms} \times 150.5\ \text{cm/ms} = 21.1\ \text{cm}$ from the reference point which was 25 cm behind the melon tip. To make the determination of a receiving point more trustworthy, it was repeated in the same manner as described above using four reference (rotation center) points: 0, 12.5, 25, and 37.5 cm behind the melon tip. The results of measurements for a sound frequency of 128 kHz are presented in Fig. 7(a) as four sets of experimental points and approximating cosine functions. Since the shift of a cosine along the ordinate scale does not bear information concerning the receiver position, these functions were arbitrarily shifted to make their zero-azimuth points coincide. Both the amplitudes and phases of the cosine functions were different for different reference points, thus indicating that the found azimuth-dependent delay variation was determined by mutual position of the reference and sound-receiving points.

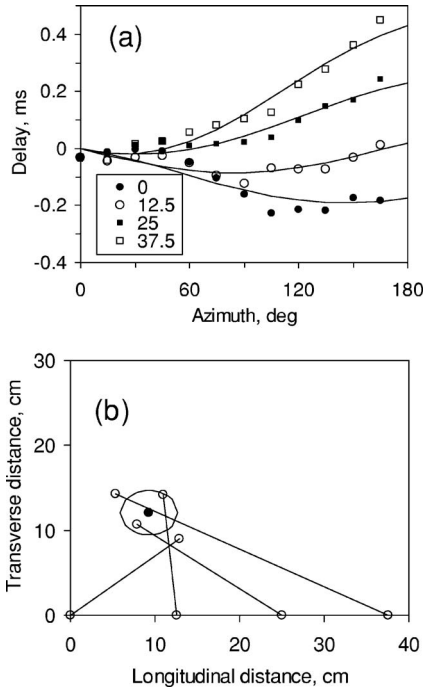


FIG. 7. (a) Determination of a sound receiving point for 128 kHz frequency. (a) Experimental data and approximating cosine function for four reference points, from 0 to 37.5 cm behind the melon tip, as indicated. (b) Vectors drawn from the four reference points according to the cosine function parameters; black solid point—mean of the four found positions; ellipse—SD area.

Parameters (amplitudes and phases) of these four cosine functions were used as lengths and directions of vectors drawn from each of the four reference points [Fig. 7(b)]. Notwithstanding some data scatter, all the vectors indicated a similar area, from 5.3 to 12.9 cm caudal and from 9.0 to 14.3 cm lateral of the melon tip. The mean of these four points was at a point 9.3 cm caudal and 12.1 cm lateral of the melon tip with longitudinal standard deviation (SD) of 3.4 cm and transverse SD of 2.6 cm.

For comparison, results of similar computation for the lowest investigated probe frequency of 16 kHz are presented in Fig. 8. Both the experimental plots and approximating cosine functions presented here obviously differed from those presented above in Fig. 7. Respectively, the vectors drawn according to the cosine parameters indicated another sound-receiving point: the mean of the four points was 21.7 cm caudal and 16.0 cm lateral of the melon tip, with longitudinal SD of 2.3 cm and transverse SD of 11.8 cm; i.e., a point different from that found at 128 kHz probe.

For brevity, we do not present in the same detail all the data obtained at other probe frequencies using the very same procedure. The results of all measurements are presented in Table I as phase (α in degrees) and amplitudes (d in ms) of approximating cosine functions. The results are presented for all the tested probe frequencies (16–128 kHz) and for four positions of the reference points (0–37.5 cm behind the melon tip) for each of the frequencies. Table I presents also the root-mean-square differences between the experimental point arrays and the approximating cosine functions (δd rms, ms). The difference values showed that the approximation

