

Research Article

Level-Dependent Changes in Perception of Speech Envelope Cues

JUDY R. DUBNO,¹ JAYNE B. AHLSTROM,¹ XIN WANG,¹ AND AMY R. HORWITZ¹

¹*Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, 135 Rutledge Avenue, MSC 550, Charleston, SC 29425-5500, USA*

Received: 17 January 2012; Accepted: 16 July 2012; Online publication: 8 August 2012

ABSTRACT

Level-dependent changes in temporal envelope fluctuations in speech and related changes in speech recognition may reveal effects of basilar-membrane nonlinearities. As a result of compression in the basilar-membrane response, the “effective” magnitude of envelope fluctuations may be reduced as speech level increases from lower level (more linear) to mid-level (more compressive) regions. With further increases to a more linear region, speech envelope fluctuations may become more pronounced. To assess these effects, recognition of consonants and key words in sentences was measured as a function of speech level for younger adults with normal hearing. Consonant–vowel syllables and sentences were spectrally degraded using “noise vocoder” processing to maximize perceptual effects of changes to the speech envelope. Broadband noise at a fixed signal-to-noise ratio maintained constant audibility as speech level increased. Results revealed significant increases in scores and envelope-dependent feature transmission from 45 to 60 dB SPL and decreasing scores and feature transmission from 60 to 85 dB SPL. This quadratic pattern, with speech recognition maximized at mid levels and poorer at lower and higher levels, is consistent with a role of cochlear nonlinearities in perception of speech envelope cues.

Keywords: basilar-membrane responses, compression, human, speech envelope, vocoder

INTRODUCTION

Speech is a broadband signal that contains varying degrees of envelope fluctuation (the “speech temporal envelope”). Recognition of speech depends on several acoustic factors, most notably signal level and audibility in specific spectral regions, and also on information carried in the speech temporal envelope (e.g., Rosen 1992; Gallun and Souza 2008). Level-dependent changes in the temporal envelope presumed to result from compression in the basilar-membrane response could play a role in the shape of the psychometric function for speech, especially when perception is dependent primarily on speech envelope cues (Glasberg and Moore 1989, 1992). However, it is unclear how speech recognition changes with increasing level when speech is processed to minimize spectral information and presented in a manner that maintains constant speech audibility.

An important consequence of the cochlear amplifier is the compressive growth in the input–output function of the basilar-membrane response, which can be estimated in humans by psychophysical procedures (for review, see Oxenham and Bacon 2003). The active mechanism has important (and interrelated) consequences for auditory function, including suppression, tuning, and temporal effects, each of which has significant implications for speech recognition. Beyond these general effects, a thorough understanding of how the basilar-membrane input–output response relates to communication in real-world environments is lacking because there is limited knowledge of the effects of nonlinear processing on complex sounds such as speech.

In recent attempts to assess the role of nonlinear effects on speech recognition, we have focused on

Correspondence to: Judy R. Dubno · Department of Otolaryngology-Head and Neck Surgery · Medical University of South Carolina · 135 Rutledge Avenue, MSC 550, Charleston, SC 29425-5500, USA. Telephone: (843) 792-7978; fax: (843) 792-7736; e-mail: dubnojr@musc.edu

level-dependent changes in masker effectiveness due to basilar-membrane response growth and presumed effects on perception of consonants in noise maskers; tonal growth-of-masking psychophysical measures were used to estimate cochlear nonlinearities (Dubno et al. 2007; Horwitz et al. 2007). Another approach was taken in the current experiment, based on an assumption that the effective magnitude of envelope fluctuations in speech changes with increasing level as a result of compression in the basilar-membrane response. This approach derived from results of a similar experiment that measured detection of gaps in broadband markers with varying depths of envelope fluctuation as a function of marker level (Horwitz et al. 2011). Gap detection is poorer for signals with larger envelope fluctuations (e.g., Glasberg and Moore 1992; Hall and Grose 1997), which is likely due to increased confusion between the imposed gap and inherent fluctuations of the signal. Compression in the basilar-membrane response would be expected to flatten the marker temporal envelope, which predicts a level-dependent change in gap detection. A similar level-dependent effect would be predicted for identification of speech features that rely on detection of brief silent intervals in speech, such as within plosive consonants (Glasberg and Moore 1989, 1992). Indeed, Glasberg and Moore (1989) found significant correlations between a measure of gap detection and a speech-in-noise task and suggested that gap detection thresholds could be useful as a predictor of speech recognition in noise.

Although most studies of speech recognition focus on information contained in speech energy in the spectral domain, substantial evidence confirms that amplitude modulations in the speech signal carry information important for communication (Rosen 1992; Shannon et al. 1995). The speech signal contains varying degrees of envelope fluctuation, which convey different types of information. The hierarchy of speech temporal information suggested by Rosen includes amplitude envelope cues with slow rates of modulation (2–50 Hz), which are associated with segment durations, rise and fall times that convey manner information and prosodic cues for syllabification and stress, and higher rates of modulation that convey periodicity cues (50–500 Hz) and temporal fine structure (500–10,000 Hz). In some studies, near-normal recognition of speech that contains little spectral information has been achieved when the speech temporal envelope is preserved in a relatively small number of frequency channels. Indeed, Shannon et al. (1995) reported that four-channel “vocoded” speech, which had very limited frequency information but relatively intact temporal envelope information, was highly intelligible.

Given the wide variety and importance of speech information carried by temporal cues, alteration of

such cues may reduce speech recognition. However, because of the redundancy of speech information, these changes may be revealed only when listeners must depend primarily on temporal envelope cues, as in the case of certain types of hearing loss or with cochlear implants (e.g., Freyman et al. 1991; Fu and Shannon 1998, 1999). For example, Freyman et al. (1991) measured consonant recognition for vowel–consonant–vowel (VCV) syllables that had been mixed at several signal-to-noise ratios (SNRs) with a speech-envelope modulated noise, which limited spectral cues. The consonant portion of the VCV syllable was amplified by 10 dB (a form of envelope alteration). Although overall consonant recognition did not change, consonant confusion patterns revealed that perception of some consonants improved with envelope alteration whereas others declined.

Fu and Shannon (1998) investigated the effect of compression and expansion on the temporal envelope of four-channel processed consonants and vowels in quiet at a conversational level of 70 dB SPL. Subjects with normal hearing achieved their best performance with no amplitude alteration; compression and expansion reduced performance. Somewhat different effects of compression and expansion were observed when 16-channel vocoded speech was presented in a background of noise (Fu and Shannon 1999). Although overall scores for vowels and (especially) consonants were poorer in noise than in quiet, scores for expanded signals were better than or equal to those with no amplitude alterations.

The effect of signal level on recognition of speech with minimal spectral information has been assessed in only a few studies (e.g., Freyman and Nerbonne 1996; Başkent 2006). Most were not designed to control for individual differences in quiet thresholds or maintain constant speech audibility as signal level increased. Similar to unprocessed speech, recognition of vocoded speech generally increases as signal level (and sensation level) increases from low to mid levels and then remains the same or decreases at higher levels. Using masking by signal-correlated noise, Freyman and Nerbonne (1996) measured the effect of changes in overall signal level and SNR on VCV syllable recognition with and without amplitude expansion. In a control condition with no expansion, as stimulus level increased from 30 to 80 dB SPL, vowel recognition improved and reached asymptote at 60 dB SPL. When the envelope of the VCV syllable was expanded by a factor of 3.0 (in terms of decibels), syllable recognition was degraded for normal-hearing subjects, but increased gradually with increasing level; vowel error patterns differed for unprocessed and expanded syllables.

Başkent (2006) examined the effects of audibility and signal level using noise-vocoded speech with

groups of subjects with normal and impaired hearing. Psychometric functions for vowels and consonants were measured in various levels of background noise as a function of the number of spectral channels. Speech was presented to normal-hearing subjects at three levels for vowels and consonants, respectively: (1) at low levels (20 and 30 dBA) selected to reduce audibility, (2) at sensation levels (35 and 50 dBA) comparable to optimal listening levels for the subjects with hearing loss in an attempt to equate audibility and comfort, and (3) at high levels (85–95 dBA) comparable to the absolute levels used by the subjects with hearing loss. For normal-hearing subjects, vowel and consonant recognition increased significantly from lower levels to comfortable levels for some conditions and reached asymptote at higher levels. It is likely that the improvement in speech recognition with increasing level was due to increasing speech sensation levels.

The current study was based on an assumption that compression in the basilar-membrane response effectively flattens the temporal envelope of fluctuating signals (in this case, vocoded speech) when more gain is applied to lower-level than higher-level components. The effective speech envelope will be “flatter” at mid levels where speech undergoes compression, but more “fluctuating” at lower and higher levels where there is less compression of the envelope. These presumed level-dependent changes in the speech temporal envelope may affect recognition of vocoded speech and, in turn, the shape of the speech psychometric function. Although not addressed in the current study, the effective magnitude of envelope fluctuations may differ for hearing-impaired subjects due to reductions in gain and compression, which could change how speech recognition changes with increasing level for these subjects.

In this experiment, younger adults with normal hearing listened to consonant–vowel (CV) syllables and sentences at a range of speech levels in the presence of a broadband noise at a fixed SNR. Speech was processed using noiseband vocoding to minimize spectral information while preserving speech envelope cues. The speech levels included low levels, at which the speech envelope is presumed to be in a more linear region. Additional speech levels were selected to represent mid levels, where the envelope is expected to be most compressed, and higher levels, where speech would be processed more linearly. CV syllables and sentences were included to determine the extent to which effects of compression on temporal envelope information differ for these stimuli. Level-dependent changes in recognition of noise-vocoded consonants and key words in sentences were assessed. For consonant recognition, analysis also included level-dependent changes in information

transmission of certain acoustic-phonetic features, including those that rely primarily on temporal envelope cues. Information transmitted for specific features is predicted to change in a level-dependent manner, to the extent that the feature is susceptible to changes in temporal-envelope information. That is, a feature that relies primarily on temporal-envelope cues would be more susceptible to changes in those cues, which in turn would result in more confusions between consonants differentiated by that feature (Hornsby et al. 2005).

