Speech identification based on temporal fine structure cues

Stanley Sheft[a)]
Parmly Hearing Institute, Loyola University Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626

Marine Ardoint and Christian Lorenzi
Laboratoire de Psychologie de la Perception (CNRS - Université Paris 5 Descartes), DEC, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France GDR CNRS 2967 GRAEC

(Received 1 June 2007; revised 9 April 2008; accepted 9 April 2008)

The contribution of temporal fine structure (TFS) cues to consonant identification was assessed in normal-hearing listeners with two speech-processing schemes designed to remove temporal envelope (*E*) cues. Stimuli were processed vowel-consonant-vowel speech tokens. Derived from the analytic signal, carrier signals were extracted from the output of a bank of analysis filters. The "PM" and "FM" processing schemes estimated a phase- and frequency-modulation function, respectively, of each carrier signal and applied them to a sinusoidal carrier at the analysis-filter center frequency. In the FM scheme, processed signals were further restricted to the analysis-filter bandwidth. A third scheme retaining only *E* cues from each band was used for comparison. Stimuli processed with the PM and FM schemes were found to be highly intelligible (50–80% correct identification) over a variety of experimental conditions designed to affect the putative reconstruction of *E* cues subsequent to peripheral auditory filtering. Analysis of confusions between consonants showed that the contribution of TFS cues was greater for place than manner of articulation, whereas the converse was observed for *E* cues. Taken together, these results indicate that TFS cues convey important phonetic information that is not solely a consequence of *E* reconstruction.

© *2008 Acoustical Society of America.* [DOI: 10.1121/1.2918540]

PACS number(s): 43.71.Gv, 43.71.Es, 43.66.Mk [KWG] Pages: 562–575

I. INTRODUCTION

Following the pioneering work of Flanagan (Flanagan and Golden, 1966; Flanagan, 1980), subsequent studies of speech intelligibility have investigated the role of two temporal features of filtered speech: Fluctuations in the envelope (*E*, the relatively slow modulations in amplitude over time), and fluctuations in the temporal fine structure [TFS, the rapid oscillations with average rate close to the center frequency (CF) of the band, or in other words, the "carrier" signal]. To assess the contribution of each temporal feature to intelligibility, speech stimuli were split into an array of contiguous frequency bands (also called analysis bands) and processed to remove either *E* or TFS cues from each band, assuming that *E* and TFS are independent components of the narrowband signal. Across studies, high levels of speech intelligibility have been obtained in quiet from normal-hearing listeners on the basis of either *E* cues (e.g., Drullman, 1995; Shannon *et al.*, 1995; Smith *et al.*, 2002; Zeng *et al.*, 2004; Xu *et al.*, 2005) or TFS cues (Gilbert and Lorenzi, 2006; Lorenzi *et al.*, 2006; Gilbert *et al.*, 2007) alone.

The results obtained with "TFS speech" (i.e., speech processed to retain only TFS information) may appear surprising because it is generally considered that, at least for nontonal languages, *E* cues carry most of the information required for speech identification in quiet (e.g., Flanagan, 1980; Shannon *et al.*, 1995; Smith *et al.*, 2002), with TFS primarily conveying pitch cues which enhance segregation of the speech signal from background sounds (e.g., Qin and Oxenham, 2003, 2006; Nelson *et al.* 2003; Stickney *et al.*, 2005; Füllgrabe *et al.*, 2006). However, the descriptive analysis of the temporal information in speech by Rosen (1992) and the intelligibility of sine-wave speech tokens (e.g., Remez *et al.*, 1981, Remez and Rubin, 1990) both suggest that TFS cues play a role in linguistic contrasts. Further work is therefore needed to assess the extent to which TFS cues *alone* can convey useful linguistic information in addition to pitch.

In recent studies investigating the intelligibility of TFS speech, processing of the narrowband speech signals, that is, the outputs of the analysis filterbank, was based upon the Hilbert transform (e.g, Smith *et al.*, 2002; Xu and Pfingst, 2003; Gilbert and Lorenzi, 2006; Lorenzi *et al.*, 2006). In each subband, the Hilbert TFS was derived as $\cos[\phi(t)]$ where $\phi(t)$ corresponds to instantaneous phase, that is, the angle of the analytic signal. The Hilbert *E*, which was obtained by taking the magnitude of the analytic signal, was discarded and TFS-speech stimuli were obtained by summing all TFS subband signals. However, potential artifacts may have influenced results. Ghitza (2001) demonstrated that despite filtering or removal, *E* cues are reconstructed at the output of peripheral auditory filters and may therefore be used by listeners. Involvement of *E* reconstruction in speech identification has been empirically confirmed using either sentences (Zeng *et al.*, 2004) or vowel-consonant-vowel (VCV) stimuli (Gilbert and Lorenzi, 2006). In these studies, TFS speech signals were passed through a bank of gammachirp auditory filters (Irino and Patterson, 1997). Envelopes reconstructed from gammachirp-filter output were then

[a)]Author to whom correspondence should be addressed. Electronic mail: ssheft@luc.edu

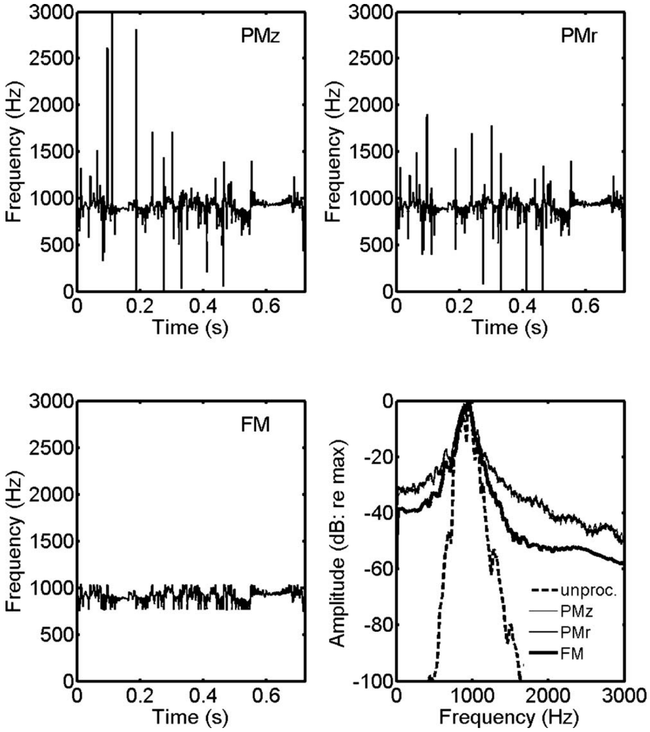

FIG. 1. For a narrowband /a s a/ speech signal, instantaneous-frequency functions with processing algorithm indicated in the upper right-hand corner of the panel, and long-term magnitude spectra (bottom-right panel). Analysis used a 0.4-octave-wide, third-order zero-phase Butterworth filter centered at 900 Hz. Magnitude spectra: Unprocessed speech (dashed line), PMz speech (continuous thin line), PMr speech (continuous line), and FM speech (continuous heavy line). Magnitude spectra were smoothed for clarity of presentation. In large part the two PM magnitude spectra overlap.

used to amplitude-modulate either noise bands or pure tones associated with the CF of the auditory filters. In agreement with Ghitza's (2001) predictions, the processed-speech stimuli were intelligible with 40–60% mean correct identification.

Several processes may contribute to this artifact originating from the use of the Hilbert transform. Ghitza (2001) pointed out that *E* reconstruction occurs because—contrary to previous assumption—the envelope and instantaneous-phase functions of band-limited signals are not independent processes (see also Papoulis, 1983). As a consequence, any manipulation of *E* will affect TFS, and *vice versa*. In addition, Gilbert and Lorenzi (2006) noted that *E* reconstruction is subsequent to processing which converts frequency modulation (FM) to amplitude modulation (AM). More precisely, the differential attenuation of auditory filtering converts the frequency excursions of TFS into dynamic variations in excitation level (that is, into *E* fluctuations). Finally, and this is a consequence of the first point, TFS subband signals extracted via the Hilbert transform typically have a greater bandwidth than the original subband signals (see Fig. 1). This results from the fact that the Hilbert *E* is not band limited. The Hilbert TFS therefore contains a wideband structure of "cancellation terms" that match and cancel the wideband content of the Hilbert *E* (Schimmel and Altas, 2005).

It follows from the above arguments that the fidelity of *E* reconstruction should be influenced by at least three factors. The first factor is the coherence between the extracted TFS and *E* signals. Because *E* and the instantaneous phase are related, any manipulation of the extracted narrowband TFS signals resulting in an improper match with the original *E* should affect the fidelity of *E* reconstruction (Schimmel and Altas, 2005). One way to achieve this kind of mismatch to reduce the fidelity of *E* reconstruction is to filter the extracted TFS signals using an all-pass filter with a random phase response. The second factor relates to effects of both analysis- and auditory-filter bandwidths. The fidelity of *E* reconstruction increases with analysis-filter bandwidth but decreases with auditory-filter bandwidth. However, the effects of changing the two bandwidths are not related by simple inversion. If the ratio of bandwidths is fixed, the fidelity of *E* reconstruction increases as auditory-filter bandwidth decreases. Gilbert and Lorenzi (2006) systematically investigated the influence of filter bandwidths on the intelligibility of TFS speech. Though results indicated that increasing the frequency resolution of the analysis filterbank did not completely abolish *E* reconstruction, the reconstructed *E* cues did not play a major role in consonant identification once the analysis bandwidth was narrower than four times the bandwidth of a normal auditory filter. The third factor is the bandwidth of the processed TFS signals. Restriction of frequency excursions to within the analysis-filter bandwidth should limit the extent of FM-to-AM conversion to degrade the fidelity of *E* reconstruction.