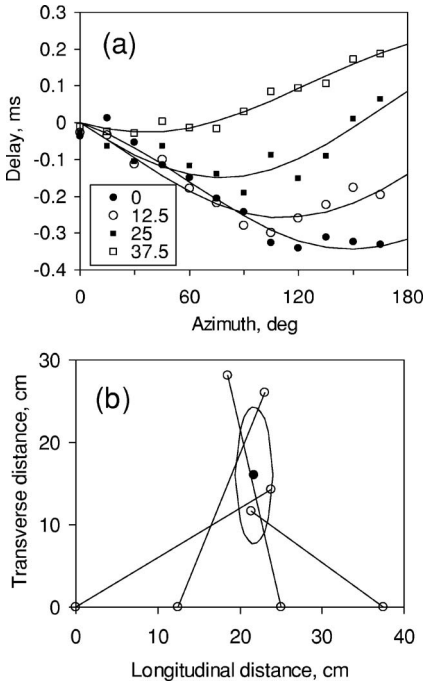


FIG. 8. The same as Fig. 7, for 16 kHz frequency.

was satisfactory enough: in majority of cases (25 of 28 cases, 89%), rms did not exceed 0.04 ms which corresponds to 6 cm distance.

The final result of conversion of the cosine parameters to sound-receiving point positions is presented in Table II and Fig. 9. In Table II, coordinates of receiving points found with the use of four different reference points (0–37.5 cm behind the melon tip) are presented together with SD of the four estimates for each of the probe frequencies (16–128 kHz). In Fig. 9, positions of sound-receiving points are presented together with their SD ellipses superimposed on a diagram of the dorsal view of the dolphin's head. Although the sound-receiving points are shown at the right side of the head, they can be equally attributed to both sides since all the data were obtained by averaging the right- and left-side measurements. The figure shows that all the points dissociated into two groups. Points determined with lower probe frequencies, 16 and 22.5 kHz, were located 21.7–26 cm caudal of the melon tip. The auditory meatus is in this approximate region. Points determined with all higher probe frequencies (32, 45, 64, 90, and 128 kHz) were concentrated in a more rostral area, from 9.3 to 13.1 cm caudal of the melon tip. This is the approximate region of the mandibular acoustic window. SD areas overlapped within each of these two groups but did not overlap between the groups.

IV. DISCUSSION

A. Agreement between the data and measurement paradigm

According to the adopted experimental paradigm, all the delay dependencies on azimuth should fall on cosine functions, and sound-receiver positions obtained at different reference points should coincide. In reality, both the experimental points somewhat deviated from cosine functions and

TABLE I. Parameters of cosine functions approximating experimental data.

F, kHz	Reference point position, cm	Cosine function parameters		
		α,deg	d, ms	δd rms, ms
16	0	149	0.185	0.023
	12.5	112	0.188	0.027
	25	77	0.193	0.035
	37.5	36	0.133	0.014
22.5	0	170	0.175	0.037
	12.5	106	0.088	0.012
	25	92	0.055	0.051
	37	27	0.060	0.040
32	0	157	0.123	0.019
	12.5	97	0.065	0.021
	25	21	0.105	0.018
	37.5	34	0.208	0.030
45	0	148	0.088	0.030
	12.5	88	0.065	0.031
	25	59	0.155	0.033
	37.5	18	0.190	0.026
64	0	144	0.085	0.020
	12.5	100	0.110	0.018
	25	18	0.135	0.047
	37.5	15	0.185	0.017
90	0	158	0.100	0.028
	12.5	95	0.073	0.043
	25	19	0.083	0.028
	37.5	17	0.205	0.037
128	0	145	0.105	0.034
	12.5	84	0.095	0.023
	25	32	0.135	0.026
	37.5	24	0.235	0.031

sound-receiving points obtained at different reference points did not precisely coincide. We suppose that moderate deviations from the theoretical predictions do not contradict the basic measurement paradigm because the method used is very error sensitive. There were at least a few factors which might reduce the precision of measurements.

(i) Background noise might deform AEP records, thus influencing the estimates of their delay. Note that a shift delay estimate by only one record sample (0.025 ms) results in a shift of receiving-point position estimates by 3.75 cm.

(ii) Compensation for response amplitude might not be very precise, both because of spontaneous variation of the response amplitude and because of the difference in latency-vs-amplitude functions at different sound-source azimuths. Note that in some cases, amplitude-dependent variation of delays was as big as around 0.2 ms (see Table I). For example, a compensation error of 10% might result in an error in the delay estimate of 0.02 ms, i.e., which results in an error of 3 cm for the receiving point position.