Consistent with the goal of relating changes in speech recognition with increasing level to effects of cochlear nonlinearities, an independent estimate of the basilar-membrane input–output response was needed for each subject. Distortion-product otoacoustic emissions (DPOAEs) are assumed to reflect the physiological nonlinear cochlear process in humans with normal outer hair cell function (Dorn et al. 2001; Neely et al. 2009). Moreover, input–output functions derived from DPOAEs and psychophysical measures provide generally similar estimates of basilar-membrane response growth, given that both are characterized by a nearly linear segment for low levels and a compressive segment for moderate levels (e.g., Dorn et al. 2001; Williams and Bacon 2005). Significant but modest correlations have been found between slopes of DPOAE input–output functions and compression estimates at higher frequencies (Williams and Bacon 2005; Johannesen and Lopez-Poveda 2008; Lopez-Poveda and Johannesen 2009), although the limited data available are characterized by large variances (Poling et al. 2011; Rodriguez et al. 2011). In this experiment, DPOAEs and DPOAE input–output functions were measured to provide estimates of basilar-membrane nonlinearities for each subject, to compare and correlate with level-dependent changes observed in recognition of vocoded speech.

METHODS

Subjects

Fifteen younger adults (mean age, 22.3 years; range, 19–28 years) with normal hearing participated in the study. Normal hearing was defined as audiometric thresholds in the right ear ≤ 10 dB HL (ANSI 2004) at octave frequencies from 0.25 to 6.0 kHz and normal immittance measures. Pure-tone thresholds were measured in quiet with a Madsen Orbiter 922 audiometer and TDH-39 headphones using a clinical adaptive psychophysical procedure (ASHA 2005). Subjects had no prior experience with the speech materials used in the experiment, vocoded speech, or the listening tasks. During practice sessions, extensive feedback was provided to subjects until they under-

stood the task and provided consistent results (more details to follow). All subjects were paid for their participation, which required approximately four to five 1.5–2.0-h sessions to complete.

Stimuli and apparatus

Speech signals. Recognition of consonants in CV syllables and key words in sentences was measured as a function of speech level (measured in overall root mean square (RMS) level). Consonant recognition was measured for 19 CV syllables formed by combining the consonants /b,t,d,f,g,k,l,m,n,p,r,s,ʃ,t,θ,v,w,j,z/ with the vowel /α/ spoken by a male talker without a carrier phrase. More information about the CV syllables is provided in Dubno and Schaefer (1992) and Dubno et al. (2003). To obtain a consonant-recognition score, each of the 19 CV syllables was repeated three times in random order for a total of 57 items. The rationale for using CV syllables with the vowel /α/ spoken by a male talker was two-fold. First, pilot testing demonstrated that recognition scores for these vocoded syllables avoided floor and ceiling and provided an appropriate range of scores across speech levels. Second, acoustic analyses of the 19 syllables showed those to be the most similar in overall RMS and peak levels, as compared to sets of vowel–consonant syllables, consonants combined with other vowels, or syllables spoken by a female talker. Given the focus of the current experiment on level-dependent changes in consonant recognition, it was important to use a set of syllables with the least variability across stimulus levels.

Recognition of key words in sentences (the final, monosyllabic word) was measured using 200 high-context sentences from the Speech Perception in Noise test (SPIN; Kalikow et al. 1977). This experiment required five lists (one for each speech level) with different sentences because it was not possible to repeat test items across the five conditions. To create five equivalent lists of 40 sentences each, recognition of key words in all 200 high-context sentences was measured in a pilot study with eight younger adults with normal hearing (none of these subjects participated in the main study). Noise-vocoded sentences were presented at 60 dB SPL (one of the speech levels used in the main study) in a spectrally shaped background noise at a +6-dB SNR; subjects repeated the final word of each sentence. Using the mean score for each sentence, the 200 sentences were distributed to form five lists of equivalent intelligibility [$F(4,156)=0.29$, $p=0.885$].

Digital noise-vocoded speech waveforms for both CV syllables and sentences were output at a sampling rate of 48.8 kHz using a Tucker-Davis Technologies

(TDT) RX6 array processor. The 48.8-kHz sampling rate was selected to be compatible with custom software and TDT equipment requirements (the original sampling rates were 33.0 kHz for nonsense syllables and 22.5 kHz for sentences). Spectral analyses were conducted on all speech signals after up-sampling to verify that the spectrum and other acoustic characteristics were not changed by this process. Amplitudes of the speech signals and background noise were controlled individually using programmable attenuators (TDT PA4). The signal was added to the noise (TDT SM3), passed through a headphone buffer (TDT HB6), and delivered to the right ear of a pair of TDH-39 headphones mounted in supra-aural cushions.

Background noise. To minimize the influence of individual differences in quiet thresholds and to maintain constant speech audibility as signal level increased, a broadband noise was present while measuring recognition of consonants and key words in sentences. The broadband noise was digitally generated and its spectrum adjusted at one-third-octave intervals (Adobe Audition, Version 1.5) to produce equivalent audibility for all subjects across all speech levels. Band levels of the background noise were set to achieve masked thresholds of 10 dB HL from 0.2 to 6.0 kHz. Given that pure-tone thresholds for all subjects were ≤ 10 dB HL, it was not necessary to customize the background noise for each subject. Figure 1 shows the spectra of vocoded CV syllables, sentences, and background noise for the five speech levels. Also included in the figure are mean quiet thresholds (to be discussed later). This figure demonstrates that, for the bandwidth of audible speech (approximately 400–5,000 Hz), speech audibility remained constant as speech level increased from 45 to 85 dB SPL. The background noise was not vocoded and was recorded onto a CD for playback. Spectral characteristics of all signals were verified using an acoustic coupler and a signal analyzer (Stanford Research SR780).

Signal processing. Speech stimuli were processed off-line with noiseband vocoding using a modified version of MATLAB software created by Qian-Jie Fu, retaining its original vocoding parameters. The overall frequency range was limited to 0.18–5.62 kHz. The unprocessed speech was first pre-emphasized using a first-order Butterworth high-pass filter with a cutoff frequency of 1.2 kHz and then band-pass filtered into four frequency bands using eighth-order Butterworth filters. Four bands were selected to effectively limit available spectral cues while still allowing for measurable speech scores. The bandwidths and cutoff frequencies for each band (0.18–0.523, 0.523–1.224, 1.224–2.666, and

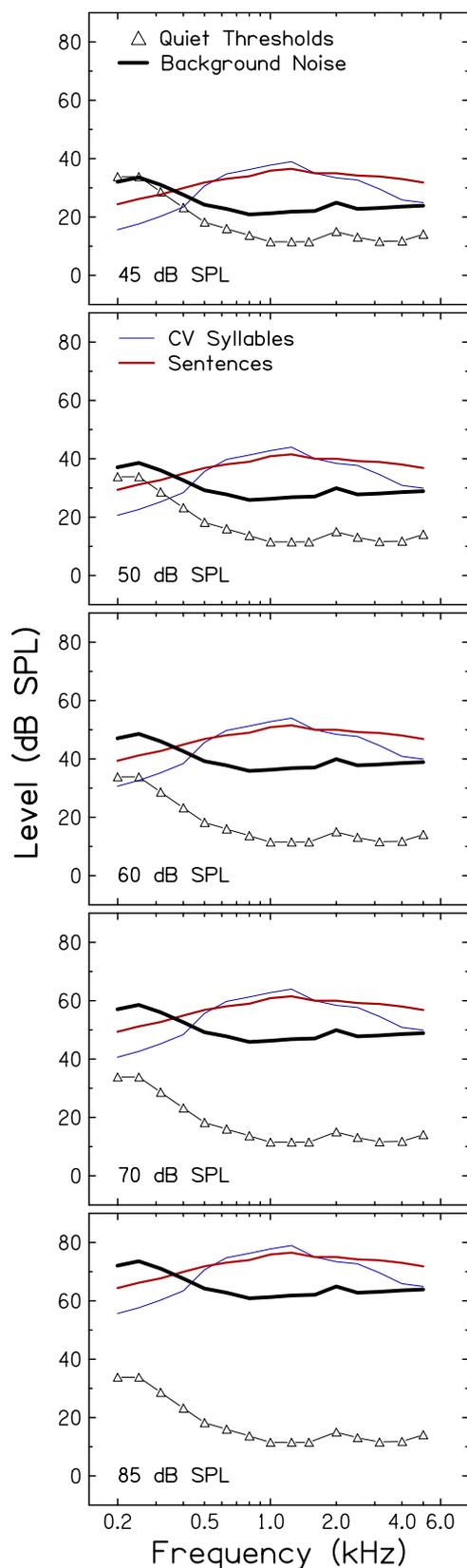


FIG. 1. One-third-octave band spectra of vocoded CV syllables (blue lines), sentences (red lines), and background noise (thick black lines) for five overall speech levels (45, 50, 60, 70, and 85 dB SPL). Mean quiet thresholds are also shown in each panel (triangles).

2.666–5.620 kHz) were based on equal basilar-membrane distance using Greenwood's formula (Greenwood 1990). The envelope in each band was then extracted by half-wave rectification and low-pass filtering (fourth-order Butterworth filter) with a 0.16-kHz cutoff frequency. The extracted envelope from each band was used to modulate a broadband noise, which was then band-pass filtered using the same analysis bands as above. The output from all bands was summed and equated in total-RMS level to the original token. After the vocoding process, the peak amplitude of each vocoded stimulus was normalized to -1 dB using Adobe Audition to minimize amplitude differences across tokens while maintaining the original envelope relationships within tokens. The total peak-to-RMS amplitude ratio for the 200 key words (23.0 dB) was somewhat larger than that for the CV syllables (17.6 dB).

Procedures

General procedures. Recognition of consonants and key words in sentences was measured at five speech levels (45, 50, 60, 70, and 85 dB SPL) in a spectrally shaped background noise. Consonant recognition was measured first for all subjects. The background noise was presented at a fixed SNR to maintain constant speech audibility as speech level increased. Due to the limited spectral information preserved in the vocoded speech, a +6-dB SNR provided a reasonable range of recognition scores (avoiding floor and ceiling) for both consonants and key words in sentences, as confirmed with pilot testing. A +6-dB SNR also provided a sufficiently large peak-to-trough range (with the trough limited by the background noise) to reveal potential compression effects on speech envelope fluctuations. To select speech levels, acoustic analyses determined that the lowest noise level must be ≥ 39 dB SPL to assure that the background noise limited audibility across the speech bandwidth (see Fig. 1). Based on this requirement and the +6-dB SNR, 45 dB SPL was selected as the lowest speech level. As described earlier, additional speech levels (50, 60, 70, and 85 dB SPL) were selected to span the range from mid (conversational) levels to higher levels.

