Consonant identification using temporal fine structure and recovered envelope cues[a)]

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

(Received 30 May 2013; revised 29 January 2014; accepted 3 February 2014)

The contribution of recovered envelopes (RENVs) to the utilization of temporal-fine structure (TFS) speech cues was examined in normal-hearing listeners. Consonant identification experiments used speech stimuli processed to present TFS or RENV cues. Experiment 1 examined the effects of exposure and presentation order using 16-band TFS speech and 40-band RENV speech recovered from 16-band TFS speech. Prior exposure to TFS speech aided in the reception of RENV speech. Performance on the two conditions was similar (~50%-correct) for experienced listeners as was the pattern of consonant confusions. Experiment 2 examined the effect of varying the number of RENV bands recovered from 16-band TFS speech. Mean identification scores decreased as the number of RENV bands decreased from 40 to 8 and were only slightly above chance levels for 16 and 8 bands. Experiment 3 examined the effect of varying the number of bands in the TFS speech from which 40-band RENV speech was constructed. Performance fell from 85%- to 31%-correct as the number of TFS bands increased from 1 to 32. Overall, these results suggest that the interpretation of previous studies that have used TFS speech may have been confounded with the presence of RENVs.
 [http://dx.doi.org/10.1121/1.4865920]

PACS number(s): 43.71.Gv, 43.71.Es [MAS] Pages: 2078–2090

I. INTRODUCTION

It is well known that many hearing-impaired (HI) listeners who have little difficulty understanding speech in quiet backgrounds experience great difficulty in backgrounds containing interfering sounds (Kochkin, 1996; Takahashi *et al.*, 2007). Understanding speech in noise, restaurants, or group situations continues to be problematic for hearing-aid users, in spite of research efforts. When the interference is temporally fluctuating, most normal-hearing (NH) individuals are able to achieve substantial gains in intelligibility while most HI listeners do not (e.g., Desloge *et al.*, 2010). Recently, a number of investigators (e.g., Lorenzi *et al.*, 2006, 2009; Hopkins and Moore, 2007; Hopkins *et al.*, 2008) have shown that this may result from an inability of HI listeners to process the temporal fine structure (TFS), as opposed to the temporal envelopes (ENVs), of speech as well as NH listeners.

The speech waveform can be characterized as the sum of bandpass signals, each comprising a slowly varying amplitude component (ENV) that modulates a rapidly varying carrier component (TFS) (e.g., Flanagan, 1980). Envelope cues have been shown to be important for speech perception in quiet when provided in as few as four to eight frequency bands (Shannon *et al.*, 1995; Zeng *et al.*, 2004). Traditionally, TFS cues have been thought to play a major role in the perception of pitch for both pure and complex tones (for reviews, see Plack and Oxenham, 2005). Results from recent psychophysical studies suggest, however, that, in addition to pitch perception, TFS cues can also contribute to speech intelligibility (Lorenzi *et al.*, 2006; Gilbert and Lorenzi, 2006; Gilbert *et al.*, 2007; Sheft *et al.*, 2008).

To evaluate the role of TFS cues for speech perception, the vocoding technique has been used to isolate the TFS component of a band-limited signal from the ENV component to create TFS speech. In this technique, speech sounds are initially split into several contiguous frequency bands. TFS cues in each band are extracted either as the phase of the Hilbert analytic signal (Hilbert, 1912) or by dividing the bandpass signal by the envelope magnitude (at each instant in time). With this processing, the signal within each band becomes a constant-amplitude, frequency-modulated signal. The band signals are then re-combined to create TFS speech (e.g., Drullman, 1995; Smith *et al.*, 2002; Gilbert and Lorenzi, 2006).

The interpretation of perceptual studies that utilize TFS speech relies on the assumption that the TFS component can be completely isolated from the ENV component. However, narrowband filtering imposes constraints on the ability to isolate a sound's TFS component from its ENV component (Zwicker, 1962; Saberi and Hafter, 1995; Ghitza, 2001; Zeng *et al.*, 2004; Heinz and Swaminathan, 2009; also see Voelcker, 1966; Rice, 1973; Logan, 1977). When broadband speech is filtered through a set of narrowband filters (such as cochlear filters), the TFS component of the broadband speech gets converted into (recovered) envelopes (RENVs) (e.g., Ghitza, 2001).

Gilbert and Lorenzi (2006) conducted a systematic perceptual and modeling study to quantify the extent of envelope recovery from TFS speech. The ability of NH listeners

[a)]Portions of this research were presented at the 36th Midwinter Meeting of the Association for Research in Otolaryngology, Baltimore, MD, February 2013.
[b)]Author to whom correspondence should be addressed. Electronic mail: jswamy@mit.edu

 0001-4966/2014/135(4)/2078/13/$30.00

to identify 16 French consonants in /a/-C-/a/ syllables was examined for TFS and RENV speech. The TFS speech was generated with 1, 2, 4, 8, or 16 analysis bands and 30-band RENV speech was then generated from each of the 5 TFS speech conditions. For TFS speech, consonant intelligibility was high and showed little decrease in performance as the number of analysis bands increased from 1 to 16 (dropping from roughly 100%- to 90%-correct over this range). The intelligibility of the RENV speech was much lower, however, and decreased with an increase in the number of TFS bands from which it was generated (decreasing from roughly 60%- to 15%-correct as the number of analysis bands increased from 1 to 8 and remaining at 15%-correct for the 16-band condition). These results suggested that the RENV cues did not play a major role in the identification of narrowband TFS speech. Specifically, it was suggested that RENV cues did not contribute to the intelligibility of TFS speech generated from signals with bandwidths less than or equal to four times the normal auditory critical bandwidths. In their study, this corresponded to the number of frequency bands being equal to or greater than 8 over the frequency range 80 to 8020 Hz.

Following this result, subsequent studies with TFS speech have used stimuli that were created using a relatively large number of bands (typically 16), and the interpretations were based on the premise that the perception of TFS speech was unlikely to be affected by RENV cues (e.g., Lorenzi *et al.*, 2006; Gilbert *et al.*, 2007; Sheft *et al.*, 2008). However, neurophysiological and modeling results have shown that RENVs remain for TFS speech created with 16 bands (e.g., Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012). It should also be noted that the high levels of 16-band TFS speech intelligibility observed for NH listeners by Lorenzi *et al.* (2006) and Sheft *et al.* (2008) required extensive training/exposure. Taken together, it is not clear whether this high level of performance is due to the use of TFS cues (as is often assumed) or due to the use of RENV cues.

The goal of the current study was to systematically evaluate the role of RENVs in the perception of TFS speech by NH listeners. Three sets of experiments were designed to address this basic question:

The first set of experiments addressed the role of exposure and test order on consonant identification for TFS and RENV speech. This aspect of the study was motivated by differences in maximal levels of TFS performance observed across earlier studies [e.g., compare Lorenzi *et al.* (2006) with Swaminathan and Heinz (2012)] and by an absence of published evidence for comparisons of TFS and RENV speech controlling for test order and training time. It was hypothesized that performance would improve with increased exposure to both TFS and RENV speech. There was no explicit hypothesis regarding the effect of the presentation order of the two types of speech stimuli.

The second experiment examined the effects of varying the number of ENV bands recovered from narrowband TFS speech. The number of RENV bands was systematically reduced in a manner that widened the individual bands to simulate the effects of sensorineural hearing loss on envelope reconstruction. A well-known consequence of sensorineural hearing loss is reduced frequency selectivity which results from the broadening of the peripheral auditory filters (e.g., Liberman and Dodds, 1984; Glasberg and Moore, 1986). In a modeling study, Heinz and Swaminathan (2009) simulated the broadening of auditory filters that typically occurs with hearing loss and showed that such broadening resulted in a reduction in the degree to which ENV cues can be recovered from TFS speech. Such broadening of peripheral auditory filters could have an effect on the envelope reconstruction from TFS speech (e.g., Heinz and Swaminathan, 2009; Lorenzi *et al.*, 2012). Thus, it was expected that performance would decrease with a reduction in the number of RENV bands.

The third experiment explored the effects of varying the number of bands in the TFS speech signals that were used for constructing 40-band RENV speech. The results of Gilbert and Lorenzi (2006) suggested that RENV cues did not play a major role in consonant perception when the bandwidth of the filters used to create the TFS speech was narrower than 4 times the bandwidth of a normal auditory filter (i.e., number of TFS bands $\geq$8 for frequencies spanning 80 to 8020 Hz). This experiment was conducted for further exploration of the results of Gilbert and Lorenzi (2006) regarding the role of RENVs in the reception of TFS speech. Specifically, it examined the effect of varying the number of bands in the TFS speech from which a 40-band RENV signal was constructed. Our choice of 40 bands for creating RENV speech leads to bands whose widths are less than 1 ERB_N (Glasberg and Moore, 1990). This choice of bandwidth is in agreement with the findings of Shera *et al.* (2002) who suggested that human cochlear filters are sharper than the standard behavioral measures.

