On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system

Shihab Shamma[a)]
Electrical and Computer Engineering Department and Institute for Systems Research, University of Maryland, College Park, Maryland 20742

Christian Lorenzi[b)]
Institut d'Etude de la Cognition, Ecole normale supérieure, Paris Sciences et Lettres, 29 rue d'Ulm, 75005 Paris, France

(Received 20 July 2011; revised 15 February 2013; accepted 26 February 2013)

There is much debate on how the spectrotemporal modulations of speech (or its spectrogram) are encoded in the responses of the auditory nerve, and whether speech intelligibility is best conveyed via the "envelope" (E) or "temporal fine-structure" (TFS) of the neural responses. Wide use of vocoders to resolve this question has commonly assumed that manipulating the amplitude-modulation and frequency-modulation components of the vocoded signal alters the relative importance of E or TFS encoding on the nerve, thus facilitating assessment of their relative importance to intelligibility. Here we argue that this assumption is incorrect, and that the vocoder approach is ineffective in differentially altering the neural E and TFS. In fact, we demonstrate using a simplified model of early auditory processing that both neural E and TFS encode the speech spectrogram with constant and comparable relative effectiveness regardless of the vocoder manipulations. However, we also show that neural TFS cues are less vulnerable than their E counterparts under severe noisy conditions, and hence should play a more prominent role in cochlear stimulation strategies.

© *2013 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4795783]

PACS number(s): 43.66.Ba, 43.64.Bt, 43.64.Sj, 43.71.An [CJP]   Pages: 2818–2833

I. INTRODUCTION

Current understanding of how speech information is represented in the early auditory system derives mostly from neurophysiological investigations of cochlear function and auditory nerve responses in animal models (Sachs and Young, 1979), as well as perceptual studies that relied primarily on "vocoders" (Dudley *et al.*, 1939) to manipulate the speech signal. Debate has often revolved around the role of the speech "temporal envelope" and "temporal fine-structure," their relative contribution to intelligibility (e.g., Drullman, 1995; Shannon *et al.*, 1995; Smith *et al.*, 2002; Zeng *et al.*, 2004, 2005; Gilbert and Lorenzi, 2006; Ardoint and Lorenzi, 2010; Swaminathan, 2010), and how that might be degraded by cochlear damage or speech processors for hearing-impaired listeners (e.g., Baskent, 2006; Lorenzi *et al.*, 2006, 2009; Hopkins *et al.*, 2008; Moore, 2008; Ardoint *et al.*, 2010; Kale and Heinz, 2010) and cochlear implantees (e.g., Shannon *et al.*, 1995; Nelson *et al.*, 2003; Zeng *et al.*, 2005). This debate, however, is premised on definitions of what constitutes the temporal envelope and the temporal fine-structure that have evolved over decades through various means and in different domains (specifically, the *acoustic* and *neural*).

Original references to these terms derived from the use of vocoder manipulations of the acoustic speech signal (Drullman, 1995). Specifically, the key first stage in a vocoder is a filterbank (often referred to as the "analysis filterbank") that mimics cochlear frequency analysis. Its outputs have traditionally been modeled as the product of the (Hilbert) envelope [or amplitude-modulation function (AM)] and a frequency-modulated (FM) sine-wave carrier at the analysis-filter center frequency. AM vocoders preserve the original AM component and discard the original FM component by replacing it by a band of noise or a tone with frequency equal to the center frequency of the analysis band (e.g., Shannon *et al.*, 1995). Conversely, FM vocoders preserve the original FM component and discard the AM (e.g., Smith *et al.*, 2002; Gilbert *et al.*, 2006). By manipulating the filter bandwidths (and hence their numbers over the audio range), one can gradually change the relative importance of the AM and FM components for conveying the intelligibility of the reconstituted speech. For example, at one extreme, when one or a few broadband filters are used, intelligibility is mostly conveyed by the carriers (FM), because by simply flattening and equalizing the envelopes (AM = 0) of all filter outputs and then summing them, one reconstitutes highly intelligible speech (e.g., Smith *et al.*, 2002; Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006). At the other extreme, if many narrowband filters are used, then the situation is reversed and intelligibility becomes severely disrupted when the AM is distorted, but is less affected by a degraded FM (Drullman, 1995; Shannon *et al.*, 1995; Smith *et al.*, 2002; Sheft *et al.*, 2008; Hopkins *et al.*, 2010). However, this is

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

[b)] Also at: UMR CNRS 8158, Laboratoire Psychologie de la Perception, Université Paris Descartes, 45 rue des saints Pères, 75006 Paris, France.

true only when speech is presented in quiet or submitted to no additional distortion. Several studies demonstrated that even when many narrowband filters are used, speech intelligibility is substantially reduced by a degraded FM when speech is presented against a complex background (e.g., Nelson *et al.*, 2003; Zeng *et al.*, 2005; Gnansia *et al.*, 2009; Hopkins *et al.*, 2008), severely filtered in the audio-frequency domain (e.g., Ardoint *et al.*, 2010), or submitted to periodic interruptions (e.g., Nelson *et al.*, 2003).