Specific procedures. Consonant recognition. Orthographic representations of the 19 consonants were displayed on a flat-panel touch-screen monitor located in the audiometric test booth. After each CV syllable presentation, subjects responded by touching the appropriate consonant symbol on the screen. For each presentation, syllables were selected randomly (with replacement) by custom MATLAB software. Prior to data collection, subjects

were familiarized with the unprocessed syllables at 60 dB SPL in quiet until consonant recognition scores $\geq 90\%$ were achieved. During this initial familiarization of the procedure and the on-screen consonant symbols, the experimenter provided verbal feedback when needed. For data collection, the set of 57 vocoded CV syllables in background noise was presented at each of the five speech levels; this procedure was repeated ten times for a total of 50 sets per subject (5 speech levels \times 10 repetitions). During these sessions, no feedback was provided. One entire set of the five speech levels was presented before a set was repeated. For the ten repetitions, each subject received a different order of speech levels based on a Latin Square design.

Key word recognition in high-context sentences. Subjects were instructed to listen to the entire sentence and repeat aloud the final monosyllabic word. Prior to data collection, 25 high-context sentences from the SPIN practice list, vocoded using the same processing, were presented twice at 45, 60, and 85 dB SPL, and once at 50 and 70 dB SPL (total of eight repetitions of the 25-word list). No feedback was provided. For data collection, the five equivalent lists of 40 sentences were presented, one list at each level. List and level order were balanced for each subject using a Latin Square design to minimize confounding effects of these factors.

Distortion-product otoacoustic emissions. DPOAEs were recorded from right ears of each of the 15 subjects using an Otodynamics Echoport ILO292 USB-II (V6.0); the “checkfit” mode ensured that the level and stimulus waveform in the ear canal were appropriate prior to recording a response. Because of differences among subjects’ ear canal shapes and sizes, the equipment occasionally adjusted signal levels at each frequency, which introduced a source of variability. Before DPOAEs were recorded, normal middle ear status was confirmed by otoscopic examination and tympanometry using a 226-Hz probe tone. DPOAEs (defined as the amplitude of the $2f_1-f_2$ response) and the level of the noise within the $2f_1-f_2$ response bin corresponding to +2 standard deviations above the mean noise level were measured in half-octave bands between 1.0 and 6.0 kHz, with L_1 and L_2 nominally fixed at 65 and 55 dB SPL, respectively. For an additional estimate of nonlinearity strength, DPOAE “sum” was computed from these levels by adding the level of the $2f_1-f_2$ response in each band and converting to dB sound pressure level. DPOAE input–output functions were measured for f_2 fixed at 1.0, 2.0, and 4.0 kHz and $f_2/$

$f_1=1.22$. L_2 ranged from 25 to 75 dB SPL in 5-dB steps. For L_2 values of 65–75 dB SPL, $L_1=L_2$. For L_2 values <65 dB SPL, L_1 was determined by the formula, $L_1=0.4 \times L_2 + 39$ dB, which has been shown to produce generally large DPOAEs in adults with normal hearing (Kummer et al. 1998). At each L_2 , averaging was stopped manually when the level of the response and the SNR were stable.

For 10 of the 15 subjects, DPOAE levels and DPOAE input–output functions were measured twice to determine test–retest reliability. Overall, responses were highly repeatable. For example, mean DPOAE summed levels were 15.7 dB SPL for the first trial and 14.1 dB SPL for the second trial ($t(9)=1.21$; $p=0.257$). Mean slope of the DPOAE input–output function at 1.0 kHz was 0.279 dB/dB for the first trial and 0.208 dB/dB for the second trial ($t(5)=1.53$; $p=0.186$). All subjects provided consistent measures of DPOAE levels. However, DPOAE input–output functions at 1.0, 2.0, and 4.0 kHz could not be reliably measured for one, five, and three subjects, respectively (no response above the noise floor); DPOAE input–output functions for these subjects were not included in data analyses.

Data analysis

Outcome measures included the following: (1) recognition scores for noise-vocoded consonants and key words in high-context sentences at each of five speech levels; (2) percent information transmitted for acoustic-phonetic features in vocoded consonants (voicing, place of articulation, and manners of articulation) at each of five speech levels; (3) slopes of DPOAE input–output functions at three frequencies, and DPOAE overall and band levels measured in the ear canal; and (4) associations among speech recognition and DPOAE measures as assessed by linear regression. Changes in recognition scores (using rau-transformed scores to stabilize the variance across conditions; Studebaker 1985) and information transmitted due to changes in speech levels were assessed by repeated-measures analysis of variance (ANOVA). For all data analyses, effects were considered significant with p values <0.05 .

RESULTS AND DISCUSSION

Perceptual learning for noise-vocoded consonants and key words in sentences

Perceptual learning of vocoded speech has been reported in several previous studies (e.g., Davis et al. 2005; Sheldon et al. 2008; Fu and Shannon 1999). To assess these effects for consonant recognition, the set

of 57 vocoded CV syllables was repeated ten times at each of five levels (total of 50 sets per subject). This procedure took approximately 0.5 h for each repetition and provided the information needed to determine when (and if) consonant-recognition scores reached asymptote. Results of a repeated-measures ANOVA, with repetition number and speech level as repeated measures (main factors), showed a significant overall increase in consonant recognition from the first to the tenth repetition [$F(9,126)=23.33, p<0.0001$] and no significant interaction with level [$F(36,504)=0.89, p=0.656$]. Post-hoc tests (Tukey HSD) revealed that scores for the final four repetitions did not change significantly [$F(3,42)=1.92, p=0.140$]. Therefore, consonant-recognition scores submitted to analyses are means of the final four repetitions for each subject at each level, representing a total of 228 responses (57 consonants \times 4 repetitions).

For key word recognition in sentences, the decision was made not to repeat test items. Therefore, to provide some practice with vocoded sentences, subjects listened to eight repetitions of the 25 practice high-context SPIN sentences. This session lasted approximately 0.5 h. From the first to the eighth presentation, mean scores across the five levels for each subject improved and ultimately stabilized. For data collection, recall that five equivalent lists of 40 items each were created from the 200 high-context SPIN sentences (one list for each level, list and level order determined by Latin Square design). Results of two repeated-measures ANOVAs with list and level order as repeated measures, revealed no significant differences in scores among lists or for different level orders [list: $F(4,56)=0.35, p=0.838$; order: $F(4,56)=0.67, p=0.614$]. Thus, ~ 3.5 h of practice with vocoded CV syllables (3.0 h) and sentences (0.5 h) occurred before recognition scores reached asymptote.

Recognition of noise-vocoded consonants and key words in sentences

Recognition of noise-vocoded consonants (top) and key words in sentences (middle) as a function of speech level is shown in Figure 2 for individual subjects (thin lines); mean scores are shown by the thick lines. To simplify comparison of results for the two speech materials, Figure 2 (bottom) also displays mean scores (± 1 standard deviation) at each level. Means and results for individual subjects show similar patterns, with scores increasing from low to mid levels and then decreasing at higher levels. A repeated-measures ANOVA was conducted with speech material and speech level as repeated measures. Analyses showed a significant main effect of speech level [$F(4,56)=34.3; p<0.001$]. Post-hoc

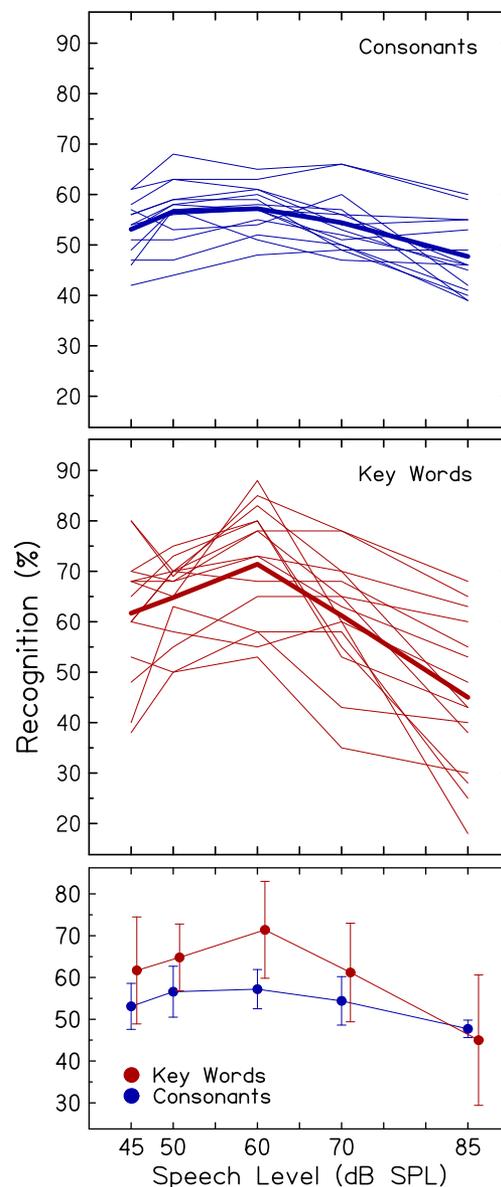


FIG. 2. Mean (thick lines) and individual (thin lines) recognition scores plotted as a function of speech level, for consonants (blue, top) and key words in sentences (red, middle). Mean scores for the two speech materials are also displayed in the bottom panel. For clarity, some data points are offset along the abscissa. Error bars indicate ± 1 standard deviation.

analyses revealed a significant quadratic (second-order) pattern [$F(1,14)=76.8; p<0.001$], which is represented by the inverted U-shaped curves for consonants and key words in Figure 2. That is, recognition improved with level increasing from 45 dB SPL, peaked around 60 dB SPL, and decreased thereafter. Additional post-hoc testing showed that mean consonant and key word recognition at 60 dB SPL was significantly higher than at lower and higher levels [consonants: $F(1,14)=69.6; p<0.001$; key words: $F(1,14)=86.8; p<0.001$]. Ranges of scores across subjects, across speech levels, and within a

speech level were substantially larger for key words in sentences than for consonants (e.g., across subjects and speech levels, ranges were 27 % for key words in sentences and 11 % for consonants). Nevertheless, scores for these two different speech materials showed a significant quadratic pattern, which demonstrates the generalizability of the findings.

