The optimal threshold for removing noise from speech is similar across normal and impaired hearing—a time-frequency masking study

Eric W. Healy[a] **and Jordan L. Vasko**
Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA
Healy.66@osu.edu, Vasko.30@osu.edu

DeLiang Wang
Department of Computer Science and Engineering, and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA
Dwang@cse.ohio-state.edu

Abstract: Hearing-impaired listeners' intolerance to background noise during speech perception is well known. The current study employed speech materials free of ceiling effects to reveal the optimal trade-off between rejecting noise and retaining speech during time-frequency masking. This relative criterion value (−7 dB) was found to hold across noise types that differ in acoustic spectro-temporal complexity. It was also found that listeners with hearing impairment and those with normal hearing performed optimally at this same value, suggesting no true noise intolerance once time-frequency units containing speech are extracted.

[Q-JF]
Date Received: April 8, 2019 **Date Accepted:** June 3, 2019

1. Introduction

It is well known that individuals having a sensorineural hearing impairment have particular difficulty understanding speech when background noise is present. Unfortunately, hearing aids and cochlear implants cannot currently remedy this common everyday difficulty. Less well understood are the mechanisms underlying this speech-in-noise deficit. The current study aims to provide some insight.

Time-frequency (T-F) masking has proven to be an effective tool to improve speech intelligibility in noise for hearing-impaired (HI) and normal-hearing (NH) listeners. In T-F masking, the speech-plus-noise mixture is first divided in time and frequency into T-F units. Units dominated by the target speech are retained whereas units dominated by background noise are attenuated. This results in vast increases in target speech intelligibility, and T-F masking has served as a foundation for several deep-learning noise-reduction algorithms (Healy *et al.*, 2013, 2015; Chen *et al.*, 2016; Zhao *et al.*, 2018). But T-F masking can also serve as a model for human perception. According to a glimpsing account of speech perception in noise, listeners extract glimpses (T-F units) of the target speech that are largely spared from the background, assemble these glimpses into a speech percept, and disregard units dominated by the background (Buus, 1985; Apoux and Healy, 2009; Healy *et al.*, 2014).

Fundamental to this T-F masking approach is the threshold for defining a T-F unit as speech or noise dominant. A trade-off exists between rejecting noise and preserving speech. At one end of the continuum, units containing even small amounts of noise are attenuated, ensuring a noise-free speech signal. But this removes a large number of units, including potentially beneficial speech-dominant units, resulting in an impoverished speech signal. At the other end of the continuum, fewer units are attenuated, requiring greater noise tolerance, but ensuring that few if any beneficial speech units are discarded. The noise-rejection threshold is termed the local criterion (LC), when expressed as a dB signal-to-noise (SNR) value, or the relative criterion (RC), when expressed relative to overall SNR (where RC = LC − SNR, all in dB units).

Prior examinations of the optimal noise-rejection threshold have generally involved sentence materials. But because T-F masking is so effective, sentence materials yield ceiling intelligibility values across a broad range of LC or RC values, obscuring the optimal threshold (Brungart *et al.*, 2006; Li and Loizou, 2008; Kjems *et al.*, 2009;

[a]Author to whom correspondence should be addressed.

Sinex, 2013; Chen, 2016). Further, some differences in optimal noise-rejection threshold have been observed across different noise-type backgrounds (e.g., Kjems *et al.*, 2009), but ceiling effects tend to also obscure these potential influences. Accordingly, the first two purposes of the current study were to (i) determine the optimal noise-rejection threshold for speech materials that do not produce ceiling effects and (ii) examine this threshold in noise backgrounds that differ in spectro-temporal complexity.

