From E-C theory to speech intelligibility in rooms

John Culling[1], Mathieu Lavandier[2], and Sam Jelfs[3]

[1] *School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, United Kingdom*

[2] *Université de Lyon, Ecole Nationale des Travaux Publics de l'Etat, Département Génie Civil et Bâtiment, Unité CNRS 1652, Rue M. Audin, 69518 Vaulx-en-Velin Cedex, France*

[3] *Welsh School of Architecture, Bute Building, King Edward VII Avenue, Cardiff, CF10 3NB, United Kingdom*

Equalization-cancellation theory accounts for the binaural unmasking of tones by assuming that the waveforms within corresponding frequency channels at the left and right ears are optimally equalized by an internal delay and then cancelled. A wide variety of experimental data have been successfully modelled using predictive equations derived from the model for each experimental design. Recently, a single equation has been developed that can make equivalent predictions for any experiment based on interaural statistics measured from the experimental stimuli, regardless of the stimulus construction. We have used this equation to predict unmasking of speech against maskers of very complex construction, one or more interfering noise sources in a reverberant room. The model assumes that effects of better-ear listening and binaural unmasking are additive within each frequency channel and weights their combined effects by the SII frequency weighting function to yield intelligibility predictions. These predictions correlate highly with speech reception thresholds measured in the same configurations. Deriving the masker statistics directly from room impulse responses, computation is sufficiently economical for the generation of intelligibility maps for a given room, spatial configuration of sources and listener orientation.

BINAURAL UNMASKING

When the interaural relations of a signal and a masker differ, there may be a release from masking compared to when they have the same interaural relationship. For instance, if a noise is identical at the two ears, but a signal is interaurally out of phase, detection threshold for the tone may be lower than when both are identical at the two ears (Hirsh, 1948). At low frequencies (e.g. 250 Hz), this difference in thresholds (the binaural masking level difference or BMLD) may be as large as 15 dB.

Binaural unmasking has been found to depend on many aspects of the stimulus configuration. Of particular interest here are the following observations. The BMLD depends on the difference in interaural phase between the signal and the masker, such that, for a diotic masker the BMLD varies cyclically with the delay/phase-shift applied to a tonal signal (Jeffress *et al*., 1952) and for a diotic signal, the BMLD

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

varies in a damped cyclical fashion as a function of the interaural delay applied to the noise (Langford and Jeffress, 1964). The BMLD also depends upon the interaural correlation of the noise. When the signal is interaurally out of phase, the BMLD is large when the correlation is high and small when the correlation is low (Robinson and Jeffress, 1963). Van der Heijden and Trahiotis (1997) pointed out that the relationship between masked threshold and masker interaural correlation is linear when threshold is plotted in linear units rather than in decibels.

EQUALIZATION-CANCELLATION (E-C) THEORY

Durlach (1963) proposed a mechanism to explain binaural unmasking that he termed Equalization-Cancellation (E-C). In this scheme, the brain was presumed to equalise the internal representations of the stimulus at each ear with the use of various transformations, and then subtract these transformed representations, one from the other. In its original formulation, any transformation was permitted in order to perform equalisation, including delays and phase or level shifts.

Durlach (1972) reviewed a wide range of data on binaural unmasking and the ability of E-C theory to predict it. He presented a revised model, based exclusively on equalisation of interaural delay, whose predictions of different data sets were assessed. This model was successful in predicting many unmasking phenomena, although a major failing was, and remains, its inability to account for the effects of differences in interaural intensity, or indeed of overall intensity. Each prediction was based on an equation, derived from E-C theory. Equation 1 (#55 in Durlach's report) predicts the effect of differences in interaural phase.

$$BMLD = \frac{k - \cos(\phi_s - \phi_n)}{k - \gamma(\phi_n / \omega_0)} \qquad \text{(Eq. 1)}$$