(iii) Small movements of the animal's head might influence the acoustic delays and therefore the estimates of receiving point positions. The range of the error might be of the same order of magnitude as the range of head movements, up to a few cm.

Taking into consideration all these factors, significant data scatter seemed inevitable. Bearing this in mind, each receiving point position was computed based on 48 delays

TABLE II. Estimates of receiving point positions.

F, kHz	Reference point position, cm	Receiving point coordinates					
		X, cm	Y, cm	Average X, cm	Average Y, cm	SD(X), cm	SD(Y), cm
16	0	23.8	14.3				
	12.5	23.0	26.1	21.7	16.0	2.3	8.3
	25	18.5	28.1				
	37.5	21.4	11.7				
22.5	0	32.9	5.8				
	12.5	16.1	12.6	26.0	6.1	7.2	3.7
	25	25.7	8.2				
	37.5	29.5	4.1				
32	0	16.9	7.2				
	12.5	13.7	9.7	13.1	8.0	2.9	5.3
	25	10.3	5.6				
	37.5	11.7	17.4				
45	0	11.1	7.0				
	12.5	12.2	9.7	11.7	9.1	1.2	5.7
	25	13.0	19.9				
	37.5	10.4	8.8				
64	0	10.3	7.5				
	12.5	15.4	16.2	10.5	7.4	3.9	4.7
	25	5.7	6.3				
	37.5	10.7	7.2				
90	0	13.9	5.6				
	12.5	13.5	10.8	12.2	5.9	2.7	8.3
	25	13.3	4.0				
	37.5	8.1	9.0				
128	0	12.9	9.0				
	12.5	11.0	14.2	9.3	12.1	3.4	2.6
	25	7.8	10.7				
	37.5	5.3	14.3				

X and Y—coordinates along the longitudinal (X) and transverse (Y) head axes, adopting the melon tip as zero; SD(X) and SD(Y)—standard deviations among four estimates of X and Y coordinates, respectively.

(12 sound-source azimuths at each of four reference points), although theoretically three delays are necessary to compute a receiver position. The deviations of experimental data from a cosine function did not exceed a few cm rms (see Table I); i.e., of the same order as may be expected due to the reasons considered above. The differences in receiving-point positions computed at different reference points did not exceed a few cm either for one and the same probe frequency (see Table II). Therefore, the obtained deviation of experimental data from true cosine functions does not contradict the basic paradigm.

B. Validity of the unilateral model

Computations described above imply that the sound is perceived by an acoustical window (windows) at only one side, ipsilateral to the sound source position. Basing on this assumption, results of left- and right-side measurements were averaged. Validity of this approach needs a comment. As shown before (Popov *et al.*, 2006), at all sound frequencies from 16 to 128 kHz, interaural intensity difference in the bottlenose dolphin reaches from 10 to 30 dB at azimuths above 15°. As Fig. 5 shows, this intensity difference resulted in big difference of AEP amplitude. Therefore, we expected that contribution of the contralateral input, if existed, little influenced temporal parameters of AEP. Assuming this suggestion, the unilateral model may be taken as valid.

C. Multiple receiving areas

The main result of the investigation presented here is the presence of at least two sound-receiving areas on the dolphin head: a rostral (9–12 cm behind the melon tip) and a caudal (21–26 cm behind the melon tip) area. The rostral sound receiving area manifested itself at probe frequencies of 32 kHz and higher, the caudal at probe frequencies of 22.5 kHz and lower. Within each of these two areas, there is no frequency-dependent regularity of receiving point positions, and SD areas of all points overlap; on the contrary, there is no SD overlap between the two areas. Thus, there is no continuous frequency-dependent shift of the sound-receiving point position but there are two separate sound-receiving areas.

The rostral sound-receiving area coincides well with the position of the mandibular acoustic window hypothesized before. Thus, the results of the present study confirm once again this hypothesis. As to the caudal sound-receiving area, its presence confirms the hypothesis of possibility of sound transmission through a near-bullar area (Popov and Supin, 1990; Popov *et al.*, 1992).