The goal of the present study was to extend the initial work of Zeng *et al.* (2004) and Gilbert and Lorenzi (2006). Conditions measured the ability of listeners to identify speech using stimulus configurations intended to vary the fidelity of *E* reconstruction. All speech-processing algorithms derived the analytic signals from the output of a bank of 0.4-octave-wide (i.e., ~2 ERB wide) analysis filters. Based on the complex envelope, one scheme estimated a phase-modulation (PM) function from the output of each analysis filter, while another derived a FM function. In both cases, the modulators were applied to a sinusoidal carrier at the analysis-filter CF. Though using a different processing sequence, the "PMz" implementation of the PM scheme was mathematically identical to the one previously used by Gilbert and Lorenzi (2006) and Lorenzi *et al.* (2006). A second PM implementation, "PMr," was obtained by simply randomizing the starting phase of the sinusoidal carriers in order to alter the match between *E* and TFS to degrade the fidelity of *E* reconstruction. Though similar, randomization of carrier starting phase is not equivalent to the procedure of Drennan *et al.* (2007) in which a random time-varying component is added to the PM function. Along with affecting the fidelity of *E* reconstruction, addition of a random component to the phase function influences the cross-spectral relationships among modulators, while carrier randomization in the PMz algorithm does not.

The FM algorithm was based on representation of the narrowband speech signals in terms of AM and FM functions (Loughlin and Tacer, 1996; Zeng *et al.*, 2004; Nie *et al.*, 2005; Stickney *et al.*, 2005). In this case, TFS cues corresponded explicitly to the complex pattern of FM extracted within each speech band. One advantage of this algorithm is that it allows for restrictions on the range of FM variations in the TFS-speech signals. In the current implementation, deviations in instantaneous frequency were restricted to the analysis-filter bandwidth. A final scheme (E) designed to retain only E cues from each band was also used for comparison.

Spectral consequences of TFS extraction are illustrated in Fig. 1 for the speech VCV /asa/. TFS signals were extracted from the output of an analysis filter centered at 900 Hz.[1] The instantaneous-frequency function (IFF)—the time rate of change of instantaneous phase—of PMz, PMr, and FM signals are shown in separate panels, with long-term spectra in the bottom right panel. For the PMz signal, the extent of IFF deviation is significantly broader than the analysis-filter passband (776–1024 Hz), and the bandwidth of the long-term magnitude spectrum is broader than that of the unprocessed VCV token. The spikes of the IFF relate to carrier phase reversals in envelope troughs of the original signal, indicating large influence of stimulus gaps on the extent of E reconstruction. Compared to PMz, PMr, and more notably FM signals exhibit smaller excursions in instantaneous frequency. This reduction in extent of IFF excursion with the FM scheme results in narrowing of the long-term spectral flanks when compared to the two PM schemes in the bottom-right panel. The spikes of the IFF can elicit an impulse response from auditory filters, with decay of this response coded as envelope. The contribution of reconstructed E cues to TFS-speech intelligibility should therefore be reduced with the PMr scheme, and even more so with the FM algorithm, when compared to the effect of PMz processing in which the IFF spikes are most prominent.

For each processing scheme (PMz, PMr, FM, and E), the intelligibility of VCV stimuli was evaluated in quiet at a comfortable listening level in seven young, normal-hearing listeners who received a moderate amount of training. Additional experiments investigated further the contribution of putatively reconstructed E cues to TFS-speech identification by (i) removing low-frequency analysis bands, (ii) increasing the frequency resolution of the analysis filterbank, and (iii) decreasing stimulus presentation level. Since these stimulus manipulations affect the second factor discussed above (namely, effects of analysis- and auditory-filter bandwidths), the fidelity of E reconstruction varied across conditions. In each experiment, identification data are considered in terms of quantitative estimates of the fidelity of E reconstruction at the output of a bank of gammatone auditory filters.

II. EXPERIMENTS

A. Method

1. Speech material

One set of 48 VCV stimuli was recorded. Speech stimuli consisted of three exemplars of 16-/aCa/ utterances (C =/p, t, k, b, d, g, f, s, ʃ, m, n, r, l, v, z, ʒ/) read by a French female speaker in quiet (mean VCV duration =648 ms; standard deviation=46 ms). The fundamental frequency of the female voice was estimated as 216 Hz using the YIN algorithm (de Cheveigné and Kawahara, 2002). Each signal was digitized via a 16 bit analog-to-digital converter at a 44.1 kHz sampling rate.

2. Speech processing

The original speech signals were processed with two different TFS algorithms referred to as PM and FM. A third scheme (E) retained only envelope rather than fine-structure modulation. Across algorithms, each VCV signal was initially bandpass filtered using third-order zero-phase Butterworth filters. The filterbank consisted of 16-0.4-octave-wide contiguous frequency or analysis bands spanning the range of 80–8020 Hz [see Gilbert and Lorenzi (2006) for additional details concerning analysis-filter characteristics].

PM conditions. To estimate the phase modulation of each analysis band, the corresponding analytic signal was shifted to baseband through use of the complex envelope. Specifically, the PM function was estimated as the angle of the product of the analytic signal and a complex exponential such that

$$\Phi_k(t) = \text{angle}\{x_{+,k}(t)\exp[-j\omega_k(t) + \theta_k]\} \tag{1}$$

with $x_{+,k}(t)$ the analytic signal of the kth analysis band, and ω_k and θ_k the angular frequency and phase, respectively, of the sinusoidal carrier used in TFS construction. As shown in Eq. (2), the subband TFS signal was generated by modulating the carrier with the derived PM function.

$$TFS_k(t) = \cos[\omega_k(t) + \Phi_k(t) + \theta_k]. \tag{2}$$

Used previously by Sheft and Yost (2001), this approach to PM estimation is similar to the approach of Schimmel and Altas (2005) for determination of the subband envelope.

In all conditions, the carrier frequency was equal to analysis-band CF. In the first implementation of the PM algorithm (PMz), θ_k was set to zero so that processing effect matched the one of Gilbert and Lorenzi (2006) and Lorenzi *et al.* (2006). In the second implementation (PMr), θ_k was selected randomly between 0 and 2π in order to degrade the relationship between the original E and TFS within each band. To compensate for the reduction in amplitude caused by E removal, TFS_k was multiplied by the root-mean-square (rms) power of the bandpass-filtered VCV. The "power-weighted" TFS signals were finally summed over all analysis bands and presented as such to the listeners.

A consequence of this weighting was that long-term spectral cues were preserved in the processed speech stimuli. The contribution of residual spectral cues to VCV identification was investigated in a pilot experiment using a 16-band analysis filterbank. For each intact VCV token, rms power was initially computed in each frequency band, and used to weight the power of a sinusoid at the band CF; both E and TFS cues were therefore removed. Stimuli were presented for identification to naive and experienced normal-hearing listeners. Identification performance of all listeners was at chance level, indicating that long-term spectral cues did not contribute to intelligibility in the experimental conditions of the present study.

FM condition. FM estimation for each analysis band was based on the time derivative of the unwrapped instantaneous-phase function $\phi(t)$, shifted to baseband as the deviation from subband CF. In this case,

$$FM_k(t) = [\phi_k'(t)/(2\pi)] - \text{CF} \quad (3)$$

with CF given in Hz. Deviations in instantaneous frequency were restricted to not exceed the analysis-filter bandwidth (defined by filter 3-dB down points). This restriction was intended to alter E reconstruction at the output of auditory filters. The (restricted) FM function was then integrated to obtain a time-dependent phase function, $\Phi_k(t)$. As in the PM schemes, the subband TFS signal was generated according to Eq. (2) by modulating the carrier with the function $\Phi_k(t)$. This algorithm for estimating FM functions is similar to the scheme used by Zeng and colleagues (e.g., Zeng *et al.*, 2005; Nie *et al.*, 2005; Stickney *et al.*, 2005).

Simulations based on correlation estimates (see below) indicated that E reconstruction at the output of gammatone auditory filters is not significantly affected by carrier starting phase (θ_k). Similar to the PMr scheme, θ_k was therefore chosen randomly between 0 and 2π. TFS_k was multiplied by the rms power of the bandpass filtered VCV, with the processed stimulus the sum over all analysis bands of the weighted subband signals.

E condition. In each analysis band, temporal envelopes were extracted as the magnitude of the analytic signal and lowpass filtered at 64 Hz using a third-order zero-phase Butterworth filter. These envelopes were used to amplitude-modulate sinusoidal carriers at frequencies corresponding to the CFs of the analysis filters with carrier starting phase randomized.

3. Quantification of E reconstruction

In each experimental condition, the TFS speech signals were passed through a bank of 32 gammatone auditory filters, each 1 ERB wide (Patterson *et al.*, 1987) with CFs uniformly spaced along an ERB scale ranging from 123 to 7743 Hz. Level-dependent implementations of filtering used the formula proposed by Glasberg and Moore (1990). In each band, the temporal envelopes were extracted using the Hilbert transform and lowpass filtered at 64 Hz with a first-order zero-phase Butterworth filter. In this case, first- rather than third-order filtering was used to better approximate the temporal modulation transfer function obtained with wideband-noise carriers (Viemeister, 1979).

For each VCV utterance, mean correlation estimates were computed between the envelopes of the original VCV stimulus and the TFS signals, both derived from gammatone-filterbank output. The correlation estimates were averaged, in terms of Fisher's z values, across the 48 VCV utterances. A high correlation estimate indicates a close resemblance between the original E and the one reconstructed at the output of auditory filters. Three types of correlation estimates were considered: (i) the correlation coefficient, as used by Gilbert and Lorenzi (2006), (ii) a depth-dependent correlation estimate, and (iii) a level-dependent correlation estimate.