Here, k is a variable that controls the effectiveness of unmasking with frequency. It's value is greater than 1. φ_s and φ_n represent the interaural phases of the signal and noise and $\gamma(\varphi_n/\omega_0)$ is a function controlling the damping effect as a function of delay of the noise (ω_0 is the signal frequency in rad/s). Equation 2 (#61 in Durlach's report) predicts the effect of masker interaural correlation, r_n, where the signal is out of phase (a different equation was needed for the case of a diotic signal),

$$BMLD = \frac{k+1}{k - r_n} \qquad \text{(Eq. 2)}$$

The present study employed a development of this approach, but here a single equation is used to make equivalent predictions for all the cases considered by Durlach, as well as facilitating predictions of novel stimulus configurations (Culling *et al*., 2005). In this equation, the noise interaural coherence, ρ_n (the maximum of the noise interaural cross-correlation function) is used in place of its interaural correlation,

$$BMLD = \frac{k - \cos(\phi_s - \phi_n)}{k - \rho_n} \qquad \text{(Eq. 3)}$$

Intuitively, Eq. 3 may be understood as relating two terms, one of which reflects how well the masker can be cancelled, while the other reflects the concomitant effect that cancellation of the masker will have on the signal. The two in combination thus determine the effective change in signal-to-noise ratio (SNR). The k-ρ_n term represents the effectiveness of noise cancellation. The interaural coherence of a noise is equivalent to the proportion of noise power that is common between the ears at the optimal equalization delay ($\omega_0\varphi_n$). Ideally, all common noise will be removed by cancellation, so the change in SNR should be inversely related to 1-ρ, but since the process is imperfect, the term k-ρ_n is used, where k>1. The k-cos(φ_s-φ_n) term represents the effect on the signal of cancellation at a delay of $\omega_0\varphi_n$. The effect on the signal varies sinusoidally as a function of cancellation delay, because one sinusoid is being subtracted from another. When φ_s and φ_n are equal, and assuming the signal has the same amplitude at each ear, the signal will be cancelled out, but when they differ by π radians, subtraction will double the signal amplitude. Thus, the effect of cancellation on the signal is proportional to 1-cos(φ_s-φ_n), which varies between 0 and 2. One can think of imperfection in the process as again being represented by replacement of 1 by k, but at this point this intuitive approach begins to break down, because that substitution yields numbers that range above 2.

Superficially, Eq. 3 appears to be a simple combination of equations 1 and 2. However, there are some subtle differences in how the equations work. This can be seen most clearly when considering the case of interaurally uncorrelated masking noise (i.e. r_n=0). According to Eq. 2, the BMLD will be (k+1)/k. However, for Eq. 3, the value of ρ_n will always be above zero, and will fluctuate as a function of time, while φ_s and φ_n will have completely random values, which will also fluctuate as a function of time. Moreover, the evaluation of ρ_n requires the definition of the method to calculate it. In particular, the range of delays over which the cross-correlation is calculated should have some sensible limit, but there is no current consensus on the appropriate range of delays. Finally, the values of ρ_n observed depend upon the width of the auditory filter. Wider filters tend to produce smaller values, and narrower filters produce larger values. The prediction of Eq. 3 can thus be rather implementation-specific for this case. Nonetheless, Fig. 1 shows that it can produce very similar predictions to Eq. 2 for stimuli of the type used by Robinson and Jeffress (1963). In this implementation, a conventional gammatone filter centred at 500 Hz was applied to the left- and right-ear signals, equalisation delays were limited to 1 ms (consistent with Durlach's revised model for a 500-Hz signal), and Eq. 3 was evaluated for each of a set of 50 stimuli of 500 ms duration. Negative BMLD values were set to zero prior to averaging, following the principle that the binaural system never impairs performance. Even after the averaging, the results were quite variable, and the error bars in Fig. 1 show the standard deviation of these averages, taken across 20 sets of 50 stimuli. For r_n=1

and -1, the predictions of Eq. 2 and Eq. 3 are identical, and the results from Eq. 3 do not vary from one stimulus to another. For r_n=0, 0.5, 0.8 and 0.9, however, the error bars are large, but the mean values are quite consistent with the predictions of Eq. 2.