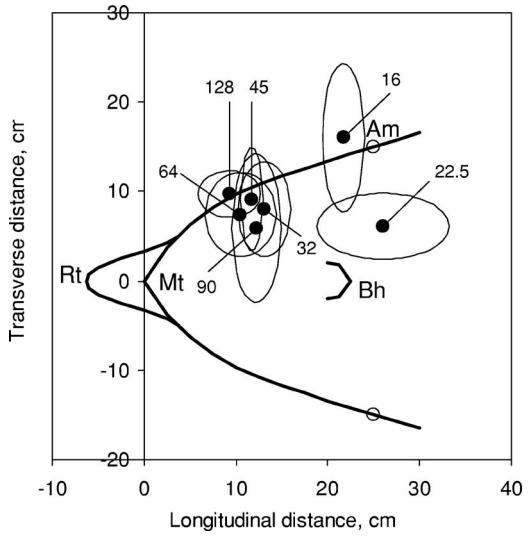


FIG. 9. Positions of sound-receiving points for frequencies from 16 to 128 kHz. Each position (solid dot) is an average of four positions found with four different reference points. Ellipses delimit SD areas for each sound-receiving point. Point positions are superimposed on a contour of the dorsal view of the dolphin's head. Rt—rostrum tip, Mt—melon tip, Bh—blowhole, Am—auditory meatus.

The results presented herein show that there is no contradiction between these two hypotheses. The two sound-receiving areas (acoustic windows) have different frequency sensitivities. In all studies of acoustic window localization performed before, wideband acoustic probes were used. So depending on stimulus parameters and experimental design, either of these two areas could be detected. A hypothesis of more than one sound-receiving area has also a morphological confirmation: a few sound-conduction pathways may be provided by a few lobes of fatty tissues connecting to the bulla (Ketten, 2000).

The significance of multiple sound-receiving areas is not clear yet. However, some speculative hypotheses can be suggested. Because of different frequency sensitivity of sound-receiving areas, the best-sensitivity axis direction may be frequency dependent (Popov *et al.*, 2006). A result of such dependence is that a perceived spectral pattern of a broadband sound stimulus depends, in turn, on the sound direction. It may provide additional cues for both localization of sound sources and sound pattern recognition.

ACKNOWLEDGMENTS

The study was supported by The Russian Ministry of Science and Education, Grant No. NSh-7117.2006.4 and The Russian Foundation for Basic Research, Grant No. 06-04-48518. The authors thank the staff of the Utrish Marine Station (supervised by Lev Mukhametov) for assistance.

Brill, R. L. (**1988**). "The jaw-hearing dolphin: Preliminary behavioral and acoustical evidence," in *Animal Sonar: Processes and Performance*, edited by P. E. Nachtigall and P. W. B. Moore (Plenum, New York), pp. 281–287.

Brill, R. L. (**1991**). "The effect of attenuating returning echolocation signals at the lower jaw of a dolphin (*Tursiops truncatus*)," J. Acoust. Soc. Am. **89**, 2851–2857.

Brill, R. L., Sevenich, M. L., Sullivan, T. J., Sustman, J. D., and Witt, R. E. (**1988**). "Behavioral evidence for hearing through the lower jaw by an echolocating dolphin, *Tursiops truncatus*," Marine Mammal Sci. **4**, 223–230.

Bullock, T. H., Grinnell, A. D., Ikezono, F., Kameda, K., Katsuki, Y., Nomoto, M., Sato, O., Suga, N., and Yanagisava, K. (**1968**). "Electrophysiological studies of the central auditory mechanisms in cetaceans," Z. Vergl. Physiol. **59**, 117–156.

Fraser, F. C., and Purves, P. E. (**1954**). "Hearing in cetaceans," Bull. British Mus. (Nat. Hist.) **2**, 103–116.

Fraser, F. C., and Purves, P. E. (**1959**). "Hearing in whales," Endeavour **18**, 93–98.

Fraser, F. C., and Purves, P. E. (**1960**). "Hearing in cetaceans: Evolution of the accessory air sacs and the structure and function of the outer and middle ear in recent cetaceans," Bull. British Mus. (Nat. Hist.) **7**, 1–140.

