Refining a model of hearing impairment using speech psychophysics

Morten L. Jepsen[a)] **and Torsten Dau**
Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, DK-2800 Kongens Lyngby, Denmark
moje@widex.com, tdau@elektro.dtu.dk

Oded Ghitza
Biomedical Engineering and Hearing Research Center, Boston University, 44 Cummington Street, Boston, Massachusetts 02215
oghitza@bu.edu

Abstract: The premise of this study is that models of hearing, in general, and of individual hearing impairment, in particular, can be improved by using speech test results as an integral part of the modeling process. A conceptual iterative procedure is presented which, for an individual, considers measures of sensitivity, cochlear compression, and phonetic confusions using the Diagnostic Rhyme Test (DRT) framework. The suggested approach is exemplified by presenting data from three hearing-impaired listeners and results obtained with models of the hearing impairment of the individuals. The work reveals that the DRT data provide valuable information of the damaged periphery and that the non-speech and speech data are complementary in obtaining the best model for an individual.

PACS numbers: 43.66.Ba, 43.66.Dc, 43.66.Sr, 43.72.Ar [QJF]
Date Received: December 4, 2013 **Date Accepted:** March 12, 2014

1. Introduction

People with similar sensitivity loss can differ greatly in their ability to understand speech, particularly in complex acoustic environments (e.g., Smoorenburg, 1992). Even with amplification and signal processing techniques provided by hearing aids, their performance can vary substantially (Plomp, 1978). It is, therefore, of great importance to obtain a better understanding of how the auditory processing of speech sounds by an individual is affected by hearing impairment. Aspects of hearing impairment which are not accounted for by sensitivity loss are referred to as supra-threshold deficits and may be associated with individual patterns of outer hair-cell (OHC) and inner hair-cell (IHC) dysfunction (e.g., Lopez-Poveda *et al.*, 2009; Jepsen and Dau, 2011; Poling *et al.*, 2012). OHC dysfunction is associated with threshold shifts due the lack of active amplification of the vibrations on the basilar membrane (BM). This active mechanism is also responsible for the compressive BM input/output function characterizing the normal system. The supra-threshold deficits associated with OHC dysfunction are, e.g., loudness recruitment and reduced frequency selectivity. IHC dysfunction also leads to a threshold shift, since the effective transformation from BM vibration to neural signals is reduced. An exemplary supra-threshold effect of IHC dysfunction is reduced temporal coding acuity. It is unclear how these deficits are reflected in psychoacoustic tasks with speech stimuli (e.g., Dubno *et al.*, 2007). Linking speech and non-speech

[a)]Author to whom correspondence should be addressed. Also at: Biomedical Engineering and Hearing Research Center, Boston University, 44 Cummington Street, Boston, Massachusetts, 02215. Current address: Widex A/S, Nymoellevej 6, DK-3540 Lynge, Denmark.

psychophysics may provide a better understanding of the underlying reasons for the observed performance of hearing-impaired (HI) listeners in speech perception tasks.

Numerous studies have been reported aiming at predicting the performance of HI listeners in a speech perception task (e.g., Jürgens and Brand, 2009; Brown *et al.*, 2010), with a peripheral model used as a front end to a conventional automatic speech recognition system, as a back end. A limiting property of such an approach is the inability to decompose the origin of errors, especially front-end versus back-end errors. Ghitza (1993a) proposed a framework based upon the Diagnostic Rhyme Test (DRT; Voiers, 1983)—a two-alternative task that minimizes the influence of cognitive and memory factors (Ghitza, 1993b).

The DRT method has been used particularly in the speech coding community to measure the intelligibility of processed speech while providing diagnostic information in terms of error patterns on an acoustic-phonetic feature space (Voiers, 1983). In the present study, a procedure was employed to improve models of peripheral hearing impairment, using a machine that mimics the DRT paradigm (Messing *et al.*, 2009). As a baseline, the computational auditory model of Jepsen and Dau (2011) was considered and adjusted to individual hearing loss based on non-speech measures. This model was iteratively adjusted by serving as a front end to a 1-bit recognizer, similar to Messing *et al.* (2009). In each iteration, DRT error patterns were obtained and the parameters of the model were adjusted by an ad hoc, knowledge-based procedure. Separating front-end induced errors from back-end induced errors, achieved by using the DRT methodology was of conceptual importance here, since this should allow the predicted errors to be associated mainly with the front end, and should provide a measure of how well the peripheral model can estimate the internal representation of speech. The hypothesis was that a peripheral model of auditory processing can be refined by considering speech data provided by the DRT framework.

