Consonant identification in noise using Hilbert-transform temporal fine-structure speech and recovered-envelope speech for listeners with normal and impaired hearing

Citation: Léger, Agnès C., Charlotte M. Reed, Joseph G. Desloge, Jayaganesh Swaminathan, and Louis D. Braida. "Consonant Identification in Noise Using Hilbert-Transform Temporal Fine-Structure Speech and Recovered-Envelope Speech for Listeners with Normal and Impaired Hearing." J. Acoust. Soc. Am. 138, no. 1 (July 2015): 389–403. © 2015 Acoustical Society of America

As Published: http://dx.doi.org/10.1121/1.4922949

Publisher: Acoustical Society of America (ASA)

Persistent URL: http://hdl.handle.net/1721.1/99898

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Consonant identification in noise using Hilbert-transform temporal fine-structure speech and recovered-envelope speech for listeners with normal and impaired hearing[a)]

Agnès C. Léger[b)]
School of Psychological Sciences, University of Manchester, Manchester, M13 9PL, United Kingdom

Charlotte M. Reed, Joseph G. Desloge, Jayaganesh Swaminathan,[c)] and Louis D. Braida
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 20 October 2014; revised 7 April 2015; accepted 11 June 2015; published online 20 July 2015)

Consonant-identification ability was examined in normal-hearing (NH) and hearing-impaired (HI) listeners in the presence of steady-state and 10-Hz square-wave interrupted speech-shaped noise. The Hilbert transform was used to process speech stimuli (16 consonants in a-C-a syllables) to present envelope cues, temporal fine-structure (TFS) cues, or envelope cues recovered from TFS speech. The performance of the HI listeners was inferior to that of the NH listeners both in terms of lower levels of performance in the baseline condition and in the need for higher signal-to-noise ratio to yield a given level of performance. For NH listeners, scores were higher in interrupted noise than in steady-state noise for all speech types (indicating substantial masking release). For HI listeners, masking release was typically observed for TFS and recovered-envelope speech but not for unprocessed and envelope speech. For both groups of listeners, TFS and recovered-envelope speech yielded similar levels of performance and consonant confusion patterns. The masking release observed for TFS and recovered-envelope speech may be related to level effects associated with the manner in which the TFS processing interacts with the interrupted noise signal, rather than to the contributions of TFS cues *per se*. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4922949]

[MSS] Pages: 389–403

I. INTRODUCTION

Many hearing-impaired (HI) listeners who have little difficulty understanding speech in quiet backgrounds experience great difficulty with speech in backgrounds containing interfering sounds. When the interference is temporally fluctuating, as in restaurants and large groups, most normal-hearing (NH) listeners achieve substantial gains in intelligibility relative to steady state interference [masking release (MR)], while many HI listeners do not (e.g., Festen and Plomp, 1990; Lorenzi *et al.*, 2006b). A number of possible factors have been proposed for the reduction or absence of MR in HI listeners. One explanation derives from the effects of reduced audibility in HI listeners. As shown by Desloge *et al.* (2010), it is as if the HI were NH listeners who experienced a background noise that raised their thresholds to those of the HI and consequently prevented them from "listening in the valleys" where the interference is relatively weak. The reduction in cochlear compression and decreased frequency selectivity that accompany sensorineural loss are another possible source of decreased MR (Moore *et al.*, 1999; Oxenham and Kreft, 2014). A reduced ability to process the temporal fine structure of speech (TFS) (the rapid fluctuations in amplitude close to the center frequency of a narrow-band signal) as well as NH individuals (Lorenzi *et al.*, 2006a; Lorenzi *et al.*, 2009; Hopkins and Moore, 2009; Hopkins *et al.*, 2008; Moore, 2014) has also been proposed as a factor in reduced MR. Support for this explanation lies in observed correlations between scores for understanding speech manipulated to degrade or convey TFS cues and the MR obtained with intact speech signals (Lorenzi *et al.*, 2006a; Lorenzi *et al.*, 2009; Hopkins and Moore, 2011; Hopkins *et al.*, 2008). More recently, it has been suggested that the lack of MR in HI listeners may be based on their having less susceptibility than NH listeners to the random amplitude fluctuations present in steady-state noise (see Stone *et al.*, 2012). According to this explanation, HI listeners perform similarly in fluctuating and steady-state background noise because they are unaffected by the modulation masking that occurs for NH listeners in a steady-state Gaussian noise due to the increased auditory bandwidths associated with cochlear hearing loss (Oxenham and Kreft, 2014).

The current study explored the TFS-based explanation for the reduced performance of HI compared to NH listeners in continuous and fluctuating background noises using the Hilbert transform to generate speech that conveyed TFS cues. Although HI listeners are generally able to make good use of the slowly varying envelope (ENV) cues of speech, they often experience a reduced ability with TFS cues

[a)]Portions of this research were presented at the 167th Meeting of the Acoustical Society of America, Providence, RI, May, 2014.
[b)]Electronic mail: agnes.leger@manchester.ac.uk
[c)]Current address: Hearing Research Center, Boston University, Boston, MA, USA.

compared to NH listeners (Lorenzi *et al*., 2006a; Hopkins and Moore, 2011). The current study sought to measure directly MR for intact speech as well as speech that was constructed to convey ENV cues or TFS cues.

The speech waveform is often characterized as the sum of slowly varying amplitudes, each of which modulates a rapidly varying carrier. In this view, each frequency channel is described by an ENV, corresponding to the slowly varying amplitude, and a TFS, corresponding to the carrier. Findings from several studies suggest that there is a dichotomy between ENV and TFS cues; however, this remains a topic of considerable interest and debate (Smith *et al*., 2002; Zeng *et al*., 2004; Gilbert and Lorenzi, 2006; Sheft *et al*., 2008; Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012; Shamma and Lorenzi, 2013; Moore, 2014). ENV cues have been shown to support robust speech identification in quiet (Smith *et al*., 2002) but they are insufficient when background sounds are present, particularly when the background is fluctuating. In a fluctuating background, speech reception of NH listeners is reduced when TFS cues are eliminated by the use of vocoders that reduce speech to slowly varying fluctuations in amplitude (Qin and Oxenham, 2003; Füllgrabe *et al*., 2006; Gilbert *et al*., 2007; Gnansia *et al*., 2008). Thus, TFS cues may play a role in enhancing the perception of NH listeners in fluctuating background noise. Moreover, recent results show that HI listeners who have difficulty in noise also have a related deficit in the ability to use TFS cues (Lorenzi *et al*., 2006a; Lorenzi *et al*., 2009; Hopkins and Moore, 2011; Hopkins *et al*., 2008).

Interpretation of the salience of TFS cues must include consideration of the fact that acoustic TFS cues can produce useful temporal coding in two ways: (1) neural responses synchronized to the stimulus TFS ("true TFS") and (2) responses synchronized to stimulus ENV [i.e., "recovered envelopes" (RENVs)] that are naturally created by narrowband cochlear filters (Ghitza, 2001). In NH listeners, it is known that performance is substantially higher for TFS cues derived from a broadband filter than for TFS cues derived from narrow band filters that encompass the same overall frequency range (Smith *et al*., 2002; Swaminathan and Heinz, 2012; Léger *et al*., 2015). This result suggests that the processing of both narrowband TFS cues themselves as well as RENVs may play a role in the effective use of TFS components of speech, which has been suggested to contribute to robust speech perception (Swaminathan, 2010; Shamma and Lorenzi, 2013; Won *et al*., 2012; Won *et al*., 2014).

Recent research has explored the role of RENVs in the perception of TFS speech. Using TFS speech generated with 1 to 16 analysis bands, Gilbert and Lorenzi (2006) filtered the signals into 30 bands from which ENVs were extracted and used to modulate the amplitude of tones at the centers of the bands. The intelligibility of the TFS signals was high and showed little effect of number of analysis bands, compared to that of the RENV speech whose intelligibility was lower and decreased with an increase in the number of TFS bands from which it was created. The results suggested that RENVs played a role in the reception of TFS speech only when the bandwidths of the channels used to generate TFS speech were sufficiently large (i.e., more than 4 times the normal auditory critical bandwidth). Lorenzi *et al*. (2012) recovered envelopes from TFS speech that had been generated within three broad frequency bands and examined the effect of the width of the analysis filters used for ENV extraction. Consonant-identification performance for NH listeners with the RENV signals was somewhat lower than that obtained with the original TFS speech, but was substantially above chance on all conditions and showed improvements with training. Overall performance decreased with an increase in the width of the analysis filters used to create the RENV speech. Swaminathan *et al*. (2014) used a similar approach to creating RENV signals from TFS speech as that of Gilbert and Lorenzi (2006), but filtered the TFS speech into 40 rather than 30 bands for ENV extraction over the same frequency range. [Their use of 40 bands was based on results of Shera *et al*. (2002), suggesting that human auditory filters are sharper than standard behavioral measures.] For NH listeners, 16-band TFS speech and 40-band RENV speech created from this TFS signal yielded similar intelligibility scores and consonant confusion patterns.

Léger *et al*. (2015) extended these results to include 40-band RENV signals created from 1- to 16-band TFS speech and to HI as well as NH listeners. For both groups of listeners, TFS speech and its corresponding RENV speech yielded highly correlated identification scores and patterns of consonant confusions. Performance on both types of speech decreased with an increase in the number of bands used to generate TFS speech. Overall scores were lower for HI than for NH listeners, and they had a larger decline in performance with an increase in the number of bands. Overall, these results provide strong support for the use of RENV cues in the reception of TFS speech by listeners with normal and impaired hearing.

Swaminathan and Heinz (2012; see also Swaminathan, 2010) used a neural model of ENV and TFS processing to predict consonant identification in NH listeners for unprocessed, TFS, and ENV speech in continuous and fluctuating background noises. Based on their results, these authors concluded that the MR observed for TFS speech may arise from the use of RENV rather than TFS. This observation agrees with recent studies that question the role of TFS cues for aiding in speech perception in the presence of fluctuating interference (Oxenham and Simonson, 2009; Freyman *et al*., 2012).

The goal of the current study was to estimate the contribution of envelope, temporal fine-structure, and recovered-envelope cues to consonant identification in various noise backgrounds for NH and HI listeners. Unprocessed speech and five types of processed speech signals were presented in backgrounds of continuous noise and square-wave interrupted noise. ENV speech was generated by extracting the ENV component of unprocessed speech in 40 adjacent frequency bands. Two types of TFS speech were generated, in which the TFS component of unprocessed speech was extracted from the wide-band speech signal or from four adjacent bands over the same frequency range. RENV speech was generated by extracting the ENV component of both types of TFS speech in 40 adjacent bands. MR was examined through comparisons of performance in continuous compared to interrupted noise for each type of speech. The role

of RENVs in the reception of TFS speech was examined by comparisons of performance on these two conditions. The major contributions of this research are its systematic investigation of the use of recovered envelope cues for understanding speech in adverse conditions, as well its estimation of the influence of sensorineural hearing loss on the ability to use recovered-envelope cues in various backgrounds through the comparison of NH and HI listeners.