These results are generally in agreement with those reported by Freyman and Nerbonne (1996) in which VCV syllables in a fixed background of signal-correlated noise (in a control condition without expansion) improved as speech level increased from 30 to 60 dB SPL. In contrast to results of the current experiment, scores remained the same at higher levels of 70 and 80 dB SPL. Large individual differences in the current study, especially for sentences, are consistent with results of Van Tasell and Trine (1996) for consonants and sentences undergoing signal-correlated noise processing to remove spectral information. In Başkent (2006), scores for normal-hearing subjects varied across stimuli, SNR, and number of spectral channels. Relevant for the current study, scores significantly increased as speech level increased from lower to comfortable levels, but generally remained the same when levels were further increased to 85 dB SPL. Although the effect of speech and noise level was addressed in these previous studies, background noise was not designed to equate audibility across subjects or maintain a constant SNR as speech level increased.

To assess recognition of vocoded speech for individual subjects, Pearson correlation coefficients and linear regressions for scores for vocoded key words and consonants were examined for each of the five speech levels and are shown in separate panels in Figure 3. Significant correlations between scores for vocoded key words and vocoded consonants were observed at 60 and 85 dB SPL ($p=0.07$ and $p=0.002$) with similar trends at 45, 50, and 70 dB SPL just missing statistical significance ($p=0.077$, 0.080, and 0.051, respectively). These results confirm that subjects with higher scores for key word recognition tended to have higher scores for consonant recognition across a range of speech levels.

Figure 3 also shows that scores were higher for vocoded key words than for vocoded consonants (mean differences for the two materials ranged from 6.8 % to 14.2 % for the first four levels), except at the highest level where consonant scores were slightly higher than key word scores (2.7 %). Results of the repeated-measures ANOVA showed significantly higher scores for key words in sentences than for consonants [$F(1,14)=13.49$, $p=0.002$] and a significant interaction of speech level and speech material [$F(4,56)=11.0$; $p<0.001$]. Post-hoc tests revealed significantly higher scores for key words than for consonants at 45, 50, 60, and 70 dB SPL ($p=0.009$, 0.001, <0.001 ,

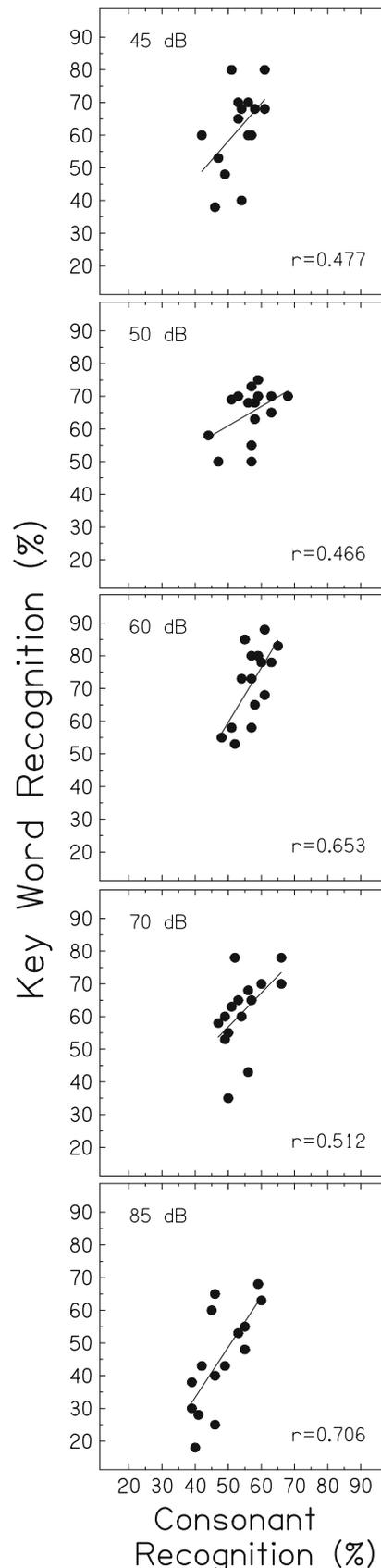


FIG. 3. Key word recognition scores plotted against consonant recognition scores, for five speech levels (top to bottom panels). Pearson correlation coefficients and linear regressions are included in each panel.

and 0.018), but no significant differences in scores at 85 dB SPL ($p=0.391$). These score differences are somewhat larger than those of Stacey and Summerfield (2008) for subjects listening to eight-channel noiseband-vocoded VCV syllables and BKB sentences at a conversational level. Prior to an auditory training program, mean scores were 40.7 % and 47.1 % for syllables and sentences, respectively. In the current study, given that the same SNR of +6 dB was used and no attempt was made to equate recognition for the two speech materials produced by different talkers, recognition scores may relate to larger inherent envelope fluctuations for sentences than for syllables. That is, the peak-to-RMS amplitude ratio for sentences was larger than that for CV syllables, which may provide more robust envelope cues, especially in the mid-level compressed region, and may result in larger changes in scores with increasing and decreasing speech level (as seen in Fig. 2).

Score differences for vocoded sentences and CV syllables may also relate to the differences in the number, variety, and availability of acoustic-phonetic and/or linguistic information. Recognition of the final word in an *unprocessed* high-context sentence typically benefits from many additional cues and redundancies, including acoustic-phonetic (features, phonemes, and syllables), lexical, semantic, and syntactical content in the other words in the sentence. However, noiseband vocoding with a small number of channels removes most of the semantic content in sentences (Ezzatian et al. 2011). Studies of perceptual learning of noise-vocoded words and sentences (Stacey and Summerfield 2008; Davis et al. 2005) demonstrate that acoustic-phonetic cues and lexical information, not semantic or syntactic information, underlie improved recognition of spectrally limited speech. Thus, it is unlikely that higher scores for recognition of key words in vocoded sentences than for vocoded consonants can be attributed exclusively to the use of syntactic or semantic content available in sentences, although the mix of lower-level and higher-level information available in spectrally limited speech is not yet clear.

Another factor that may contribute to score differences between vocoded sentences and consonants is their duration (hundreds of milliseconds for sentences, tens of milliseconds for consonants). That is, listeners have more opportunities with a longer speech signal to sample and detect information required for correct recognition, a process termed “multiple looks” (Viemeister and Wakefield 1991). The multiple-looks model has been used primarily to explain decreases in detection thresholds with increases in signal duration (temporal integration), but has also been applied to discrimination of amplitude modulated signals of different durations (Lee and Bacon 1997) and perception of multiple

presentations (“glimpses”) of speech stimuli (Holt and Carney 2005; Wang and Humes 2010). If listeners are primarily relying on envelope information in noise-vocoded speech, greater numbers of modulation cycles (more samples of the peaks and valleys in the speech waveform) are available in sentences than in consonants, which may contribute to improved recognition. More research using spectrally limited speech with a range of durations is needed to verify this assumption.

Finally, poorer performance for vocoded CV syllables than sentences could relate to differences in the spectrum of the speech and background noise; the noise was shaped to achieve equivalent thresholds of 10 dB HL across frequency for all subjects (see Fig. 1). Thus, the background noise was shaped to be parallel to the normal audibility curve, rather than following the more common practice of shaping the noise to be parallel to the speech spectrum. At a given speech level, speech audibility for syllables and sentences was generally similar in lower and mid-frequency regions, whereas speech audibility was less for syllables than sentences in higher frequency regions. Although fine-grained spectral information was limited as a result of four-channel vocoding, this difference in speech audibility could have reduced the availability of higher frequency speech cues important for consonant recognition and resulted in poorer scores. Nevertheless, given that audibility across the audible speech bandwidth remained constant as speech level increased as a result of a fixed SNR, differences in speech audibility in certain regions would not explain level-dependent changes in speech scores for consonants and words, which is the primary focus of this experiment.

Figure 3 shows that individual differences in scores at a particular level were much larger for key words in sentences than for consonants, which are also seen in Figure 2. Friesen et al. (2001) found a similar pattern in normal-hearing subjects listening to processed consonants and HINT sentences. Standard deviations, averaged across noise conditions and number of channels, were larger for sentences (7.24 %) than consonants (3.58 %). In the current experiment, larger individual differences for key words than for consonants may be attributed to differences in the nature of the two recognition tasks. That is, the consonant recognition task featured a closed-set response format and thousands of trials, whereas key word recognition used an open set and verbal response format and hundreds of trials. Larger individual differences at a particular level for vocoded key words than consonants may also reflect individual differences in the ability to benefit from limited linguistic information in vocoded sentences.

Also clearly shown in Figure 2 is that *changes* in scores (increases then decreases) with increasing speech level were much larger for key words (27.3 %) than for consonants (11.1 %). These effects, which are especially relevant to the focus of the current study on level-dependent changes, may be due to acoustic-phonetic factors and/or subject-related factors. First, larger *changes* in scores with increases in speech level may be a consequence of the larger speech peaks (more fluctuating temporal envelopes) for sentences than for syllables. If changes in scores with increasing speech level reflect the nonlinearity of the basilar-membrane response, level-dependent differences in the amount of compression may have larger effects on signals with more prominent peaks, which could result in greater level-dependent changes in recognition of vocoded sentences than vocoded consonants. Second, larger individual differences for key words in sentences than for consonants could relate to individual differences in the basilar-membrane response (even among these otherwise homogeneous subjects with very similar pure-tone thresholds) and in subsequent effects on level-dependent changes in speech recognition. These associations will be discussed in more detail in a later section.