2. Experimental methods

Temporal masking curves (TMC; Nelson *et al.*, 2001) and speech discrimination data (DRT) were obtained from three HI listeners. Following the methodology proposed by Jepsen and Dau (2011), a "phase-1" peripheral model was derived for each individual based on the TMC data and the audiogram. Next, DRT acoustic-phonetic error patterns were derived for the phase-1 model, using the methodology proposed by Messing *et al.* (2009). Finally, obeying biological constraints, parameters of the phase-1 model were adjusted, resulting in a "phase-2" model, with error patterns closer to those measured behaviorally.

Three listeners with mild-to-moderate sensorineural hearing loss participated in this study. Listeners S1, S2, and S3 were 21, 45, and 27 yr old, respectively. Only one ear of each listener was measured in the speech and non-speech tasks. The audiograms of the measured ears are shown in Fig. 1 (open symbols). The listeners were recruited based on the differences in their audiograms, to potentially produce different results in the speech task.

TMCs have been suggested as a useful method for behaviorally estimating the cochlear input-output function in humans (e.g., Nelson *et al.*, 2001), even though several limitations and drawbacks of the method have recently been reported (e.g., Wojtczak and Oxenham, 2009; Lopez-Poveda and Johannesen, 2012). Forward masking of a fixed-level brief tone was measured as a function of the signal-masker interval. The signal was a pure tone with duration of 20 ms, including a Hanning window applied over its entire duration. The signal frequency (f_{sig}) was either 1 or 4 kHz. The signal was presented at 10 dB sensation level (SL). The masker was a pure tone with a duration of 200 ms including 5-ms raised-cosine on- and off-ramps. The masker frequency was equal to f_{sig} (on-frequency condition) or 0.6 f_{sig} (off-frequency condition). The masker level was adjusted by the adaptive procedure to reach masked signal

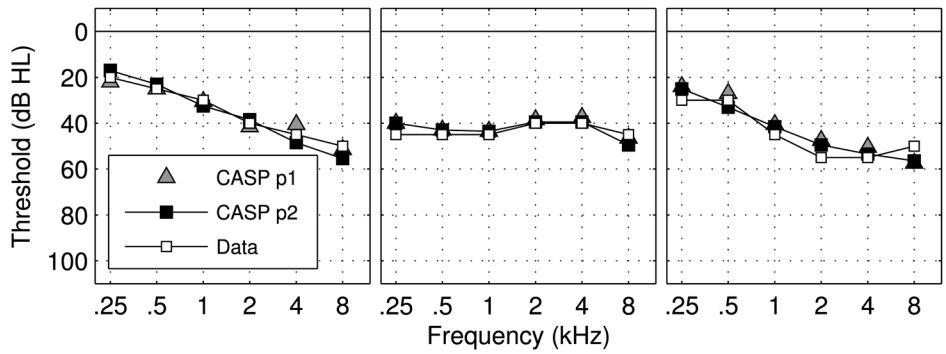

Fig. 1. Audiograms of the measured ears of the three HI listeners. Pure-tone thresholds are plotted in dB hearing level (HL). Open symbols indicate measured thresholds, while filled symbols indicate simulated thresholds by the corresponding models. Gray triangles show thresholds for the model after phase 1. Black symbols indicate the thresholds of the individually fitted model after phase 2.

threshold. A three-interval three-alternative forced choice paradigm with a two-up one-down rule was applied.

The DRT database consists of 96 minimal word pairs spoken in isolation. Words in a pair differ only in their initial consonants. The feature classification follows the binary system suggested by Jakobson *et al.* (1952). The six dimensions, used by Voiers, are Voicing, Nasality, Sustention, Sibilation, Graveness, and Compactness (denoted VC, NS, ST, SB, GV, and CM, respectively). To minimize back-end induced errors, *synthetic* DRT stimuli were used. The synthetic stimuli were generated using HLSyn (Hanson and Stevens, 2002), a modification of the Klatt speech synthesizer. The experimental paradigm was a one-interval two-alternative forced-choice experiment (to assure a task with minimum cognitive load). The masker was an additive stationary speech-shaped noise. Data were obtained at a speech presentation level of 70 dB sound pressure level and signal-to-noise ratios (SNR) of 0 and 10 dB.