II. METHODOLOGY

A. Participants

The experimental protocol for testing human subjects was approved by the internal review board of the Massachusetts Institute of Technology. All testing was conducted in compliance with regulations and ethical guidelines on experimentation with human subjects. All listeners provided informed consent and were paid for their participation in the experiments. All listeners were native speakers of American English.

1. Listeners with normal hearing

Eight NH listeners participated in the study. To reflect the age range of the HI listeners, two age ranges were included: under 25 years for younger listeners (2 male and 2 female, mean age of 20.75 years, range of 18–22 years), and over 60 years for older listeners (1 male and 3 female, mean age of 67.0 years, range of 60–70 years). A clinical audiogram was conducted on the subject's first visit to screen for normal hearing in at least one ear. For the younger listeners, the criteria were 15 dB hearing level (HL) or better at the octave frequencies in the range of 250 to 8000 Hz. For the older listeners, the criteria were 20 dB HL or better at octave frequencies in the range of 250 to 4000 Hz and 30 dB HL at 8000 Hz. A test ear that met these criteria was selected for each NH listener. This was the right ear in seven of the eight listeners.

2. Listeners with hearing impairment

Nine listeners with bilateral sensorineural hearing loss participated in the study. Each listener was required to have had a recent clinical audiological examination to verify that the hearing loss was of cochlear origin on the basis of air- and bone-conduction audiometry, tympanometry, speech-reception thresholds, and word-discrimination scores. On the listener's first visit to the laboratory, an audiogram was re-administered for comparison with the listener's most recent evaluation from an outside clinic. In all cases, good correspondence was obtained between these two audiograms.

A description of the HI listeners is provided in Table I, which contains information on sex, test ear, audiometric thresholds, five-frequency pure-tone average (over the five octave frequencies in the range of 250 to 4000 Hz; five-frequency PTA), etiology of loss, and hearing-aid use. The listeners ranged in age from 19 to 73 years. A test ear was selected for monaural listening in the experiments (shown in Table I). Typically, this was the ear with better average thresholds across test frequencies. Hearing losses ranged from mild/moderate to moderate/severe across listeners and in general were bilaterally symmetric. With only one exception, the difference in loss between the ears of a given subject at a given frequency was in the range of 0–20 dB. The exception occurred for HI-2 at 500 Hz where the left ear threshold was 30 dB better than that of the right ear.

TABLE I. Description of hearing-impaired subjects in terms of sex, audiometric thresholds in dB HL in left and right ears at six frequencies, pure-tone average (PTA) over five octave frequencies between 250 and 4000 Hz, history/etiology of loss, and age in years. For each subject, the test ear is denoted by bold lettering. The final two columns provide information about the speech presentation level (prior to NAL amplification) and the SNR used in the consonant testing for each subject.

			Audiometric thresholds in dB HL specified for frequencies in kHz										
Subject	Sex	Ear	0.25	0.50	1.0	2.0	4.0	8.0	Five-Frequency PTA	Etiology	Age	Speech level (dB SPL)	SNR (dB)
HI-1	M	L	15	20	25	35	40	35	27	Hereditary	31	68	−8
		R	**15**	**20**	**15**	**40**	**35**	**25**	25				
HI-2	F	**L**	**15**	**20**	**60**	**50**	**15**	**10**	32	Congenital: unknown cause	21	65	−6
		R	15	50	65	55	25	20	42				
HI-3	M	**L**	**20**	**15**	**30**	**45**	**60**	**90**	34	Hereditary: Stickler syndrome	21	65	−2
		R	25	20	30	40	65	90	36				
HI-4	F	**L**	**15**	**20**	**25**	**35**	**60**	**85**	31	Congenital: maternal rubella	66	65	−4
		R	20	25	30	40	65	105	36				
HI-5	M	**L**	**15**	**15**	**55**	**60**	**70**	**85**	43	Hereditary: Stickler syndrome	19	65	−2
		R	20	25	55	65	70	90	47				
HI-6	F	**L**	**20**	**30**	**50**	**60**	**70**	**80**	46	Congenital: premature birth	64	60	+2
		R	35	35	40	60	80	80	50				
HI-7	F	**L**	**45**	**50**	**60**	**65**	**65**	**80**	57	Early-childhood fistulas	24	70	−2
		R	60	55	60	70	70	75	63				
HI-8	M	L	55	65	70	65	85	95	68	Hereditary: Alport syndrome	73	75	+5
		R	**60**	**65**	**65**	**70**	**75**	**75**	67				
HI-9	M	**L**	**70**	**80**	**85**	**80**	**70**	**55**	77	Congenital: maternal rubella	20	75	+1
		R	60	75	85	90	90	75	80				

The HI listeners are listed in Table I in roughly increasing order of severity of loss based on the five-frequency pure-tone average. The audiometric configurations observed across the hearing losses of these listeners included: (i) sloping high-frequency loss (HI-1, HI-3, HI-4, HI-5, and HI-6; listeners HI-3 and HI-5 are siblings who share the same hereditary condition), (ii) cookie-bite loss with near-normal thresholds at 250, 4000, and 8000 Hz and moderate loss in the mid-frequency range (HI-2); and (iii) relatively flat loss with no more than a 20-dB difference between adjacent audiometric frequencies (HI-7, HI-8, and HI-9). With the exception of HI-1, participants made regular use of bilateral hearing aids at the time of entry into the study.

B. Absolute-detection thresholds

Measurements of absolute-detection thresholds for pure tones were obtained for each NH and HI listener at frequencies of 250, 500, 1000, 1500, 2000, 3000, 4000, and 8000 Hz in the left and right ears. Threshold measurements were obtained using a three-interval, three-alternative, adaptive forced-choice procedure with trial-by-trial correct-answer feedback. Tones were presented with equal *a priori* probability in one of the three intervals and the listener's task was to identify the interval containing the tone. Each interval was cued on the visual display during its 500-ms presentation period with a 500-ms inter-stimulus interval. Tones had a 500-ms total duration with a 10-ms Hanning-window ramp on and off (yielding a 480-ms steady-state portion). During the experimental run, the level of the tone was adjusted adaptively using a one-up, two-down rule to estimate the stimulus level required for 70.7% correct. The step size was 8 dB for the first two reversals, 4 dB for the next two reversals, and 2 dB for the remaining six reversals. The threshold was estimated as the mean level in dB sound pressure level (SPL) across the final six reversals. Listeners had unlimited response time and were provided with visual trial-by-trial feedback following each response.

Thresholds were measured in a block of 16 runs where the 8 test frequencies were measured in random order first in one ear (selected at random) and then in the other ear. These measurements were obtained on the first day of testing.

C. Speech-in-noise measurements

1. Speech stimuli

The speech stimuli consisted of recordings of monosyllables in /a/-C-/a/ format with 16 values of C = /p, t, k, b, d, g, f, s, ʃ, v, z, dʒ, m, n, r, l/. These recordings were taken from the corpus of Shannon *et al.* (1999). The complete set of stimuli consisted of one utterance of each of the 16 syllables from four male and four female speakers for a total of 128 stimuli. The speech stimuli were divided into a 64-item training set and a 64-item test set, with each set composed of the 16 recordings of two male and two female talkers. The recordings were digitized with 16-bit precision at a sampling rate of 32 kHz.

2. Background noise conditions

Speech-shaped noise (spectrally shaped to match the average of the spectra of the complete set of VCV stimuli) was added to the speech stimuli before further processing described below in Sec. II C 3. Three background noises were studied.

a. Baseline condition. Speech presented in a background of continuous noise with a presentation level of 30 dB SPL. This relatively low-level noise was added to all signals based on the suggestion of Hopkins *et al.* (2010) for the use of low-noise-noise to limit the amplification of low-level portions of the speech signal in TFS processing.

b. Continuous noise. Baseline condition (30 dB SPL continuous noise) plus additional continuous background noise scaled to achieve an overall (baseline plus continuous) RMS presentation level selected to yield roughly 50%-Correct consonant identification for the two NH groups and for individual HI listeners for unprocessed speech.

c. Interrupted noise. Baseline condition (30 dB SPL continuous noise) plus additional square-wave interrupted noise at a rate of 10 Hz and a duty cycle of 50%. The overall RMS level of the baseline-plus-interrupted noise was adjusted to be equal to that of the continuous noise.

The levels of speech and noise employed in the testing for the NH groups and for the individual HI listeners are shown in the final two columns of Table I. For the NH listeners, the speech level was 60 dB SPL and testing was conducted with a signal-to-noise ratio (SNR) of −10 dB. For the HI listeners, amplification was applied to the speech-plus-noise stimulus using the NAL-RP formula (Dillon, 2001). A comfortable level for listening to speech in the baseline condition was established for each HI listener and an SNR was selected to yield roughly 50%-Correct identification of consonants in unprocessed speech. Psychometric functions showing consonant identification scores as a function of SNR are provided in the Appendix for a group of young NH listeners and for five of the HI listeners. The functions shown for unprocessed speech support the selection of the SNRs employed in the experiment for 50%-Correct performance. For HI listeners, the SNR used was strongly correlated with the five-frequency PTA (Pearson correlation, $\rho = 0.87$, $p < 0.01$). The same speech levels and SNRs were used in all test conditions for a given listener.