Because of their everyday difficulties in noise, it is widely believed that HI listeners are more sensitive to, or less tolerant of, noise. This is certainly true in everyday environments. But what is not known is whether this noise intolerance leads HI listeners to perform best at a different position on the noise-rejection continuum—whether they prefer to reject more noise at the expense of losing some speech information. Prior work on this topic has primarily involved NH listeners and so the noise-rejection threshold for HI listeners relative to that of NH listeners has not been established. Accordingly, the third purpose of the current study was to establish this comparison.

2. Methods

There were three groups of 12 subjects each. A first group consisted of listeners with bilateral sensorineural hearing impairment who wore bilateral hearing aids. These individuals were recruited from The Ohio State University Speech-Language-Hearing Clinic to represent typical patients. Audiograms collected on day of test are displayed in Fig. 1, where subject numbers, ages in years, and genders are also provided. Air- and bone-conduction audiometric thresholds helped establish the sensorineural nature of hearing loss. Two additional groups consisting of young NH listeners were recruited from courses at The Ohio State University. NH was defined as audiometric thresholds of 20 dB Hearing Level (HL) or below at octave frequencies from 250 to 8000 Hz in both ears on day of test. The ages for the first NH group were 19 to 25 yr (mean = 20.9) and all were female, and those for the second group were 19 to 31 yr (mean = 22.0) and 7 were female. All subjects received a cash incentive or course credit for participating.

The speech materials consisted of standard recordings of the CID W-22 monosyllabic words in the carrier phrase, "Say the word___." Two background noises were employed. First was a speech-shaped noise (SSN), which involved Gaussian noise shaped to match the long-term average amplitude spectrum of all concatenated W-22 words used for testing. The second noise was cafeteria noise from an Auditec CD. This noise consisted of three overdubbed recordings in a busy hospital-employee

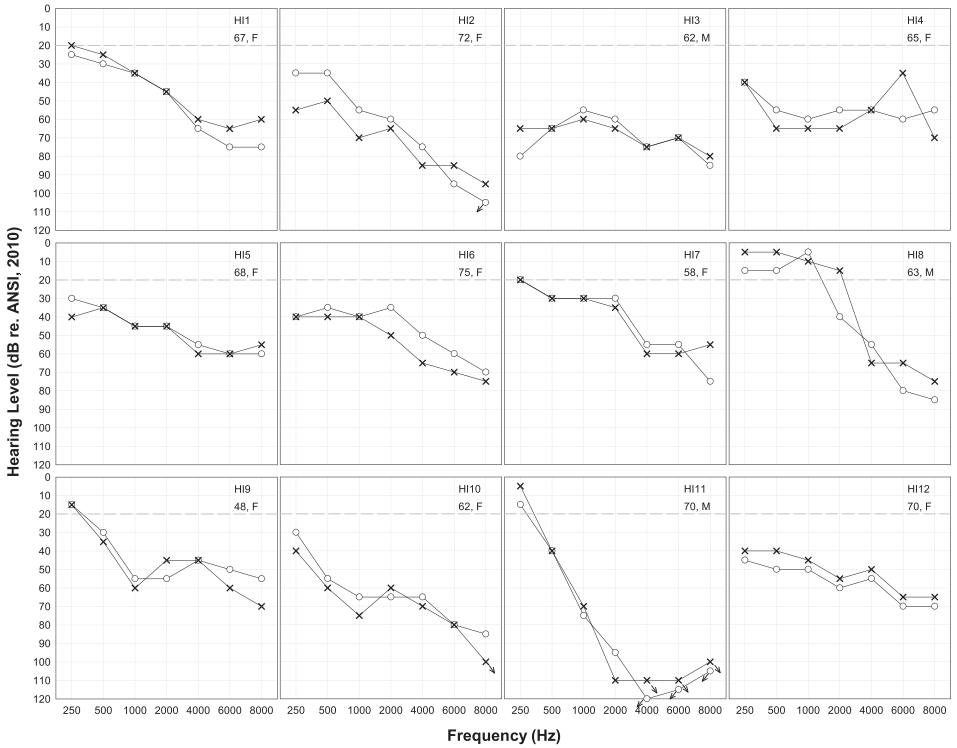

Fig. 1. Pure-tone air-conduction audiometric thresholds for the listeners with sensorineural hearing impairment. Right ears = circles; left ears = X's. The dashed horizontal line at 20 dB HL represents the NH threshold. Arrows indicate thresholds beyond audiometer limits. Listener ages (years) and genders are also given.