More recently, cochlear implants have injected a measure of urgency in this research and helped broaden its focus to the representation of the spectrotemporal patterns (or the *spectrogram*) in the early auditory stages [see Wilson and Dorman (2008) for a review]. One key mystery is how the spectrotemporal modulations of speech (and implicitly, the auditory cues crucial for its intelligibility) remain robustly received despite the limited dynamic range of the average firing rates (Sachs and Young, 1979), loss of neural phase-locking at high frequencies (e.g., Johnson, 1980), and cochlear nonlinearities such as inner and outer hair cell compression. With these questions, the intelligibility of the vocoded signals has become closely intertwined with the encoding of the AM and FM vocoded speech components on the auditory nerve via the average or/and phase-locked responses in the auditory nerve (often referred to as the *place versus time* dichotomy as in Sachs and Young, 1979).

Because of the multiple and different references to the temporal envelope and temporal fine-structure or carrier in the acoustic and neural signals, we shall reserve the terms AM and FM to describe the vocoder signal's components (that is, the "acoustic" speech cues within each analysis frequency band), while the symbols "E" and "TFS" will exclusively refer to *neural temporal envelope* and *neural temporal fine-structure* that are defined with the help of a simplified model of auditory processing described in the next section. These definitions will minimize confusion as we proceed to explore how the AM and FM speech components are reflected in the E and TFS of the auditory nerve, and how the neural E and TFS contribute to the intelligibility of the speech signal. Another important term to clarify is *phase-locked responses,* which refers to the ability of the neural response to represent faithfully the instantaneous phase of the cochlear filter outputs. In most mammals, phase-locking is accurate on the auditory nerve up to about 2 kHz and then gradually declines so it becomes no longer detectable at about 5–6 kHz (e.g., Johnson, 1980). Since the envelopes of the cochlear outputs are usually relatively slow (<500 Hz for speech), then neural phase-locking to the temporal envelope remains accurate even on carriers that exceed 6 kHz which are themselves represented simply by an average (non-phase locked) responses (Joris and Yin, 1992; Kale and Heinz, 2010). Therefore, we shall restrict the use of acronym TFS to the neural phase-locked encoding of the *carrier* of the cochlear outputs, whereas E will refer to the encoding of the short-term average response rate which may also fluctuate (but only relatively slowly) in time.

Because of the peculiarities of cochlear analysis, transduction, and subsequent auditory processing, the mapping of the (acoustic) AM and FM components of a vocoder speech signal to the evoked (neural) E and TFS cues is often complex and dependent on the auditory model (see Heinz and Swaminathan, 2009; Ibrahim and Bruce, 2010; Swaminathan, 2010). Nevertheless, because the AM has an intuitive interpretation as a spectrotemporal pattern, it has commonly been assumed to map to the average firing rate of the auditory nerve (or neural E). By contrast, the FM has been thought of as the carrier of the cochlear filter responses, or the neural TFS, but it is usually left unclear just how one can assess from this neural TFS its contribution to the encoding of the spectrotemporal modulations of speech. This is problematic because the association of the acoustic AM and FM components with the neural E and neural TFS cues is, in fact, quite complicated, with changes in one affecting the expression of the others, making it difficult to assess exactly what auditory information is conveyed by each cue. This intrinsic difficulty derives from the fact that for a band-limited signal, the AM and FM components are not independent and information about one can be extracted from the other (Voelcker, 1966; Rice, 1973; Logan, 1977). As a consequence, the AM (envelope) can be recovered (to within a scale factor) from the FM carrier, or theoretically, from the zero-crossings of the band-limited signal. One way to demonstrate this reconstruction is by analyzing the FM carrier by a cochlear-like filterbank which converts the frequency excursions of the FM into AM fluctuations, a process which is presumed to occur at the output of the cochlear filters (Ghitza, 2001; Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006; Sheft *et al.*, 2008; Heinz and Swaminathan, 2009; Ibrahim and Bruce, 2010; Swaminathan, 2010). Additional psychophysical, modeling, and electrophysiological studies (e.g., Sheft *et al.*, 2008) have corroborated these transformations (Voelcker, 1966; Rice, 1973; Logan, 1977; Yang *et al.*, 1992), and confirmed the absence of a one-to-one mapping between AM and neural E on the one hand, and FM and neural TFS on the other hand.