Overall, it was hypothesized that the interpretation of previous results that have used TFS speech may have been confounded by the presence of RENV cues.

II. GENERAL METHODOLOGY

A. Subjects

A total of 30 young NH subjects (12 males, 18 females) who were native speakers of American English were employed across the three experiments. Subjects provided informed consent, and a clinical audiogram was obtained to screen for normal hearing, defined as 15 dB hearing level (HL) or better at octave frequencies in the range of 250 to 8000 Hz. They ranged in age from 18 to 25 yrs with a mean age of 19.9 yrs. All testing took place in the right ear, except for one subject who was tested in her left ear due to a threshold of 20 dB HL at 8000 Hz in the right ear. All subjects were paid for their participation in the study.

B. 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, j, m, n, r, l/. These recordings were taken from the corpus of Shannon *et al.* (1999). The stimulus set employed in all the experiments consisted of one utterance of each of the 16 syllables from two male and two female

speakers for a total of 64 stimuli. The recordings were digitized with 16-bit precision at a sampling rate of 32 kHz and presented at a level of either 68 dB sound pressure level (SPL) (Experiment 1A) or 70 dB SPL (Experiments 1B, 2, and 3).

C. Experimental procedure

Subjects were tested on their ability to identify the set of 16 consonants using a one-interval 16-alternative forced-choice procedure without correct-answer feedback. On each trial of the experiment, one of the stimuli from the set of 64 syllables was selected and processed according to one of the stimulus processing conditions described below. 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 appeared on a computer monitor following each stimulus presentation and the response was selected using a computer mouse. No time limit was imposed on the subjects' responses. Each experimental run consisted of 64 trials derived from a different random-order presentation (without replacement) of the 64 syllables in the stimulus set with all stimuli processed according to the same stimulus-processing condition. Each run lasted roughly 4 to 7 min depending on the subject's response time. The experiments consisted of multiple runs (between 21 and 62, depending upon the experimental condition) conducted under varying stimulus-processing conditions. Test sessions lasted 2 h including breaks, and each experiment required multiple sessions. Additional procedural details specific to each experiment are provided in Secs. III (Experiment 1), IV (Experiment 2), and V (Experiment 3) below.

Experiments were controlled by a desktop PC equipped with a high-quality, 24-bit PCI sound card (E-MU 0404 by Creative America, Milapita, CA). The level-calibrated speech stimuli were played out using MATLAB™ (Mathworks, Natick, MA); passed through a Tucker-Davis (TDT, Alachua, FL) 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 (Old Lyme, CT) 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.

D. Stimulus processing

Prior to presentation to the listener, the speech stimuli were processed according to one of the following three conditions:

(1) *Intact speech*. Intact speech stimuli were created by using the unmodified samples directly.

(2) *TFS speech*. TFS speech stimuli were created according to the methods described in Gilbert and Lorenzi (2006) and Lorenzi *et al*. (2006). This involved bandpass filtering the unmodified samples into N_{TFS} bands of equal bandwidth on a log frequency scale spanning 80 to 8020 Hz. The Hilbert transform (Hilbert, 1912) was used to decompose each bandpass signal into envelope (i.e., the magnitude of the Hilbert analytic signal) and fine-structure (i.e., the cosine of the phase of the Hilbert analytic signal) components. The envelope component was discarded and the fine-structure component was normalized to the long-term average energy of the original bandpass signal. The resulting normalized fine-structure components for all bands were then summed to yield the TFS speech. The number of bands, N_{TFS}, was fixed at 16 for all TFS-only conditions and will be referred to throughout the paper as TFS(16) speech.

It should be noted that the current study has chosen this method for generating TFS speech in order to facilitate comparisons with studies conducted using the same method (e.g., Lorenzi *et al*., 2006; Gilbert and Lorenzi, 2006). Other researchers (e.g., Hopkins *et al*., 2010) have identified a potential problem with this method in that low-level portions of the speech signal can be amplified by high gain in order to achieve the uniform amplitude of TFS speech. Such amplification can negatively affect intelligibility, e.g., through excessive spectral and temporal masking. Hopkins *et al*. (2010) suggested the addition of low-noise-noise before TFS processing to limit this amplification. Although this modified processing might improve intelligibility both on the TFS speech itself as well as RENV speech generated from this TFS speech (see below), we chose to follow the original method of Lorenzi *et al*. (2006).

(3) *RENV speech*. RENV speech stimuli were created by first generating a TFS speech stimulus according to the method described above. This TFS speech stimulus was then bandpass filtered into N_{RENV} 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 RENV component was estimated by full-wave rectification followed by processing with a 300-Hz low-pass filter (sixth order Butterworth) and this was then used to modulate a tone carrier at the center frequency of the band. 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 number of TFS bands, N_{TFS}, ranged from 1 to 32 and the number of RENV bands, N_{RENV}, ranged from 8 to 40 depending upon the experimental condition. These processing conditions will be referred to using the notation RENV(N_{TFS}, N_{RENV}). For example, RENV(16,40) speech refers to RENV speech created by recovering 40 bands of envelopes from 16-band TFS speech.

For the RENV(16,40) condition, care was taken to ensure that no fine structure was introduced in the RENV signals by filter ringing. Zeng *et al*. (2004) suggested that such filter ringing artifacts (in which the TFS component leaks into the ENV component) may occur with narrowband processing for a large number of bands (e.g., 64). The neural metrics developed by Heinz and Swaminathan (2009) were used to compute the similarity in TFS coding between TFS(16) and RENV(16,40) for auditory-nerve fiber frequencies ranging from 200 Hz to 2 kHz. Across all frequencies, the cross-correlation in TFS between the TFS(16) and

RENV(16,40) was minimal (∼0.1) consistent with no leakage of TFS into the RENV signals.

E. Data analysis

For each subject, a percent-correct score was calculated for each 64-trial run where chance performance on the 16-item set was 6.25%-correct. Stimulus-response confusion matrices were generated for each run and added across the final ten runs (Experiment 1A) or across the final five runs (Experiments 1B, 2, and 3) for each subject and each experimental condition. For each experiment and condition, these matrices were also added across subjects to compute overall percent-correct scores and measures of relative overall and feature information transfer (Miller and Nicely, 1955; Wang and Bilger, 1973; Houtsma, 1983). The consonant features were voicing (voiced versus unvoiced), manner of articulation (constriction versus non-constriction), place of articulation (front versus middle versus back), and nasality (nasal versus non-nasal) as defined by Swaminathan and Heinz (2012, Table I). The confusion matrices were also subjected to a form of metric multidimensional scaling analysis (Braida, 1991).

Repeated-measures analyses of variance (ANOVAs) were conducted for each of the experiments using arcsin-transformed percent-correct scores of individual subjects on each test condition included in a given experiment.

III. EXPERIMENT 1

A. Experiment 1A

Procedure. Experiment 1 examined the role of exposure and test order on the reception of TFS(16) and RENV(16,40) speech. Ten subjects (4 male, 6 female, mean age of 20.6 yrs) participated in the first stage of this experiment (Experiment 1A). All subjects began the experiment by completing one 64-trial test run with Intact speech for familiarization with the test procedure. Five subjects (Group 1) then completed 20 runs of the TFS(16) condition followed by 20 runs of the RENV(16,40) condition. The remaining 5 subjects (Group 2) were tested in the order of 20 runs of RENV(16,40), followed by 20 runs of TFS(16). The experiment typically required two 2-h sessions to complete. In addition, the Group 2 subjects completed a third 2-h session during which they were re-tested on 20 runs of the

TABLE I. Summary of experimental conditions and results.