Acoustic-phonetic features in noise-vocoded consonants

Analysis of acoustic-phonetic features in vocoded consonants and changes with increasing speech level was conducted to provide additional information on the potential contribution of cochlear nonlinearities to level-dependent changes in consonant recognition. To further investigate the effect of speech level on perception of envelope cues, consonant confusion matrices were analyzed, focusing on the features of voicing, place of articulation, and manner of articulation. The latter category was further subdivided to assess frication, nasality, and plosive manners. Correct and incorrect responses to each CV syllable were pooled across subjects to create a single consonant confusion matrix for each speech level, which was then submitted to the Feature Information Transfer (FIX) program¹ to compute the proportion of transmitted information (Miller and Nicely 1955; Wang and Bilger 1973). Speech envelope cues primarily convey voicing and manner features, whereas perception of place of articulation primarily relies on spectral cues (e.g., Van Tasell et al. 1987; Shannon et al. 1995; Xu et al. 2005; Başkent 2006). Few data are available describing level-dependent changes in fea-

ture transmission, especially when SNR and speech audibility are held constant. Hornsby et al. (2005) examined confusion matrices for 22 unprocessed CV syllables presented in speech-shaped noise at overall levels from 65 to 100 dB SPL. They found that higher signal levels had the greatest negative effect on the features of duration and place of articulation.

Figure 4 presents information transmitted as a function of speech level for three acoustic-phonetic features in the top panel (voicing, place of articulation, and manner of articulation); information transmitted for three manner subcategories (frication, nasality, and plosive) are in the bottom panel. In the top panel, differences in the overall proportion of information transmitted for each feature are clear, with the largest proportion for voicing and the smallest proportion for place, as would be expected for these signals with limited fine-grained spectral information.

Of particular interest for the current study is that *changes* in information transmitted with increasing speech level for most, but not all, of these acoustic-

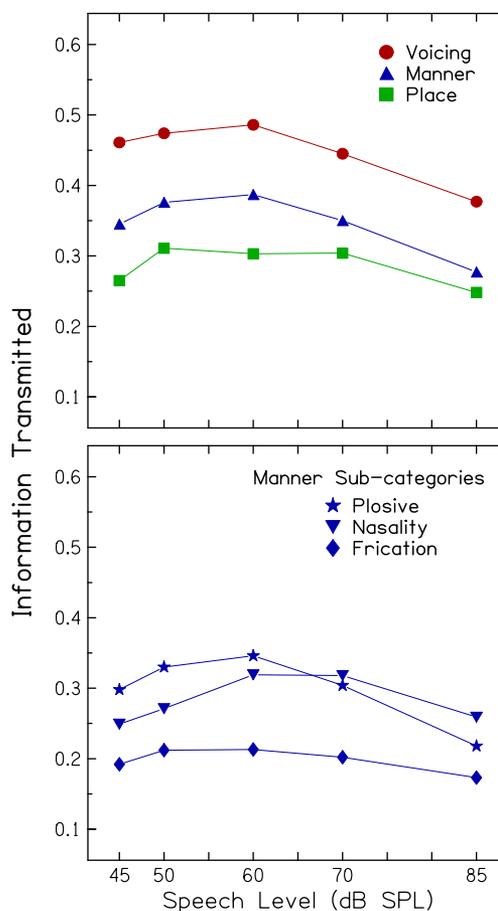


FIG. 4. *Top:* mean information transmitted plotted as a function of speech level for the three acoustic-phonetic features of voicing, manner of articulation, and place of articulation. *Bottom:* same as *top panel*, but for three sub-categories of manners of articulation (plosive, nasality, and frication).

¹ The Feature Information Transfer (FIX) program was obtained from the University College London website: <http://www.phon.ucl.ac.uk/resource/software.html>.

phonetic features are generally similar to changes in overall consonant recognition. A repeated-measures ANOVA with voicing, place, and three manner categories showed significant main effects of speech level [$F(4,56)=9.5$; $p<0.001$] and feature [$F(4,56)=40.8$; $p<0.001$] and a significant interaction between level and feature [$F(16,224)=2.7$; $p<0.001$]. Similar to results for overall consonant recognition, post-hoc tests revealed significant quadratic patterns for all features, except for place of articulation. That is, information transmitted for the features of voicing, frication, nasality, and plosive increased significantly when speech level increased from 45 dB SPL, reached a maximum at 60 dB SPL, and decreased at higher levels. In contrast, information transmitted for the place feature did not vary significantly in this pattern with increasing speech level from 45 to 85 dB SPL. These findings support and extend the results for level-dependent changes observed for overall consonant recognition. They further suggest that the nonlinear pattern in consonant recognition is consistent with the assumption of changes in “effective” temporal envelope cues with increasing level, cues that are important for the transmission of voicing and certain manner features.

The *change* of information transmitted as speech level increased also differed across acoustic-phonetic feature. Of the features with significant quadratic patterns, the plosive feature had the largest change across speech level (12.8 %), which may result from the abrupt change in amplitude typical for this feature. As speech level increased above a compressive mid-level into a more linear region, envelope peaks may increase, resulting in more rapidly declining feature transmission. In contrast, the frication feature had the smallest change with speech level (4.0 %). This may be due, in part, to the influence of the steady-state background noise, which is most similar to the frication feature, a finding also reported by Freyman et al. (1991).

Are basilar-membrane nonlinearity effects revealed in recognition of vocoded speech?

Level-dependent changes in recognition of consonants and key words in sentences. A goal of the current experiment was to assess the contribution of basilar-membrane nonlinearities to recognition of speech, under conditions where recognition was determined primarily by temporal envelope cues and level-dependent changes in recognition were not confounded by changes in speech audibility. It was assumed that, as a result of compression in the basilar-membrane response, “effective” speech envelope fluctuations would be reduced as speech level increased from lower (more linear) to mid-level

(more compressive) regions. With further increases to a more linear region, speech envelope fluctuations would become more pronounced. If listeners store templates of speech stimuli in long-term memory (based on a lifetime of experience listening to conversational speech) for comparison to internal representations of ongoing speech (e.g., Moore 2003), cochlear nonlinearities may provide optimal effective envelope cues for perception of speech at these conversational levels. Flatter speech envelopes may also be optimal for detection of silent periods in speech to the extent that they reduce the confusion between temporal gaps in speech (such as in plosive consonants and between words) and inherent fluctuations of the signal. Speech signals with less variance would also allow greater sensitivity to detecting changes in the temporal envelope. As a result of these effects, benefit of speech envelope cues may be maximized at mid levels where compression effects are expected to be strongest (e.g., Glasberg and Moore 2000) and reduced at lower and higher levels.

To test these working hypotheses, a second-order polynomial fit was applied to each subject’s five consonant recognition scores and another polynomial fit applied to each subject’s five key word recognition scores. Slopes at the lowest and highest speech levels (45 and 85 dB SPL) were then computed from these nonlinear functions. Steeper slopes (i.e., more change in recognition) at these levels were presumed to reflect more compression in the basilar-membrane response. That is, more compression would result in “flatter” effective speech envelopes in the mid-level compressed region and greater changes to more “fluctuating” envelopes at lower and higher speech levels, where less compression of the envelope is expected. Greater envelope changes would predict steeper declines in recognition at lower and higher levels for subjects with more compression (Poling et al. 2011; Horwitz et al. 2012).

Figure 5 displays scatterplots of slope values computed at 45 and 85 dB SPL, showing significant negative correlations for both consonants and key words ($p<0.001$ and $p=0.001$, respectively). That is, subjects with steeper increasing slopes at the lowest speech level (more improvement in recognition with increasing level, perhaps reflecting more compression) had steeper decreasing slopes at the highest speech level (greater declines with increasing level). Figure 6 displays scatterplots showing significant negative correlations between this same slope at 85 dB SPL and the range of recognition scores (highest score *minus* lowest score), for both consonants and key words (both $p<0.001$). That is, subjects with steeper slopes at 85 dB SPL (greater declines in recognition with increasing level) had larger *changes*

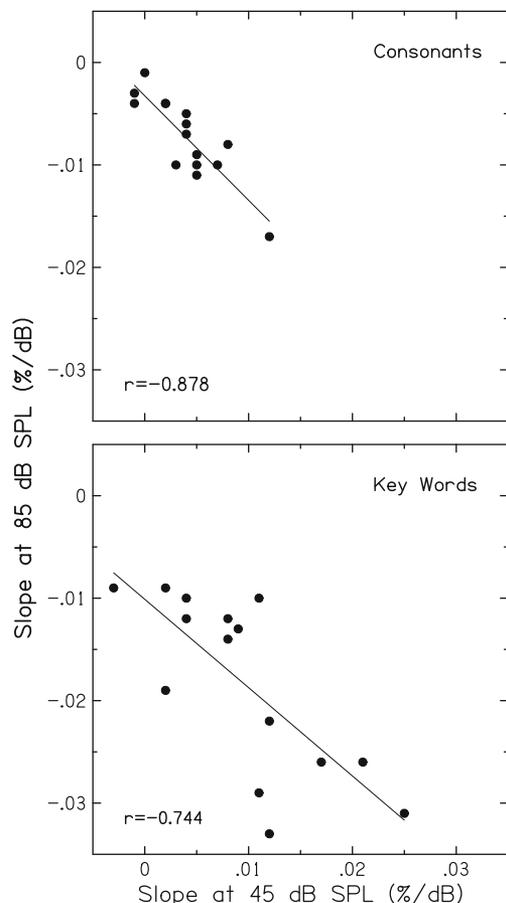


FIG. 5. Slope (percent per dB) at the highest speech level (85 dB SPL) plotted against slope at the lowest speech level (45 dB SPL), for recognition of consonants (*top*) and key words in sentences (*bottom*). Slopes for each speech material were computed from the polynomial fit applied to the score-level function for each subject. Pearson correlation coefficients and linear regression functions are included in each panel.

in recognition scores with increasing level. Thus, level-dependent *changes* in recognition of consonant and key words may also reflect the effects of basilar-membrane compression on speech envelopes.