3. Stimulus processing

Prior to presentation to the listener, the speech-plus-noise stimuli were processed to yield six different types of speech that included an unprocessed (U) and an envelope (E) condition as well as two temporal-fine structure (T1 and T4) and two recovered-envelope (R1 and R4) conditions.

a. Unprocessed speech (U). The unmodified V-C-V waveforms were scaled to the desired level and played out directly.

b. Envelope speech (E). Unprocessed speech was bandpass-filtered into 40 bands of equal bandwidth on a log frequency scale spanning 80 to 8020 Hz (see Gilbert and Lorenzi, 2006 for a description of the frequency bands), where the filters were sixth order Butterworth processed both forward and backward temporally to yield 72 dB/oct rolloff. The ENV within each band was computed as the modulus of the Hilbert analytic signal and was followed by lowpass filtering at 64 Hz using a sixth order Butterworth filter (72 dB/oct rolloff). The ENVs were used to modulate tone carriers with frequencies equal to the center frequency of each band and random starting phases. The bands were then combined to yield the processed waveform.

c. Temporal fine-structure speech (T). TFS speech stimuli were created from Unprocessed speech according to the methods described in Gilbert and Lorenzi (2006) and Lorenzi *et al*. (2006a). This involved bandpass filtering the unmodified samples into bands of equal bandwidth on a log frequency scale spanning 80 to 8020 Hz, where the filters were sixth order Butterworth processed both forward and backward temporally to yield 72 dB/oct rolloff. The Hilbert transform was used to decompose each bandpass signal into ENV (i.e., the magnitude of the Hilbert analytic signal) and TFS (i.e., the cosine of the phase of the Hilbert analytic signal) components. The ENV component was discarded and the TFS component was normalized to the long-term average energy of the original bandpass signal. The resulting normalized TFS components for all bands were then summed to yield the TFS speech. Two TFS conditions were created which differed by the number of bands used to filter the original speech signal. One condition (T1) was derived from one band covering the entire frequency range of 80 to 8020 Hz. The second condition (T4) involved the use of four bands with equal bandwidth on a logarithmic scale over the same frequency range.

d. Recovered-envelope speech (R). RENV speech stimuli were created from the T1 and T4 signals described above. The TFS speech stimulus was first bandpass filtered into 40 bands of equal bandwidth on a log frequency scale spanning 80 to 8020 Hz, where the bandpass filters were created using the auditory chimera package for MATLAB™ (Smith *et al*., 2002). For each bandpass signal, the (recovered) ENV component was estimated by full-wave rectification followed by processing with a 300-Hz lowpass filter (sixth order Butterworth, 72 dB/oct rolloff). This (recovered) ENV was then used to modulate a tone carrier at the center frequency of the band and random starting phase. Each resulting band signal was re-filtered through the corresponding bandpass filter to eliminate spectral splatter, and the final processed band signals were summed to yield the RENV stimulus.[1] The RENV signal created from the T1 signal is referred to as R1 and that created from the T4 signal as R4.

4. Experimental procedure

Consonant identification was measured using a one-interval 16-alternative forced-choice procedure without correct-answer feedback. On each trial of the experiment, one of the 64 syllables from either the training or test set was selected and processed according to one of the six speech conditions with one of the three noises. This processed stimulus was then presented and the subject was instructed to identify its medial consonant. A 4×4 visual display of the response alternatives was displayed on a computer monitor and the response was selected using a computer mouse. No time limit was imposed on the subjects' responses.

The speech conditions were tested in the order of U, T1, R1, T4, R4, and E. This order was chosen to optimize training based on Swaminathan *et al*. (2014). For each speech condition, performance was measured first for the baseline condition and then with the continuous and interrupted noise in random order. An individual experimental run consisted of 64 trials derived from a different random-order presentation (without replacement) of the 64 syllables in the stimulus set (training/test) with all stimuli processed according to the same stimulus/noise processing condition. Eight 64-trial runs were obtained for each speech plus noise condition with one exception: only one run was obtained (using the set of test stimuli) for U speech in the baseline noise condition to limit listeners' exposure to unprocessed signals. For the remaining 17 speech/noise conditions, the first three runs were conducted using the training set of stimuli and the final five runs were conducted using the test set. Each run lasted roughly 4 to 6 min depending on the subject's response time. Test sessions lasted 2 h including breaks and 5 to 7 sessions were required to complete the testing depending on the subject. A complete set of data was obtained on all listeners with the exception of HI-9 who was unable to perform sufficiently above chance levels on the T4 and R4 conditions.

Experiments were controlled by a desktop PC equipped with a high-quality, 24-bit PCI sound card (E-MU 0404 by Creative Professional). The level-calibrated speech-plus-noise stimuli were played using MATLAB™; passed through a Tucker-Davis (TDT) PA4 programmable attenuator and a TDT HB6 stereo headphone buffer; and presented monaurally to the subject in a soundproof booth via a pair of Sennheiser HD580 headphones. The primary experimental engine used to generate and present stimuli and to record responses was the AFC Software Package for MATLAB™ provided by Stephan Ewert and developed at the University of Oldenburg, Germany. A monitor, keyboard, and mouse located within the sound-treated booth allowed interaction with the control PC.

5. Data analysis

The three runs obtained with the training stimulus set and the first run obtained with the test set were considered practice on a given condition and were discarded. The final four runs obtained using the test set were retained for analysis. Stimulus-response confusion matrices were generated for each run, added across the final four runs for each subject and each experimental condition, and used to calculate percent-correct scores (where chance performance on the 16-item set was 6.25%-Correct).

The magnitude of the MR was examined for each type of speech by subtracting the percent-correct score obtained in interrupted noise from that obtained in continuous noise. A normalized measure of MR (NMR) was also employed to take into account the variation in intelligibility in the baseline condition and in continuous noise across listeners. The NMR was defined as follows:

$$\mathrm{NMR} = \frac{\%\text{–Correct}_{\mathrm{Interrupted}} - \%\text{–Correct}_{\mathrm{Continuous}}}{\%\text{–Correct}_{\mathrm{Baseline}} - \%\text{–Correct}_{\mathrm{Continuous}}}. \quad (1)$$

For further statistical analysis, the percent-correct scores were converted into rationalized arcsine units (RAU) (Studebaker, 1985) for each subject and on each test condition. Repeated-measures analyses of variance (ANOVAs) were conducted using those RAU scores. Due to an incomplete set of data on HI-9, his results were not included in the ANOVA tests.

For each speech type and noise condition, the stimulus-response confusion matrices were added across the eight NH listeners and across the nine HI listeners for additional analyses. These included computations of feature information transfer (Miller and Nicely, 1955; Wang and Bilger, 1973) and a form of metric multidimensional scaling analysis (Braida, 1991). In these analyses, the data of HI-9 were included for the four speech conditions on which he was tested.

III. RESULTS

A. Absolute threshold measurements

Detection thresholds obtained with 500-ms signals are plotted in dB SPL as a function of frequency in Fig. 1. Mean thresholds across the test ears of the four younger and four older NH listeners are shown in the top left panel. The remaining nine panels show left- and right-ear measurements of each of the nine HI listeners. The HI listeners are arranged in increasing order of mean pure-tone threshold across the frequencies of 250, 500, 1000, 2000, and 4000 Hz. This five-frequency PTA in dB SPL is provided in the panel of each HI listener. The 500-ms threshold data shown in Fig. 1 confirm the audiometric configurations and symmetry of the hearing losses as shown in the clinical results of Table I.

B. Consonant-identification scores

Consonant-identification scores in %-Correct are shown in Fig. 2. Each panel represents results with one of the six types of speech. Within each panel, scores for each noise background are shown as means across the NH and HI groups (the two leftmost bars within each panel) and are also provided for each individual HI listener. The gray bars between the continuous and interrupted data points are provided for visual guidance and represent MR in percentage points.

The results of a three-way ANOVA on consonant-identification scores (in RAU) with group (NH or HI; note that HI-9 was excluded from the analyses because of missing values) as a between-subject factor, and processing type (U, E, T1, R1, T4, and R4) and noise background (baseline, continuous, interrupted) as within-subject factors are reported below. Mean results across listeners in each of the two groups in RAU (excluding those of HI-9) are provided for each of these factors in Table II. Scores of NH and HI listeners were different [$F(1,14) = 19$, $p < 0.001$], with the average score of the HI listeners being lower than that of the NH listeners. For both NH and HI listeners, scores varied as a function of the processing type [$F(5,70) = 189$, $p < 0.001$], with the best scores being generally obtained in the U condition, and the worst in the T4 and R4 conditions. For both NH and HI listeners, scores also varied as a function of the noise background [$F(2,28) = 442$, $p < 0.001$], with the best scores

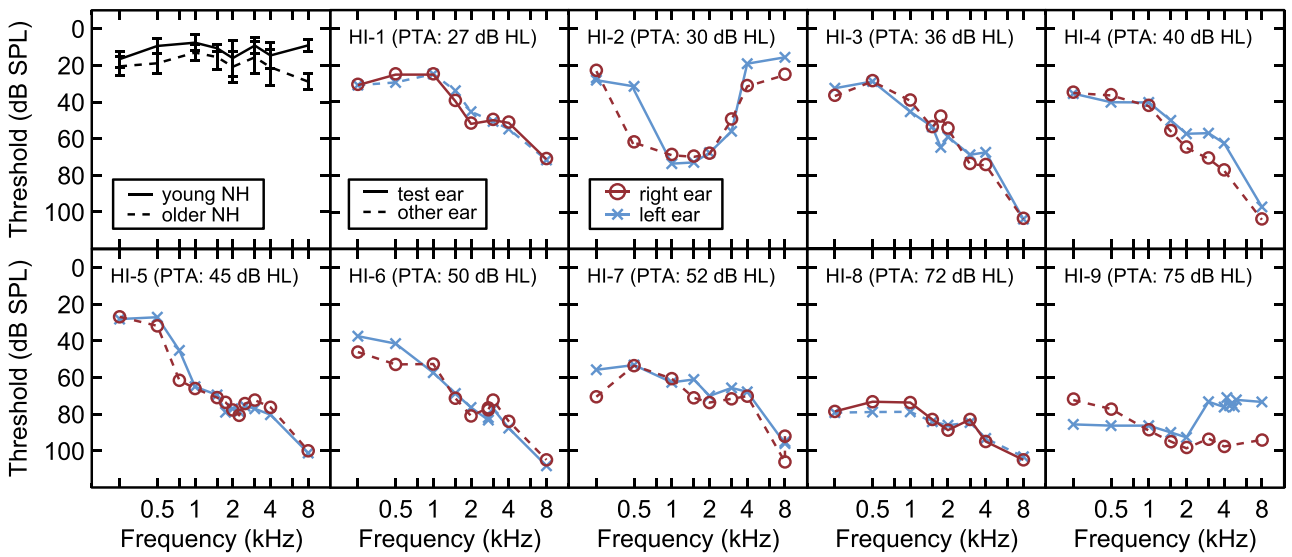

FIG. 1. (Color online) Detection thresholds estimated using 500 ms tones (in dB SPL) as a function of the frequency (in kHz) for NH and HI listeners. In the top left panel, detection thresholds of NH listeners are averaged over four younger NH (filled line) and four older NH listeners (dotted line). The error bars represent the standard deviation about the mean. Individual detection thresholds are reported in the remaining nine panels for each of the nine HI listeners. The right ear is represented by circles and the left ear by ×. The test ear is represented by the solid line and the non-test ear by the dotted line. The five-frequency PTA is reported for each HI listener; see text for details.