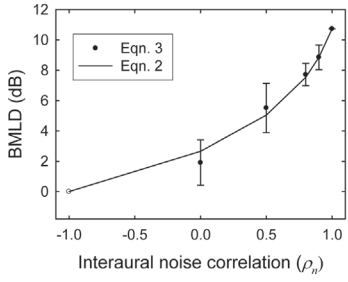

Fig. 1: Comparison of Eq. 2 and Eq. 3 for prediction of thresholds for out of phase signals for various levels of masker interaural correlation.

Although the analysis above indicates that Eq. 3 can be a relatively clumsy predictive tool, requiring the synthesis of stimuli and repeated measurements to overcome the stochastic nature of its predictions, this is not always the case, and it has the advantage of being much more general. It can be used in place of all the cases separately considered by Durlach, as well as for new combinations of phase shift, delay and masker coherence values, regardless of their construction. This feature is exploited below in order to predict the intelligibility of speech in maskers of very complex construction, produced by room simulation. For this purpose, it is more convenient to use parameters measured from the target and masker stimuli, or even directly from room impulse responses, rather than attempt to base them on the method of stimulus construction.

SPEECH INTELLIGIBILITY IN ROOMS

Reverberation can be detrimental to speech understanding. Some reduction in intelligibility results directly from distortion of the speech signal, which becomes temporally smeared (Houtgast and Steeneken, 1985). However, Lavandier and Culling (2008) showed that there is also an effect of reverberation on speech perception in noise, in which listeners' ability to binaurally unmask the speech is impaired. This effect was shown to occur at lower levels of reverberation than the temporal-smearing effect. Here, we show that empirical measurements of binaural intelligibility thresholds can be accurately predicted for a wide variety of simulated listening situations designed to probe the limits of the technique.

Experimental data

Speech reception thresholds (50% keyword intelligibility) were measured against speech-shaped noise for IEEE sentences such as "HOLD the HAMMER near the END to DRIVE the NAIL" (keywords in capitals). These stimuli were convolved with

binaural room impulse responses (BRIRs) generated from a rectangular room model (Allen and Berkley, 1979) in order to create different virtual listening situations. Four different sizes of room were modelled. In three of the rooms, the coherence of the interfering noise was manipulated by placing it at two different distances. For one of the rooms, the coherence of the noise was manipulated by using four different absorption coefficients for the internal surfaces. For one room, the interferer was placed either symmetrically or asymmetrically with respect to the room and the listening position, creating masking stimuli that were either reverberant but diotic, or reverberant and interaurally incoherent. Finally, in one room, the left- and right-ear BRIRs were convolved by independent noises in order to artificially force interaural coherence close to zero. The distance of the target speech source was always 2 m. The target and interfering noise were always placed on a different bearing, separated by 60°. In order to specifically examine the effect of reverberation on cancellation of the noise, the simulation included no model of the head and the target speech was presented anechoically. Following this processing, the root-mean-square (rms) power of target and masker were equalised at each ear independently. The circumstances leading to a larger or smaller degree of binaural unmasking at a constant spatial separation were thus extensively explored using a mixture of both realistic and contrived listening scenarios. There were 16 conditions in all, labelled A-P.

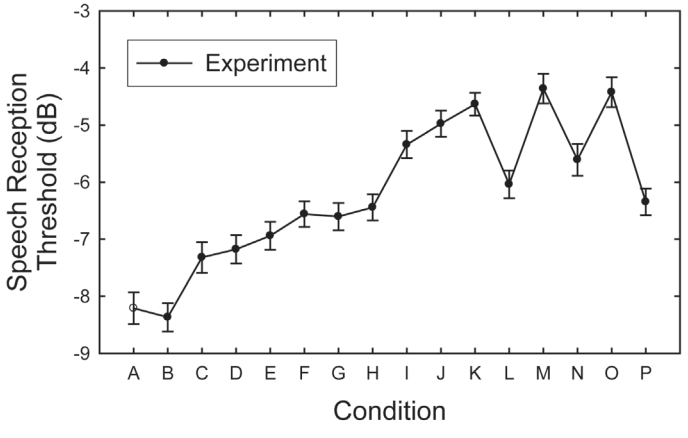

Fig. 2: Mean speech reception thresholds for each of 16 listening scenarios. Error bars are one standard error of the mean.