Insensitive to both the carrier level in each subband and the modulation depth of reconstructed E cues, the correlation coefficient may overestimate E reconstruction. To introduce these sensitivities into measures of envelope correlation, correlation coefficients were weighted to derive separate depth- and level-dependent estimates. In both cases, weighting functions were based on envelope cross-spectral power across gammatone-filterbank channels. If ac coupled, envelope cross-spectral power varies in the same linear manner for overall level and depth, independent of cross correlation. For depth-dependent estimates, channel outputs were normalized to a constant overall level for each stimulus. Following normalization, the cross-spectral power of the ac-coupled envelopes of the original and processed speech tokens was determined for each channel. The correlation coefficients determined for each channel were then scaled by the relative cross-spectral power associated with that channel. With the exception of manner of normalization, a similar approach was used to derive a level-dependent measure. In this case, channel outputs were normalized to result in a constant *ratio* of envelope ac-rms to dc-level while retaining original differences in overall level among channel outputs.

4. Procedure

All stimuli were generated using a 16-bit digital-to-analog converter at a 44.1-kHz sampling rate, and delivered monaurally to the right ear of Sennheiser HD 212 Pro headphones. The rms values of the stimuli were equalized for an average level of 80 dB(A). Listeners were tested individually in a double-walled soundproof booth. In a typical experimental session, a complete set of the 48 VCV utterances corresponding to a given experimental condition was presented at random. Each listener was instructed to identify the presented consonant. The 16 possible choices were presented on the computer monitor, and the listener entered their response with the computer mouse. Feedback was not provided. Percent-correct identification was calculated and a confusion matrix was built from the responses to the 48 VCV utterances. Reception of the phonetic features of voicing, manner of articulation (occlusive versus constrictive), place of articulation, and nasality was determined by means of information-transmission analysis (Miller and Nicely, 1955) on the individual confusion matrices (see Table I for the assignment of consonant features in French).

In the first experiment, all listeners participated in ten sessions for each of the following conditions: (1) unprocessed speech, (2) PMz TFS speech, (3) PMr TFS speech, and (4) FM TFS speech. Forty-eight-trial blocks for each of the four conditions were interleaved with the order of presentation randomized across listeners. All listeners participated in four sessions in the E-speech condition upon completion of the 40 TFS-speech sessions.

5. Listeners

Data were collected from seven normal-hearing listeners. Their ages ranged from 21 to 32 years (mean age: 24 years; standard deviation: four years); all were native French speakers. The same listeners participated in experiments 1 through 4, except for one listener would did not run in experiment 4. In accordance with the Helsinki declaration (2004), all listeners were fully informed about the goal of the present study and provided written consent before their participation.

TABLE I. Phonetic features of the 16 French consonants used in this study (Martin 1996).

Consonant	Voicing	Manner	Place	Nasality
/p/	unvoiced	occlusive	front	non nasal
/t/	unvoiced	occlusive	middle	non nasal
/k/	unvoiced	occlusive	back	non nasal
/b/	voiced	occlusive	front	non nasal
/d/	voiced	occlusive	middle	non nasal
/g/	voiced	occlusive	back	non nasal
/f/	unvoiced	constrictive	front	non nasal
/s/	unvoiced	constrictive	middle	non nasal
/ʃ/	unvoiced	constrictive	back	non nasal
/v/	voiced	constrictive	front	non nasal
/z/	voiced	constrictive	middle	non nasal
/ʒ/	voiced	constrictive	back	non nasal
/l/	voiced	constrictive	middle	non nasal
/r/	voiced	constrictive	back	non nasal
/m/	voiced	occlusive	front	nasal
/n/	voiced	occlusive	middle	nasal

B. Results

1. Experiment 1: Basic comparison between processing schemes

The fidelity of *E* reconstruction was first estimated for the conditions of experiment 1 [16-band analysis filterbank, 80 dB(A) level]. This experimental condition is labeled 16B in subsequent references. Results obtained with the three correlation algorithms described in Sec. II A are presented in separate panels of Fig. 2. As a function of gammatone-filter CF, the dependent variable is the correlation between envelopes of the original and TFS-processed stimuli, with processing scheme the parameter. The number in parentheses below each panel title is the mean correlation across gammatone-filter channels estimating *E* fidelity in the *E* condition. Numbers between parentheses in the figure legends correspond to the mean correlation estimate in the respective TFS condition. Overall, the three estimates of *E* reconstruction in the TFS conditions are relatively low ($r < 0.4$) in most auditory channels, except for the low-frequency channels encompassing the fundamental frequency of the talker's voice. When compared to results from the *E* condition shown in each panel title, TFS correlation estimates drop by a factor that ranges from roughly 4.5 to 13.6. Note that much lower estimates of *E* reconstruction are observed with correlation algorithms dependent on either carrier level or envelope depth. To the extent that subband audibility and modulation depth are important, contingent correlations suggest limited utility of reconstructed *E* cues.

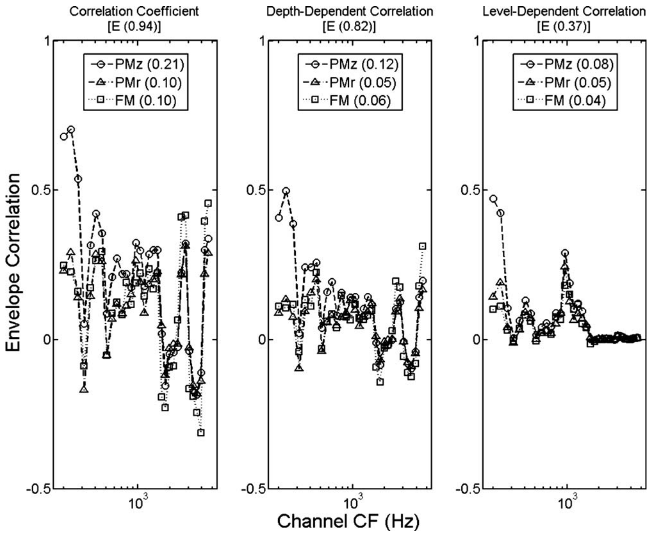

FIG. 2. Assessment of envelope reconstruction in condition 16B. Mean correlation estimates between the original speech envelopes and the envelopes of the stimuli in the PMz (circles), PMr (triangles), and FM (squares) conditions are shown a function of gammatone-filter CF. Left panel: correlation coefficient; middle panel: depth-dependent correlation estimate; right panel: level-dependent correlation estimate. The number in parentheses below each panel title is the mean correlation across gammatone-filter channels estimating *E* fidelity in the *E* condition. In each figure legend, numbers between parentheses correspond to the mean correlation estimate across filter channels in the respective TFS condition.

Across auditory channels, the three correlation estimates are generally similar for the three processing schemes, except for the lowest channels where correlation is substantially smaller, and similar, for the PMr and FM schemes compared to PMz. Overall, correlation data indicate that (i) poor reconstruction of *E* cues occurs in most auditory channels except for the lowest ones, and (ii) randomizing the starting phase of the sinusoidal carriers (as in the PMr scheme) and restricting excursions in instantaneous frequency to the passband of analysis filters (as in the FM scheme) affect mostly, and to the same extent, *E* reconstruction in the lowest auditory channels. Finally, since analysis filters were zero phase, correlation results are not appreciably altered with change in approach from use of the correlation coefficient to the maximum value of the cross-correlation function, that is, with consideration of values other than those obtained at an analysis lag of zero.

The top left panel of Fig. 3 shows mean identification scores averaged across listeners as a function of session number in the 16B conditions. The parameter is speech-processing scheme. The remaining four panels of Fig. 3 show mean percent of information received for each phonetic feature with feature indicated at the top of the panel. Mean

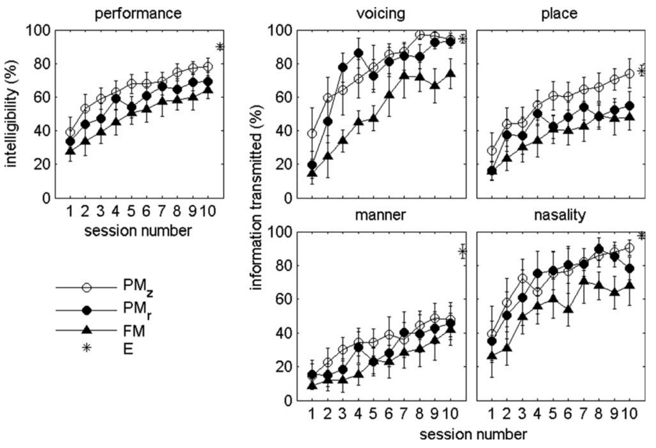

FIG. 3. Mean identification performance (left panel) and percent of information received for each phonetic feature (middle and right panels) as a function of session number for the 16-band PMz (open circles), PMr (filled circles), and FM (filled triangles) TFS-speech conditions. Mean performance averaged across four repeated session in the 16-band E-speech condition is indicated with stars on the right side of each panel. Error bars represent one standard error of the mean. A score of 6.25% corresponds to chance identification performance.

identification scores for unprocessed speech (not shown) were 100% correct for each session. All listeners showed little effect of training on the identification of speech stimuli in the 16-band *E*-speech condition. For the *E* condition, identification scores and percent of information received for each phonetic feature were therefore averaged across the four repeated sessions, with results indicated on the right side of each figure panel. With 16 possible consonants across the 48 VCV stimuli, 6.25% correct represents chance performance in measures of overall intelligibility.