cafeteria and contained a variety of sound sources, including the speech of multiple talkers and the sounds of dishes. The SSN was used for the first group of NH listeners, and the cafeteria noise was used for the second group of NH listeners and the HI listeners. The 44 kHz, 16-bit signals were downsampled to 16 kHz for processing.

The T-F mask employed currently was the Ideal Binary Mask (IBM; Hu and Wang, 2001; Wang, 2005), because it has a single SNR threshold to divide T-F units into speech or noise. The IBM was created by first dividing the speech-plus-noise mixture into T-F units using 64 equal ERB_N-width gammatone frequency channels with center frequencies spanning 50 to 8000 Hz, and 20-ms time frames with 10-ms overlap (where ERB_N is the equivalent rectangular bandwidth for NH listeners). Each T-F unit was assigned a value of 1 if its SNR exceeded a criterion value and a value of 0 otherwise. Because the mask was the ideal version, SNR was based on the separate speech and noise signals. The result was a T-F matrix of 1's and 0's that was multiplied with the speech-plus-noise mixture cochleagram (spectrogram) to retain units dominated by speech and discard units dominated by noise. Six noise-rejection thresholds were examined. These RC values were −20, −15, −10, −5, 0, and +5 dB and represent dB SNR values relative to the overall SNR of −8 dB. There was also one unprocessed condition in which speech and noise were mixed but not subjected to the IBM. Figure 2 displays the results of IBM processing at different RCs. First shown are the target speech (a) and the speech-plus-noise mixture (b). The remaining panels display this mixture subjected to the IBM having different values of RC. At RC = −20 dB, little if any potentially beneficial speech information is removed, but considerable noise remains. At RC = +5 dB, noise is largely absent, but the speech appears slightly impoverished. Also apparent in the figure is the ability of the IBM to accurately extract target speech from obscuring noise (e.g., compare RC = 0 dB to speech in quiet).

Listeners heard 25 words in each of the 7 conditions in random order. The word list-to-condition correspondence was randomized so that, across listeners, each set of words was heard in each condition an approximately equal number of times. Listeners were tested while seated with the experimenter in a double-walled audiometric booth. Signals were played from a PC using Echo Digital Audio (Santa Barbara, CA) Gina 3G digital-to-analog converters and presented diotically over Sennheiser HD 280 headphones (Wedemark, Germany). The presentation level at each earphone was 65 dBA for NH listeners and 65 dBA plus frequency-specific gains customized for