Associations among distortion-product otoacoustic emissions and recognition of vocoded speech. To provide an independent estimate of basilar-membrane nonlinearities for each subject, DPOAE input-output functions were measured for f_2 frequencies of 1.0, 2.0, and 4.0 kHz. Figure 7 (*top*) shows individual DPOAE input-output functions for f_2 at 1.0 kHz. The responses generally have a linear segment at lower levels, a compressive region at mid-levels, and return to linear at higher levels. Slopes were estimated with a straight-line fit of the DPOAE levels for L_2 between 40 and 65 dB SPL (data plotted with solid lines in Fig. 7, *top*), similar to procedures used by Kummer et al. (1998) and Williams and Bacon (2005). Computed in this manner, mean slopes of input-output functions

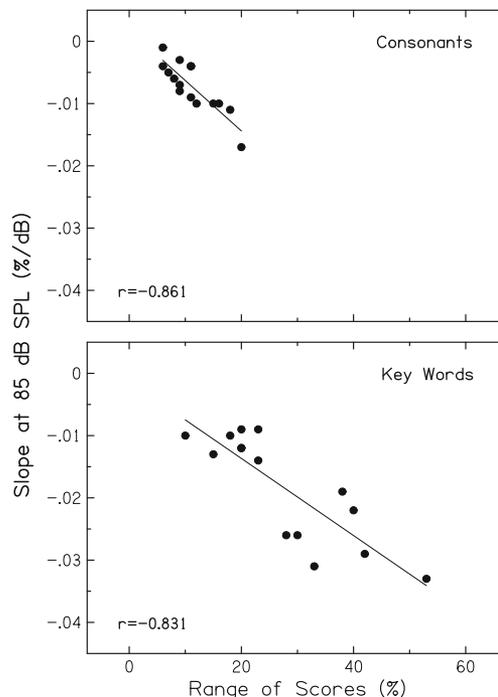


FIG. 6. Slope (percent per dB) calculated from speech scores at the highest level (85 dB SPL) plotted against the range of scores for recognition of consonants (*top*) and key words in sentences (*bottom*). Pearson correlation coefficients and linear regression functions are included in each panel.

were 0.19, 0.23, and 0.28 dB/dB at 1.0, 2.0, and 4.0 kHz, respectively. These slopes are comparable to those reported previously for younger adults with normal hearing (e.g., Dorn et al. 2001; Williams and Bacon 2005), where slopes of DPOAE input-output functions at similar frequencies ranged from 0.11 to 0.30 dB/dB. DPOAE levels were also measured in half-octave bands between 1.0 and 6.0 kHz. In the current study, mean DPOAE level was 17.1 dB SPL, which is comparable to the range of values (13–24 dB SPL) reported for younger adults with normal hearing (e.g., Kummer et al. 1998; Mills et al. 2007).

Of interest in the current study are individual differences in basilar-membrane nonlinearities and the extent to which these relate to individual differences in level-dependent changes in speech recognition, described earlier. In the current study with younger adults with similar average pure-tone thresholds (range=8.8 dB), individual differences were considerable. For example, DPOAE summed levels ranged from 3.2 to 24.7 dB SPL, and DPOAE input-output slopes at 1.0 kHz ranged from 0.01 to 0.50 dB/dB. To determine the consistency with which these values reflect the strength of cochlear nonlinearities in individual subjects with normal hearing, associations between DPOAE summed and band levels, and input-output slopes were analyzed for DPOAEs at 1.0,

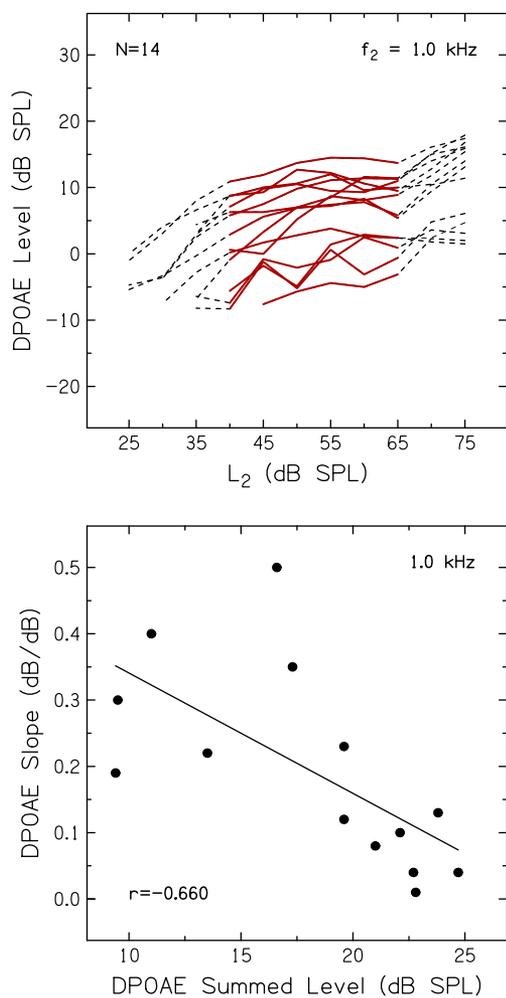


FIG. 7. *Top:* DPOAE levels plotted as a function of L_2 for f_2 of 1.0 kHz. DPOAE input–output function slopes were computed from DPOAE levels recorded for L_2 between 40 and 65 dB SPL (red lines). *Bottom:* slopes of DPOAE input–output functions for f_2 at 1.0 kHz (from the top panel) plotted against DPOAE summed levels. The Pearson correlation coefficient and linear regression function are also included.

2.0, and 4.0 kHz; results for DPOAE summed levels and the input–output function at 1.0 kHz are shown in Figure 7 (bottom). As expected, a significant negative correlation was found ($p=0.005$), supporting the assumption that subjects with larger DPOAE summed levels (indicating more robust basilar-membrane nonlinearities) had shallower DPOAE input–output slopes. Similar negative correlations were observed for DPOAE summed levels and input–output functions at 2.0 and 4.0 kHz.

To determine if the presumed effects of basilar-membrane nonlinearities are consistent with level-dependent changes in recognition of vocoded consonants and sentences, correlations were computed between DPOAEs (estimate of the strength of nonlinearities) and various measures reflecting changes in recognition scores with increasing speech level. As

described earlier, score changes with increases in speech level may be a marker of the effects of basilar-membrane compression on speech envelopes. That is, larger ranges of scores reflect larger changes in the “effective” temporal envelope due to the more linear response at lower and higher levels. Consistent with this assumption and shown in Figure 8, DPOAE input–output slopes at 2.0 kHz were negatively correlated with the range of scores for key words in sentences. Although the correlation just missed statistical significance for this small group of subjects ($N=9$; $p=0.066$), subjects with shallower DPOAE input–output slopes tended to have larger ranges of scores. As described earlier (Fig. 2), the highest scores for consonants and key words in sentences were achieved at 60 dB SPL, a level within the compressive region of the basilar-membrane response. When speech levels vary from this mid level in either direction, compression-related changes in the temporal envelope are predicted to be detrimental to speech recognition. We further hypothesized that subjects with more robust nonlinearities would demonstrate larger level-dependent changes. Thus, the negative correlation between DPOAE input–output slopes and score ranges is consistent with the assumption that the change in speech recognition with increasing level reflects the effects of basilar-membrane nonlinearities.

Finally, associations were examined between DPOAE measures and the change in recognition scores as speech level increased from 60 dB SPL (compressed envelopes) to 70 dB SPL (more fluctuating envelopes); for most subjects, scores declined as speech level increased (see Fig. 2). As shown in Figure 9, positive correlations were observed between the strength of cochlear nonlinearities (DPOAE slope) and the magnitude of this score change, for consonants ($r=0.483$, $p=0.081$) and key words in sentences ($r=0.628$, $p=0.050$). Subjects with shallower

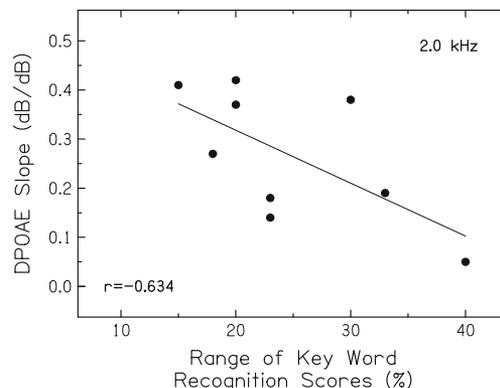


FIG. 8. Slopes of DPOAE input–output functions for f_2 of 2.0 kHz plotted against the range of key word recognition scores. The Pearson correlation coefficient and linear regression function are also included.

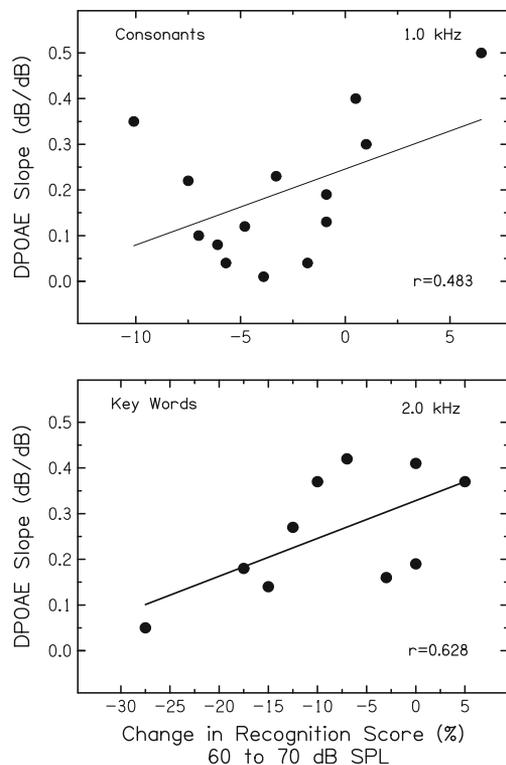


FIG. 9. *Top:* Slopes of DPOAE input–output functions for an f_2 of 1.0 kHz plotted against the change in consonant recognition scores with speech level increasing from 60 to 70 dB SPL. *Bottom:* slopes of DPOAE input–output functions for an f_2 of 2.0 kHz plotted against the change in key word recognition scores with speech level increasing from 60 to 70 dB SPL. Pearson correlation coefficients and linear regression functions are included in each panel.