Figure 3 shows that identification scores improved regularly with training for each TFS-speech condition. Error bars reveal important between-listener differences during training sessions, consistent with previous studies (e.g., Lorenzi *et al.*, 2006). However, high levels of consonant identification were globally achieved at the end of the ten repeated sessions. For individual listeners, mean scores computed across the two best sessions varied between 50 and 90% correct. Across listeners, performance was generally best in the PMz condition with 80% correct identification based on the two best sessions, and poorest in the FM condition in which the mean identification score was 65% correct. An intermediate level of performance (70% correct) was achieved in the PMr condition. This ordering of results with performance better in the PMz than PMr and FM conditions is roughly consistent with the ordering of mean correlation values in the analyses shown in Fig. 2. In the PMz and PMr conditions, mean listener scores correspond to nearly stable performance, but a trend observable in Fig. 3 suggests that intelligibility in the FM condition may have continued to improve with additional training.

Data from the PMz condition approximate the 87% correct reported by Lorenzi *et al.* (2006) from similar conditions. It is noteworthy that similar training effects were observed in the current work with the alternative TFS-processing schemes of PMr and FM. The current results also showed that with much less training (four repeated sessions only), the highest mean identification score was obtained for *E* speech (90%), consistent with results from previous studies comparing the intelligibility of *E* and TFS speech processed through a 16-band analysis filterbank (Lorenzi *et al.*, 2006; Gilbert *et al.*, 2007).

Two repeated-measures analyses of variance (ANOVA) with factor processing condition confirmed these observations. Percent-correct identification scores averaged across the two best sessions were transformed into rationalized arcsine units prior to statistical analysis. The first analysis compared the *E* condition to the average of the TFS ones, and revealed a significant effect of processing scheme [$F_{(1,6)} = 13.28$, $p = 0.01$]. The second analysis was more detailed and involved all four processing schemes (*E*, PMz, PMr, and FM) without averaging across TFS results. As in the first analysis, the main effect of processing scheme was significant [$F_{(3,18)} = 10.79$, $p < 0.001$]. *Post-hoc* analyses (Tukey HSD) indicated that identification scores for *E* speech were significantly greater than those measured for PMr and FM ($p < 0.005$), but did not differ from those measured with PMz speech ($p = 0.1$). Identification scores for PMz speech were significantly greater than those obtained with FM speech ($p < 0.05$), but did not differ from the PMr-speech results ($p = 0.3$). Identification scores for PMr and FM speech also did not differ ($p = 0.67$).

The results of information-transmission analyses revealed that for each TFS-speech condition, the amount of information received for voicing, manner, place, and nasality improved with training. Different patterns of information reception were observed across speech-processing conditions at the end of training sessions. For the PMz, PMr, and FM conditions, greatest information was received for voicing and nasality, less information for place of articulation, and least for manner. For most phonetic features, greatest information was generally received in the PMz condition, and least with the FM scheme. As with the TFS-speech conditions, greatest information was also received for voicing and nasality in the *E* condition. However, less information was received for manner, and least information for place in the *E* condition, opposite the ordering obtained with TFS speech for these two features. Thus, TFS-and *E*-coding schemes differed in the ranking of phonetic information transmitted regarding the occlusive-constrictive and front-middle-back distinctions.

These observations are consistent with the results of two repeated-measures ANOVAs conducted on percent-information-received calculated across the two best sessions. The first analyses included percent-information-received for the *E* and for the mean of the three TFS conditions. Analysis showed significant main effects of factors processing scheme [$F_{(1,6)} = 14.27$, $p < 0.01$] and phonetic feature [$F_{(3,18)} = 19.90$, $p < 0.0001$], along with a significant interaction [$F_{(3,18)} = 3.93$, $p < 0.05$]. To obtain a more detailed view of the interaction between processing scheme and phonetic feature, the second analysis was performed without averaging across TFS conditions. As in the first analysis, there was a significant main effect of processing scheme [$F_{(3,18)} = 10.39$, $p < 0.001$], phonetic feature [$F_{(3,18)} = 23.4$, $p < 0.00001$], and a significant interaction [$F_{(9,54)} = 3.10$, $p < 0.005$]. *Post-hoc* analyses (Tukey HSD) indicated that for each TFS-processing scheme (PMz, PMr, and FM), reception of voicing and nasality did not differ significantly ($p \sim 1$). The same was true for reception of place and manner ($p=0.5-0.9$). Reception of voicing and nasality was significantly greater than reception of place ($p < 0.01$) and manner ($p < 0.01$), except for the FM condition where they did not differ from reception of place ($p \sim 0.1$). For each phonetic feature, reception of information did not significantly vary with TFS-processing scheme ($p=0.8-1.0$), except for voicing and nasality where significantly greater information was received in the PMz then FM condition ($p < 0.05$). Additional Tukey-HSD tests indicated that for the *E* condition, reception of voicing, nasality, and manner did not differ ($p=0.4-1.0$); reception of voicing and nasality were significantly greater than reception of place ($p < 0.05$), but reception of place and manner did not differ ($p=0.59$).

The high levels of intelligibility (>65% correct) obtained with the TFS-processing schemes contrasts with the large drop in mean correlation-based estimates of *E* fidelity shown in Fig. 2, especially in terms of depth- and level-dependent correlation estimates. Indicating poor mean fidelity of *E* reconstruction, this comparison suggests at best a limited basis for TFS-speech recognition. Manipulations—via the use of different TFS-processing schemes—of the extent of *E* reconstruction at the output of auditory filters had clear but relatively modest effects on the intelligibility of TFS speech: (i) for the PM algorithms, randomizing the starting phase of the sinusoidal carriers did not significantly affect speech intelligibility, and (ii) restricting the excursion of instantaneous frequency within the analysis-filter passband affected the reception of voicing and nasality with the FM scheme, but left reception of place and manner unchanged. Moreover, high levels of information reception (>70%) were observed for voicing and nasality despite the reduction of reconstructed *E* cues across all forms of TFS speech. Finally, analysis of information transmission suggested that different phonetic information regarding manner was received with TFS versus *E* speech. This difference between TFS and *E* speech is in line with the fact that a number of specific *E* features such as transients, silent gaps, overall duration, and rise time can signal manner of articulation, especially the distinction between plosives and fricatives (e.g., Rosen, 1992).

Taken together, results indicate that identification of TFS speech does not rely solely on reconstructed *E* cues. Results also emphasize the notion that TFS and *E* speech do not convey the same acoustic/phonetic cues.

2. Experiment 2: Effect of removing low-frequency analysis bands

The correlation data of Fig. 2 show that for the present stimuli, potential *E* reconstruction is greatest in auditory channels centered below about 340 Hz, especially with the PMz scheme. Two factors contribute to this result. First, the low-frequency region encompasses the region of the fundamental frequency (216 Hz) of the female voice used in the current experiments. The second aspect relates to the fact that the fidelity of *E* reconstruction is inversely related to cochlear-filter bandwidth. Greater reconstruction of *E* cues is therefore expected in the low-frequency region where auditory filters are narrower.

Estimates of weighting or importance functions that indicate contribution to speech perception by CF of analysis band are not uniform. While the function of French and Steinberg (1947) has an initial highpass slope which reduces contribution from the low frequencies (and would thereby minimize anticipated effect of the stimulus manipulation of experiment 2), the 1980 Speech Transmission Index (STI) of Steeneken and Houtgast [see Steeneken and Houtgast (1999)] is relatively flat [for discussion of these differences, see Humes *et al.* (1986)]. Subsequent work by Steeneken and Houtgast (2002) did show a highpass segment when evaluating errors within a phoneme class. Effect of spectral content on within-class error is anticipated in the work of Miller and Nicely (1955). In that work, the contrast between results obtained with either lowpass or highpass filtering of stimuli showed greater predictability of error in the lowpass conditions. Miller and Nicely attributed the result to greater redundancy of speech information in the lowpass conditions. A similar emphasis on low-frequency content was reported by Turner *et al.* (1998) who measured weighting functions in broadband conditions. Turner and his co-workers interpreted their result as indicating a facilitative contribution to speech perception of redundant low-frequency content. In that demonstrations of cross-spectral utilization of redundancy are most often based on modulation [e.g., comodulation masking release, and see Sheft (2008)], and that modulation processing is the basis of the algorithms of the present work, a basis for the stimulus manipulation of experiment 2 is seen in past studies.

The goal of the second experiment was to assess the contribution of low-frequency analysis bands to TFS-speech intelligibility. In the second experiment (termed HF, for "High Frequency"), the VCV stimuli of experiment 1 were processed using the PMz, PMr, and FM schemes, but with the lowest five analysis filters centered below 340 Hz removed from the synthesis process. In the "unprocessed-speech" condition, the original VCV stimuli were simply passed through the analysis filterbank with the lowest five filters omitted in signal synthesis. Listeners were tested over four repeated sessions for each HF TFS-speech condition, and just once for the unprocessed HF speech. The three HF TFS-speech conditions were interleaved with order of presentation randomized across listeners. The apparatus, procedure, and presentation level were identical to those used in experiment 1.