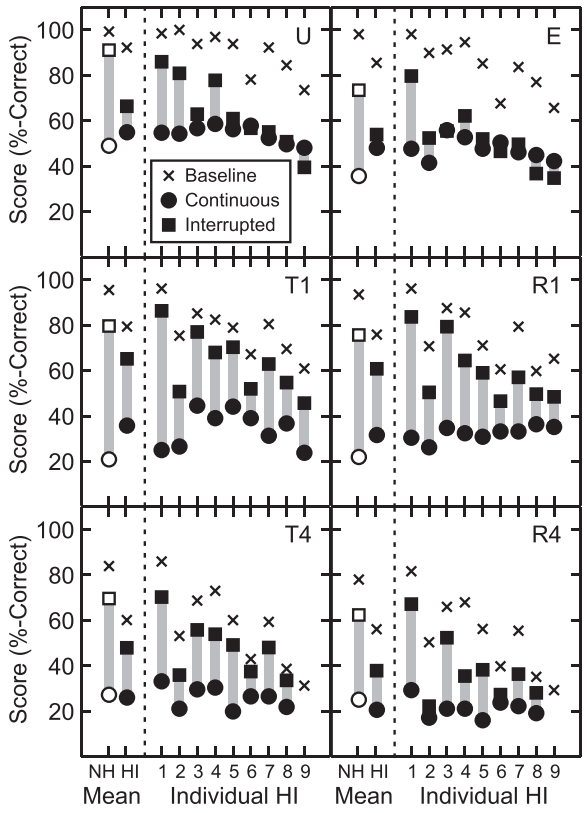

FIG. 2. Consonant identification scores in %-Correct for each of the six types of speech (indicated at the top right of each panel: U, E, T1, R1, T4, and R4). Within each panel, mean results are first shown for the NH and HI groups (left-most bars, unfilled and filled symbols, respectively) and then for each of the individual HI listeners. In each panel, scores are reported for each of the three backgrounds of noise: baseline (×), continuous (●), and interrupted (■). The gray bars between the continuous and interrupted data points are provided for visual guidance and represent MR in percentage points.

being generally obtained in the baseline condition, and the worst in continuous noise.

The effect of processing type and noise background was different for NH and HI listeners [interaction between group and processing type: $F(5,70) = 5$, $p < 0.001$; and interaction between group and noise background: $F(2,28) = 40$, $p < 0.001$]. Furthermore, for both groups, the effect of processing type varied as a function of noise background [interaction between processing type and noise background: $F(10, 140) = 33$, $p < 0.001$] and this effect was different for the two listener groups [interaction between group, processing type and noise background: $F(10, 140) = 4$, $p < 0.001$]. Compared to the U condition, the E condition generally led to poorer scores in noise for NH but not for HI listeners. Both groups showed poorer performance on the T and R conditions compared to U, although some between-group differences were observed as a function of noise type. For NH listeners, all T and R conditions were worse than U in continuous noise but reductions in scores in baseline and in interrupted noise were observed only for T4 and R4. For HI listeners, the T and R conditions led to poorer scores than U in both baseline and in continuous noise; in interrupted noise, however, scores for T1 and R1 were similar to U and only modestly lower than U for T4 and R4. For both NH and HI listeners, scores obtained with T1 and T4 were similar to those obtained with R1 and R4, respectively; see further analysis of this effect below (Sec. III D).

TABLE II. Identification scores (in RAU) for the six types of speech (rows: U, E, T1, R1, T4, and R4; the row "mean" shows scores averaged over the different types of speech) for NH and HI listeners (columns; the column "mean" shows scores averaged over the two groups). Scores are shown for each of the three backgrounds of noise (baseline, continuous, and interrupted) as well as averaged over the different backgrounds ("all noises"). Results of an ANOVA (see Sec. III B) indicated that all global effects and interactions were significant. (Note that results of HI-9 were excluded from the ANOVA and from this table as well.)

	Baseline			Continuous			Interrupted			All noises		
	NH	HI	Mean	NH	HI	Mean	NH	HI	Mean	NH	HI	Mean
U	115	99	107	52	55	54	92	66	79	87	73	80
E	111	89	100	39	48	44	77	54	65	75	64	70
T1	105	80	93	24	37	30	87	65	76	72	61	66
R1	103	77	90	23	33	28	81	61	71	69	57	63
T4	87	60	74	30	27	28	73	48	61	63	45	54
R4	84	56	70	25	22	23	67	39	53	59	39	49
Mean	101	77	89	32	37	35	79	56	67	71	56	

C. Normalized masking release (NMR)

NMR is shown in Fig. 3 for each of the six processing types as a function of the SNR tested. These data are replotted in Fig. 4 to show comparisons in NMR across speech conditions for NH and HI listeners. In both figures, NMR is shown as the mean of the NH listeners and for each of the nine individual HI listeners. Horizontal lines indicate a NMR of zero. NMR below 0 shows that the scores were lower for interrupted noise than for continuous noise. In Fig. 3, the NMR shows a tendency to decrease with an increase in the test SNR for the U and E conditions, but not for the T and R conditions. Note that the test SNR was highly correlated with the five-frequency PTA. As shown in Fig. 4, the magnitude of the NMR was in the range of roughly 0.6 to 0.8 across conditions for the NH listeners (first row), and three basic patterns of NMR were observed among the HI listeners. HI-1, HI-2, and HI-4 (second row) showed positive values of NMR across conditions; HI-3, HI-5, and HI-7 (third row) generally had small positive values of NMR for U and E and much larger values for the T and R conditions; and HI-6, HI-8, and HI-9 generally had negative values of NMR for U and E together with positive values for the T and R conditions.

The results of a two-way ANOVA on NMR (in RAU) with group (NH or HI) as a between-subject factor and processing type (U, E, T1, R1, T4, and R4) as a within-subject factor are reported below. NMR was different for NH and HI listeners [$F(1,14) = 31$, $p < 0.001$], with the average NMR of the HI listeners being lower than that of the NH listeners. NMR globally varied as a function of the processing type [$F(5,70) = 32$, $p < 0.001$], with the best NMR being obtained in T1 and R1, on average. The effect of the processing type on NMR varied between groups [interaction between processing type and group: $F(5,70) = 7$, $p < 0.001$]. For NH listeners, NMR was high and similar across processing types. For HI listeners, however, large differences between

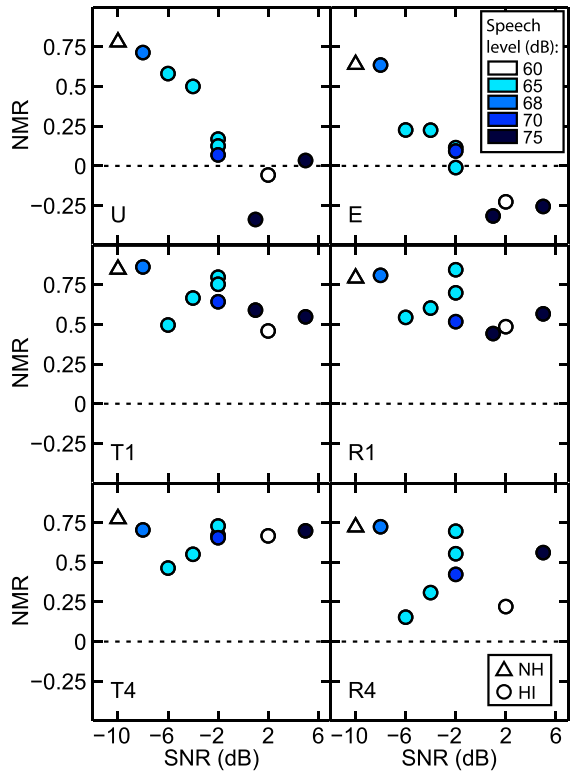

FIG. 3. (Color online) Normalized masking release (NMR) in %-Correct as a function of the SNR tested (in dB, see Table I) for each processing condition (U, E, T1, R1, T4, and R4). Each panel shows averaged NMR for NH listeners and individual NMR for HI listeners (note that no NMR was computed for HI-9 in T4 and R4). The speech presentation level used is indicated by color (shade). The horizontal dotted line shows NMR of 0.

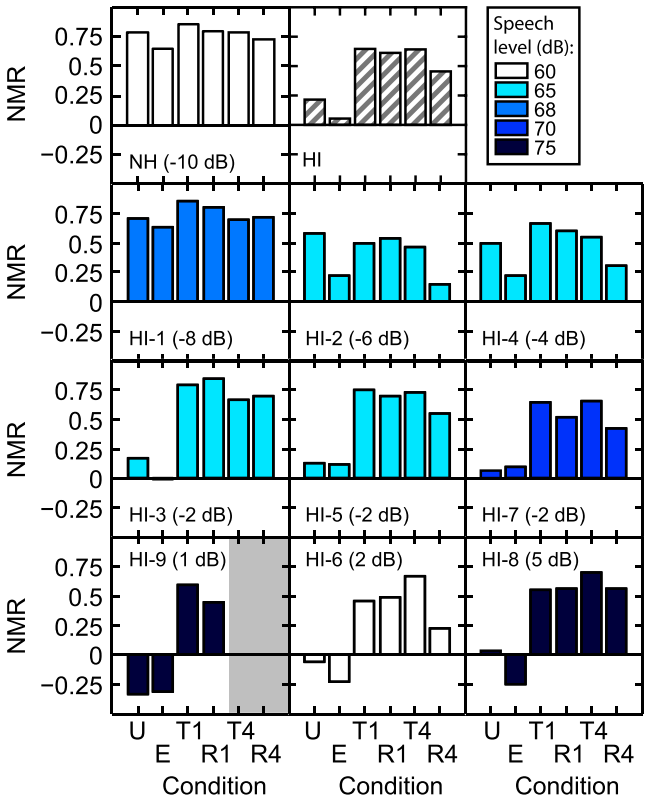

FIG. 4. (Color online) Averaged normalized masking release (NMR) in %-Correct for the NH and HI listeners (top panels) as a function of the processing condition (U, E, T1, R1, T4, and R4). NMR for individual HI listeners is also shown (middle and bottom panels; note that no NMR was computed for HI-9 in T4 and R4, as highlighted by the gray area). The speech presentation level used is indicated by color (shade). Note that averaged NMR for HI listeners was computed across different presentation levels (as indicated by the hashed fill). HI listeners are ordered based on the SNR at which they were tested, which is reported in each panel. The horizontal line shows NMR of 0.

processing conditions were observed. For U and E, large variability was observed in NMR among HI listeners, and those variations were related to the SNR tested (see trends in Fig. 3). For T and R speech, less variability was observed across HI listeners and/or SNR tested (except for R4). NMR was positive for all HI listeners and generally higher than for U and E.