To circumvent these difficulties, and to critically evaluate the validity of the assumptions that have motivated a multitude of vocoder-based experiments, we shall rely on a simplified biologically plausible model of early auditory processing to assess the representation of the acoustic spectrogram of speech in E and TFS. The key conclusion we arrive at is that the E and TFS cues convey roughly comparable and relatively stable representations of the acoustic spectrogram *regardless* of any vocoder manipulations. We shall consequently argue that the relative contributions of the neural E and TFS cues to overall intelligibility are not readily accessible or alterable psychoacoustically via different vocoder schemes, and that instead, vocoder manipulations that implicitly attempt to modify the neural E or TFS cues (by changing the acoustic AM and FM components) in fact always alter both cues in comparable ways, thus failing to adequately assess either. The model also sheds light on the effects of various cochlear pathologies and cochlear prosthetic stimulation on the encoding of neural E and TFS cues. It finally suggests that the neural TFS cues are generally more robust under severe noisy conditions and hence are likely to play a more important role in hearing in a broad range of acoustic environments and tasks.

II. METHODS

A. Encoding and manipulating the speech signal

Different versions of a single speech signal taken from the IEEE sentence database were generated as follows. This sentence (“The birch canoe slid on the smooth planks”) was left as such (clean version), or added to a steady pink noise at different signal-to-noise *ratios* (SNRs). The clean or noisy versions of this speech signal were sampled at a 44.1-kHz sampling frequency. They were left as such in the “unprocessed” condition. Thus, these signals contained both AM and FM information. These signals were also bandpass filtered using zero-phase, Butterworth filters (36 dB/oct roll-off) into either 1, 8, or 16 adjacent frequency bands spanning the range 80–8020 Hz. The cutoff frequencies used and technical details regarding stimulus generation are given in Gilbert and Lorenzi (2006). These bandpass filtered signals were then processed in two ways. In the first (referred to as “AM”), the AM component was extracted in each frequency band, using the Hilbert transform followed by low-pass filtering with a Butterworth filter (cutoff frequency = 64 Hz, 36 dB/oct roll-off). The filtered AM function was used to amplitude modulate a broadband noise carrier (as in Shannon *et al.*, 1995). The modulated noise carriers were then frequency-limited by filtering with the same bandpass filters used in the original analysis filterbank. The resulting modulated noises from each analysis band were finally combined. Thus, these signals contained AM information only. In the second (referred to as “FM”), the Hilbert transform was used to decompose the signal in each frequency band into its AM and FM components. The AM component was discarded. The FM component in each band was multiplied by a constant equal to the root-mean-square (RMS) power of the bandpass filtered signal (as in Gilbert and Lorenzi, 2006). The “power-weighted” FM signals were then summed over all frequency bands. Thus, these signals contained FM information only. In all conditions, the global (RMS) power value of each stimulus was equalized.

B. The early auditory model

In normal auditory nerve responses, the neural E and TFS cues are completely intermingled. However, to explore the separate contributions of these cues, we shall construct two extreme idealizations of the responses that we shall refer to as the “E-route” and the “TFSroute.” The model depicted in Fig. 1 begins with a “cochlear filterbank” of 128 highly overlapping filters (24 filters/octave over a 5.2-octave range), designated as $h_s(t)$, where s is the location of the cochlear filter along the tonotopic axis. The filters are assumed to have constant Q-tuning that is broader ($Q_{3\,\mathrm{dB}} = 4$) than the commonly used gammatone filter (see Fig. 2 of Carlyon and Shamma, 2003, for more details). In order to represent separately and explicitly the E and TFS information, the filter