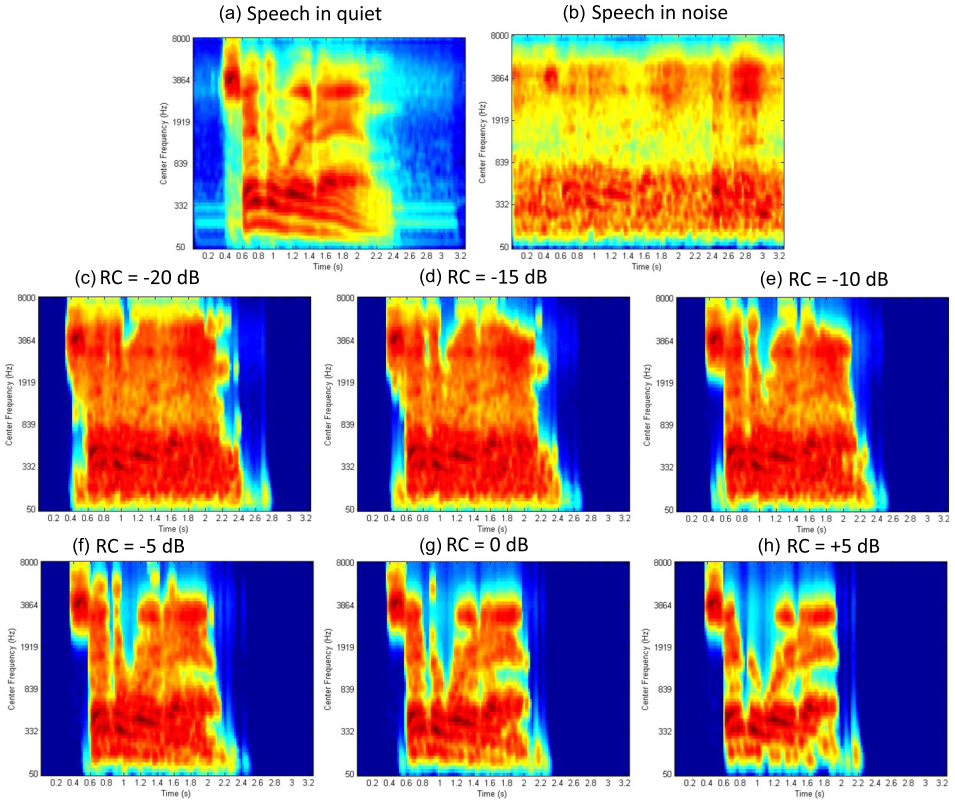

Fig. 2. (Color online) Spectrogram representations of, "Say the word *an*." (a) In quiet and (b) mixed with noise from a busy cafeteria at -8 dB SNR. Shown in (c)–(h) are the speech-plus-noise mixture in (b) after IBM processing to extract the target speech using six different RC values (RC = dB rel. to overall SNR).

each HI listener using the NAL-RP hearing-aid gain formula. Presentation level was calibrated using a flat-plate coupler and type I sound-level meter (Larson Davis AEC 101 and 824) and NAL-RP gains were provided using a Rane (Cumberland, RI) DEQ 60L digital equalizer. At the start of the session, listeners heard a brief practice consisting of 29 words from the consonant-nucleus-consonant corpus, each in the carrier phrase "Ready, ___." Five words were presented in quiet, followed by six words in each of the following conditions: unprocessed, RC = −15 dB, RC = 0 dB, and the first-heard condition. Feedback was provided during practice but not during testing. Listeners were instructed to report the target word in each trial and encouraged to guess if unsure. The experimenter controlled the presentation of words and recorded responses. One HI subject (HI7) was mistakenly run at a constant level of 77 dBA, rather than 65 dBA plus NAL-RP gains, but her data were retained because of her relatively flat hearing-loss configuration and scores well within the range of the other subjects.

3. Results

Figure 3 (left panel) displays group-mean word recognition scores and standard errors of the mean for each condition and each group of listeners. Because the HI listeners were recruited without strict audiogram restrictions, their data were averaged to represent typical hearing-aid wearing patients. As can be seen, the use of W-22 word lists rather than sentence materials produced scores free of ceiling effects. It is first notable that the IBM improved speech recognition considerably in all conditions—scores improved relative to the unprocessed condition by 23 percentage points in the poorest condition and by 62 percentage points in the best condition. Speech recognition accuracy peaked at an RC value of −5 dB for all listener groups and in both noise types. To obtain a more exact measure of the optimal RC value, a quadratic fit was made to each function in the left panel of Fig. 3. The peaks of these curves were at −7.0, −6.6, and −7.1 dB for the HI, NH SSN, and NH cafeteria-noise groups, respectively.