Ketten, D. R. (**1990**). "Three-dimensional reconstructions of the dolphin ear," in *Sensory Abilities of Cetaceans. Laboratory and Field Evidence*, edited by J. A. Thomas and R. A. Kastelein (Plenum, New York), pp. 81–105.

Ketten, D. R. (**1992a**). "The cetacean ear: Form, frequency, and evolution," in *Marine Mammal Sensory Systems*, edited by J. A. Thomas, R. A. Kastelein, and A. Ya. Supin (Plenum, New York), pp. 53–75.

Ketten, D. R. (**1992b**). "The marine mammal ear: Specialization for aquatic audition and echolocation," in *The Evolutionary Biology of Hearing*, edited by D. Webster, R. R. Fay, and A. N. Popper (Springer, New York), pp. 717–754.

Ketten, D. R. (**1997**). "Structure and function in whale ears," Bioacoustics **8**, 103–135.

Ketten, D. R. (**2000**). "Cetacean ears," in *Hearing by Whales and Dolphins*, edited by W. W. L. Au, A. N. Popper, and R. R. Fay (Springer, New York), pp. 43–108.

McCormick, J. G., Wever, E. G., Palin, G., and Ridgway, S. H. (**1970**). "Sound conduction in the dolphin ear," J. Acoust. Soc. Am. **48**, 1418–1428.

McCormick, J. G., Wever, E. G., Ridgway, S. H., and Palin, G. (**1980**). "Sound reception in the porpoise as it relates to echolocation," in *Animal Sonar Systems*, edited by R. G. Busnel and J. F. Fish (Plenum, New York), pp. 449–467.

Møhl, B., Au, W. W. L., Pawloski, J., and Nachtigall, P. E. (**1999**). "Dolphin hearing: Relative sensitivity as a function of point of application of a contact sound source in the jaw and head region," J. Acoust. Soc. Am. **105**, 3421–3424.

Norris, K. S. (**1968**). "The evolution of acoustic mechanisms in odontocete cetaceans," in *Evolution and Environment*, edited by E. T. Drake (Yale University, New Haven), pp. 297–324.

Norris, K. S. (**1969**). "The echolocation of marine mammals," in *The Biology of Marine Mammals*, edited by H. J. Andersen (Academic, New York), pp. 391–424.

Norris, K. S. (**1980**). "Peripheral sound processing in odontocetes," in *Animal Sonar System*, edited by R.-G. Busnel and J. F. Fish (Plenum, New York), pp. 495–509.

Popov, V. V., and Supin, A. Ya. (**1990**). "Localization of the acoustic window at the dolphin's head," in: *Sensory Abilities of Cetaceans: Laboratory and Field Evidence*, edited by J. A. Thomas and R. A. Kastelein (Plenum, New York), pp. 417–426.

Popov, V. V., Supin, A. Ya., and Klishin, V. O. (**1992**). "Electrophysiological study of sound conduction in dolphins," in *Marine Mammal Sensory Systems*, edited by J. A. Thomas, R. A. Kastelein, and A. Ya. Supin (Plenum, New York), pp. 269–276.

Popov, V. V., Supin, A. Ya., Klishin, V. O., and Bulgakova, T. N. (**2006**). "Monaural and binaural hearing directivity in the bottlenose dolphin: Evoked-potential study," J. Acoust. Soc. Am. **119**, 636–644.

Supin, A. Y., Popov, V. V., and Mass, A. M. (**2001**). *The Sensory Physiology of Aquatic Mammals* (Kluwer, Boston), 332 pages.

Varanasi, U., and Malins, D. C. (**1971**). "Unique lipids of the porpoise (*Tursiops gilli*): Difference in triacylglycerols and wax esters of acoustic (mandibular and melon) and blubber tissues," Biochim. Biophys. Acta **23**, 415–418.

Varanasi, U., and Malins, D. C. (**1972**). "Triacylglycerols characteristics of porpoise acoustic tissues: Molecular structures of diisovaleroylglycerides," Science **176**, 926–928.

Wilson, W. D. (**1960**). "Equation for the speed of sound in sea water," J. Acoust. Soc. Am. **32**, 1357.