D. Analysis of confusion matrices

1. Feature information analysis

Feature analyses were conducted using a set of consonant features (defined in Table III) that included nasality, approximant, strident, sonorant, voicing, continuancy, and place of articulation. The conditional information transfer on this set of features was calculated using a fixed-order sequential feature information analysis (the SINFA technique of Wang and Bilger, 1973, as implemented in the FIX program of the Department of Phonetics and Linguistics, University College London). The SINFA analysis was employed to remove redundancies among the features whose relative feature information transfer (IT) was examined in the fixed order listed in Table II. This fixed order was determined based on the mean rank of features obtained from unrestricted SINFA analyses conducted on the HI confusion matrices. In Fig. 5, relative conditional feature IT is plotted for NH and HI listeners on each of the seven features for each noise type and speech condition. The top two panels show results obtained for the baseline noise for NH (upper plot) and HI (lower plot), the middle panels for interrupted noise, and the bottom panels for continuous noise. The six bars shown for each speech feature within each panel represent different types of speech and are ordered as U, E, T1, R1, T4, and R4.

In the baseline condition, NH feature performance was similar (and high) for U, E, T1, and R1 with a modest decrease in performance across features for the T4 and R4 conditions. Across features and conditions, HI feature scores were lower than NH scores and showed a larger drop in performance for the T4 and R4 conditions particularly for the

TABLE III. Classification of the 16 consonants on a set of seven phonetic features. (Approximant is abbreviated as "approx." and continuant as "contin.")

	/p/	/t/	/k/	/b/	/d/	/g/	/f/	/s/	/ʃ/	/v/	/z/	/dʒ/	/m/	/n/	/r/	/l/
Nasality	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
Approx.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Strident	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0
Sonorant	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1
Voicing	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1
Contin.	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1
Place	0	1	2	0	1	2	0	1	2	0	1	1	0	1	2	1

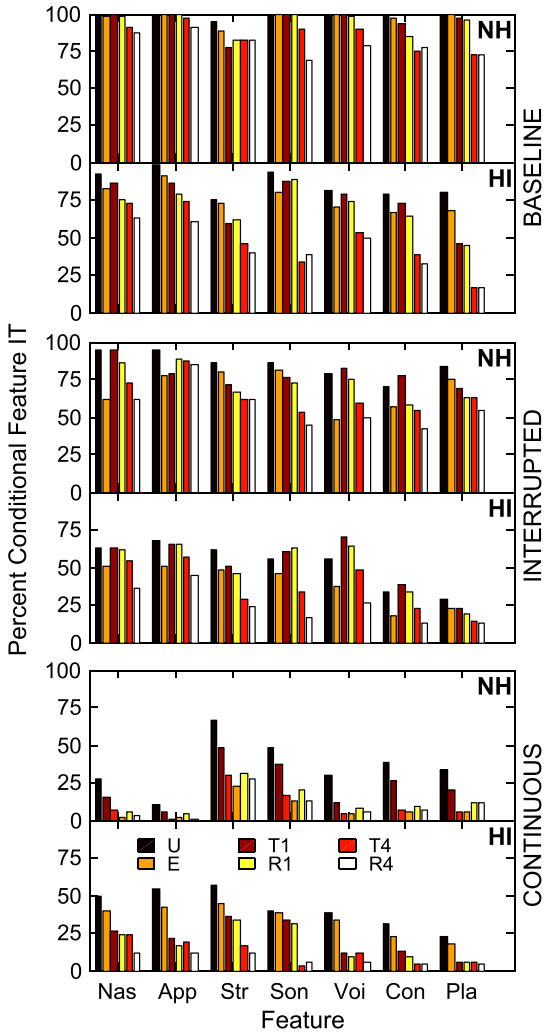

FIG. 5. (Color online) Relative conditional feature information transfer (IT) for the NH and HI listener groups on seven features for each of three noise types and six speech conditions. The upper two panels show results for the baseline noise condition, the middle two panels for interrupted noise, and the lower two panels for continuous noise. For each noise type, NH data are plotted above the HI data. The seven features shown along the abscissa are nasality (NAS), approximant (APP), strident (STR), sonorant (SON), voicing (VOI), continuancy (CON), and place (PLA). For each feature and in each panel, bars show results for six speech conditions ordered as U, E, T1, R1, T4, and R4.

features sonorant and place. Both groups of listeners had a roughly 25 percentage-point decrease in feature scores in interrupted noise compared to the baseline condition; however, the features of continuancy and place showed greater decreases for the HI listeners. In the continuous background noise, nasality and approximant were better received by the HI than by the NH group; other than that feature reception was similar for the two groups. Stridency and sonorant were the best-perceived features in the T and R conditions. Across speech conditions and for both groups of listeners, performance on any given feature was highly similar for the T1 and R1 conditions and for the T4 and R4 conditions.

2. Metric multidimensional scaling analysis

To compare confusion patterns for different speech conditions and listener groups, we used a form of metric multidimensional scaling (Braida, 1991). In each speech and noise condition, consonants are assumed to be identified on the basis of the sample value of a four-dimensional vector of cues $\vec{c} = \langle c_1, c_2, c_3, c_4 \rangle$. When a consonant is presented, the components of $\vec{c}$ are independent identically distributed Gaussian random variables with means ($\vec{X}_j = \langle X_{j1}, X_{j2}, X_{j3}, X_{j4} \rangle$) and a common variance $\sigma^2 = 1.0$. Each consonant is thus associated with a *stimulus center* specified by the mean value of the cue vector for that consonant. The listener is assumed to assign a response by determining the identity of the *response center* $\vec{R}_k = \langle R_{k1}, R_{k2}, R_{k3}, R_{k4} \rangle$ that is closest to the cue vector on a given stimulus presentation. Stimulus and response centers were estimated from the confusion matrices and the overall structure of the confusion matrix was represented by the set of values $d'(i,j)$ calculated for each pair (i,j) of stimuli $d'(i,j)^2 = \Sigma_{k=1}^{4}(X_{ik} - X_{jk})^2$. This allowed comparison of the structures of the confusion matrices between different speech processing conditions. For example, Fig. 6 shows results comparing performance on T4 versus R4 speech for NH and HI listeners for continuous and interrupted noise backgrounds. For each condition and group, the set of $d'_{T4}(i,j)$ is plotted against the set of $d'_{R4}(i,j)$. These data indicate a high degree of correlation between the two conditions (results of the Pearson product-moment correlation: ρ^2 in the range of 0.72 to 0.94) across both noise backgrounds and listener groups. The comparison of the T1 and R1 conditions also yielded high values of correlation with ρ^2 in the range of 0.59 to 0.94.

These comparisons were extended to include correlations of the sets of pairwise d' for all combinations of U, E, T1, and T4 speech in each of the three noise backgrounds (yielding 66 pairs of comparisons) for NH and HI listeners. Because T and R conditions yield extremely similar results for both NH and HI listeners, correlations were not computed on R scores. The results of these analyses (ρ^2) are shown in Fig. 7(A), where the rows and columns are labeled with the twelve noise/processing conditions and each cell presents the strength of the correlation between the two conditions represented by that cell. The correlations are coded on a color/shading scale in terms of the value of ρ^2: darker shadings correspond to higher correlations.

The upper half of the diagonal in Fig. 7(A) represents the NH data. In continuous noise, the confusion patterns were strongly correlated within speech conditions (all $\rho^2 > 0.73$) for NH listeners. In interrupted noise, the strength of the correlations within speech conditions was much lower (all $\rho^2 < 0.55$). Weak to modest correlations were also observed between continuous and interrupted noise (all $\rho^2 < 0.53$). All correlations with baseline noise conditions were weak (all $\rho^2 < 0.27$), which is likely due to the high levels of performance.

The lower half of the diagonal in Fig. 7(A) represents the HI data, which shows a larger number of modest to strong correlations than was observed for NH listeners. In continuous noise, strong correlations were observed within speech conditions similar to those seen for NH listeners (all $\rho^2 > 0.74$). However, moderate-to-strong correlations were also observed in the interrupted noise within speech conditions (all $\rho^2 > 0.52$) as well as between continuous and interrupted noise (all $\rho^2 > 0.49$, except for T1 versus T1 and T4).

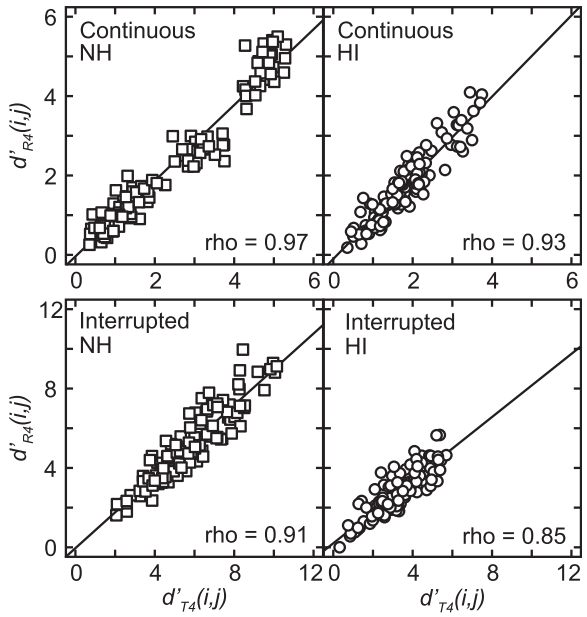

FIG. 6. Comparison of the confusion matrices summed (separately) across 8 NH listeners (left panels) and 8 to 9 HI listeners (right panels; the data of HI-9 were included only for the four speech conditions on which he was tested) obtained using a metric multidimensional scaling. Values of d' and the results (ρ) of Pearson correlation analyses are reported for R4 as a function of T4 in continuous (top panels) and interrupted (bottom panels) noise. Note that the scale is not the same for the top and bottom panels.