The results of the experiment are shown in Fig. 2. The manipulations of masker position etc. produced thresholds that varied over a range of about 4 dB. The conditions are ordered in descending sequence of masker interaural coherence. Broadly speaking, higher thresholds are observed on the right-hand side of the figure, where masker interaural coherence is low. However, at low values of coherence (generally conditions in which there is greater reverberation) the thresholds do not follow a monotonic sequence. We believe that this effect is produced by room colouration (see below).

Modelling the data

A model was developed using Eq. 3. The parameters of the model were measured from the stimuli. Target and masker BRIRs were both convolved with noise and filtered into different frequency channels by a gamma-tone filterbank (Patterson *et al*., 1987, 1988). Each channel was subjected to interaural cross-correlation: ρ_n was taken from the peak cross-correlation value based on the masker BRIRs, φ_n was derived from the cross-correlation delay at which this peak occurred and φ_s was taken from a similar measurement based on the target BRIRs.

As noted above, in high levels of reverberation, masker coherence does not appear to be the single controlling variable for these data. High levels of reverberation introduce substantial colouration, which will differ for sound sources at different positions. Colouration is the product of the vector summation of many superimposed copies of the same sound, sometimes cancelling, sometimes reinforcing different frequencies. Colouration changes in an erratic way, depending on the precise position of the source and receiver within a room, and may differ both between different source locations and across the ears. If important frequencies for speech were relatively attenuated for the masking noise, then an improvement in intelligibility will occur. Equally a worsening of intelligibility may occur if the reverse is true. This effect may occur despite our post-processing equalization of overall rms level at each ear, because the colouration of the masker will emphasise different frequencies to different degrees; the overall level may be fixed, but the level in the range most important for speech may be quite varied.

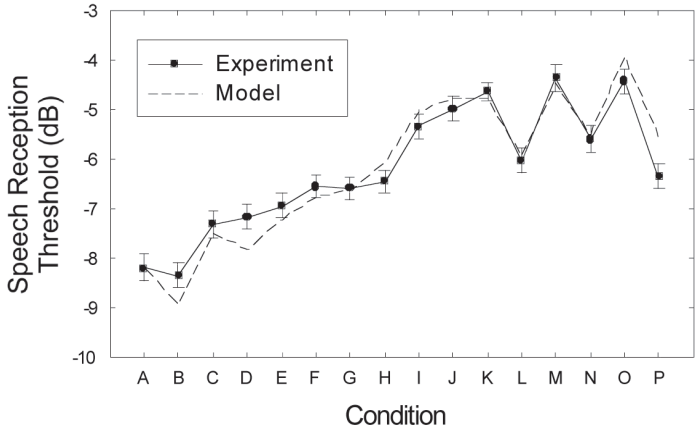

Fig. 3: As Fig. 2, but with model predictions (dashed line) added.

In order to account for the colouration, a simple model of better-ear listening was implemented. In this model, a further contribution to the effective SNR in each frequency channel was taken to be the better of the two actual SNRs observed at each ear. Thus the overall effective SNR in each frequency channel was the better of the two ratios at each ear, plus the predicted binaural unmasking effect from Eq.

3. Adding this component should also, in principle, allow the model to predict the effects of head-shadow.

Finally, the importance of different frequency channels for speech understanding was modelled by applying the SII weightings (ANSI, 1997) to the effective SNRs in each frequency channel. The weighted SNRs were then summed to give the predicted binaural intelligibility level differences.

Figure 3 shows the predicted thresholds (dashed line), superimposed upon the empirical data. The correlation between the two is 0.97. In order to align the predictions and data, the mean value of the predictions has been set to equal that of the data. This manipulation makes no difference to the correlation value and is equivalent to adjusting the value of SII needed for 50% intelligibility with our materials. Further work in our laboratory has shown that the method also works with multiple interfering noise sources and with real-room BRIRs, which include the effects of head-shadow.