In Fig. 4, bars corresponding to the HF condition show the mean identification scores and percent of information received for each phonetic feature calculated across listeners and sessions. Consistent with previous work (e.g., French and Steinberg, 1947), the mean identification score for unprocessed HF speech was 100% correct. For comparison, mean data from the two best sessions of experiment 1 (labeled 16B) are shown on the left of each panel. For each speech-processing condition, similar results were obtained in the 16B and HF conditions. Removing information in the low audio-frequency range (<340 Hz) encompassing the re-

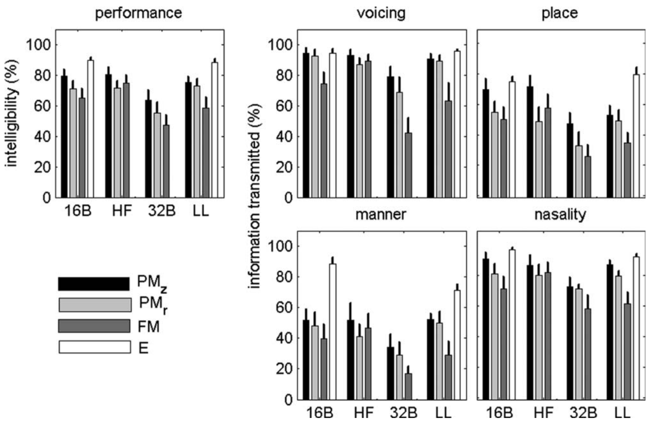

FIG. 4. Mean identification performance (left panel) and percent of information received for each phonetic feature (middle and right panels) calculated in each speech-processing condition: PMz (black bars), PMr (light gray bars), FM (dark gray bars), *E* (open bars). In each panel, data are from the 16B [16-band analysis filterbank, 80 dB(A)], HF [16-band analysis filterbank with the five lowest bands removed, 80 dB(A)], 32B [32-band analysis filterbank, 80 dB(A)], and LL [16-band analysis filterbank, 45 dB(A)] conditions. Error bars represent one standard error of the mean.

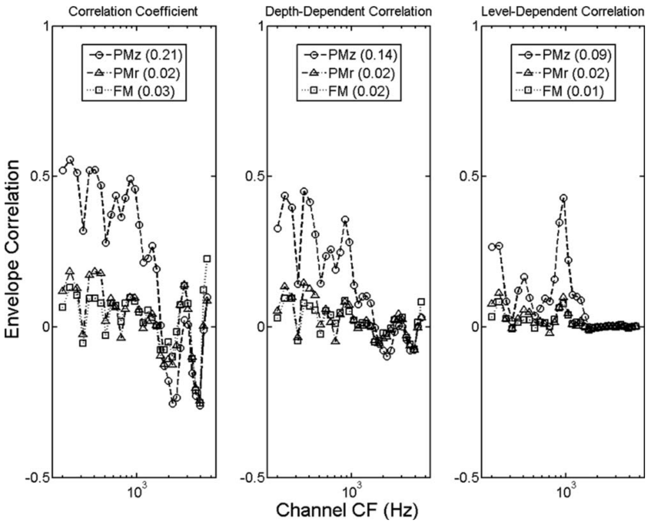

FIG. 5. For experiment 3, assessment of envelope reconstruction in condition 32B. Stimuli were processed with a 32-band analysis filterbank. Except for the omission of the mean correlation estimates for the *E* condition which was not run in experiment 3, otherwise as in Fig. 2.

gion of the stimulus fundamental frequency did not significantly affect identification scores or phonetic-feature reception, as shown by two repeated-measures ANOVAs and *post-hoc* tests.[2] These results demonstrate that *E* cues potentially reconstructed in the low audio-frequency range where auditory filters are narrowest do not contribute to the intelligibility of TFS speech.

3. Experiment 3: Effect of increasing analysis-filterbank frequency resolution

Results shown in Fig. 2 indicate that the reconstruction of *E* cues at the output of most auditory filters is not abolished with an analysis-filter bandwidth of roughly 2 ERB. Additional simulations were run to assess the effect of increasing analysis-filter resolution on *E* reconstruction. Analysis-filterbank parameters increased the number of channels from 16 to 32 while reducing filter bandwidth in half to approximately 1 ERB. Results of correlation analysis are presented in Fig. 5. Compared to the 16-channel results (see Fig. 2), correlation estimates from the 32-band condition (termed 32B) decrease markedly in most auditory channels with the PMr and FM algorithms. For the PMz scheme, correlation estimates are either unaffected or elevated in auditory channels tuned between roughly 300 and 1000 Hz. The increase in correlation observed for the PMz scheme is not dependent on the specific values of the gammatone-filter CFs.

The goal of the third experiment was to assess the effect of adjusting the bandwidths of the analysis filterbank to approximate normal auditory frequency resolution. In the 32B condition, the VCV stimuli were processed with the PMz, PMr, and FM schemes, using a 32-band analysis filterbank. In the unprocessed-speech condition, the original VCV stimuli were unaltered. Since previous studies indicate that consonant recognition asymptotes with 16-band resolution (e.g., Shannon *et al.*, 1995; Xu *et al.*, 2005), the *E* condition was not included. Listeners were tested over four repeated sessions for each 32B TFS-speech condition, and just once with unprocessed speech. The four 32B conditions were interleaved with order of presentation randomized across listeners. Apparatus, procedure, and presentation level were identical to those described in experiment 1.

In Fig. 4, bars corresponding to the 32B condition show the mean identification scores and percent of information received for each phonetic feature calculated in each TFS-speech condition. The mean identification score for unprocessed speech was 100% correct. Compared to the 16B results, consonant identification dropped by roughly 20 percentage points in every 32B condition. The greatest change in reception of phonetic information was for voicing and place in the PMr and FM conditions. It is, however, important to note that identification scores and reception of nasality remained greater than 50% in all conditions. These observations were confirmed by two repeated-measures ANOVAs and *post-hoc* analysis.[2]

Results indicate that increasing analysis-filterbank resolution to approach normal auditory resolution degrades the intelligibility of TFS speech with *all* TFS processing schemes. Although predicted on theoretical grounds, this finding is only partially consistent with the correlation data reported in Fig. 5. *E* reconstruction at the output of auditory filters is notably degraded by the increase in analysis-filter resolution for PMr and FM speech, and left unchanged (and sometimes elevated) with PMz speech. Moreover, despite correlation estimates close to zero with either the PMr or FM scheme, speech identification scores were greater than 50%. Thus, the discrepancy between correlation analysis and psychophysical data provide additional evidence that TFS-speech intelligibility is not reliant on reconstructed *E* cues. The degradation in TFS-speech intelligibility in the 32B condition remains to be explained. One possibility is that increasing frequency resolution to 32 bands not only affects the fidelity of *E* reconstruction, but also disrupts the trans-

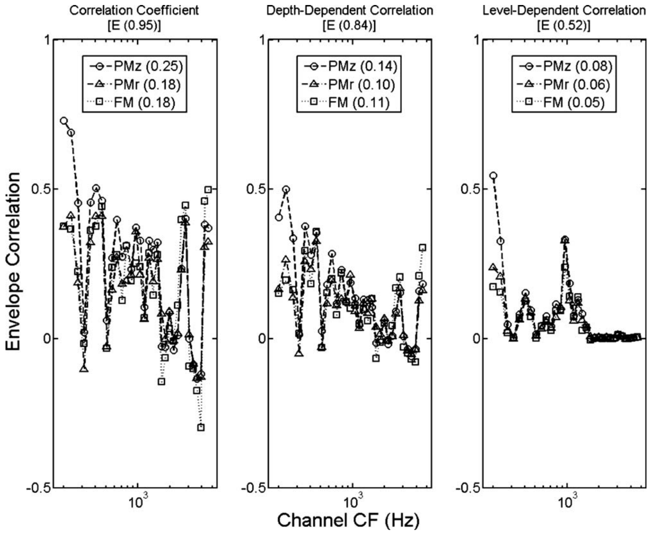

FIG. 6. For experiment 4, assessment of envelope reconstruction in condition LL. Stimulus level was 45 dB SPL, otherwise as in Fig. 2.

mission of TFS cues *per se*. This hypothesis is explored in Sec. III.

4. Experiment 4: Effect of decreasing stimulus level

Experiments 1–3 were run with stimuli presented at 80 dB(A). A number of physiological (e.g., Rhode, 1971; Robles *et al.*, 1986) and psychophysical studies (e.g., Glasberg and Moore, 2000; Rosen and Stock, 1992; Hicks and Bacon, 1999; Bernstein and Oxenham, 2006) indicate that frequency selectivity improves at low levels, at least for frequencies of 1 kHz and above. As a consequence, reducing stimulus presentation level to about 40–45 dB sound pressure level (SPL) should enhance the fidelity of *E* reconstruction for TFS speech. This predicted effect of auditory-filter resolution is opposite the one associated with analysis-filter resolution, explored in experiment 3.

Predictions concerning the effect of auditory-filter bandwidth were obtained by comparing *E* reconstruction at the output of gammatone filters for 80 and 45 dB SPL presentation levels. Results obtained at 45 dB SPL are shown in Fig. 6. For this analysis, gammatone-filter bandwidth was narrowed by a factor of 1.5 to simulate the effect of level reduction on auditory frequency resolution (Glasberg and Moore, 2000). Overall, the correlation data indicate that *E* reconstruction increases with reduction in level. The predicted changes in *E* reconstruction are rather modest. However, these changes are consistent across auditory channels and TFS-processing schemes.

The goal of the fourth experiment was to assess the effect of decreasing stimulus level on TFS-speech intelligibility. In this final experimental condition (labeled LL, for "Low Level"), the same VCV stimuli were processed using the PMz, PMr, FM, and *E* schemes described above, using the 16-band analysis filterbank. In the unprocessed-speech condition, the original VCV stimuli were unaltered. Listeners were tested over four repeated sessions for each of the four LL TFS-speech conditions and with *E* speech; data were collected from only one session with unprocessed LL speech. The TFS- and *E*-speech conditions were interleaved with order of presentation randomized across listeners. The apparatus, procedure, and presentation level were identical to those described in experiment 1, except that stimuli were presented at an average level across listeners of 45 dB SPL. This level was initially determined individually as the minimum presentation level yielding both 100% identification with unprocessed speech and an *E*-speech identification score similar to the one obtained at 80 dB(A). This lower level varied across the six listeners between 40 and 49 dB(A).