3. Modeling approach

3.1 Peripheral model as front end

In the peripheral model, the computational auditory signal processing and perception model (CASP; Jepsen *et al.*, 2008; Jepsen and Dau, 2011), the acoustic stimuli are first processed by the outer and middle ear filters, followed by the dual-resonance nonlinear (DRNL) filterbank (Lopez-Poveda and Meddis, 2001) simulating BM processing. The processing of the subsequent stages is carried out in parallel in the frequency channels. Inner hair-cell transduction is modeled roughly by half-wave rectification followed by a first-order lowpass filter with a cut-off frequency at 1 kHz. The expansion stage transforms the output of the IHC stage into an intensity-like representation by applying a squaring expansion. The adaptation stage simulates dynamic changes in the gain of the system in response to changes in the input level. The output of the adaptation stage is processed by a modulation filterbank, which is a bank of bandpass filters tuned to different modulation frequencies. The output of the preprocessing stages, termed the internal representation (denoted IR), was generated using DRNL filters in the range from 0.1 to 8 kHz and six modulation filters with center frequencies logarithmically spaced and ranging from 0 to 46 Hz.

3.2 The back end

The back end (Messing *et al.*, 2009) is based on template-matching. A template-match operation comprises measuring the "distance" of the unknown token to the templates and labeling the unknown token as the template with the smaller distance. Hence, template matching is defined by the distance measure and the choice of templates. Here,

the distance measure was the Euclidean distance. The templates were the IRs of each word, in quiet. New template IRs were obtained after each iteration, using the current adjusted model. For a given test word, the corresponding IR (IR_x) was calculated at the prescribed SNR, and the Euclidean distance (i.e., the mean-squared-errors, MSE) between IR_x and the two templates were calculated across time, frequency and modulation frequency. The Euclidean distance was considered here as representing the *perceptual* distance between the test IR and the template. The detector chose the template that produced the smallest MSE as representing the input word. In the present study, a probabilistic decision criterion (a *soft* decision) was introduced which reflects internal noise in the model.

3.3 Iterative procedure

Following Jepsen and Dau (2011), the parameters of the cochlear stages of the peripheral model were adjusted in order to estimate degraded processing due to hair-cell loss. The input-output (I/O) behavior of the DRNL filterbank was adjusted to correspond to the basilar-membrane (BM) I/O functions estimated behaviorally in the TMC experiment for the three HI listeners, in terms of the compression exponent and the knee point. After parameters were estimated at 1 and 4 kHz, linear interpolation and extrapolation were used to obtain parameter-sets for a range of filter center frequencies (0.1 to 8 kHz). The suggested procedure also provided estimates of the effects of OHC and IHC losses with respect to sensitivity. OHC loss was estimated from the fitted I/O functions. The IHC loss was then considered as the difference between the total sensitivity loss and the OHC loss. The loss of sensitivity due to IHC loss was simulated as a linear attenuation at the output of the hair-cell transduction stage. The baseline model of normal hearing was denoted MNH, and the models fitted to listeners S1, S2, and S3 were assigned subscript "p1" (for phase-1) and denoted $M1_{p1}$, $M2_{p1}$, and $M3_{p1}$, respectively. In order to evaluate the sensitivity of the p1 models, the individual audiograms were simulated (see gray triangles in Fig. 1). This simulation was done using the "optimal detector" back-end framework and procedure presented in Jepsen and Dau (2011).

Next, the DRT stimuli were processed by the phase-1 models. The responses were analyzed by the 1-bit recognizer described above, generating error patterns along the Jakobsonian acoustic-phonetic space (Jakobson *et al.*, 1952). The interpretation of the error patterns steered the adjustment of the cochlear stage to better match the human error patterns. This was done in an *ad hoc* procedure based on knowledge about the acoustic correlates of the Jakobsonian dimensions. In the peripheral models, simulated OHC dysfunction reduces the temporal and spectral resolution of the model output, as well as its amplitude, while simulated IHC loss reduces the amplitude of the IR. A combined IHC/OHC dysfunction can lead to an attenuation of parts of the speech representation below the (simulated) absolute threshold.