	Grp.	No. subj.	Test order	Condition	Range of mean %-C across subjects		Mean (std dev.) (%-C)
					Min.	Max.	
Experiment 1A	1	5	1	TFS(16)	40.8	68.0	51.8 (10.2)
			2	RENV(16,40)	41.4	62.6	50.5 (8.7)
	2	5	1	RENV1(16,40)	8.0	38.4	19.4 (15.0)
			2	TFS(16)	45.3	74.2	60.6 (12.0)
			3	RENV2(16,40)	47.8	65.5	53.8 (7.1)
Experiment 1B	1	3	Alternating TFS / RENV	TFS(16)	38.1	56.3	46.8 (9.1)
				RENV(16,40)	34.7	55.0	42.8 (10.7)
	2	3	Alternating RENV / TFS	RENV(16,40)	43.4	56.3	48.6 (6.7)
				TFS(16)	49.1	60.0	52.9 (6.1)
Experiment 2	1	4	1	TFS(16)	45.3	63.8	56.0 (9.0)
			2	RENV(16,8)	7.2	12.2	9.6 (2.2)
			3	RENV(16,16)	6.6	11.6	9.1 (2.6)
			4	RENV(16,32)	6.6	42.5	17.9 (16.7)
			5	RENV(16,40)	5.3	61.9	29.1 (26.8)
	2	4	1	TFS(16)	28.4	54.1	40.9 (10.5)
			2	RENV(16,40)	33.1	48.1	39.4 (6.3)
			3	RENV(16,32)	11.9	40.9	28.0 (12.8)
			4	RENV(16,16)	6.6	18.8	14.3 (5.4)
			5	RENV(16,8)	7.5	14.4	10.3 (2.9)
Experiment 3	1	3	1	TFS(16)	38.1	56.9	44.8 (10.5)
			2	RENV(1,40)	78.8	87.8	84.2 (4.8)
			3	RENV(2,40)	73.4	83.8	79.8 (5.6)
			4	RENV(4,40)	65.9	80.9	75.3 (8.2)
			5	RENV(8,40)	54.7	81.9	70.3 (14.0)
			6	RENV(16,40)	46.3	60.6	54.7 (7.5)
			7	RENV(32,40)	21.9	45.9	34.5 (12.1)
	2	3	1	TFS(16)	55.0	77.5	64.5 (11.7)
			2	RENV(32,40)	20.6	33.1	27.9 (6.5)
			3	RENV(16,40)	47.8	60.3	54.5 (6.3)
			4	RENV(8,40)	65.5	73.4	70.8 (4.5)
			5	RENV(4,40)	77.2	82.5	80.5 (2.9)
			6	RENV(2,40)	90.3	95.6	93.2 (2.7)
			7	RENV(1,40)	93.1	95.9	94.5 (1.4)

RENV(16,40) condition. The first set of RENV(16,40) data obtained on Group 2 is referred to as RENV1(16,40) and the second set as RENV2(16,40). See Table I for a summary of the subject groups and conditions tested.

Results. The results of Experiment 1A are shown in Fig. 1 for the five subjects (Group 1) who were tested in the order of TFS(16) followed by RENV(16,40). Individual-subject percent-correct scores are plotted in Fig. 1 as a function of run number for TFS and RENV conditions as well as the average across subjects. In addition to the individual-run performance, each panel shows the average percent-correct score across the final ten runs. In general, performance improved over the course of the first five to ten runs of the TFS condition and stabilized for the remaining TFS runs. Performance began at a higher level for the RENV condition and stabilized in fewer runs.

Summary data are also provided in Table I which gives the range, mean, and standard deviation (s.d.) of the percent-correct scores across the Group 1 subjects. Mean scores were quite similar for the two conditions: 51.8% (s.d. of 10.2%) for TFS(16) and 50.5% (s.d. of 8.7%) for RENV(16,40). A one-way repeated-measures ANOVA to test for the effect of condition indicated no significant difference between TFS(16) and RENV(16,40) scores [$F(1,4) = 0.184$, $p = 0.69$].

The results of Experiment 1A for the five Group 2 subjects, tested in the order of RENV1(16,40) followed by TFS(16), are plotted in Fig. 2 and summarized in Table I. A different pattern of behavior was observed for these subjects compared to those tested in the reverse order shown in Fig. 1. For each of the listeners shown in Fig. 2, RENV1 performance was lower than TFS performance. Furthermore, for three of these subjects, RENV1 scores remained roughly at chance (6.25%) levels throughout the 20 runs of exposure. When these same listeners were re-tested on the RENV2(16,40) condition (following exposure to TFS speech), however, the test scores improved dramatically and were similar to those obtained under the TFS condition. Mean scores over the final ten runs of each condition were 19.4% (s.d. of 15.0%) for RENV1, 60.6% (s.d. of 12.0%) for TFS, and 53.8% (s.d. of 7.1%) for RENV2 averaged across subjects. A repeated-measures ANOVA conducted on the RENV1, TFS, and RENV2 scores indicated a significant effect of condition [$F(2,8) = 14.44$, $p = 0.002$, effect size $\eta^2 = 0.73$]. A *post hoc* Tukey-Kramer test indicated that the TFS and RENV2 conditions were not significantly different from each other and had significantly higher scores than the RENV1 condition.

A formal comparison of the Group 1 results for TFS(16) and RENV(16,40) with the Group 2 results for TFS(16) and

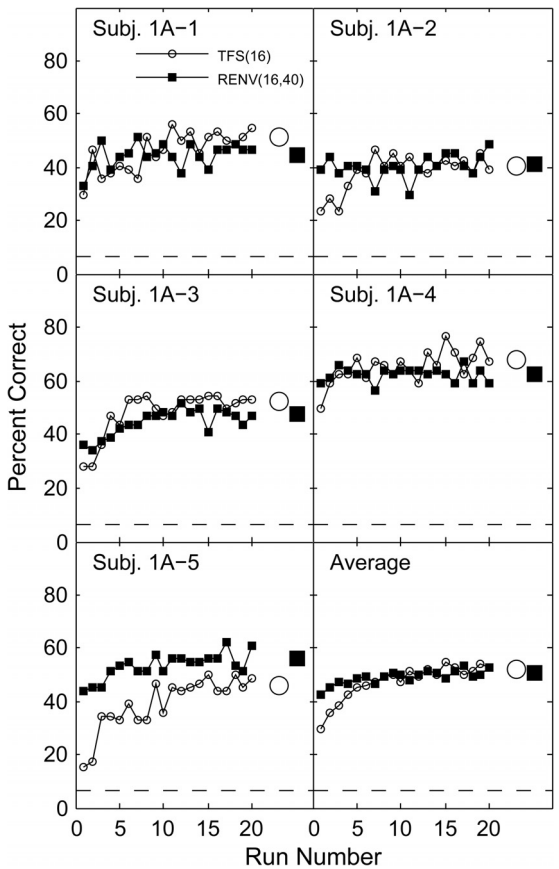

FIG. 1. Percent-correct score plotted as a function of run number for five individual subjects and the average across subjects in Experiment 1A, Group 1: TFS(16) followed by RENV(16,40) speech. For each subject, the mean performance for the final ten runs of each processing type is plotted to the right of the individual-run data. Chance level (1/16) is indicated by the dashed line.

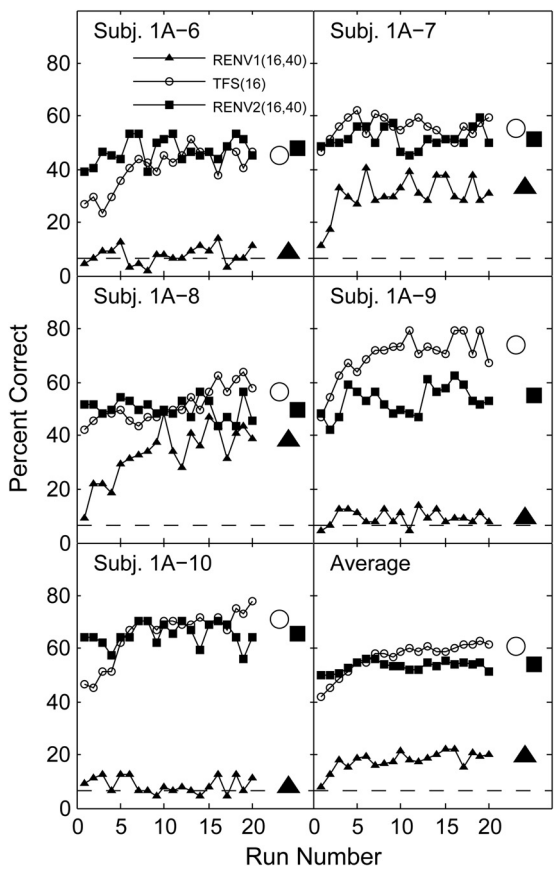

FIG. 2. Percent-correct score plotted as a function of run number for five individual subjects and the average across subjects in Experiment 1A, Group 2: RENV1(16,40) followed by TFS(16) followed by RENV2(16,40) speech. For each subject, the mean performance for the final ten runs of each processing type is plotted to the right of the individual-run data. Chance level (1/16) is indicated by the dashed line.