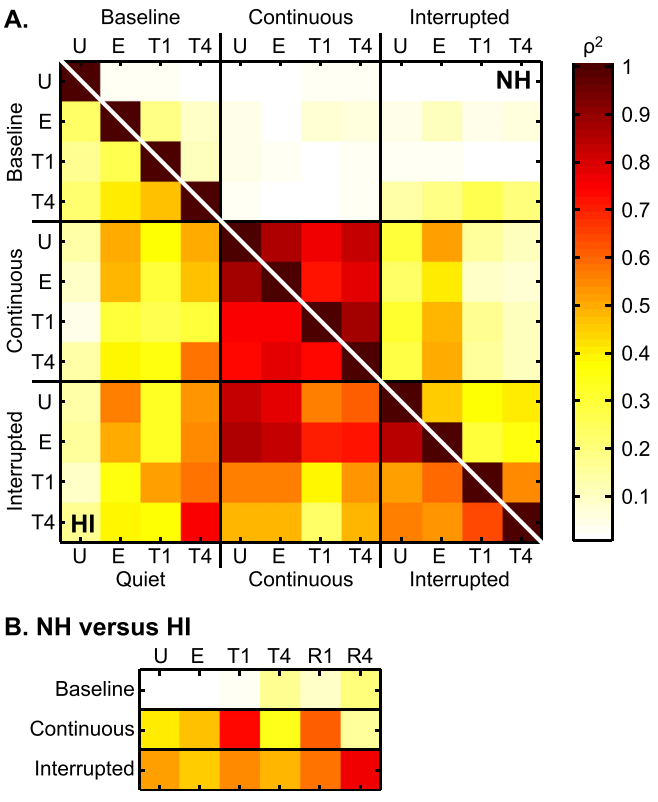

FIG. 7. (Color online) (A) Summary of the comparison of the sets of d' of NH (upper half of the diagonal) and HI (bottom half of the diagonal) listeners for all combinations of U, E, T1, and T4 speech in each of the three noise backgrounds. Because T and R conditions yield extremely similar results for both NH and HI listeners, correlations were not computed on R scores. Each cell presents the strength (color/shade coded, see legend) of the correlation (ρ^2) between the two conditions represented by that cell. (B) As above, but showing correlations of d' between NH and HI listeners for three noises and six speech types.

Furthermore, some modest correlations were observed for baseline versus continuous and interrupted noise, such as with T4 in baseline noise (all $\rho^2 > 0.47$, except for versus T1 in continuous noise) and E in baseline noise versus U and E in continuous and interrupted noise (all $\rho^2 > 0.49$).

Correlations of the sets of pairwise d' between the NH and HI groups were examined within noise condition and are shown in Fig. 7(B). The confusion patterns of the two groups were modestly to strongly correlated for several conditions in continuous noise ($\rho^2 > 0.47$ for E, T1, and R1) and for all conditions in interrupted noise (all $\rho^2 > 0.45$) but showed very little correlation for any type of speech in the baseline condition (all $\rho^2 < 0.22$).

IV. DISCUSSION

A. Overall performance

The results obtained with the HI listeners were inferior to those of the NH listeners as evidenced by lower levels of performance in the baseline condition across speech-processing types and by the need for higher levels of SNR to achieve a given level of performance.

In the baseline condition (low-level continuous background noise), both groups exhibited scores ordered as U $\approx$ E > T1 $\approx$ R1 > T4 $\approx$ R4. In Fig. 2, it can be seen that NH listeners demonstrated a comparatively small drop in performance across all conditions, which suggests an ability to use both ENV and TFS cues in isolation. Compared to U, the HI listeners demonstrated a larger decrease in performance for T1/R1 and T4/R4 than for E. These results indicate that the HI listeners were less affected by the removal of TFS cues in the E condition than they were by the removal of ENV cues in the T1 and T4 conditions, which suggests a greater reliance on the use of ENV compared to TFS cues for consonant recognition.

In the continuous background noise, the performance of the NH and HI listeners for U speech was equivalent at roughly 50%-Correct, as intended by the selection of SNR for the NH group and for individual HI listeners. Testing the remaining speech conditions with the same SNR resulted in T4 and R4 scores that were similar for NH and HI listeners; however, for the E, T1, and R1 conditions, the NH group had scores that were roughly 10 percentage-points lower than those of the HI listeners. Thus, when the SNR was adjusted to yield similar performance for NH and HI listeners for U speech, the continuous noise proved to be more effective in masking E, T1, and R1 for the NH listeners. (These differences are discussed further in Sec. IV C.)

In the interrupted noise background, performance for NH listeners was always intermediate between baseline and continuous scores and generally tended to be closer to the baseline. The largest difference between the interrupted-noise and baseline scores was observed for E speech. Performance for HI listeners was similarly intermediate between baseline and continuous scores (and closer to baseline) for T1, R1, T4, and R4 speech. Their performance was closer to continuous noise scores for U and E speech, and five listeners (HI-3, HI-5, HI-7, HI-8, and HI-9) had higher performance with T1 than with U in interrupted noise. A

possible explanation of this effect is discussed in Sec. IV D. The similarity in performance for HI listeners on U and E speech provides additional support for their dependence on ENV cues for speech reception. Performance in continuous and interrupted noise backgrounds is compared in Sec. IV D which discusses NMR.

For all noise conditions and for both listener groups, E scores were always higher than R scores. The E speech was obtained directly from the U speech (using the Hilbert transform), while the R speech was obtained indirectly, through the recovery of ENV cues from the T speech. A difference in scores between E and R could indicate that reconstructed ENV cues do not convey as much information as the original speech ENV cues, which could indicate that ENV recovery is not perfect. However, there is an alternate explanation as to why R scores were worse than E scores. This could arise because the ENV cues of the intact signal were disrupted during the generation of T speech via the Hilbert transform (either because ENV cues were removed, or due to distortions introduced by the processing; see Sec. IV C). Furthermore, the frequency cutoff over which ENV cues were extracted was different for E (<64 Hz) and R (<300 Hz) speech, which may have influenced intelligibility. Because of these confounds, it is not possible to make any strong inferences regarding the source of the poorer performance on R compared to E speech.

B. Consonant-confusion patterns

For both groups of listeners, the structure of the confusion matrices was highly correlated across the different types of speech processing in the continuous-noise background (see Fig. 7). This result is supported by the feature scores shown in the bottom two panels of Fig. 5, which indicate that the relative strength of the various features remained fairly constant across speech types for each of the two listener groups. Thus, within each group of listeners, the presence of a continuous background noise led to the same types of errors regardless of the type of speech processing. For the NH group, no strong correlations were observed among any of the other conditions which may be due in part to the relatively high levels of performance (and thus a lack of off-diagonal entries in the confusion matrices) in the baseline and interrupted noise backgrounds. For the HI group, on the other hand, correlations were observed between certain conditions in continuous and interrupted noise (both U and E in interrupted noise were fairly well-correlated with each of the other speech types in continuous-noise). This result suggests that the perception of the speech sounds in interrupted noise was similar to that in continuous noise for U and E (where there was little NMR) but not for T1 and T4 (where a larger NMR was observed). This provides further support for the conclusion that, when listening to U speech, HI listeners were relying primarily on ENV cues.

There was little correlation of the confusion patterns between the NH and HI groups except for T1 and R1 in continuous noise and R4 in interrupted noise [see Fig. 7(B)]. The feature analysis shown in Fig. 5 provides some insight into this lack of correlation in continuous noise backgrounds based on differences in reception of the features of nasality and approximant which were much better received by the HI group. Perhaps the higher overall SNRs at which the HI were tested compared to the NH group may have led to greater saliency of the relatively low-frequency cues associated with these two features. In interrupted noise, the much better reception of the continuant and place features by the NH compared to HI listeners may be responsible for the lack of correlation.

C. Role of RENVs in perception of TFS speech

Consistent with other studies (Gilbert and Lorenzi, 2006; Swaminathan, 2010; Lorenzi *et al.*, 2012; Swaminathan *et al.*, 2014; Léger *et al.*, 2015), our results suggest that the reception of TFS speech is based on the use of RENVs for both NH and HI listeners. In the current study, the reception of one-band and four-band TFS speech was compared to speech signals created through the extraction of 40 bands of ENVs from each type of TFS speech. Overall levels of performance as well as patterns of confusion were highly similar for T1 versus R1 and T4 versus R4 speech for both the NH and HI listener groups. The high degree of resolution in extracting the recovered envelopes (40 bands) may have contributed to the similarity in scores between T and R speech. However, previous results (Swaminathan *et al.*, 2014) indicate that the use of fewer bands (e.g., 32 instead of 40) would have only led to a slight decrease in performance for R speech. Referring to Fig. 6, it can be seen that there is a high degree of correlation in the structure of the confusion matrices for both interrupted and continuous noise for both listener groups, even though overall levels of performance were higher for the NH group (as evidenced by higher d' values). Thus, it appears that the perceptual processing of both types of speech is nearly indistinguishable and relies on the same types of cues.

One possible explanation of the reduced ability of the HI listeners to understand TFS speech is that they have a deficit in ENV recovery. However, providing TFS cues as narrow-band RENV cues in the R1 and R4 conditions did not benefit the HI listeners over the original TFS stimuli. Several possible implications arise from these results. One is that ENV recovery from TFS speech was already optimal for the HI listeners and thus the artificial-recovery of the ENVs did not provide any further benefit. Second, it is possible that the RENVs were internally smeared due to the broadened auditory bands associated with cochlear hearing loss and thus led to fewer independent usable ENVs (e.g., Baskent, 2006; Swaminathan and Heinz, 2012; Lorenzi *et al.*, 2012). Third, the RENV cues may have been corrupted by the amplification of noise during the generation of TFS speech (see Apoux *et al.*, 2013). Future work will explore strategies for presenting artificially recovered ENV cues to HI listeners in a manner that is more accessible. One such strategy may be to present alternative bands of artificially recovered ENV cues dichotically to the HI listener to achieve some degree of separation in the auditory filterbank.

D. Masking release

For the NH listeners, both MR and NMR were substantial for each of the six types of speech. The results obtained here with unprocessed speech can be compared to other studies that have employed square-wave or sine-wave amplitude modulations in the range of 8 to 16 Hz (e.g., Füllgrabe *et al.*, 2006; Lorenzi *et al.*, 2006b; Gnansia *et al.*, 2008; Gnansia *et al.*, 2009). The MR of 47 percentage points (corresponding to a NMR of 79%) observed in the current study for U speech is consistent with the range of MR values reported across the studies cited above (roughly 30 to 65 percentage points). Similar to the values obtained with U speech, MR and NMR for T1, R1, T4, and R4 speech averaged 52 percentage points and 78%, respectively. Swaminathan (2010) examined MR for broadband and 16-band TFS speech in consonants using an 8-Hz sinusoidal modulation of speech-shaped noise. Maximal MR was observed at an SNR of +10 dB for both the broadband (25 percentage point MR) and 16-band (10 percentage points) TFS conditions. The higher values observed in the current study for broadband TFS speech may be related to differences in experimental parameters such as the use of a square-wave rather than sinusoidal amplitude modulator. Substantial values of MR (39 percentage points) and NMR (65%) were also obtained with 40-band E speech, but these were the lowest among the six speech types. Although previous studies employing a relatively small number of wideband vocoder channels have reported negligible MR (e.g., Füllgrabe *et al.*, 2006 with a 4-channel vocoder and Gnansia *et al.*, 2009 with an 8-channel vocoder), other experimental conditions employing 16- or 32-channel vocoders have reported MR in the range of roughly 10 to 25 percentage points (Füllgrabe *et al.*, 2006; Swaminathan, 2010; Gnansia *et al.*, 2008; Gnansia *et al.*, 2009). The larger MR reported here for E speech may be due to differences in experimental conditions including our use of square-wave modulation versus the sine-wave modulation employed in these previous studies, a 40-channel vocoder, and a more adverse SNR of −10 dB.