For example, the overestimated number of errors in the NS dimension is interpreted as the result of reduced energy at low-frequencies of the model output. This can be compensated for by an increase in the simulated IHC loss, while remaining inside the uncertainty range of the measured audiogram. The models fitted to listeners S1, S2, and S3 were denoted $M1_{p2}$, $M2_{p2}$, and $M3_{p2}$, respectively.

4. Results

The results from the psychoacoustic data used to fit model parameters, i.e., the TMCs and DRT data, are not shown explicitly here. Some general observations were that BM compression was found for listener S2 at 1 and 4 kHz, and for S3 at 4 kHz, but for all listeners the slopes of the BM I/O function were larger than for NH listeners, i.e., above about 0.25 dB/dB. In the DRT, listener S1 exhibited many more errors than the NH listeners at both SNRs and had the worst overall performance among the HI listeners of this study. S2 showed a high error rate (above 30%) at 0 dB SNR in all dimensions except NS. At an SNR of 10 dB, there were substantially fewer errors for

this listener. S3 had the best performance at an SNR of 0 dB among the HI listeners and, similar to the other listeners, benefitted from a better SNR.

The predicted error patterns are shown in Fig. 2 as error differences, in percent, relative to the measured DRT data. The zero line reflects a perfect match to the human data. Negative values indicate that the model produced fewer errors than the corresponding listener. Bars show the error difference and boxes indicate one standard deviation of the data. The "+" sign stands for the attribute that is present and the "−" sign for the attribute that is absent. (For example, the pair meat-beat: "meat" is NS+, "beat" is NS−.) Matches were considered good if they were within the boxes. The results obtained with MNH_{p1} generally showed good matches. The model of normal hearing could account reasonably well for the measured error patterns, suggesting that the peripheral model is able to reflect aspects of auditory processing relevant for speech discrimination.

The differences between predicted and measured error rates for the three HI listeners are shown in the two left columns of Fig. 2, marked $M1_{p1}$, $M2_{p1}$, and $M3_{p1}$. Specifically, model $M1_{p1}$ could account for the errors of listener S1 in the dimensions NS, ST−, SB, GV+, and CM+, while errors were underestimated in VC, ST+, and CM−. Model $M2_{p1}$ accounted for the measured errors of listener S2 in the dimensions ST−, GV, and CM at the SNR of 0 dB, and VC, SB−, and CM at the SNR of 10 dB.