RENV1(16,40) was conducted using a repeated-measures ANOVA with a between-subjects variable of group and a within-subjects variable of condition. The results indicated that the group effect barely missed significance [$F(1,8) = 5.26$, $p = 0.051$, partial $\eta^2 = 0.40$], although significance was achieved for the effects of condition [$F(1,8) = 15.56$, $p = 0.004$, partial $\eta^2 = 0.66$], and the interaction between condition and subjects within groups [$F(1,8) = 13.86$, $p = 0.006$, partial $\eta^2 = 0.63$]. The interaction effect arises from the fact that Group 2 performed better on TFS than Group 1 (61%- versus 52%-correct) but worse on RENV (19%- versus 50%-correct). Overall, the results suggest that: (1) presentation order had a significant effect on the perception of TFS and RENV speech and (2) prior exposure to TFS speech aids in the perception of RENV speech.

B. Experiment 1B

Procedure. To follow up the results indicating an effect of the order in which TFS and RENV signals were presented in Experiment 1A, additional data were obtained in Experiment 1B with a new set of six subjects (3 male, 3 female, mean age of 20.3 yrs). Following one initial run with Intact speech, these listeners completed a total of 20 additional runs that alternated between individual runs of TFS(16) and RENV(16,40) conditions. Three of the subjects began with TFS speech (Group 1) while the other three began with RENV speech (Group 2). The experiment required one 2-h session to complete (except for one subject who required two sessions). See Table I for a summary of the subject groups and conditions tested.

Results. The results of Experiment 1B are shown in Fig. 3 and summarized in Table I. These results indicate a carryover in performance from one run to the next (regardless of processing condition) over the first five to ten runs with performance stabilizing over the final ten runs. For Group 1, the mean over the final five runs of each condition was 46.8%-correct (s.d. of 9.1%) for TFS(16) and 42.8% (s.d. of 10.7%) for RENV(16,40). For Group 2, these means were 52.9% (s.d. of 6.1%) for TFS(16) and 48.6% (s.d. of 6.7%) for RENV(16,40). A repeated-measures ANOVA was conducted using a within-subjects variable of group (test order) and a between-subjects variable of condition. A significant effect of condition was observed [$F(1,4) = 19.60$, $p = 0.01$, partial $\eta^2 = 0.83$] but not group/test order [$F(1,4) = 0.79$, $p = 0.42$] or the interaction between condition and group [$F(1,4) = 0.02$, $p = 0.89$]. This result suggests that further exposure to RENV speech (i.e., more than the ten runs provided) may have been required to improve performance levels to those observed on TFS speech (as was seen in Experiment 1A where the number of RENV runs ranged from 20 to 40).

C. Discussion

The results of Experiment 1A indicate that prior exposure to TFS speech aids in the ability to perceive RENV speech. All listeners in this experiment were able to identify TFS speech at levels substantially greater than chance; however, their performance on RENV speech showed a dependence on prior exposure to TFS speech. Those subjects receiving TFS prior to RENV speech performed comparably on both conditions following roughly 1.5 h exposure to each stimulus type, while those subjects who received RENV first performed substantially worse compared to their subsequent performance on TFS speech. In fact, three of the five subjects who were tested in the order of RENV followed by TFS were never able to advance beyond chance levels of performance on RENV speech. Although no overall group effect was observed in comparing the TFS and RENV scores of Group 1 with the TFS and RENV1 scores of Group 2, there was a significant interaction between condition and group arising from the better performance of Group 2 compared to Group 1 on TFS speech and the opposite pattern on RENV speech. When RENV was retested in Group 2 following exposure to TFS speech, however, performance on RENV speech increased dramatically and was not

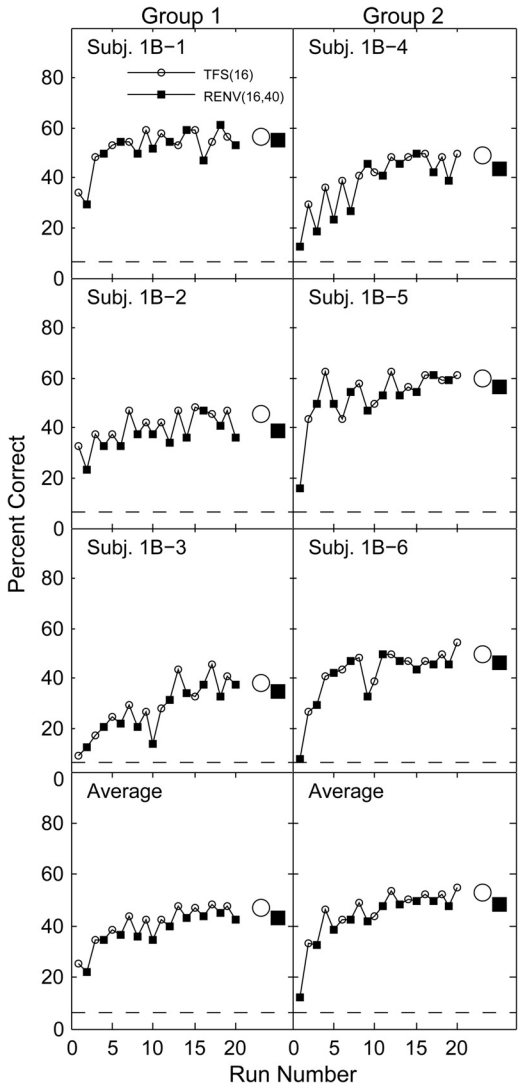

FIG. 3. Percent-correct score plotted as a function of run number for subjects tested in Experiment 1B with alternating runs of TFS(16)/RENV(16,40) speech (left column) and with alternating runs of RENV(16,40)/TFS(16) speech (right column). Average performance across subjects is also shown for each test order. For each subject, the mean performance for the final five runs for each processing type is plotted to the right of the individual-run data. Chance level (1/16) is indicated by the dashed line.

significantly different from that obtained on TFS speech. It should be noted, however, that the power of the statistical tests employed here to detect potentially small but significant differences between the TFS and RENV speech conditions is limited based on the relatively small number of subjects (three to five) tested within each group. The results of Experiment 1B demonstrate continuity in the learning process for TFS and RENV speech when individual runs were alternated between the two conditions. Taken together, these results suggest that exposure to TFS speech can prime listeners for making use of cues present in RENV speech, consistent with the findings of Swaminathan and Heinz (2012) that indicated an interaction between TFS and RENV cues.

IV. EXPERIMENT 2

A. Procedure

Experiment 2 was designed to investigate the effect of the number of RENV bands recovered from 16-band TFS speech on the intelligibility of RENV speech. A decrease in the number of RENV bands (accompanied by a subsequent increase in bandwidth) may reflect the increased width of auditory critical bands observed in listeners with sensorineural hearing loss. This experiment employed 8 subjects (3 male, 5 female, mean age of 19.6 yrs). Following an initial familiarization run with Intact speech, subjects completed ten runs of the TFS(16) condition and then proceeded to ten runs of each RENV(16,N_{RENV}) condition where $N_{RENV} = 8$, 16, 32, and 40. Four subjects were tested in increasing order of N_{RENV} (Group 1) and the remaining four subjects were tested in decreasing order of N_{RENV} (Group 2). The experiment required two to four 2-h test sessions to complete. The subject groups and conditions are summarized in Table I.

B. Results

The results of Experiment 2 are summarized in Fig. 4 and in Table I. The top panel of Fig. 4 shows the RENV(16, N_{RENV}) percent-correct scores for individual subjects (and means across subject) as functions of N_{RENV}. Also provided in this panel (at the far right) are the percent-correct scores for TFS(16) speech. The mean data indicate a decrease in RENV performance with a decrease in N_{RENV}. Averaged across the Group 1 subjects, the mean TFS(16) score was 56.0%-correct (s.d. of 9.0%) and RENV scores ranged from 9.6%-correct (s.d. of 2.2%) for RENV(16,8) to 29.9%-correct (s.d. of 26.8%) for RENV(16,40). Averaged across the Group 2 subjects, the mean TFS(16) score was 40.9%-correct (s.d. of 10.5%) and RENV scores ranged from 10.3%-correct (s.d. of 2.9%) for RENV(16,8) to 39.4% (s.d. of 6.3%) for RENV(16,40). On average across the eight subjects, performance decreased from 34.2%- to 22.9%- to 11.7%- to 10.0%-correct as N_{RENV} decreased from 40 to 32 to 16 to 8. Two of the subjects (both from Group 1), however, performed at chance on all of the RENV conditions despite an ability to perform the TFS listening task comparably to the other subjects. For the remaining six subjects who did show some ability to use RENVs, the mean score on the