INTELLIGIBILITY MAPS FOR NOVEL ROOMS

Having validated a method for predicting speech intelligibility in rooms, it should be possible to make confident predictions for novel rooms. Architectural design software often now includes auralisation modules that can generate accurate BRIRs. These could potentially be used to predict intelligibility of speech against any pattern of noise interferers from plan for any part of the room. Moreover, direct cross-correlation of the impulse responses can provide a more compulationally efficient and non-stochastic prediction, making this an ideal application of Eq. 3. Here we illustrate the possibilities using our simple rectangular room model (i.e. with no head-shadow effects).

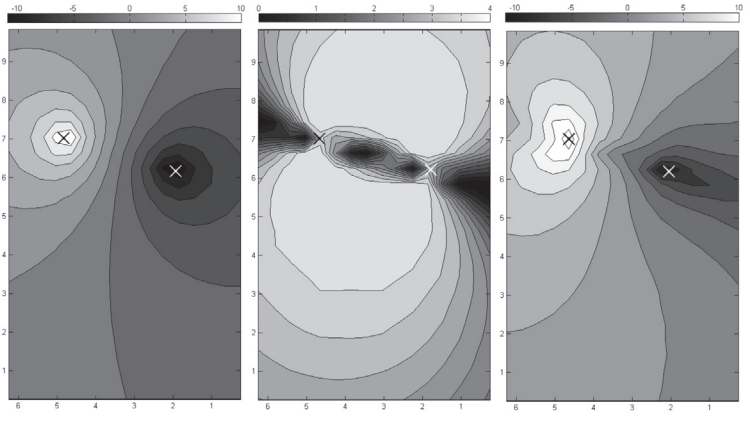

Fig. 4: Intelligibility maps for an anechoic room. Left: signal-to-noise ratio (dB) for all points in the room for speech located at the black cross and noise located at the white cross. Middle: binaural unmasking effect predicted for all points in the same situation: Right panel: overall intelligibility, the sum of the first two panels.

Figure 4 shows the derivation of such a room map for the simple case of an anechoic room. The room is 10 m × 6.4 m and 2.5 m high. Measured points are in a 0.25 m grid at a height of 1.5 m. The left- and right-ear BRIRs for each point in space have been passed directly into a gammatone filterbank. They were then cross-correlated in order to determine ρ_n, φ_n and φ_s from the height and delay of the cross-correlation maxima. The rms power of each filter output was also measured. The left panel shows SNR, taken from the relative rms power of the target and masker BRIRs, at every point in the room for a voice target located at the black cross and a noise masker located at the white cross. Since the room is anechoic and there is no head-shadow, SNR simply reflects the relative distance to the two sources. The SNR in each gammatone filterbank channel is weighted by the appropriate SII weightings and then summed, but, because the SNR is the same at all frequencies in this example, this has no effect on the pattern of predictions. The middle panel shows the predicted binaural intelligibility level difference from Eq. 3. Here, the SII weighting is more important and, combined with the limited range of frequencies over which binaural unmasking is effective, it limits the best binaural intelligibility level differences to 4 dB or so. The two ears are located 20 cm apart and are always orientated towards the target speech, as though the listener were facing the source of the voice. It can be seen that locations on the axis linking the two sources provide little binaural unmasking effect, because in this area, both sources have a similar interaural time delay (hence φ_s-φ_n is close to zero). The overall intelligibility map (right panel) is the sum of the other two panels.