For HI listeners, individual differences were observed in terms of their ability to take advantage of the temporal interruptions in the noise to improve consonant reception for the U and E conditions; however, all HI listeners exhibited a substantial NMR for the T1, R1, T4, and R4 conditions. For any given HI listener, performance on U and E speech was similar in terms of NMR, as seen in Fig. 4. HI-1, HI-2, and HI-3, with mild-to-moderate PTA (and tested at SNR in the range of −4 to −8 dB), showed positive NMR for both conditions, while the remaining listeners (tested at SNR in the range of −2 to +5 dB) had either no appreciable NMR or slightly negative values (indicating better performance on continuous than interrupted noise). This pattern may be interpreted as an indication that testing at higher SNRs leads to a reduction in NMR. However, for several of the HI listeners for whom psychometric functions were obtained in the supplementary experiment, overlaps were seen between the continuous and interrupted noise curves for U as well as for E (see Fig. 8). This observation implies that testing these listeners at lower values of SNR would not have led to substantial changes in estimates of their NMR. All listeners, on the other hand, showed positive NMR for the T and R conditions.

For both NH and HI listeners, when MR is observed in the U and E conditions, it arises due to increased performance in interrupted noise. However, for the T and R conditions, the observed MR arises due to a combination of a drop in continuous-noise performance as well as to higher performance in interrupted-noise. The decrease in continuous-noise performance arises from a combination of decreased baseline performance and an increased negative effect of continuous noise on TFS speech. The experiment could have been reconfigured to equalize the performance in continuous noise by adjusting SNR for each speech type independently instead selecting a fixed SNR to yield 50% correct on U speech in continuous noise. If this had been the case, the continuous- and interrupted-noise psychometric functions for T1 shown in Fig. 8 indicate that MR would still have been observed. This arises due to the shallow slope of the interrupted-noise relative to the continuous-noise functions.

One explanation for the lack of MR for HI listeners for U and E speech lies in the modulation masking theory of Stone and colleagues (e.g., see Stone *et al.*, 2012; Stone and Moore, 2014). They postulate that the MR in NH listeners arises due to a release from modulation (rather than energetic) masking effects of Gaussian noise. When the Gaussian noise is replaced with low-noise noise, which exhibits minimal modulation masking, the MR disappears (Stone and Moore, 2014). If HI listeners are less susceptible to modulation masking (e.g., due to increased auditory bandwidths, see Oxenham and Kreft, 2014), then this would explain their lack of MR in the presence of Gaussian noise maskers. A reduced susceptibility to modulation masking may also account for the higher scores of the HI compared to NH listeners in continuous speech-shaped Gaussian noise noted above on the E, T1, and R1 conditions using an SNR selected to equate performance in U speech.

A possible explanation for the presence of MR for HI listeners with T1 and T4 speech may arise from NH studies that show more robust MR for speech containing TFS cues compared to vocoded speech where such cues are removed (e.g., Hopkins and Moore, 2009). However, the importance of TFS cues for "listening in the gaps" has been called into question by several studies exploring this hypothesis (Oxenham and Simonson, 2009; Freyman *et al.*, 2012). For example, Oxenham and Simonson (2009) did not find a greater MR for low-pass filtered speech (which would contain TFS cues) compared to high-pass filtered speech (where TFS cues for resolved harmonics would be absent). In studies using whispered speech (which removes natural TFS cues but does not alter the spectral details of the vocal-tract resonance), Freyman *et al.* (2012) observed a substantial MR that was constant across a wide range of SNR, as well as a small MR for a vocoded signal created from the whispered speech.

Although all listeners in the current study (both NH and HI) demonstrated an MR for the T1 and T4 conditions, there are several observations to suggest that this was not due primarily to the use of TFS cues *per se*. One such observation

is that the size of the MR and NMR measured in the R1 and R4 conditions was roughly equivalent to that of the T1 and T4 conditions, respectively. Because the ENV-recovery method used to create the R conditions effectively eliminated any fine-structure cues, it is unlikely that TFS can account for this performance. These results, along with those of other studies demonstrating MR for ENV-based speech under certain conditions (including factors such as number of vocoder channels and modulation rate and depth of the interrupted noise), call into question the role of TFS cues in explaining the MR observed for NH listeners.

An alternative explanation for the presence of MR for the HI listeners with T1, R1, T4, and R4 speech may be related to the manner in which TFS processing (which removes some ENV cues) interacts with the interrupted-noise condition as the SNR decreases. Consider a stimulus consisting of speech plus interrupted noise. Assuming that the noise is square-wave modulated with a duty cycle of 50% and 100% modulation depth, the stimulus alternates between equal-duration blocks of speech-plus-noise (outside of the gaps) and speech only (in the gaps). As SNR decreases, the unprocessed stimulus alternates between blocks of essentially noise only (i.e., speech plus noise at a very low SNR) and speech-only, with the speech-only blocks contributing much less energy to the total stimulus. In a given band, TFS-processing removes global amplitude modulations and, in doing so, amplifies the energy of the speech-only blocks to match the energy of the noise-dominated blocks. This amplification of the speech-only blocks increases the time-average SNR of the entire stimulus. This increase in SNR may render the TFS-processed, interrupted-noise stimulus more intelligible than the equivalent-SNR continuous noise stimulus, which does not benefit from a similar amplification of speech-only blocks. This may explain the observed HI-listener MR for the T1 and T4 conditions as well as the derivative R1 and R4 conditions. It may also explain the fact that, in interrupted noise, five of the ten HI listeners exhibited performance in T1 that actually exceeded the performance in U while being tested at the same SNR. In other words, this suggests that the increase in MR demonstrated by HI listeners with T and R speech (relative to U or E speech) might be the consequence of an increase in the audibility of the speech in the valleys of the interrupted noise.

V. SUMMARY AND CONCLUSIONS

- The performance of the HI listeners was inferior to that of the NH listeners in terms of lower consonant-identification scores on the baseline conditions and the need for higher levels of SNR to achieve a given level of performance in continuous background noise.
- Patterns of performance suggest that HI listeners rely primarily on ENV cues for consonant recognition.
- For both NH and HI listeners, the intelligibility of TFS speech can be accounted for by RENVs on the basis of similar levels of performance and highly similar consonant-confusion patterns. However, providing TFS cues as narrow-band RENV cues did not benefit the HI listeners.
- NH listeners exhibited substantial benefits in the presence of the fluctuating compared to continuous noise maskers (masking release) for the unprocessed speech stimuli as well as for stimuli processed to retain mostly ENV or TFS cues.
- HI listeners exhibited masking release for the TFS and RENV speech types (T1, T4, R1, and R4) but generally not for the unprocessed and ENV speech stimuli. The masking release observed for TFS and RENV speech may be related to level effects associated with the manner in which the TFS processing interacts with the interrupted noise signal, rather than to the presence of TFS itself.

ACKNOWLEDGMENTS

This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award No. R01DC000117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. During this research, the leading author was a Postdoctoral Associate at the Research Laboratory of Electronics at MIT.

APPENDIX: PSYCHOMETRIC FUNCTIONS FOR U, E, AND T1 SPEECH

Additional data were obtained examining the effect of SNR on consonant recognition scores for three of the speech conditions (U, E, and T1) in continuous and interrupted noise. The methodology for this experiment was identical to that described for the main experiment with the following exceptions.

1. Subjects

Four young NH listeners participated in this experiment (mean age of 20.8 years, standard deviation of 2.2 years, including three of the young NH listeners from the main study and one additional new listener) as well as five of the HI listeners (HI-1, HI-2, HI-5, HI-8, and HI-9).

2. Signal-to-noise ratios

In addition to the SNR tested for each speech condition in the main experiment, consonant recognition scores were obtained with at least two other values of SNR at each condition. The supplemental data were obtained after the main experiment was completed.

3. Speech materials

The supplemental experiment employed the "test" set of /a/-C-/a/ speech stimuli. Five runs were presented using random-order presentation of the 64 syllables in the test set. The first run was discarded and the final four runs were used to calculate percent-correct performance on a given speech/noise condition.

4. Data summary and analysis

Percent-correct scores were plotted as a function of SNR and sigmoidal functions were fit to the data. The lower end of the function was limited by chance performance on the 16-alternative forced-choice procedure (i.e., 6.25%-Correct) and the higher end of the function was assumed to reach asymptote at the level of performance measured for the baseline condition in the main experiment.

5. Results

The results are shown in Fig. 8 for the NH group [mean results over the four listeners, panel (a)] and each of the five HI listeners [panels (b)–(f)]. The sigmoidal fits to the data points are shown for each of the six conditions (3 speech conditions × two noises) in each panel. For the NH listeners, the shape of the sigmoidal functions was similar for U, E, and T1 speech in continuous background noise with a shift to higher SNR values for E compared to U and for T1 compared to U and E. The slopes of the U and T1 functions in interrupted noise were extremely shallow and that of the E function was somewhat shallower than in continuous noise. The results of HI-1 were similar to those of the NH group with a several dB shift to higher SNR values. With the exception of HI-2, the remaining HI listeners had the highest performance on T1 in interrupted noise and the lowest performance on T1 in continuous noise with close spacing of the remaining functions. The results for HI-2 indicated similar functions for all conditions with the exception of a shallower slope for U in interrupted noise and least sensitive performance on T1 in continuous noise.

6. Discussion

Previous studies have noted the dependence of MR on the SNR at which is measured (e.g., Bernstein and Grant, 2009) and this effect can be observed in the psychometric functions shown in Fig. 8. Specifically, when the slopes of the functions for continuous and interrupted noise for a given type of speech are not the same, the difference in performance between the two curves will vary as a function of SNR. The choice of SNR at which to calculate MR is somewhat arbitrary, although it is obvious that an SNR should be selected at which performance is reasonably above the lower bound and below the upper bound of the psychometric function, at least for continuous noise.