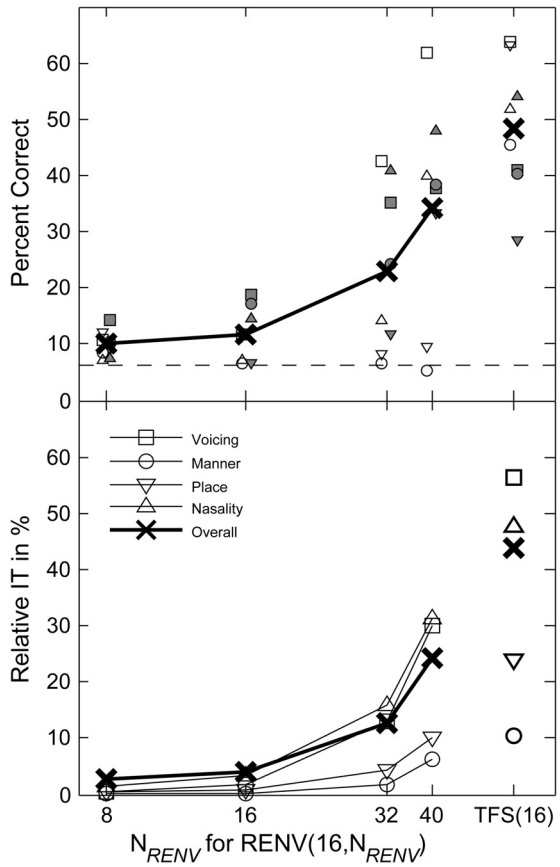

FIG. 4. Top panel: Mean percent-correct score across the final five runs of testing for subjects in Experiment 2 plotted as a function of N_{RENV} for RENV(16, N_{RENV}) speech. Mean scores obtained with TFS(16) speech are also provided on the right. Four individual subjects tested in order of increasing N_{RENV} are shown by unfilled data points and four individual subjects tested in order of decreasing N_{RENV} are shown by the offset filled data points. Mean performance across all eight subjects is also shown by the large X symbols which are connected by thick black lines. Chance level (1/16) is indicated by the dashed line. Lower panel: Voicing, manner, place, and nasality feature information transfer and relative overall information transfer across all eight subjects plotted as a function of N_{RENV}.

RENV(16 40) condition (43.2%-correct) was similar to that on the TFS(16) condition (46.6%-correct).

A repeated-measures ANOVA was conducted with the between-subjects variable of group/test order and the within-subjects variable of condition. The ANOVA indicated a significant effect of number of bands in the RENV speech condition [$F(3,18) = 9.44$, $p = 0.0006$, partial $\eta^2 = 0.61$] but not group [$F(1,6) = 1.47$, $p = 0.27$] or the interaction between group and condition [$F(3,18) = 0.51$, $p = 0.68$]. *Post hoc* Tukey-Kramer tests were conducted on repeated-measures ANOVAs of each group separately. For Group 1, there was no main effect of condition and thus no differences between any pairs of conditions. For Group 2, a significant effect of condition was observed and the Tukey-Kramer test indicated significant differences in scores between the $N_{RENV} = 40$ and the $N_{RENV} = 16$ and 8 conditions and between the $N_{RENV} = 32$ and the $N_{RENV} = 16$ and 8 conditions. No other pairwise comparisons reached significance.

The bottom panel of Fig. 4 summarizes the results of overall and relative unconditional feature information-transfer

(IT) analyses (Miller and Nicely, 1955) on each of four consonant features: voicing, manner, place, and nasality. These IT analyses were performed based on the results of the final five runs for each condition summed over the eight subjects. These analyses show a decrease in relative overall IT from 34.3% to 10.0% as N_{RENV} decreased from 40 to 8 bands. For RENV(16,40) speech, the scores for voicing and nasality were similar (roughly 50% relative feature IT) and higher than those for manner and place (roughly 10% to 20%). As N_{RENV} decreased to 32, voicing and nasality feature IT was reduced to 20%–25% while manner and place scores both dropped to below 10%. For N_{RENV} set to 16 and 8, no appreciable IT was observed either for overall performance or on any of the consonant features. In general, IT scores for RENV(16,40) speech were similar to those observed for TFS(16) speech.

C. Discussion

As the bandwidth used to recover envelopes from 16-band TFS speech increased (with a corresponding decrease in the number of recovered bands themselves), there was a rapid drop in the ability to understand RENV speech. Performance with RENV(16,40) speech was for most subjects similar to their performance on TFS(16) speech but was little better than chance for an RENV(16,16) signal. Thus, the ability to use RENV cues suffered with an increase in the bandwidth used for envelope recovery from TFS(16) speech. The effect shown here may be related to the difficulty experienced by listeners with sensorineural hearing loss in understanding TFS speech (e.g., see Lorenzi *et al.*, 2006, 2009; Hopkins and Moore, 2007; Hopkins *et al.*, 2008). If NH listeners make use of RENVs when listening to TFS speech, and if this ability is related to the filter bandwidth used for envelope recovery, then this suggests that the broadened cochlear filters of HI listeners may limit their ability to extract RENV cues from TFS speech.

V. EXPERIMENT 3

A. Procedure

This experiment, which examined the effect of varying the number of bands in the TFS speech from which a 40-band RENV signal was constructed, employed six subjects (2 male, 4 female; mean age of 19.0 yrs). Following an initial familiarization run with Intact speech, subjects completed 10 runs of the TFS(16) condition and then proceeded to 10 runs of RENV(N_{TFS}, 40) conditions for $N_{TFS} = 1$, 2, 4, 8, 16, and 32. Three subjects were tested in increasing order of N_{TFS} (Group 1) and the remaining three subjects were tested in decreasing order of N_{TFS} (Group 2). The experiment required three or four 2-h sessions to complete. The subject groups and conditions are summarized in Table I.

B. Results

The results of Experiment 3 are summarized in Fig. 5 and in Table I. The top panel of Fig. 5 shows the RENV(N_{TFS}, 40) percent-correct scores for individual subjects (and means across subject) as a function of N_{TFS}. Also

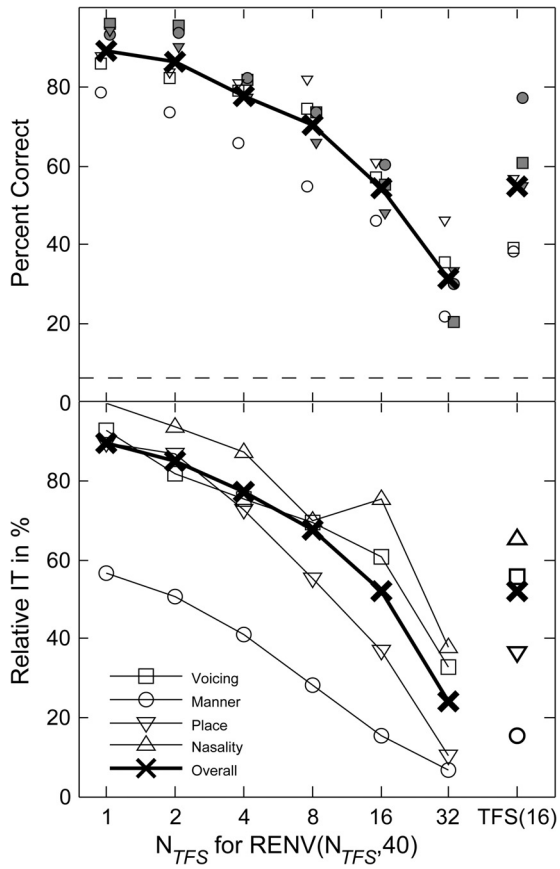

FIG. 5. Top panel: Mean percent-correct score across the final five runs of testing for subjects in Experiment 3 plotted as a function of N_{TFS} for RENV(N_{TFS}, 40) speech. Mean scores obtained with TFS(16) speech are also provided on the right. Three individual subjects tested in order of increasing N_{TFS} are shown by unfilled data points and three individual subjects tested in order of decreasing N_{TFS} are shown by the offset filled data points. Mean performance across all six subjects is also shown by the large X symbols which are connected by thick black lines. Chance level (1/16) is indicated by the dashed line. Lower panel: Voicing, manner, place, and nasality feature information transfer and relative overall information transfer across all subjects plotted as a function of N_{TFS}.

shown, at the far right, are the scores for TFS(16) speech. The results indicate a systematic decrease in performance with an increase in N_{TFS} accompanied by a more rapid drop in performance when N_{TFS} exceeds 8. Effects were similar across the six subjects. For the Group 1 subjects tested with increasing order of N_{TFS}, mean percent-correct scores ranged from 84.2% (s.d. of 4.8%) for $N_{TFS} = 1$ to 34.5%-correct (s.d. of 12.2%) for $N_{TFS} = 32$. For the Group 2 subjects tested in the opposite order, mean percent-correct scores ranged from 94.5% (s.d. of 1.4%) for $N_{TFS} = 1$ to 27.9%-correct (s.d. of 6.5%) for $N_{TFS} = 32$. Means across all subjects were 89.3%-, 86.5%-, 77.9%-, 70.6%-, 54.6%-, and 31.1%-correct for $N_{TFS} = 1$, 2, 4, 8, 16, and 32, respectively. Mean performance on TFS(16) speech (54.6%-correct) was identical to that obtained on the RENV(16,40) speech (54.6%).