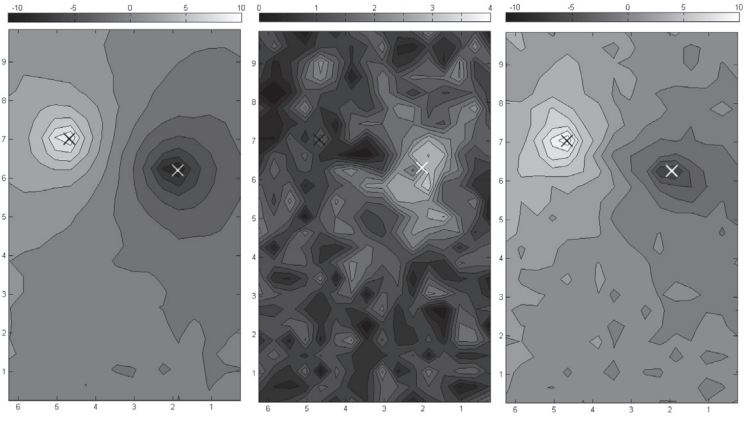

Fig. 5: Intelligibility maps, similar to Fig. 3, but for a reverberant room. The absorption coefficient for all surfaces is 0.5.

Figure 5 shows similar maps, but for a reverberant room (same dimensions), in which all surfaces have an absorption coefficient of 0.5. Although the room is reverberant, the pattern of SNRs (left panel) is fairly similar to that for the anechoic room. However, the binaural unmasking effect (middle panel) is radically different. The sizes of effect observed are generally smaller, reflected by darker shades of grey. This

is caused by the reduction in ρ_n produced by the reverberation. Only close to the noise source, where the direct sound from the noise dominates, is the effect of binaural unmasking substantial. Because binaural unmasking has been largely suppressed, when the two elements are combined to produce an overall intelligibility map, the pattern is largely dominated by the SNR effect.

The computational demands of creating such maps could be large, but is considerably ameliorated by working directly from the BRIRs and using a simple predictive equation to evaluate the effect of unmasking. The present method is equivalent to, but much more computationally efficient than, the method developed by Beutelmann and Brand (2006). We anticipate that it may ultimately form the basis of useful architectural tools.

CONCLUSIONS

Equalisation-cancellation theory is a powerful method for predicting binaural unmasking, which fails substantially only when relatively large differences in interaural level are introduced to the stimuli. Within this constraint, the method of generating E-C predictions presented here allows predictions for any stimulus configuration, regardless of its construction. The method is particularly suited to predicting the effect of convolutive distortions such as those introduced by room reverberation. Combined with effects of signal-to-noise ratio and with the SII weightings, it can produce accurate, non-stochastic predictions of speech intelligibility in noise in different room locations. The method is sufficiently computationally efficient to facilitate the generation of intelligibility maps from room designs.

RFERENCES

Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943-950.

ANSI (**1997**). "Methods for calculation of the speech intelligibility index," ANSI S3.5-1997 (American National Standards Institute, New York).

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331-342.

Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (**2005**). "Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," J. Acoust. Soc. Am. **118**, 552.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking level differences," J. Acoust. Soc. Am., **35**, 1206-1218.

Durlach, N. I. (**1972**). "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory* edited by J. Tobias, volume II (Academic, New York) pp. 371-462.

Jeffress, L. A., Blodgett, H. C., and Deatherage, B. H. (**1952**). "The masking of tones by white noise as a function of the interaural phases of both components. I. 500 cycles," J. Acoust. Soc. Am. **35**, 523-527.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069-1077.

Langford, T. L., and Jeffress, L. A. (**1964**). "Effect of noise crosscorrelation on binaural signal detection," J. Acoust. Soc. Am. **36**, 1455-1458.

Lavandier, M., and Culling, J. F. (**2008**). "Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer," J. Acoust. Soc. Am. **123**, 2237-2248.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). "An efficient auditory filterbank based on the gammatone function," presented to the Institute of Acoustics speech group on auditory modelling at the Royal Signal Research Establishment.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "Spiral VOS final report, Part A: The auditory filterbank," Cambridge Electronic Design, Contract Report (APU 2341).

Robinson, D. E., and Jeffress, L. A. (**1963**). "Effect of varying the interaural noise correlation on the detectability of tonal signals," J. Acoust. Soc. Am. **35**, 1947-1952.

van der Heijden, M., and Trahiotis, C (**1997**). "A new way to account for binaural detection as a function of interaural noise correlation," J. Acoust. Soc. Am. **101**, 1019-1022.