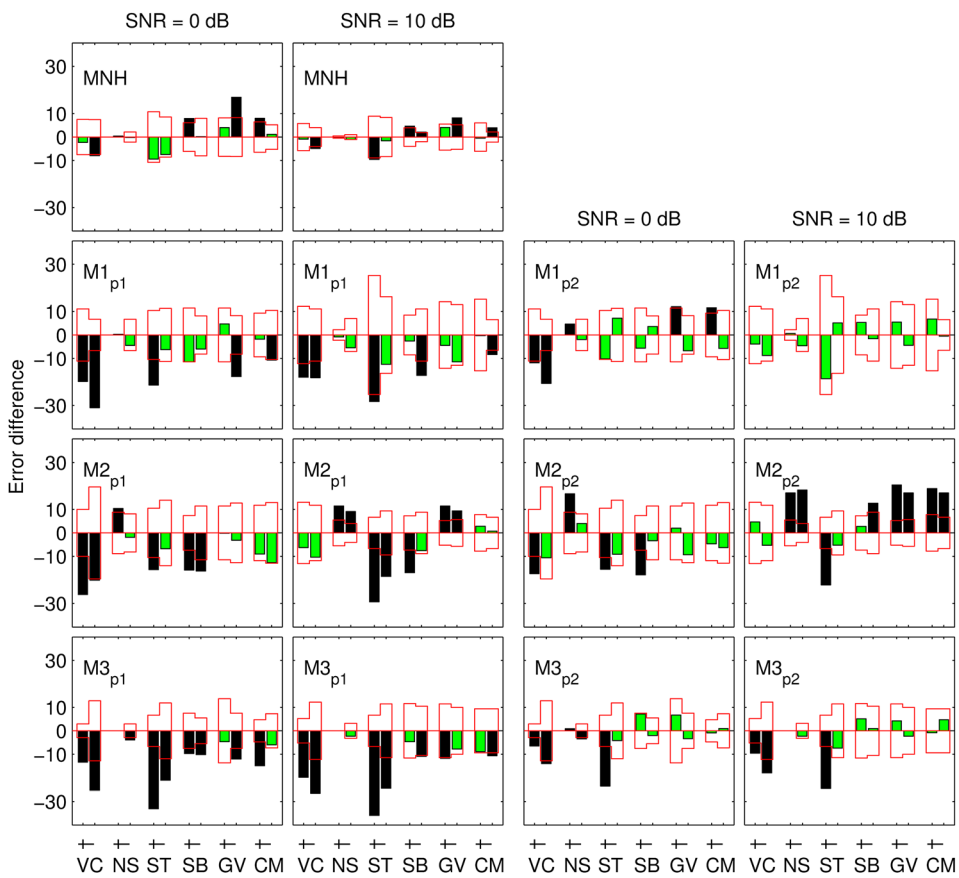

Fig. 2. (Color online) Error patterns for a model of NH listeners and models of individual HI listeners. Subscripts p1 and p2 indicate phase-1 and phase-2 models, respectively. Errors are presented as error differences between the predicted and the measured DRT data (model - human), in percent. The zero line reflects a perfect match. The boxes represent the standard deviation of the data. When the predicted error rate was within 1 standard deviation, the bars are indicated by a green (gray) color.

Model $M3_{p1}$ provided good matches to listener S3 in the dimensions NS and GV+ at both SNRs, while the error rates were substantially underestimated in the remaining dimensions.

The DRT error differences between the human data and the phase-2 model predictions are shown in the two right panels of Fig. 2, marked as $M1_{p2}$, $M2_{p2}$, and $M3_{p2}$. The IHC loss parameter was adjusted *ad hoc* and was guided by the following criteria. Too few model errors in the dimension VC indicates that the model might be too sensitive in the mid-frequency range. This is, for example, the case for $M1_{p1}$. Too few model errors in the dimension ST also suggests that the model may be too sensitive in the mid-frequency range. It was thus considered to reduce the sensitivity at mid frequencies to improve the predictions of the DRT error patterns by increasing the IHC loss component. $M2_{p1}$ generally underestimated the error rates, except for the dimensions NS and GV, and this suggests that the OHC gain was too large and the model's IR of the stimuli was too detailed. Too many model errors in the NS dimension (as in the case for $M2_{p1}$) indicates that the sensitivity at low frequencies should be slightly increased, since the primary cue for NS is at low frequencies. For model $M3_{p2}$, a better match was obtained by slightly reducing the sensitivity at the mid- to high frequencies. This was so since the $M3_{p1}$ model had too few errors along the dimensions ST, SB, and CM, which can be associated with the mid- and high-frequency information. Reducing the sensitivity in phase 2 affected the errors differently along all six dimensions. To summarize, for model $M1_{p2}$, the matches were within 1 standard deviation in 19 of 24 attributes. $M2_{p2}$ produced closer matches to the data than $M2_{p1}$. $M3_{p2}$ produces an error within 1 standard deviation in 16 of 24 attributes, which is an improvement compared to $M3_{p1}$ that only reached within 1 standard deviation in 7 of 24 attributes. For the evaluation of the sensitivity of the p2 models, associated audiograms were simulated (see the black squares in Fig. 1).

5. Discussion

It was shown that the DRT methodology resulted in a characterization of the hearing impairment different from that which was obtained using the TMC data. This observation appears relevant since the ultimate goal of many models of the damaged auditory periphery is to form a basis for speech intelligibility prediction.

For NH listeners, the phase-1 model produced fairly good predictions of the measured DRT data. For the HI listeners, the individual phase-1 models produced fewer DRT errors than their corresponding listeners. To obtain a closer match of the behavioral DRT data, sensitivity needed to be reduced; this was achieved by increasing the value of HL_{IHC} in individual frequency channels while keeping the estimated amount of HL_{OHC}. The resulting phase-2 models generated better matches of the DRT error patterns, without affecting the predictions of the TMC data. Interestingly, the adjusted values of HL_{IHC} reflected a linear reduction of the effective gain at the output of the cochlea but did not affect the amount of predicted compression or the frequency tuning of the peripheral filters.