A repeated-measures ANOVA was conducted with a between-subjects variable of group/test order and a within-subjects variable of condition on the data of the six subjects. A significant effect was found for condition [$F(3,18) = 9.44$, $p < 0.0001$, partial $\eta^2 = 0.98$] and for the interaction between group and condition [$F(5,20) = 8.32$, $p < 0.0001$, partial

$\eta^2 = 0.67$] but not for group [$F(1,4) = 1.11$, $p = 0.35$]. The interaction effect can be seen in Fig. 5: Group 1 subjects (who were tested in order of increasing N_{TFS}) did worse than Group 2 subjects (tested in the reverse order) at the lowest values of N_{TFS} but better at the highest values. *Post hoc* Tukey-Kramer testing of comparisons among conditions was conducted on each group separately based on a repeated-measures ANOVA. For the Group 1 subjects tested in increasing order of N_{TFS}, all pairs of conditions were significantly different with the following exceptions: $N_{TFS} = 1$, 2; $N_{TFS} = 2$, 4; $N_{TFS} = 2$, 8; and $N_{TFS} = 4$, 8. For the Group 2 subjects tested in decreasing order of N_{TFS}, all pairs of conditions were significantly different with the exception of $N_{TFS} = 1$, 2 and $N_{TFS} = 4$, 8. Thus, performance was significantly worse in both groups for $N_{TFS} = 32$ compared to $N_{TFS} = 16$, and for $N_{TFS} = 16$ compared to $N_{TFS} = 8$.

The bottom panel of Fig. 5 summarizes the results of overall and relative unconditional feature IT analyses on each of four consonant features: voicing, manner, place, and nasality. These IT analyses were performed based on the results of the final five runs for all six subjects. These analyses indicate a decrease in relative overall IT from 90% to 34% as N_{TFS} increased from 1 to 32. For the broadband condition ($N_{TFS} = 1$), voicing, place, and nasality were well received (relative feature IT > 90%), whereas manner was received at a level of only 58% relative IT. The negative effect of increasing N_{TFS} was greater for place than for voicing and manner with nasality intermediate between these two.

C. Discussion

The results of Experiment 3 may be compared with those of Gilbert and Lorenzi (2006). Their study employed 16 French consonants in /a/-C-/a/ syllables and used $N_{RENV} = 30$ (compared to $N_{RENV} = 40$ in the current study) in recovering envelopes from TFS speech with N_{TFS} in the range of 1 to 16. In both of these studies, the RENV cues fell off with an increase in the number of TFS bands even though the total bandwidth (80 to 8020 Hz) remained the same for all conditions. However, our data indicate that RENV cues persist even for N_{TFS} equal to 32, which is in contrast to the conclusion of Gilbert and Lorenzi (2006) that RENV cues are abolished for $N_{TFS} \geq 8$. These conflicting conclusions may reflect the use of a larger number of RENV bands in our study. This explanation is supported by the results from Experiment 2 which show a systematic decrease in performance with a decrease in the number of RENV bands extracted from 16-band TFS speech (see Fig. 4). Overall, the results from this study suggest that it is difficult to completely abolish RENV cues, even when the TFS speech is created over 32 narrow channels. Hence, extra care should be taken when interpreting TFS-speech intelligibility data from such vocoder-based studies designed to isolate TFS cues from ENV cues.

VI. COMPARISON BETWEEN RENV AND TFS SPEECH CONDITIONS

The TFS(16) condition and the RENV(16,40) condition, which were included in each of the four experiments, yielded similar levels of overall performance (where mean performance was roughly 50%-correct for both conditions). Further comparisons of these two conditions were undertaken to determine whether similar cues were used in understanding TFS and RENV speech. The comparisons were based on analyses of stimulus-response confusion matrices for the TFS(16) and RENV(16,40) conditions. These matrices were compiled from each of the two subject groups within Experiments 1A (using the RENV2 results from Group 2), 1B, 2, and 3 [yielding eight matrices for TFS(16) speech and eight matrices for RENV(16,40) speech]. Two approaches were used to analyze the matrices and to correlate performance between the TFS(16) and RENV(16,40) conditions. The first approach made use of metric multidimensional scaling to compute a measure of d' for the 120 possible pairs of consonant stimuli. The second approach made use of a sequential information analysis (SINFA; Wang and Bilger, 1973) to examine performance on a set of four consonantal speech features.

A. Metric multidimensional scaling analysis

To compare the confusion matrices we used a form of metric multidimensional scaling (Braida, 1991). In each experiment (TFS and RENV) 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, according to the method developed in Braida (1991), from each of the confusion matrices that resulted from each of the eight TFS(16) and eight RENV(16,40) experiments (two groups each for Experiments 1A, 1B, 2, and 3). Because the locations of the response centers estimated from any given confusion matrix are thought to reflect the observer's expectations, the availability of feedback, etc., the overall structure of the confusion matrix may be represented by the set of values $d'(i, j)$ calculated for each pair (i, j) of stimuli $d'(i,j)^2 = \sum_{k=1}^{4} (X_{ik} - X_{jk})^2$. This allowed comparison of the structures of the confusion matrices for the TFS and RENV speech by comparison of the set of $d'_{TFS}(i, j)$ with the set of $d'_{RENV}(i, j)$.

The results for the listeners of Group 2 in Experiment 1B are shown in Fig. 6. There appears to be a linear relationship between them with $d'_{RENV} = k d'_{TFS} + err$ (where *err* is the noise component in the data) with $k = 0.878$ and a correlation coefficient $\rho = 0.878$. Table II summarizes the correlations between $d'_{RENV}(i, j)$ and $d'_{TFS}(i, j)$ for all conditions tested. Generally the values of k are less than 1.0, which indicates better performance on identifying the TFS versions of the stimuli. An exception to this is the value $k = 1.088$

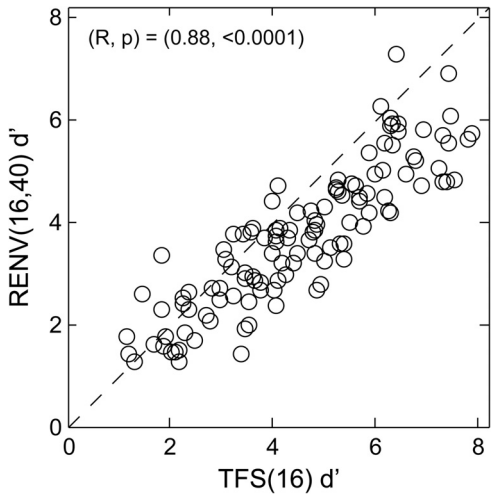

FIG. 6. Plot of the TFS(16) versus RENV(16,40) $d'(i, j)$ values representing the structure of the confusion matrices obtained from Experiment 1B, Group 2 as calculated from the stimulus/response centers estimated from these matrices. The correlation between these values is indicated.

observed for the listeners of Group 1 in Experiment 3 who identified the RENV(16,40) stimuli after receiving extensive practice on other RENV conditions (see Table I). Each of the correlations is significantly different from zero (ranging from 0.772 to 0.924). Taken together these two results indicate that the confusion matrices had a similar structure: Consonant pairs that were easily distinguished when presented as TFS were also easily distinguished when presented as RENV, and vice versa.

The smallest slope k and correlation coefficient ρ were observed for the set of Group 1 subjects in Experiment 2. As observed previously, this may be due to the near-chance performance of two of the subjects in identifying the RENV consonants. When the data from these two subjects are omitted (Group 1*) the value of k increases from 0.421 to 0.866 and ρ increases from 0.772 to 0.813 roughly the same as that for the Group 2 subjects in Experiment 2. This tends to confirm the observation that the Group 1 subjects consisted of two subgroups who responded differently to the RENV stimuli.

TABLE II. Values of the slope $k\left(d'_{\mathrm{RENV}} = kd'_{\mathrm{TFS}}\right)$ and correlation coefficient ρ for the identification experiments that used TFS(16) and RENV(16,40) speech. Also shown are the number of times each of the 16 consonants was presented to each listener ("Pres."). See text for further details.