In Fig. 4, bars corresponding to the LL condition show the mean identification scores and percent of information received for each phonetic feature in each speech-processing condition (PMz, PMr, FM, and *E*). As indicated above, the mean identification score across listeners obtained for unprocessed LL speech was 100% correct. Identification scores were not significantly affected by reduction in presentation level in any of the processed-speech conditions.[3] In principle, decreasing stimulus level could affect performance for reasons unrelated to the fidelity of *E* reconstruction. However, the absence of a significant level effect on mean performance with *E* speech argues against involvement of such factors counteracting potential level effects on *E* reconstruction in the TFS conditions. In contrast to mean performance, information received for all phonetic features was significantly affected by level. Compared to results from experiments 1–3, the maximum decrease in information received due to the reduction in level was observed for place reception in the PMz-speech condition, and manner reception in the *E*-speech condition.

Results show that when expressed in terms of identification performance, TFS and *E* speech are relatively robust to change in level (and therefore audibility). The data also indicate that moderate modifications in auditory-filter bandwidth due to change in level, and that subsequent changes in *E* reconstruction, especially in the high-frequency region where level effects on frequency resolution are greater, do not affect TFS-speech intelligibility. This result indicates that changes in audibility and frequency resolution are an unlikely basis of the deficits in TFS intelligibility reported by Lorenzi *et al.* (2006) for listeners with mild-to-moderate cochlear hearing loss.

III. DISCUSSION

The goal of the present research was to evaluate the contribution of subband TFS cues to consonant identification in experimental conditions intended to reduce the reconstruction of *E* cues, an artifact resulting from the use of the Hilbert transform to extract TFS. Overall, results indicated that all factors manipulated in order to alter the fidelity of *E* reconstruction had little to no effect on TFS-speech intelligibility. Decreasing the bandwidth of analysis filters to approach normal auditory resolution and restricting the bandwidth of the processed TFS signal to the analysis-filter bandwidth in the FM scheme had greater effects on TFS-speech intelligibility than randomizing the starting phase of the subband TFS signals in the PMr scheme. Nevertheless, moderate to high levels of consonant identification and

TABLE II. The mean envelope correlation between outputs of adjacent gammatone-filterbank channels. Each column is for a different processing scheme, and each row for a separate experimental condition with 16B, 32B, and LL referring to conditions from experiments 1, 3, and 4, respectively.

	E	PMz	PMr	FM
16B	0.99	0.67	0.68	0.60
32B	[a]	0.72	0.70	0.67
LL	0.99	0.38	0.41	0.28

[a] The *E* processing scheme was not used in experiment 3.

phonetic-information reception were achieved despite such stimulus manipulations. In addition, two manipulations intended to assess the contribution of low- and high-frequency regions to *E* reconstruction (removal of frequency information in the F0 region and change in presentation level, respectively) left TFS-speech intelligibility relatively unchanged. Finally, in certain experimental conditions (e.g., condition 32B), predictions based on *E* reconstruction were not compatible with PMz-speech identification data. Taken together, results suggest that the intelligibility of TFS consonants is not dependent on reconstruction of envelope cues by auditory filtering.

Signal analysis has so far only considered local or within-channel characteristics to estimate the fidelity of *E* reconstruction in the TFS-speech conditions. Crouzet and Ainsworth (2001) measured the between-channel envelope correlations of natural speech, finding high inter-channel correlations, especially between adjacent channels. Crouzet and Ainsworth argued that these correlations in natural speech aid consonant recognition. Importance of cross-channel correlation was recognized by Steeneken and Houtgast (1999) who incorporated a redundancy correction into their STI algorithm. To further explore the distinction between *E* and reconstructed *E*, adjacent-channel envelope correlations were calculated for the stimuli of experiments 1,3, and 4 with results shown in Table II. Each table entry is the mean envelope correlation between outputs of adjacent gammatone-filterbank channels. As in the other correlation calculations, correlation estimates were averaged in terms of Fischer's *z* values across the 48 VCV utterances. Comparison of results obtained in the *E* conditions to those from the three TFS-speech conditions (PMz, PMr, and FM) indicates large drops in adjacent-channel envelope correlations if based on *E* reconstruction, most notably in the low-level conditions of experiment 4. Thus, not only is there a loss of local or within-channel fidelity with *E* reconstruction (i.e., the results of Figs. 2 and 6), there is also a loss of a more global or cross-channel fidelity implicated in speech perception.

Signal analysis also so far has been based on a single processing structure utilizing a gammatone filterbank. To evaluate the effect of model parameters, all simulations were rerun using three additional model structures. In the first, the approach of Oxenham and Moore (1997) was added to the gammatone model to incorporate the effect of basilar membrane compression. The remaining two additional model structures used gammachirp filters (Irino and Patterson, 2006), with one the static and the other the dynamic realization. Across all conditions, effects of model structure on measured correlations were relatively small, with consistently the largest effect of model type obtained in the PMz conditions. PMz results from all model structures are shown in Table III. Compared to the initial gammatone results, the largest effect of model modification on the fidelity of *E* reconstruction was obtained with the dynamic gammachirp filterbank. With fidelity of *E* reconstruction dependent on filter passband characteristics (see Sec. I) and the dynamic model temporally modifying these characteristics, some effect on *E* reconstruction is anticipated. However, the largest effects, those from the PMz conditions, are still relatively small and do not alter the interpretation of insufficient basis by *E* reconstruction to fully account for listener performance.

On first pass, the result of relatively small effect of model structure may seem surprising, especially in light of other work indicating larger effects [e.g., Stone and Moore (2007)]. However, the contrast of result is due to analysis structure. For example, Stone and Moore evaluated the effect of compression by comparing the original to the processed stimuli. In the present work, the central question is the effect of stimulus signal processing (i.e., the PMz, PMr, and FM algorithms) as estimated at some level of the auditory system. To answer this question, comparisons involve stimuli that have undergone processing through the same auditory-model structure. In other words, the analysis is not directly

TABLE III. For the PMz processing scheme, mean correlation estimates across simulated auditory filterbank channels assessing fidelity of *E* reconstruction. Each column is for a different model structure (see text for additional details). Each row is for a separate combination of experimental condition (16B, 32B, and LL referring to conditions from experiments 1, 3, and 4, respectively) and correlation metric. The values from the first column are also shown in the figure legends of Fig. 2.

	Gammatone	Compressive Gammatone	Static Gammachirp	Dynamic Gammachirp
16B Correlation Coefficient	0.21	0.23	0.24	0.34
16B Depth-Dependent Correlation	0.12	0.12	0.13	0.16
16B Level-Dependent Correlation	0.08	0.08	0.10	0.11
32B Correlation Coefficient	0.21	0.22	0.23	0.32
32B Depth-Dependent Correlation	0.14	0.14	0.14	0.15
32B Level-Dependent Correlation	0.09	0.09	0.10	0.11
LL Correlation Coefficient	0.25	0.26	0.24	0.27
LL Depth-Dependent Correlation	0.14	0.14	0.13	0.14
LL Level-Dependent Correlation	0.08	0.08	0.10	0.11

evaluating the effects of various model structures, but rather determining if use of various accepted model structures leads to differences in effect of speech-processing scheme.

The decrease in TFS-speech intelligibility with increase in number of analysis-filterbank channels from 16 to 32 is not accounted for by degradation in reconstructed E cues. Specifically, correlation analyses from experiment 3 predicted an effect of TFS-processing scheme not obtained in the subject data. Also, performance levels in the PMr and FM conditions were higher than would be anticipated with a correlation of reconstructed E of close to zero.

Alternative to involvement of E reconstruction, increasing frequency resolution may disrupt the transmission of TFS cues. Figure 1 illustrates that IFF spikes are one characteristic of the fine structure of filtered speech signals; increasing analysis-band frequency resolution progressively distorts this aspect of TFS speech. More generally, any attempt to alter the fidelity of E reconstruction should, in turn, affect the fidelity of TFS transmission because E and TFS are related. This hypothesis was explored by assessing the fidelity of TFS transmission. An approach similar to the one used to assess E reconstruction was applied to quantify the fidelity of TFS transmission. Unprocessed speech signals and TFS speech generated using the PMz, PMr, and FM schemes were passed separately through the bank of 32 gammatone auditory filters described above. Using the FM-processing scheme, TFS signals were extracted from the output of each gammatone filter. FM functions were restricted in order to keep deviations in instantaneous frequency within a 1 ERB passband. Flanagan and Golden (1966) observed similar effects on intelligibility when lowpass filtering the E and FM functions of vocoded speech. Consequently, FM functions were lowpass filtered at 64 Hz, using the same filtering applied in the estimation of E reconstruction. For each VCV utterance, mean correlation coefficients were computed between the FM functions of the unprocessed VCV stimuli and the corresponding TFS signals. Two types of correlation estimates were considered: (i) the correlation coefficient, and (ii) a level-dependent correlation estimate. A high correlation estimate indicates a close resemblance between the original TFS and that of the processed stimuli at the output of auditory filters.