It should be noted that the hearing loss estimates were based on data collected at only 1 and 4 kHz (and the audiogram at frequencies between 0.25 and 8 kHz). Similar data could be obtained at more frequencies to gain more confidence in the used cochlear model parameters. The use of interpolation and extrapolation of the parameters at other DRNL filterbank frequencies may prove to be a too crude assumption. Furthermore, OHC and IHC losses have in the present study only been considered in terms of associated sensitivity losses but not in terms of temporal coding which was outside the scope.

Future research programs to extend and strengthen the approach presented may address the following issues: (1) the usage of test data separate from the fitting/training data (this should lead to a more rigorous evaluation of the predictive power of the model) and (2) the development of an automated fitting/optimization procedure for phase 2.

Finally, the suggested framework may be applied to the development of signal processing algorithms for hearing aids. It is suggested to use speech test results as an integral part of the hearing aid design in an iterative procedure, with the goal of minimizing the phonetic confusions produced by the hearing aid connected in tandem with a model of the patient's peripheral hearing impairment. The confusions should be minimized across the acoustic-phonetic features and across different environmental conditions.

References and links

Brown, G. J., Ferry, R. T., and Meddis, R. (**2010**). "A computer model of auditory efferent suppression: Implications for the recognition of speech in noise," J. Acoust. Soc. Am. **127**, 943–954.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (**2007**). "Estimates of basilar-membrane nonlinearity effects on masking of tones and speech," Ear. Hear. **28**, 2–17.

Ghitza, O. (**1993a**). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acoust. Soc. Am. **93**, 2160–2171.

Ghitza, O. (**1993b**). "Processing of spoken CVCs in the auditory periphery. I. Psychophysics," J. Acoust. Soc. Am. **94**, 2507–2516.

Hanson, H. M., and Stevens, N. (**2002**). "A quasiarticulatory approach to controlling acoustic source parameters in a klatt-type formant synthesizer using HLsyn," J. Acoust. Soc. Am. **112**, 1158–1182.

Jakobson, R., Fant, C. G. M., and Halle, M. (**1952**). "Preliminaries to speech analysis: The distinctive features and their correlates," Tech. Rep. **13** (Acoustic Laboratory, MIT, Cambridge, MA).

Jepsen, M. L., and Dau, T. (**2011**). "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **129**, 262–281.

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**, 422–438.

Jürgens, T., and Brand, T. (**2009**). "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," J. Acoust. Soc. Am. **126**, 2635–2648.

Lopez-Poveda, E. A., and Johannesen, P. T. (**2012**). "Behavioral estimates of the contribution of inner and outer hair cell dysfunction to individualized audiometric loss," J. Assoc. Res. Otolaryngol. **13**, 485–504.

Lopez-Poveda, E. A., Johannesen, P. T., and Merchán, M. A. (**2009**). "Estimation of the degree of inner and outer hair cell dysfunction from distortion product otoacoustic emission input/output functions," Audiolog. Med. **7**, 22–28.

Lopez-Poveda, E. A., and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," J. Acoust. Soc. Am. **110**, 3107–3118.

Messing, D. P., Delhorne, L., Bruckert, E., Braida, L. D., and Ghitza, O. (**2009**). "A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise," Speech Commun. **51**, 668–683.

Nelson, D. A., Schroder, A. C., and Wojtczak, M. (**2001**). "A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **110**, 2045–2064.

Plomp, R. (**1978**). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am. **63**, 533–549.

Poling, G. L., Horwitz, A. R., Ahlstrom, J. B., and Dubno, J. R. (**2012**). "Individual differences in behavioral estimates of cochlear nonlinearities," J. Assoc. Res. Otolaryngol. **13**, 91–108.

Smoorenburg, G. (**1992**). "Speech reception in quiet and in noisy conditions by individuals with noise induced hearing loss in relation to their tone audiogram," J. Acoust. Soc. Am. **91**, 421–437.

Voiers, W. D. (**1983**). "Evaluating processed speech using the diagnostic rhyme test," Speech Technol. **1**, 30–39.

Wojtczak, M., and Oxenham, A. J. (**2009**). "Pitfalls in behavioral estimates of basilar-membrane compression in humans," J. Acoust. Soc. Am. **125**, 270–281.