DPOAE slopes (more compression) tended to have larger declines in scores as speech level increased from 60 to 70 dB SPL. Consistent with these results, a significant positive correlation ($r=0.753$, $p=0.016$, not shown) was found between DPOAE input–output slopes and consonant recognition scores at 70 dB SPL; subjects with shallower slopes (more compression) had poorer scores. A significant negative correlation ($r=-0.542$, $p=0.035$, not shown) was also found between DPOAE summed levels and consonant recognition scores at 70 dB SPL; subjects with larger DPOAEs had poorer scores. Taken together (along with Fig. 8), these results are consistent with the notion that, for individuals with more robust basilar-membrane nonlinearities, larger declines in recognition occur when speech transitions from compressed to more fluctuating.

Although the associations reported here are generally in line with hypothesized effects, they should be viewed with caution. Comparisons of estimates of basilar-membrane compression from behavioral and DPOAE measures are necessarily qualitative because of possible influences of the interference between multiple generating sources of the DPOAE, which is

not optimal for assessment of compression. This and other factors unrelated to compression may affect the shape of the DPOAE input–output function, contribute to variability in the results, and impact the strength of correlations. Moreover, the activation of the medial olivocochlear (MOC) reflex during measures of speech recognition in noise may decrease cochlear gain under certain conditions and complicate interpretation of behavioral and objective measures of compression (e.g., Jennings et al. 2011; Wojtczak and Oxenham 2010). Additional research with larger subject groups having wider ranges of thresholds is necessary to confirm these initial findings, using procedures that optimize measurement of DPOAE input–output functions and minimize MOC reflex activity during measures of speech recognition in noise.

GENERAL DISCUSSION

This experiment was designed to measure recognition of consonants and key words in sentences that were processed to preserve envelope information while minimizing spectral cues. Consonant and key word recognition showed significant changes with increasing level, revealing a quadratic pattern (see Fig. 2). Specifically, recognition improved from 45 dB SPL, peaked at ~ 60 dB SPL, and then declined at higher levels. Information transmitted for voicing and manner, but not for place of articulation, revealed the same level-dependent quadratic pattern. Given that SNR and speech audibility were constant and subjects were relying primarily on temporal envelope information, poorer speech recognition at lower and higher levels relative to mid levels may reflect corresponding changes in speech envelope fluctuations due to basilar-membrane nonlinearities. The quadratic pattern is consistent with the findings of Fu and Shannon (1998), where external compression and expansion was used to vary the temporal envelope of four-channel vocoded consonants and vowels presented at a fixed speech level, thus avoiding potential confounds from level changes. As would be predicted from a compression-related hypothesis, highest scores for normal-hearing subjects were achieved with no amplitude alteration, whereas compression and expansion of temporal envelopes reduced performance.

Consistent with larger fluctuations in temporal envelopes in sentences than CV syllables, larger changes in scores with changes in level were observed for key words than for consonants. Although subjects showed generally similar quadratic patterns, large individual differences in scores and level-dependent changes in scores were observed. Correlations between speech recognition measures and objective

measures of cochlear nonlinearities using DPOAEs were consistent with compression-related effects.

Although level-dependent changes in speech envelopes related to basilar-membrane nonlinearities provide a reasonable explanation for the quadratic pattern in the recognition scores, alternative explanations should be considered. First, improvements in speech recognition with increases in lower level speech are often attributed to sensation-level effects. That is, speech recognition improves with increases in speech audibility (increased speech sensation levels). However, in the current study, speech audibility and the audible speech bandwidth remained constant as signal level increased from 45 dB SPL due to the use of background noise at a fixed SNR (see Fig. 1). Acoustical analyses assured that, for all subjects and at all speech levels, background noise and not detection thresholds limited audibility across the audible speech bandwidth. Under these conditions, speech recognition should remain constant with increases in signal levels (e.g., Dubno et al. 2005a, b). Thus, improvements in speech recognition from lower to mid levels in the current experiment cannot be attributed to increased speech audibility.

Second, declines in recognition of unprocessed speech at higher than conversational levels for subjects with normal and impaired hearing have been reported in several studies (e.g., Studebaker et al. 1999). Although a clear understanding of the factors responsible for these declines has yet to be established, explanations generally relate to presumed changes in frequency selectivity at higher signal levels. The consequences of these changes at higher levels include wider auditory filters and increased upward spread of masking, resulting in higher masked thresholds and reduced speech audibility (Dubno et al. 2005a, b). Given that the perceptual consequences of changes in frequency selectivity at higher levels relate primarily to the use of spectral and not temporal information, it is unlikely that these changes can explain declines in recognition of speech that has been processed to preserve envelope information while minimizing spectral cues. For example, Hornsby et al. (2005) analyzed changes in information transmitted for unprocessed speech presented at 65–100 dB SPL, focusing on distortion of spectral cues. In contrast to results of feature analyses in the current study, the voicing feature was most resistant, and the place feature was among the most susceptible to the effects of increasing speech level.

Third, level-dependent changes in certain spectral cues likely contribute to the shape of the speech psychometric function for unprocessed (non-vocoded) speech. For example, spectral discrimination has been shown to be a nonmonotonic function

of level (e.g., Alves-Pinto and Lopez-Poveda 2005), such that discrimination is poorer at 70 dB SPL and better at lower and higher levels. By removing this spectral information in vocoded speech, confounding spectral cues at mid levels may no longer be available, which could result in better performance.

Fourth, it is possible that some processes related to the encoding of speech in the auditory nerve could, independent of basilar-membrane compression, underlie the level-dependent quadratic pattern of scores. For example, declines in speech recognition at high levels (“rollover”) have been attributed in the clinical literature to changes in neural synchrony or phase locking associated with auditory nerve or brainstem pathology (e.g., Dirks et al. 1977; Bess et al. 1979), although these effects would not be relevant for subjects in the current study who are younger adults with normal hearing. As another example, changes with signal level in auditory nerve discharge rates for different spontaneous rate fiber groups have been related to increases and decreases in discriminability of intensity increments (e.g., Kawase et al. 1993). That is, for lower-level signals nearer threshold, discriminability would be relatively poor and then improve to some peak value with increasing signal level and fiber discharge rate (at the steepest portion of the rate-level function). As signal levels increase further, discharge rates would approach saturation and discriminability would decline again, thus mapping out a quadratic pattern (best performance at mid levels and poorer performance at lower and higher levels). These predicted level-dependent changes in discriminability of intensity increments (Chintanpalli et al. 2012) have not been applied to the encoding of speech in the auditory nerve at a range of levels (e.g., Young and Sachs 1979; Sachs and Young 1979, 1980; Delgutte 1980, 1995). Therefore, further studies are needed to determine if the shape of auditory nerve rate-level functions and a resultant optimal range or “sweet spot” of neural encoding may relate to level-dependent improvements and declines in perception of envelope cues in speech that are unrelated to cochlear nonlinearities.

Thus, we are unaware of a parsimonious explanation that accounts for both increasing and decreasing overall scores and envelope-dependent acoustic-phonetic feature transmission with increases in speech levels, individual differences in level-dependent changes in scores, and differences in recognition for consonants and key words in high-context sentences. As a further test of these working hypotheses, additional experiments are underway whereby effects on speech recognition of external amplitude compression and expansion are examined for subjects with normal hearing and mild-to-moderate sensorineural hearing loss.

CONCLUSIONS

1. For younger adults with normal hearing, recognition scores for consonants and key words in sentences with limited spectral information increased significantly from lower to mid levels and then decreased significantly at higher levels. Changes in scores with increasing speech levels and individual differences in scores at a particular speech level were larger for key words than for consonants, despite subjects' very similar pure-tone thresholds. Poorer recognition of vocoded speech at lower and higher levels relative to mid (conversational) levels are consistent with presumed changes in speech envelope fluctuations due to cochlear compression.
2. Recognition scores were higher for noise-vocoded key words in sentences than for vocoded consonants, which may relate to larger inherent envelope fluctuations for sentences than consonants, the availability of acoustic-phonetic and some linguistic information in sentences, differences in durations of speech stimuli, and differences in audible speech in higher frequency regions.
3. Information transmitted for acoustic-phonetic features that convey temporal envelope cues (voicing and manner) revealed the same level-dependent quadratic pattern as overall consonant recognition scores. Information transmitted for place of articulation (a feature that relies primarily on spectral cues) did not vary significantly with speech level.
4. Level-dependent changes in speech recognition and their associations with the strength of basilar-membrane nonlinearities (estimated from DPOAE levels and input-output slopes) provide support for a role of compression in the basilar-membrane input-output response in processing of complex sounds such as speech.
5. Although the possibility remains that the quadratic pattern observed in recognition scores was not related to cochlear nonlinearities, we are unaware of an alternative explanation that accounts for both increasing and decreasing overall scores, information transmitted in certain consonant features, and patterns of individual differences.

ACKNOWLEDGMENTS

This work was supported (in part) by research grants R01 DC00184 and P50 DC00422 from NIH/NIDCD and by the South Carolina Clinical and Translational Research (SCTR) Institute, with an academic home at the Medical University

of South Carolina, NIH/NCRR Grant number UL1 RR029882. This investigation was conducted in a facility constructed with support from Research Facilities Improvement Program Grant Number C06 RR14516 from the National Center for Research Resources, National Institutes of Health. The authors thank Qian-Jie Fu and Emily Buss for providing MATLAB code for signal processing and stimulus presentation, and Magdalena Wojtczak, Skyler Jennings, and Reviewer Enrique Lopez-Poveda for helpful suggestions. Assistance with data collection and analysis from Fu-Shing Lee, Bryant Mersereau, Gayla Poling, and Jenny West is gratefully acknowledged.