A two-way mixed analysis of variance (6 RC values × 3 listener groups) on rationalized arcsine units was performed to examine potential differences in the functions displayed in the left panel of Fig. 3. Of primary interest was the potential interaction between RC value and subject group, because this would suggest differences in the shapes of the three curves. This interaction was nonsignificant [$F(10,165) = 0.5$, $p = 0.89$], suggesting a similar RC function for both SSN and cafeteria noise and for both HI and NH listeners. Of secondary interest was the main effect of RC value, which was significant [$F(5,165) = 62.9$, $p < 0.001$]. *Post hoc* testing using the Holm–Sidak method indicated that scores at the −5 dB peak of the function did not differ from scores at −10 dB ($p > 0.05$), but they did differ from scores at other values of RC ($p < 0.05$). Finally, the main effect of listener group was significant [$F(2,33) = 4.0$, $p < 0.05$], and Holm–Sidak *post hoc* tests indicated that overall scores for the NH SSN group differed ($p < 0.05$) from those for the HI group, but that other comparisons were nonsignificant.

Another analysis was conducted on the acoustic stimuli to determine the proportion of retained T-F units in each condition. The right panel of Fig. 3 displays these values, which are represented as means and standard deviations for the 25 words in each condition. Apparent is the smaller proportion of units retained as the RC value becomes larger and as more noise is rejected at the expense of potential speech

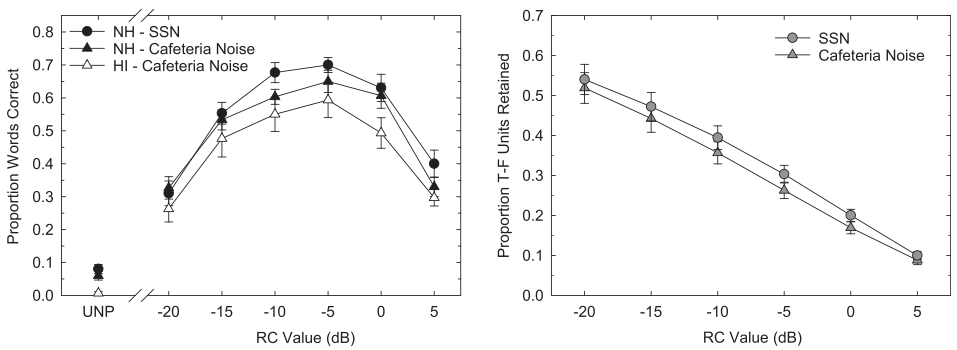

Fig. 3. (Left panel) Group-mean word recognition scores (and standard errors) for listeners hearing speech subjected to ideal binary masking at six different noise-rejection thresholds (RC). Scores in the same speech-in-noise condition prior to binary masking are also displayed (unprocessed, UNP). One group of HI listeners and two groups of NH listeners were employed, one for each noise type. SSN = speech-shaped noise. (Right panel) Proportion of T-F units retained in the IBM at different values of RC for CID W-22 words in SSN and cafeteria noise. Displayed are means and standard deviations across the 25 words used for testing.

information. At the optimal RC value of −7 dB, the proportion of retained units is approximately 0.35 for the noise having less acoustic spectro-temporal fluctuation (SSN) and 0.30 for the noise having greater acoustic spectro-temporal fluctuation (cafeteria noise).

4. Discussion

Prior work investigating the LC or RC has revealed that the IBM produces ceiling intelligibility across a broad range of values. These values can be as broad as RC = −20 to +5 dB (Brungart *et al.*, 2006; Li and Loizou, 2008; Kjems *et al.*, 2009; Roman and Woodruff, 2013; Sinex, 2013). The current use of word lists allowed speech perception to be assessed without these ceiling effects. Even with ceiling effects absent, the current curves remained somewhat broad, but over a smaller span of roughly 5 to 10 dB (scores did not differ significantly from −10 to −5 dB and were generally similar from −10 to 0 dB). But the current functions do allow the optimal RC value to be established with accuracy. This value was found via curve fitting to be −7 dB, for both SSN and cafeteria-noise backgrounds. This noise-rejection criterion resulted in speech composed of 30% to 35% of its original T-F units, with the remaining 65% to 70% of units absent.