Expt.	Group	Pres.	k	ρ
1A	1	200	0.881	0.892
1A	2	200	0.816	0.924
1B	1	60	0.803	0.852
1B	2	60	0.878	0.878
2	1	80	0.421	0.772
2	1*	40	0.866	0.813
2	2	80	0.714	0.813
3	1	60	1.088	0.906
3	2	60	0.791	0.853

B. SINFA

A SINFA (Wang and Bilger, 1973 as implemented in the FIX program of the Department of Phonetics and Linguistics, University College London) was conducted to compute conditional information transfer on the features of voicing, manner, place, and nasality (Swaminathan and Heinz, 2012). In this technique (which removes redundancies among the features), relative unconditional feature information is first computed for each of the four features and the feature with the highest relative IT is held constant on the next iteration of the analysis. The feature with the highest relative conditional IT is then also held constant in computing feature transmission in the next iteration of the analysis, and so on, until a set of relative conditional feature IT scores has been obtained for the full set of features. Each of the 16 matrices under consideration (as described above) was subjected to the SINFA analysis to determine the hierarchical order in which the features were extracted. Only 2 of the 24 possible orders occurred: (1) nasality, voicing, place, and manner or (2) voicing, nasality, place, and manner. The first of these two orders was observed in 9 of the 16 analyses. This predominant feature order was then used to conduct a second set of SINFA analyses on each of the 16 matrices using a fixed order for obtaining conditional relative feature IT (i.e., the features were analyzed in the order of nasality, voicing, place, and manner).

The results of the fixed-order SINFA are shown in Fig. 7 where the relative conditional IT score for each of the four features from each of the eight groups of subjects for TFS(16) is plotted versus that group's score on RENV(16,40). A correlation coefficient was computed between the IT measures for TFS(16) and RENV(16,40)

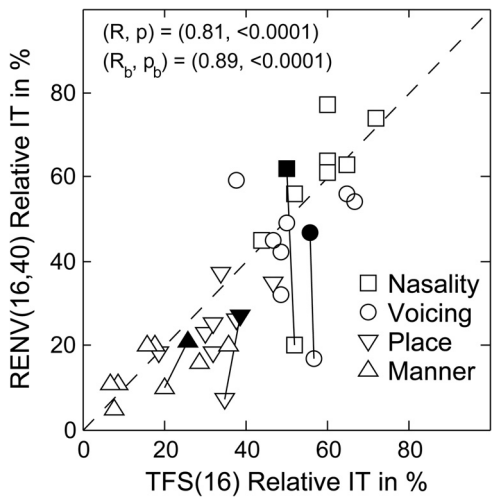

FIG. 7. Relation between feature (voicing, manner, place, and nasality) conditional relative information transfer for TFS(16) speech versus RENV(16,40) speech. Information transfer values (open symbols) were calculated using SINFA analysis for each of eight sets of subjects consisting of the two test sub-groups for each experiment. The correlation (R,p) between TFS(16) and RENV(16,40) is indicated. Information transfer values were also calculated for Experiment 2, Group 2 using only the two subjects who were able to perform at above chance on the RENV conditions (filled symbols). When these values replaced the values calculated using all four subjects in Group 2 (indicated by the lines in the figure), the correlation values (R_b, p_b) were obtained.

speech shown in the upper panel of Fig. 6, indicating $R = 0.81$ and $p < 0.0001$. When data for the two subjects from Experiment 2, Group 1 (who were unable to perform above chance on the RENV conditions) were removed from the analysis, the correlation between the two conditions increased to $R = 0.89$, $p < 0.0001$. This result supports the use of similar speech cues for understanding TFS(16) and RENV(16,40) speech. In both cases, nasality and voicing were better perceived (scores of 17% to 77% relative conditional feature IT) than place and manner (scores of 5% to 38%).

The strong correlations observed between TFS(16) and RENV(16,40) speech in both the multidimensional-scaling and SINFA analyses suggest that similar cues are being used in the perception of these two types of stimuli, and that envelope recovery may play a role in the perception of TFS(16) speech by NH listeners. Hence, any interpretation of results with such TFS speech stimuli should factor in the contributions of RENV cues in addition to TFS cues.

VII. GENERAL DISCUSSION

A. RENVs can contribute to the intelligibility of TFS speech

The goal of this study was to assess the role of RENV cues to the perception of TFS speech. For the two speech conditions that were included in all the experiments [TFS(16) and RENV(16,40)] significant correlations were observed for the pairwise distance measures (Table II, Fig. 6) as well as for conditional feature performance (Fig. 7). These experimental results strongly suggest that RENVs were not completely abolished and could have contributed to the perception of TFS speech by our NH listeners. The basis for such recovery of envelopes from TFS speech signals that were supposedly stripped of envelope cues through Hilbert processing is explained by Ghitza (2001). When the TFS speech signal is passed through a bank of sufficiently narrow filters (a criterion that is met by the normal auditory filters of NH listeners), envelope cues re-emerge.

For the TFS(16) speech, reception of the features of voicing and nasality was somewhat better than for manner and place (Figs. 4 and 5). This pattern differs from that observed by Gilbert and Lorenzi (2006) for narrowband TFS speech who showed high reception of voicing and place with somewhat lower reception of manner. Their overall performance, however, was substantially higher than that observed in the current study. A possible explanation for the observed differences in feature reception with TFS speech between the current study and that of Gilbert and Lorenzi (2006) may lie in the language difference (English versus French consonants). Swaminathan and Heinz (2012) reported similar overall performance to the current study on the TFS(16) condition and their feature results were similar to those observed here in the poor reception of manner and intermediate perception of place. One difference is that nasality was much better received than voicing in the Swaminathan and Heinz (2012) data. Our results for the RENV(16,40) condition show the same pattern of feature reception as observed for the TFS(16) condition. This pattern differs from that reported in a previous study of RENV speech by Gilbert and Lorenzi (2006) who reported much lower feature reception for RENV compared to TFS speech. Overall, the significant correlations observed in the reception of speech features and on the metric multidimensional scaling analysis between the two types of speech (Fig. 7) suggest that listeners were relying on the same types of cues in the perception of both the TFS and RENV speech signals.

B. Training and token variability effects

Even after several hours of exposure sufficient for stable levels of performance, it should be noted that the intelligibility of TFS(16) speech reported here was substantially lower than the scores reported in previous studies (e.g., Lorenzi *et al.*, 2006; Gilbert and Lorenzi, 2006; Sheft *et al.*, 2008). Possible factors contributing to this difference may lie in the language difference (French versus English) and in differences in token variability. The French studies employed a set of 48 /a/-C-/a/ syllables spoken by one female French talker made up of 3 utterances of each of 16 different values of C. The current study, on the other hand, employed a set of 64 /a/-C-/a/ tokens produced by two male and two female speakers of American English (each speaker contributing 1 token for each of 16 values of C). The overall larger number of tokens employed in each run and the use of both male and female talkers may have contributed to the greater difficulty of the task in the current study. Previous research has demonstrated that for a small-to-moderate number of speech stimuli (such as the 16 consonants employed here), performance decreases as the number of tokens per speech stimulus increases from 1 to 4 but remains stable in the range of 4 to 16 tokens (Uchanski and Braida, 1998; Uchanski *et al.*, 1992).

Our data also indicate substantial variability among individual subjects in the ability to understand both TFS(16) and RENV(16,40) speech. Across subjects and experiments, mean TFS(16) scores ranged from 28.4%- to 77.5%-correct and mean RENV(16,40) scores ranged from 5.3%- to 65.5%-correct. Such a wide variability in scores following similar amounts of exposure to the stimuli indicates different learning strategies and/or abilities among subjects in using the cues available in the stimuli. In these experiments, the total amount of exposure to the stimuli was limited to either 20 runs (Experiment 1) or 10 runs (Experiments 2 and 3) per condition. Some subjects may have required additional training to attain maximum performance. Such inter-subject variability has also been observed by Lorenzi *et al.* (2006) in learning curves reported for individual NH listeners with TFS speech. In their Fig. 2(A), it can be seen that some subjects achieved asymptotic performance levels of roughly 90%-correct on TFS speech within the first 5 runs of training, while other subjects required as many as 15 runs to reach similar levels of performance.

C. Implications for hearing impairment and CI signal processing

Although this study did not involve the testing of HI or cochlear implant (CI) listeners, it is still possible to consider what these findings imply for improving hearing aid and/or CI signal processing strategies.