Results are shown in Figs. 7 and 8 for stimuli generated with 16- and 32-band analysis filterbanks, respectively. Overall, the fidelity of TFS transmission (as estimated by both correlation indexes) is notably degraded across most auditory channels when the frequency resolution of the analysis filterbank is increased from 16 to 32 bands. This reveals that E reconstruction is not the only factor affected by change in the analysis filterbank. Additional simulations were therefore run to assess the fidelity of TFS transmission in the other experimental conditions. To relate listener performance to signal analysis, mean identification scores are plotted in Fig. 9 as a function of the fidelity of E reconstruction (top panels) and TFS transmission (bottom panels) for each experimental condition (16B, HF, 32B, LL) and TFS-speech processing scheme (PMz, PMr, FM). The variance in identification scores accounted for by each correlation estimate (R^2) is much greater for TFS fidelity (e.g., 69% of variance with the level-dependent correlation estimate) than for E reconstruction (e.g., 35% of variance with the level-dependent estimate). To evaluate the significance of the five regressions of Fig. 9, separate ANOVAs were performed on the data of each panel. Level of significance was reduced from 0.05 to 0.01 to correct error rate for the use of five analyses based on the same listener data. Results confirmed that regressions were only significant for fidelity of TFS transmission (the bottom panels of Fig. 9) and not for the fidelity of E reconstruction (the top panels). For TFS results based on the correlation coefficient, analysis had $F_{(1,10)}=11.43$, $p=0.007$, and for the level-dependent results, $F_{(1,10)}=22.15$, $p=0.001$. Simulations thus confirm that altering the fidelity of E reconstruction also affects the fidelity of TFS transmission. More importantly,

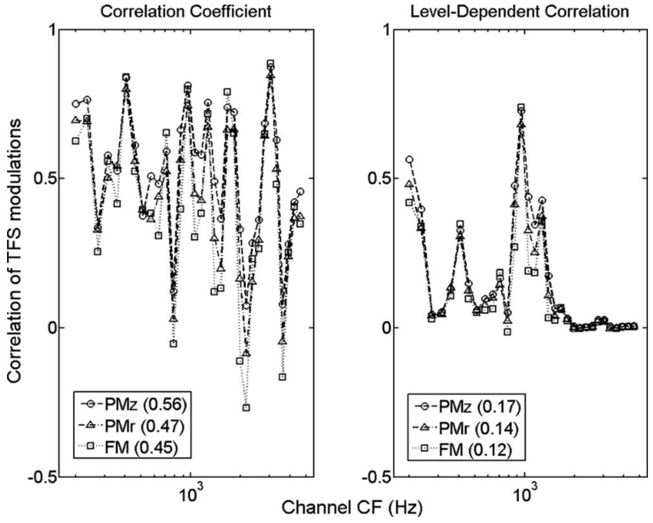

FIG. 7. Assessment of fidelity of TFS transmission in condition 16B. Mean correlation estimates computed between the TFS of the original and processed speech stimuli in the PMz (open circles), PMr (open triangles), and FM (open squares) conditions are shown a function of gammatone-filter CF. Left panel: correlation coefficient; right panel: level-dependent correlation estimate. In each figure legend, numbers between parentheses correspond to the mean correlation estimate computed across gammatone-filter channels.

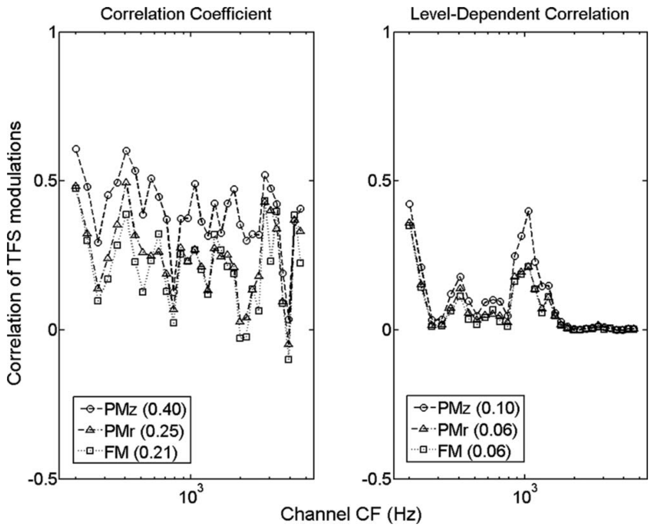

FIG. 8. Assessment of fidelity of TFS transmission in condition 32B. Stimuli were processed with a 32-band analysis filterbank, otherwise as in Fig. 7.

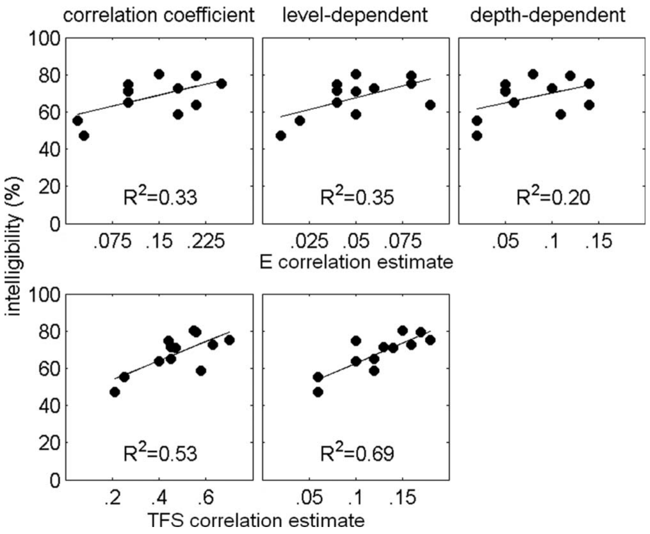

FIG. 9. Mean identification scores across listeners as a function of fidelity of *E* reconstruction (top panels) and TFS transmission (bottom panels). Each panel corresponds to a given correlation estimate. In each panel, individual symbols correspond to a given experimental condition (16B, HF, 32B, LL) and TFS-speech processing scheme (PMz, PMr, FM). Variance of identification scores accounted for by each correlation estimate (R^2) is shown in each panel.

they support the interpretation that variation in the fidelity of TFS transmissions was a stronger determinant of listener performance in the present study than was the extent of *E* reconstruction.

Despite significance of regression, low absolute values of correlation in the TFS conditions, especially in the level-dependent case, call for further comment. In general, significance of regression indicates capturing of trend despite approximation of metric. By definition, correlation averages across stimulus duration to diminish contribution of local distinctive features. As noted in the Introduction, spikes of the IFF, a local distinctive feature, are a significant contributor to *E* reconstruction. IFF spikes may also represent a significant feature in terms of fine structure, with perceptual validity established in the report of Jeffress (1968). For speech stimuli, spiking of the IFF is not simply an artifact, but rather an intrinsic stimulus aspect coding envelope troughs, onsets, offsets, and rapid phase transitions due to articulation. As such, they may convey relevant information for speech perception. Current work is evaluating the contributions of specific features of TFS to speech perception with the intention that results may allow for refinement of analysis metrics.

In a seminal study describing the temporal information present in speech, Rosen (1992) suggested that *E* cues signaled mainly segmental cues to manner and voicing, whereas TFS contributed primarily to segmental cues to place, and to a smaller extent voicing and nasality. The results of the present work are consistent with the outcome of Rosen's analysis, except that the contribution of TFS cues was greater for voicing than place. Nasality was shown to be extremely well transmitted by both *E* and TFS cues, presumably because all nasals are voiced in French. The association between manner and voicing features permits signaling of manner information by voicing patterns (cf. Rosen, 1992). The current findings also establish that TFS and E cues do not signal segmental cues in identical ways, especially in the case of manner and place. Overall, the differences in feature-transmission scores between TFS and *E* speech demonstrate that two kinds of information can be extracted from the analytic signal. Thought *E* reconstruction at the output of cochlear channels cannot be fully eliminated from the processing of TFS signals, results from the present work support the interpretation that *E* and TFS cues can make separate and distinct contributions to speech perception.

IV. CONCLUSIONS

Taken together, the results indicate that:

(1) Moderate to high levels of consonant identification can be obtained on the basis of speech fine-structure cues extracted using two different speech-processing techniques with either 16 or 32 analysis-frequency bands. Moreover, compared to temporal envelope cues, fine-structure conveys different phonetic information regarding manner and place of articulation. These data support the results of two recent studies conducted by Lorenzi *et al.* (2006) and Gilbert *et al.* (2007) suggesting that temporal fine-structure cues carry specific information useful for speech identification.

(2) Consistent with the results of Gilbert and Lorenzi (2006), envelope cues reconstructed at the output of auditory filters do not contribute substantially to consonant identification when 1- or 2-ERB-wide analysis filters are used. A signal-processing technique which extracts the FM-speech patterns within each analysis band and restricts deviations in instantaneous frequency within the analysis-filter bandwidth minimizes the contribution of such reconstructed envelope cues. However, correlation analysis indicates that such a scheme also degrades TFS speech cues.

(3) The identification of consonants on the basis of speech fine-structure cues is robust to variation in stimulus level and auditory frequency resolution. This suggests that the large deficits in TFS-speech intelligibility previously reported for listeners with mild-to-moderate cochlear hearing loss are unlikely due to changes in audibility and frequency selectivity.

Allowing for measurement of speech perception based on temporal fine-structure cues in normal-hearing and hearing-impaired listeners, results from the present procedures may help in the design of prosthetic devices. In addition, the correlation analysis suggests a role for quantitative measures of the fidelity of TFS transmission in device assessment.

ACKNOWLEDGMENTS

This research was supported by NIDCD Grant No. DC005423 to S. S., a MENRT/Paris Descartes grant to M. A., and a research fellowship from the Parmly Hearing Institute of Loyola University Chicago to C.L. This research was also supported by the GDR CNRS 2967 GRAEC. The authors thank C. Micheyl, A. Oxenham, and N. Viemeister for valuable discussions on the results of this study.

[1] The analysis filter was a third-order zero-phase Butterworth filter. The filter CF (900 Hz) was chosen to encompass formant transitions of the unprocessed vowel signal. The low and high 3-dB-cutoff frequencies of the analysis filter were 766 and 1024 Hz, respectively. In the current example, the starting phase of the PMr sinusoidal carrier was 90°, contrasting with the 0° of the PMz scheme.