REFERENCES

- ALVES-PINTO A, LOPEZ-POVEDA EA (2005) Detection of high-frequency spectral notches as a function of level. *J Acoust Soc Am* 118:2458–2469
- AMERICAN NATIONAL STANDARDS INSTITUTE (2004) Specification for audiometers. ANSI S3.6-2004, American National Standards Institute, New York
- AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION (2005) Guidelines for manual pure-tone threshold audiometry. American Speech-Language-Hearing Association, MD
- BASKENT D (2006) Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels. *J Acoust Soc Am* 120:2908–2925
- BESS FH, JOSEY AF, HUMES LE (1979) Performance-intensity functions in cochlear and eighth nerve disorders. *Am J Otol* 1:27–31
- CHINTANPALLI A, JENNINGS SG, HEINZ MG, STRICKLAND EA (2012) Modeling the anti-masking effects of the olivocochlear reflex in auditory nerve responses to tones in sustained noise. *J Assoc Res Otolaryngol* 13:219–235
- DAVIS MH, JOHNSRUDE IS, HERVAIS-ADELMAN A, TAYLOR K, MCGETTIGAN C (2005) Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen* 134:222–241
- DELGUTTE B (1980) Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. *J Acoust Soc Am* 68:843–857
- DELGUTTE B (1995) Physiological models for basic auditory percepts. In: Hawkins HL, McMullen TA, Popper AN, Fay RR (eds) *Auditory computation*. Springer, New York, pp 157–220
- DIRKS DD, KAMM C, BOWER D, BETSWORTH A (1977) Use of performance-intensity functions for diagnosis. *J Speech Hear Disord* 42:408–415
- DORN PA, KONRAD-MARTIN D, NEELY ST, KEEFE DH, CYR E, GORGA MP (2001) Distortion product otoacoustic emission input/output functions in normal-hearing and hearing-impaired human ears. *J Acoust Soc Am* 110:3119–3131
- DUBNO JR, SCHAEFFER AR (1992) Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners. *J Acoust Soc Am* 91:2110–2121
- DUBNO JR, HORWITZ AR, AHLSTROM JB (2003) Recovery from prior stimulation: masking of speech by interrupted noise for younger and older adults with normal hearing. *J Acoust Soc Am* 113:2084–2094
- DUBNO JR, HORWITZ AR, AHLSTROM JB (2005A) Word recognition in noise at higher-than-normal levels: decreases in scores and increases in masking. *J Acoust Soc Am* 118:914–922
- DUBNO JR, HORWITZ AR, AHLSTROM JB (2005B) Recognition of filtered words in noise at higher-than-normal levels: decreases in scores with and without increases in masking. *J Acoust Soc Am* 118:923–933

- DUBNO JR, HORWITZ AR, AHLSTROM JB (2007) Estimates of basilar-membrane nonlinearity effects on masking of tones and speech. *Ear Hear* 28:2–17
- EZZATIAN P, LI L, PICHORA-FULLER MK, SCHNEIDER BA (2011) The effect of priming on release from informational masking is equivalent for younger and older adults. *Ear Hear* 32:84–96
- FREYMAN RL, NERBONNE GP (1996) Consonant confusions in amplitude-expanded speech. *J Speech Hear Res* 39:1124–1137
- FREYMAN RL, NERBONNE GP, COTE HA (1991) Effect of consonant-vowel ratio modification on amplitude envelope cues for consonant recognition. *J Speech Hear Res* 34:415–426
- FRIESEN LM, SHANNON RV, BASKENT D, 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
- FU QJ, SHANNON RV (1998) Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners. *J Acoust Soc Am* 104:2570–2577
- FU QJ, SHANNON RV (1999) Recognition of spectrally degraded speech in noise with nonlinear amplitude mapping. *Proc IEEE Int Conf Acoust Speech Signal Process* 1999:369–372
- GALLUN F, SOUZA P (2008) Exploring the role of the modulation spectrum in phoneme recognition. *Ear Hear* 29:800–813
- GLASBERG BR, MOORE BCJ (1989) Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. *Scand Audiol Suppl* 32:1–25
- GLASBERG BR, MOORE BCJ (1992) Effects of envelope fluctuations on gap detection. *Hear Res* 64:81–92
- GLASBERG BR, MOORE BCJ (2000) Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *J Acoust Soc Am* 108:2318–2328
- GREENWOOD DD (1990) A cochlear frequency-position function for several species-29 years later. *J Acoust Soc Am* 87:2592–2605
- HALL JW, GROSE JH (1997) The relation between gap detection, loudness, and loudness growth in noise-masked normal-hearing listeners. *J Acoust Soc Am* 101:1044–1049
- HOLT RF, CARNEY AE (2005) Multiple looks in speech sound discrimination in adults. *J Speech Lang Hear Res* 48:922–943
- HORNSBY BW, TRINE TD, OHDE RN (2005) The effects of high presentation levels on consonant feature transmission. *J Acoust Soc Am* 118:1719–1729
- HORWITZ AR, AHLSTROM JB, DUBNO JR (2007) Speech recognition in noise: estimating effects of compressive nonlinearities in the basilar-membrane response. *Ear Hear* 28:682–693
- HORWITZ AR, AHLSTROM JB, DUBNO JR (2011) Level-dependent changes in detection of temporal gaps in noise markers by adults with normal and impaired hearing. *J Acoust Soc Am* 130:2928–2938
- HORWITZ AR, AHLSTROM JB, DUBNO JR (2012) Individual and level-dependent differences in masking for adults with normal and impaired hearing. *J Acoust Soc Am* 131:EL323–EL328
- JENNINGS SG, HEINZ MG, STRICKLAND EA (2011) Evaluating adaptation and olivocochlear efferent feedback as potential explanations of psychophysical overshoot. *J Assoc Res Otolaryngol* 12:345–360
- JOHANNESSEN PT, LOPEZ-POVEDA EA (2008) Cochlear nonlinearity in normal-hearing subjects as inferred psychophysically and from distortion-product otoacoustic emission input/output functions. *J Acoust Soc Am* 124:2149–2163
- KALKOW DN, STEVENS KN, ELLIOTT LL (1977) Development of a test of speech intelligibility in noise using test material with controlled word predictability. *J Acoust Soc Am* 61:1337–1351
- KAWASE T, DELGUTTE B, LIBERMAN MC (1993) Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones. *J Neurophysiol* 70:2533–2549
- KUMMER P, JANSSEN T, ARNOLD W (1998) The level and growth behavior of the $2f_1/2$ distortion product otoacoustic emission and its relationship to auditory sensitivity in normal hearing and cochlear hearing loss. *J Acoust Soc Am* 103:3431–3444
- LEE J, BACON SP (1997) Amplitude modulation depth discrimination of a sinusoidal carrier: effect of stimulus duration. *J Acoust Soc Am* 101:3688–3693
- LOPEZ-POVEDA EA, JOHANNESSEN PT (2009) Otoacoustic emission theories and behavioral estimates of human basilar membrane motion are mutually consistent. *J Assoc Res Otolaryngol* 10:511–523
- MILLER GA, NICELY PE (1955) An analysis of perceptual confusions among some English consonants. *J Acoust Soc Am* 27:338–352
- MILLS DM, FEENEY PM, GATES GA (2007) Evaluation of cochlear hearing disorders: normative distortion product otoacoustic emission measurements. *Ear Hear* 28:778–792
- MOORE BCJ (2003) Temporal integration and context effects in hearing. *J Phon* 31:563–574
- NEELY ST, JOHNSON TA, KOPUN J, DIERKING DM, GORGA MP (2009) Distortion-product otoacoustic emission input/output characteristics in normal-hearing and hearing-impaired human ears. *J Acoust Soc Am* 126:728–738
- OXENHAM AJ, BACON SP (2003) Cochlear compression: perceptual measures and implications for normal and impaired hearing. *Ear Hear* 24:352–366
- POLING GL, AHLSTROM JB, HORWITZ AR, DUBNO JR (2011) Individual differences in behavioral estimates of cochlear nonlinearities. *J Assoc Res Otolaryngol* 13:91–108
- RODRIGUEZ J, NEELY ST, JESTEADT W, TAN H, GORGA MP (2011) Comparison of distortion-product otoacoustic emission growth rates and slopes of forward-masked psychometric functions. *J Acoust Soc Am* 129:864–875
- ROSEN S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci* 336:367–373
- SACHS MB, YOUNG ED (1979) Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. *J Acoust Soc Am* 66:470–479
- SACHS MB, YOUNG ED (1980) Effects of nonlinearities of speech encoding in the auditory nerve. *J Acoust Soc Am* 68:858–875
- SHANNON RV, ZENG FG, KAMATH V, WYGONSKI J, EKELID M (1995) Speech recognition with primarily temporal cues. *Science* 270:303–304
- SHELDON S, PICHORA-FULLER MK, SCHNEIDER BA (2008) Effect of age, presentation method, and learning on identification of noise-vocoded words. *J Acoust Soc Am* 123:476–488
- STACEY PC, SUMMERFIELD AQ (2008) Comparison of word-, sentence-, and phoneme-based training strategies in improving the perception of spectrally distorted speech. *J Speech Lang Hear Res* 51:526–538
- STUDEBAKER GA (1985) A 'rationalized' arcsine transform. *J Speech Hear Res* 28:455–462
- STUDEBAKER GA, SHERBECOE RL, McDANIEL DM, GWALTNEY CA (1999) Monosyllabic word recognition at higher-than-normal speech and noise levels. *J Acoust Soc Am* 105:2431–2444
- VAN TASELL DJ, TRINE TD (1996) Effects of single-band syllabic amplitude compression on temporal speech information in nonsense syllables and in sentences. *J Speech Hear Res* 39:912–922
- VAN TASELL DJ, SOLI SD, KIRBY VM, WIDIN GP (1987) Speech waveform envelope cues for consonant recognition. *J Acoust Soc Am* 82:1152–1161
- VIEMEISTER NF, WAKEFIELD GH (1991) Temporal integration and multiple looks. *J Acoust Soc Am* 90:858–865
- WANG MD, BILGER RC (1973) Consonant confusions in noise: a study of perceptual feature. *J Acoust Soc Am* 54:1248–1266
- WANG X, HUMES L (2010) Factors influencing recognition of interrupted speech. *J Acoust Soc Am* 128:2100–2111

- WILLIAMS EJ, BACON SP (2005) Compression estimates using behavioral and otoacoustic emission measures. *Hear Res* 201:44–54
- WOJTCZAK M, OXENHAM AJ (2010) Recovery from on- and off-frequency forward masking in listeners with normal and impaired hearing. *J Acoust Soc Am* 128:247–256
- XU L, THOMPSON CS, PFINGST BE (2005) Relative contributions of spectral and temporal cues for phoneme recognition. *J Acoust Soc Am* 117:3255–3267
- YOUNG ED, SACHS MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of auditory-nerve fibers. *J Acoust Soc Am* 66:1381–1403