A number of studies using TFS speech suggest that listeners with sensorineural hearing loss have a reduced ability to use TFS cues (Lorenzi *et al*., 2006; Moore *et al*., 2006; Hopkins and Moore, 2007; Hopkins *et al*., 2008; Lorenzi *et al*., 2009), although their ability to use ENV cues is not degraded. It is unlikely that the inability of HI listeners to process TFS speech stems from a degradation in the ability of auditory-nerve fibers to phase lock to TFS. Recent neurophysiological evidence indicates that phase locking in quiet is not degraded following noise-induced hearing loss (Kale and Heinz, 2010), although it may be degraded in background noise (Henry and Heinz, 2012). Although this implies that phase-locking may not be degraded with noise-induced hearing loss, other etiologies of hearing loss may have an effect on the encoding of TFS.

The inability of HI listeners to process TFS speech may be related to broadened cochlear tuning which can lead to less effective recovery of ENV cues from TFS (Lorenzi *et al*., 2012). Indeed, the results of Experiment 2 indicate that as the bandwidth used to recover envelopes from 16-band TFS speech increased (with a decrease in the number of recovered bands themselves), there was a rapid drop in the ability to understand RENV speech. Performance with RENV(16,40) speech was for most subjects similar to their performance on TFS(16) speech (see Fig. 4) but was little better than chance for an RENV(16,16) signal. The performance of NH subjects with RENV(16,16) was largely consistent with the scores obtained from HI subjects in the study of Lorenzi *et al*. (2006). The mechanism by which the recovery of ENV cues from TFS speech are disrupted following filter broadening is not clearly understood. Hopkins *et al*. (2010) suggested that the deficits in the processing of TFS speech observed in HI listeners could be related to the amplification of low-level portions of the speech signal. Such amplification would be more detrimental to HI subjects than NH subjects because of their increased susceptibility to temporal and spectral masking. If the intelligibility of TFS speech is conveyed as RENVs, any disruptions introduced due to the TFS processing schemes will also have a comparable effect on the RENV signals (Apoux *et al*., 2013).

The potential for RENV cues to contribute to the perceptual salience of acoustic TFS has important implications for auditory prosthesis design. An obvious approach to signal-processing would be to design speech-processing schemes that convert acoustic TFS to RENVs (Won *et al*., 2012), similar to the processing that occurs in a normal cochlea. Such novel schemes would convey both "true" and "recovered" ENVs to HI listeners in an effort to improve speech perception in degraded listening conditions. Further research is needed to determine the role of the broadened critical bands that typically accompany sensorineural hearing loss (e.g., Glasberg and Moore, 1986; Dubno and Schaefer, 1992; Desloge *et al*., 2012) on the ability to use RENVs.

VIII. SUMMARY AND CONCLUSIONS

(1) In NH listeners, after sufficient exposure, providing 40 bands of envelopes recovered from 16-band TFS speech gives roughly the same intelligibility as 16-band TFS speech and similar patterns of speech confusions. This suggests that envelope cues were not completely abolished in studies that have used vocoded 16-band TFS speech stimuli.
(2) Presentation order had a significant effect on the perception of TFS speech and RENV speech, suggesting that prior exposure to TFS speech facilitates performance on RENV speech.
(3) Even after sufficient exposure, the intelligibility scores obtained with 16-band TFS speech in this study were substantially lower than the scores reported in previous studies (~50% versus ~90%). This suggests that greater speech-token variability and stimulus complexity can have a large impact on the intelligibility of TFS speech.
(4) Reducing the number of RENV bands in a manner that widened the individual filter bands led to decreased performance on RENV speech. This suggests that the inability of HI listeners to process TFS speech may be related to broadened cochlear tuning which can lead to less effective recovery of envelope cues from TFS.
(5) Analyses of consonant confusions suggest that similar cues are being used in the perception of TFS and RENV speech; thus envelope recovery may play a role in the perception of TFS speech by NH listeners.

ACKNOWLEDGMENTS

This research was supported by Grant Nos. R01 DC00117 and R43 DC013006 from the National Institutes of Health, NIDCD.

[1]Note that the 300 Hz cutoff frequency of the lowpass filter used to estimate the signal envelope can cause aliasing when the envelope is subsequently used to modulate tone carriers—even after re-filtering with the original bandpass filter. For the 64 /a/-C-/a/ utterances and for all combinations of N_{TFS} and N_{RENV} used in these experiments, 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.
Braida, L. D. (**1991**). "Cross modal integration in the identification of consonant segments," Q. J. Exp. Psychol. **43**, 647–677.
Desloge, J. G., Reed, C. M., Braida, L. D., and Delhorne, L. A. (**2012**). "Auditory-filter characteristics for listeners with real and simulated hearing impairment," Trends Amplif. **16**, 19–39.
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.
Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.
Dubno, J. R., and Schaefer, A. B. (**1992**). "Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners," J. Acoust. Soc. Am. **91**, 2110–2121.
Flanagan, J. L. (**1980**). "Parametric coding of speech spectra," J. Acoust. Soc. Am. **68**, 412–419.
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.
Glasberg, B. R., and Moore, B. C. J. (**1986**). "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," J. Acoust. Soc. Am. **79**, 1020–1033.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hearing Res. **47**, 103–138.
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.
Henry, K. S., and Heinz, M. G. (**2012**). "Diminished temporal coding with sensorineural hearing loss emerges in background noise," Nat. Neurosci. **15**, 1362–1364.
Hilbert, D. (**1912**). *Grundzüge einer Allgemeinen Theorie der Linearen Integralgleichungen (Basic Principles of a General Theory of Linear Integral Equations)* (B. G. Teubner, Leipzig), pp. 1–282.
Hopkins, K., and Moore, B. C. J. (**2007**). "Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information," J. Acoust. Soc. Am. **122**, 1055–1068.
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 addition of low-level, low-noise noise on the intelligibility of sentences processed to remove temporal envelope information," J. Acoust. Soc. Am. **128**, 2150–2161.
Houtsma, A. J. M. (**1983**). "Estimation of mutual information from limited experimental data," J. Acoust. Soc. Am. **74**, 1626–1629.
Kale, S., and Heinz, M. G. (**2010**). "Envelope coding in auditory nerve fibers following noise-induced hearing loss," J. Assoc. Res. Otolaryngol. **11**, 657–673.
Kochkin, S. (**1996**). "Customer satisfaction and subjective benefit with high performance hearing aids," Hear. Rev. **3**, 16–26.
Liberman, M. C., and Dodds, L. W. (**1984**). "Single-neuron labeling and chronic cochlear pathology. III. Sterocilia damage and alterations of threshold tuning curves," Hear. Res. **16**, 55–74.
Logan, B. F., Jr. (**1977**). "Information in the zero crossings of bandpass signals," Bell Syst. Tech. J. **56**, 487–510.
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. (**2006**). "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., Wallaert, N., Gnansia, D., Leger, 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., Glasberg, B. R., and Hopkins, K. (**2006**). "Frequency discrimination of complex tones by hearing-impaired subjects: Evidence for loss of ability to use temporal fine structure," Hearing Res. **222**, 16–27.
Plack, C. J., and Oxenham, A. J. (**2005**). "The psychophysics of pitch," in *Pitch Perception*, edited by C. J. Plack, R. R. Fay, A. J. Oxenham, and A. N. Popper (Springer, New York), pp. 7–55.
Rice, S. O. (**1973**). "Distortion produced by band limitation of an FM wave," Bell Syst. Tech. J. **52**, 605–626.
Saberi, K., and Hafter, E. R. (**1995**). "A common neural code for frequency- and amplitude-modulated sounds," Nature **374**, 537–539.
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.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
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.
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.
Takahashi, G., Martinez, C. D., Beamer, S., Bridges, J., Noffsinger, D., Sugiura, K., and Bratt, G. W. (**2007**). "Subjective measures of hearing aid benefit and satisfaction in the NIDCD/VA follow-up study," J. Am. Acad. Audiol. **18**, 323–349.
Uchanski, R. M., and Braida, L. D. (**1998**). "Effects of token variability on our ability to distinguish between vowels," Percept. Psychophys. **60**, 533–543.
Uchanski, R. M., Millier, K. M., Reed, C. M., and Braida, L. D. (**1992**). "Effects of token variability on resolution for vowel sounds," in *The Auditory Processing of Speech: From Sounds to Words*, edited by M. E. H. Schouten (Mouton de Gruyter, Berlin), pp. 291–302.
Voelcker, H. B. (**1966**). "Towards a unified theory of modulation. I. phase-envelope relationships," Proc. IEEE **54**, 340–354.
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.
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.
Zwicker, E. (**1962**). "Direct comparisons between the sensations produced by frequency modulation and amplitude modulation," J. Acoust. Soc. Am. **34**, 1425–1430.