[2] Two repeated-measures ANOVAs were conducted on the data collected across the seven listeners of experiments 1, 2, and 3. The first ANOVA was conducted on identification scores transformed into rationalized arcsine units prior to statistical analysis with factors experimental condition (three levels: 16B, HF, and 32B) and processing scheme (three levels: PMz, PMr, and FM). The second ANOVA was conducted on arcsine-transformed percent-information-received with factors experimental condition (three levels), processing scheme (three levels), and phonetic feature (four levels: voicing, nasality, manner, and place). The first ANOVA showed significant main effects of experimental condition [$F_{(2,12)}=20.41$, $p<0.0005$] and processing scheme [$F_{(2,12)}=21.18$, $p<0.0005$], without significant interaction between factors [$F_{(4,24)}=2.68$, $p=0.06$]. *Post-hoc* analyses (Tukey HSD) indicated that identification scores differed significantly between the 16B and 32B conditions only ($p<0.005$). The second ANOVA showed significant main effects of experimental condition [$F_{(2,12)}=18.77$, $p<0.0005$], processing scheme [$F_{(2,12)}=46.70$, $p<0.0001$], and phonetic feature [$F_{(3,18)}=30.07$, $p<0.0001$]; all interactions between factors were significant at the 0.05 level except for interaction between factors experimental condition and phonetic feature, and the triple interaction between factors experimental condition, scheme, and phonetic feature. Interaction between factors experimental condition and scheme restricted to the 16B and 32B conditions was also significant [$F_{(4,24)}=5.46$, $p<0.005$].

[3] Two repeated-measures ANOVAs were conducted on the data collected across the six common listeners of experiments 1 and 4. The first ANOVA was conducted on identification scores with factors experimental condition (two levels: 16B and LL) and processing scheme (four levels: PMz, PMr, FM, and E). The second ANOVA was conducted on percent-information-received with factors experimental condition (two levels), processing scheme (four levels), and phonetic feature (four 1evels). The first ANOVA showed that the main effect of experimental condition (i.e., stimulus level) on identification scores was not significant [$F_{(1,5)}=1.21$; $p=0.32$]. The main effect of processing scheme was significant [$F_{(3,15)}=13.41$, $p<0.001$], but the interaction between factors experimental condition and scheme was not [$F_{(3,15)}=1.25$, $p=0.33$]. The second ANOVA showed significant main effects of experimental condition [$F_{(1,5)}=9.14$, $p<0.05$], phonetic feature [$F_{(3,15)}=40.87$, $p<0.0001$], and scheme [$F_{(3,15)}=18.79$, $p<0.0001$]. The interactions between factors experimental condition and processing scheme [$F_{(3,15)}=0.23$, $p=0.87$] and between factors experimental condition and phonetic feature [$F_{(3,15)}=1.06$, $p=0.39$] were not significant. As in the l6B condition alone, the interaction between factors processing scheme and phonetic feature was significant [$F_{(9,45)}=3.02$, $p<0.01$]. This pattern of results was the same across experimental conditions because the interaction between factors experimental condition, processing scheme, and phonetic feature was not significant [$F_{(9,45)}=1.57$, $p=0.15$].

Bernstein, J. G. W., and Oxenham, A. J. (**2006**). “The relationship between frequency selectivity and pitch discrimination: Effects of stimulus level,” J. Acoust. Soc. Am. **120**, 3916–3928.

Crouzet, O., and Ainsworth, W. A. (**2001**). “On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation,” in *Workshop on Consistent and Reliable Cues for Sound Analysis* (Aalborg, Denmark).

de Cheveigné, A., and Kawahara, H. (**2002**). “YIN, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Am. **111**, 1917–1930.

Drennan, W. R., Won, J. H., Dasika, V. K., and Rubenstein, J. T. (**2007**). “Effects of temporal fine structure on the lateralization of speech and on speech understanding in noise,” J. Assoc. Res. Otolaryngol. **8**, 373–383.

Drullman, R. (**1995**). “Temporal envelope and fine structure cues for speech intelligibility,” J. Acoust. Soc. Am. **97**, 585–592.

Flanagan, J. L. (**1980**). “Parametric coding of speech spectra,” J. Acoust. Soc. Am. **68**, 412–419.

Flanagan, J. L., and Golden, R. M. (**1966**). “Phase vocoder,” Bell Syst. Tech. J. **45**, 1493–1509.

French, N. R., and Steinberg, J. C. (**1947**). “Factors governing the intelligibility of speech sounds,” J. Acoust. Soc. Am. **19**, 90–119.

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

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.

Glasberg, B. R., and Moore, B. C. J. (**1990**). “Derivation of auditory filter shapes from notched-noise data,” Hear. Res. **47**, 103–138.

Glasberg, B. R., and Moore, B. C. J. (**2000**). “Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise,” J. Acoust. Soc. Am. **108**, 2318–2328.

Hicks, M. L., and Bacon, S. P. (**1999**). “Psychophysical measures of auditory nonlinearities as a function of frequency in individuals with normal hearing,” J. Acoust. Soc. Am. **105**, 326–338.

Humes, L. E., Dirks, D. D., Bell, T. S., Ahlstrom, C., and Kincaid, G. E. (**1986**). “Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners,” J. Speech Hear. Res. **29**, 447–462.

Irino, T., and Patterson, R. D. (**1997**). “A time domain, level dependent auditory filter: The gammachirp,” J. Acoust. Soc. Am. **101**, 412–419.

Irino, T., and Patterson, R. D. (**2006**). “A dynamic compressive gammachirp auditory filterbank,” IEEE Trans. Audio, Speech, Lang. Process. **14**, 2222–2232.

Jeffress, L. A. (**1968**). “Beating sinusoids and pitch changes,” J. Acoust. Soc. Am. **43**, 1464.

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.

Loughlin, P. J., and Tacer, B. (**1996**). “On the amplitude and frequency decomposition of signals,” J. Acoust. Soc. Am. **100**, 1594–1601.

Martin, P. (**1996**). *Éléments de phonétique, avec application au français (Elements of Phonetics, with Application to French)* (Les Presses de l’Universite Laval, Sainte-Foy).

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

Nelson, P. B., Jin, S., Carney, A. E., and Nelson, D. A. (**2003**). “Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners,” J. Acoust. Soc. Am. **113**, 961–968.

Nie, K., Stickney, G., and Zeng, F.-G. (**2005**). “Encoding frequency modulation to improve cochlear implant performance in noise,” IEEE Trans. Biomed. Eng. **52**, 64–73.

Oxenham, A. J., and Moore, B. C. J. (**1997**). “Modeling the effects of peripheral nonlinearity in listeners with normal and impaired hearing,” in *Modeling Sensorineural Hearing Loss*, edited by W. Jesteadt (Erlbaum, Mahwah, N.J.), pp. 273–288.

Papoulis, A. (**1983**). “Random modulation: A review,” IEEE Trans. Acoust., Speech, Signal Process. **31**, 96–105.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). “An efficient auditory filterbank based on the gammatone function,” in *Paper Presented at a Meeting of the IOC Speech Group on Auditory Modeling at RSRE*, Malvern, England, December 14–15.

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.

Qin, M. K., and Oxenham, A. J. (**2006**). “Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech,” J. Acoust. Soc. Am. **119**, 2417–2426.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). “Speech perception without traditional speech cues,” Science **212**, 947–949.

Remez, R. E., and Rubin, P. E. (**1990**). “On the perception of speech from time-varying acoustic information: Contributions of amplitude variations,” Percept. Psychophys. **48**, 313–325.

Rhode, W. S. (**1971**). “Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique,” J. Acoust. Soc. Am. **49**, 1218–1231.

Robles, L., Ruggero, M. A., and Rich, N. C. (**1986**). “Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input-output functions, tuning curves and phase responses," J. Neurophysiol. **41**, 692–704.

Rosen, S. (**1992**). "Temporal information in speech: acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.

Rosen, S., and Stock, D. (**1992**). "Auditory filter bandwidths as a function of level at low frequencies (125-1 kHz)," J. Acoust. Soc. Am. **92**, 773–781.

Schimmel, M., and Altas, L. E. (**2005**). "Coherent envelope detection for modulation filtering of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, Philadelphia, PA, pp. 221–224.

Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Sheft, S. (**2008**). "Envelope processing and sound-source perception," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 233–280.

Sheft, S., and Yost, W. A. (**2001**). "Auditory abilities of experienced signal analysts," AFRL Prog. Rep. No. 1, contract SPO700-98-D-4002.

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

Steeneken, H. J. M., and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility," Speech Commun. **28**, 109–123.

Steeneken, H. J. M., and Houtgast, T. (**2002**). "Phoneme group specific octave-band weights in predicting speech intelligibility," Speech Commun. **38**, 399–411.

Stickney, G. S., Nie, K., and Zeng, F.-G. (**2005**). "Contribution of frequency modulation to speech recognition in noise," J. Acoust. Soc. Am. **118**, 2412–2420.

Stone, M. A., and Moore, B. C. J. (**2007**). "Quantifying the effects of fast-acting compression on the envelope of speech," J. Acoust. Soc. Am. **121**, 1654–1664.

Turner, C. W., Kwon, B. J., Tanaka, C., Knapp, J., Hubbartt, J. L., and Doherty, K. A. (**1998**). "Frequency-weighting functions for broadband speech as estimated by a correlational method," J. Acoust. Soc. Am. **104**, 1580–1585.

Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.

Xu, L., and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Am. **114**, 3024–3027.

Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). "Relative contribution of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.

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.

Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.