The current study employed only a single overall SNR. But motivation for this decision comes from Kjems *et al.* (2009), who varied overall SNR by over 50 dB and found that the mask pattern and the intelligibility it produced were similar as long as SNR and LC covaried, and noise-rejection threshold was expressed as RC. Accordingly, because RC is considered to control for and hold across a range of overall SNR values (and because the same value was found for different noise types) this optimal RC value of −7 dB may be considered quite general.

This optimal RC value was also found to be common to both HI and NH listeners. This is true despite the use of typical (older) HI and ideal (younger) NH listeners. As mentioned in Sec. 1, the everyday speech-perception difficulties of HI listeners in noise are well established, but the underlying mechanisms are less clear. It is not well understood whether the HI deficit reflects true noise intolerance or an inability to locate and extract glimpses of clean speech in the mixture. The current results suggest that the speech-perception difficulty of HI listeners in noise is almost entirely attributable to their limited ability to resolve the speech-plus-noise mixture into small T-F units and extract desired units. This conclusion is supported by the similarity across HI and NH listeners in noise tolerance as well as similarity in the level of performance once clean speech units were extracted and delivered to listeners (most conditions within 5–6 percentage points across HI and NH in the same noise type). This limited ability of HI listeners to resolve the mixture and extract clean speech likely stems primarily from broad auditory tuning and perhaps from poor temporal resolution consequent to limited audible bandwidth and listening at low sensation levels (see Moore, 2007 for a review of these issues). HI listeners cannot resolve the speech-plus-noise mixture into T-F units small enough to be relatively noise free, and instead all of their available units contain some noise, hindering everyday performance.

5. Conclusions

(1) The optimal noise-rejection threshold is RC = −7 dB and holds for background noises that differ widely in acoustic spectro-temporal fluctuation characteristics.
(2) This threshold is also the same for both HI and NH listeners.
(3) The difficulty faced by HI listeners perceiving speech in background noise appears almost entirely attributable to the limited ability to extract T-F regions of the mixture containing relatively clean speech, and does not arise from a true noise intolerance.
(4) When T-F masking is employed as a basis for noise-reduction algorithms, similar noise-rejection criteria appear to be optimal for both HI and NH listeners.

Acknowledgments

This work was supported in part by the National Institute on Deafness and other Communication Disorders (Grant No. R01 DC015521 to E.W.H. and Grant No. R01 DC012048 to D.L.W.) and by a summer graduate fellowship from The Ohio State University Center for Cognitive and Brain Sciences (J.L.V.).

References and links

Apoux, F., and Healy, E. W. (**2009**). "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," Hear. Res. **255**, 99–108.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Buus, S. (**1985**). "Release from masking caused by envelope fluctuations," J. Acoust. Soc. Am. **78**, 1958–1965.

Chen, F. (**2016**). "Representing the intelligibility advantage of ideal binary masking with the most energetic channels," J. Acoust. Soc. Am. **140**, 4161–4169.

Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., and Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604–2612.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660–1669.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.

Healy, E. W., Youngdahl, C. L., and Apoux, F. (**2014**). "Evidence for independent time-unit processing of speech using noise promoting or suppressing masking release(L)," J. Acoust. Soc. Am. **135**, 581–584.

Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2nd Ed. (Wiley, Chichester), pp. 45–91.

Roman, N., and Woodruff, J. (**2013**). "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," J. Acoust. Soc. Am. **133**, 1707–1717.

Sinex, D. G. (**2013**). "Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters," J. Acoust. Soc. Am. **133**, 2390–2396.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Zhao, Y., Wang, D. L., Johnson, E. M., and Healy, E. W. (**2018**). "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," J. Acoust. Soc. Am. **144**, 1627–1637.