In the main experiment, the decision was made to measure performance for each listener in all speech/noise conditions at a single SNR selected to yield approximately 50%-Correct performance for unprocessed speech in continuous noise. As a result, MR was measured at different points along the respective psychometric functions for the different types of speech. However, as shown in Fig. 8, the selected SNR still meets the criterion of yielding performance that is

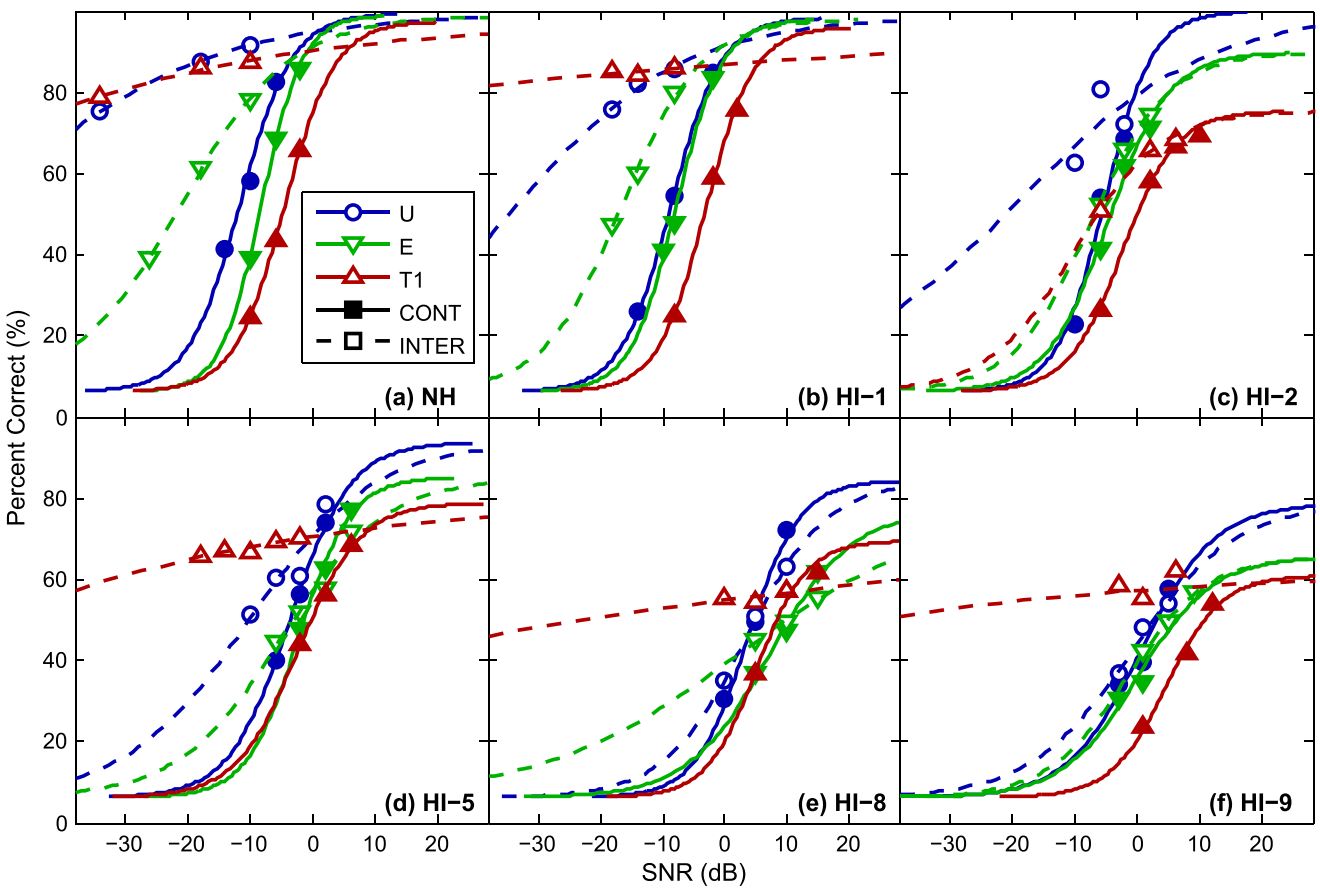

FIG. 8. (Color online) %-Correct scores plotted as a function of SNR in dB for U, E, and T1 speech in continuous and interrupted noise. (a) Mean data across four NH listeners. (b)–(f) Data for five individual HI listeners. The standard deviations of the individual %-Correct data points shown in these plots were all less than 11%. Also shown are sigmoid fits (see text for details) to the scores for each speech type and noise type.

above chance and below saturation for each speech type which confirms the merit of the MR and NMR results reported here.

[1]Note that the 300 Hz cut-off frequency of the low-pass filter used to estimate the signal envelope can cause aliasing when the envelope is subsequently used to modulate low-frequency tone carriers—even after re-filtering with the original bandpass filter. For the 64 VCVs and for all numbers of bands used, this level of the aliased component was always at least 36.5 dB below the non-aliased envelope component.

Apoux, F., Yoho, S. E., Youngdahl, C. L., and Healy, E. (**2013**). "Can envelope recovery account for speech recognition based on temporal fine structure?," POMA **19**, EL050072–EL050077.

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.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. **43A**, 647–677.

Desloge, J. G., Reed, C. M., Braida, L. D., Perez, Z. D., and Delhorne, L. A. (**2010**). "Speech reception by listeners with real and simulated hearing impairment: Effects of continuous and interrupted noise," J. Acoust. Soc. Am. **128**, 342–359.

Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York), pp. 239–247.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R. L., Griffin, A. M., and Oxenham, A. J. (**2012**). "Intelligibility of whispered speech in stationary and modulated noise maskers," J. Acoust. Soc. Am. **132**, 2514–2523.

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (**2006**). "Masking release for consonant features in temporally fluctuating background noise," Hear. Res. **211**, 74–84.

Ghitza, O. (**2001**). "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," J. Acoust. Soc. Am. **110**, 1628–1640.

Gilbert, G., Bergeras, I., Voillery, D., and Lorenzi, C. (**2007**). "Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues," J. Acoust. Soc. Am. **122**, 1336–1339.

Gilbert, G., and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444.

Gnansia, D., Jourdes, V., and Lorenzi, C. (**2008**). "Effect of masker modulation depth on speech masking release," Hear. Res. **239**, 60–68.

Gnansia, D., Pean, V., Meyer, B., and Lorenzi, C. (**2009**). "Effects of spectral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am. **125**, 4023–4033.

Heinz, M. G., and Swaminathan, J. (**2009**). "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," J. Assoc. Res. Otolaryngol. **10**, 407–423.

Hopkins, K., and Moore, B. C. J. (**2009**). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acoust. Soc. Am. **125**, 442–446.

Hopkins, K., and Moore, B. C. J. (**2011**). "The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise," J. Acoust. Soc. Am. **130**, 334–349.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2008**). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," J. Acoust. Soc. Am. **123**, 1140–1153.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2010**). "The effects of the addtiion of low-level, low-noise noise on the intelligibility of sentences processed to remove temporal envelope information," J. Acoust. Soc. Am. **128**, 2150–2161.

Léger, A. C., Desloge, J. G., Braida, L. D., and Swaminathan, J. (**2015**). "The role of recovered envelope cues in the identification of temporal-fine-structure speech for hearing-impaired listeners," J. Acoust. Soc. Am. **137**, 505–508.

Lorenzi, C., Debruille, L., Garnier, S., Fleuriot, P., and Moore, B. C. J. (**2009**). "Abnormal processing of temporal fine structure in speech for frequencies where absolute thresholds are normal," J. Acoust. Soc. Am. **125**, 27–30.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006a**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

Lorenzi, C., Husson, M., Ardoint, M., and Debruille, X. (**2006b**). "Speech masking release in listeners with flat hearing loss: Effects of masker fluctuation rate on identification scores and phonetic feature reception," Int. J. Audiol. **45**, 487–495.

Lorenzi, C., Wallaert, N., Gnansia, D., Léger, A. C., Ives, D. T., Chays, A., Garnier, S., and Cazals, Y. (**2012**). "Temporal-envelope reconstruction for hearing-impaired listeners," J. Assoc. Res. Otolaryngol. **13**, 853–865.

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Moore, B. C. J. (**2014**). *Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss* (World Scientific, Hackensack, NJ), pp. 81–102.

Moore, B. C. J., Peters, R. W., and Stone, M. A. (**1999**). "Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips," J. Acoust. Soc. Am. **105**, 400–411.

Oxenham, A. J., and Kreft, H. A. (**2014**). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," Trends Hear. **18**, 1–14.

Oxenham, A. J., and Simonson, A. M. (**2009**). "Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am. **125**, 457–468.

Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.

Shamma, S., and Lorenzi, C. (**2013**). "On the balance of envelope and tmeporal fine structure in the encoding of speech in the early auditory system," J. Acoust. Soc. Am. **133**, 2818–2833.

Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.

Sheft, S., Ardoint, M., and Lorenzi, C. (**2008**). "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, 562–575.

Shera, C. A., Guinan, J. J., Jr., and Oxenham, A. J. (**2002**). "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," Proc. Natl. Acad. Sci. U.S.A. **99**, 3318–3323.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.

Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**, 317–326.

Stone, M. A., and Moore, B. C. J. (**2014**). "On the near-existence of 'pure' energetic masking release for speech," J. Acoust. Soc. Am. **135**, 1967–1977.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Lang. Hear. Res. **28**, 455–462.

Swaminathan, J. (**2010**). "The role of envelope and temporal fine structure in the perception of noise degraded speech," Ph.D. dissertation, Purdue University, West Lafayette, IN.

Swaminathan, J., and Heinz, M. G. (**2012**). "Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise," J. Neurosci. **32**, 1747–1756.

Swaminathan, J., Reed, C. M., Desloge, J. G., Braida, L. D., and Delhorne, L. A. (**2014**). "Consonant identification using temporal fine structure and recovered envelope cues," J. Acoust. Soc. Am. **135**, 2078–2090.

Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.

Won, J. H., Lorenzi, C., Nie, K., Li, X., Jameyson, E. M., Drennan, W. R., and Rubinstein, J. T. (**2012**). "The ability of cochlear implant users to use temporal envelope cues recovered from speech frequency modulation," J. Acoust. Soc. Am. **132**, 1113–1119.

Won, J. H., Shim, H. J., Lorenzi, C., and Rubinstein, J. T. (**2014**). "Use of amplitude modulation cues recovered from frequency modulation for cochlear implant users when original speech cues are severely degraded," J. Assoc. Res. Otolaryngol. **15**, 423–439.

Zeng, F. G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y. Y., and Chen, H. (**2004**). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," J. Acoust. Soc. Am. **116**, 1351–1354.