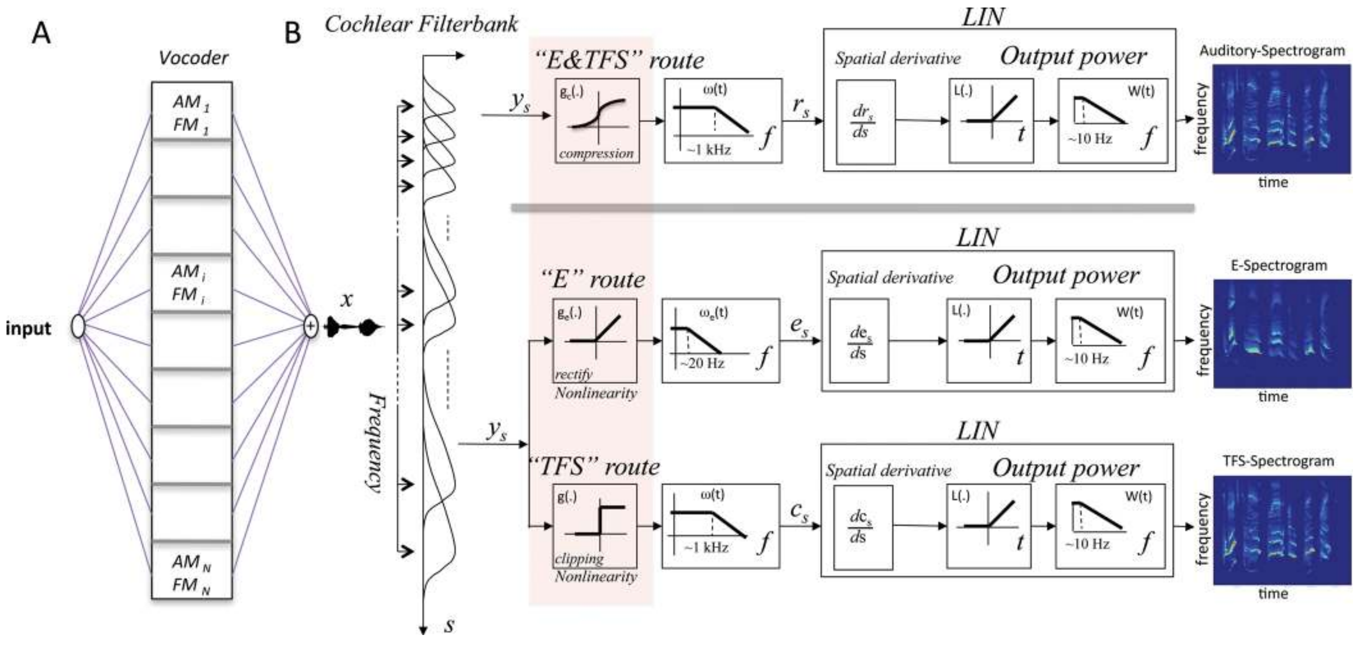

FIG. 1. Conceptual diagram of E and TFS cues in the early stages of the auditory pathway. (A) Vocoded speech signals are usually computed prior to presentation in psychoacoustic experiments using vocoders and cochlear prosthetic devices. Signals are analyzed into multiple frequency bands after which each is subjected to AM or FM of its carrier. The outputs from all bands are then summed to produce the “vocoded” signal $x(t)$. (B) Simplified stages of early auditory processing, beginning with a cochlear frequency analysis stage, and followed by inner hair-cell transduction into neural responses on the auditory nerve. The top “E&TFS” route depicts the “normal” processing in which the inner hair-cell applies a compressive nonlinearity followed by membrane low-pass filtering that gradually attenuates phase-locked responses. In the bottom pathways, the hair cell nonlinearity is modified to highlight the encoding due to two extremes: an “E” route in which only the threshold nonlinearity is retained to rectify the cochlear outputs (i.e., with no saturation or an infinite dynamic range), followed by a low-pass filter that demodulates the cochlear outputs and preserves only their slow envelopes (see Sec. II for more details); a “TFS” route in which the nonlinearity has a 1-bit dynamic range (i.e., is infinitely clipped) that preserves only the zero-crossings of the carrier of the cochlear outputs, followed by the hair cell low-pass filter producing the c_s signal. The three different forms of the hair cell nonlinearity are highlighted by the pink-shadowed box. Common to all pathways is a final LIN to extract the final spectrogram representation from all response cues. The final *auditory*, E, and TFS spectrograms due to the various forms of the hair cell nonlinearity are roughly equivalent.

outputs are fed into two parallel routes that include hair-cell transduction nonlinearities, threshold $g_e(\cdots)$ and saturation $g(\cdots)$, and subsequent low-pass filtering. The model also includes a lateral inhibitory network that extracts a final spectrogram from the E and TFS responses, as described in Shamma (1985a,b) and Lyon and Shamma (1996). The analytical implications of these simulated operations are elaborated upon below separately for the E and TFS routes.

1. The E route

For the E route, the slow envelopes of the filter outputs are extracted by half-wave rectification [referred to also as a pure threshold nonlinearity, $g_e(\cdots)$], followed by low-pass filtering [$\omega_e(t)$] with a relatively long time-constant (3 dB cut-off ~20 Hz) to preserve only slow modulations known to be important for speech intelligibility (Drullman *et al.*, 1994). The envelope pattern is given by $e_s = g_e(x * h_s) * \omega_e = g_e(y_s) * \omega_e$, where the cochlear filter output is y_s $(= x * h_s)$, $x(t)$ is the input acoustic signal, h_s is the impulse response of the sth location along the tonotopic axis, and $*$ is the temporal convolution operator. Cochlear responses are then sharpened by a lateral-inhibitory network (LIN) module (Shamma, 1985a,b) that effectively applies a derivative across the tonotopically ordered cochlear channels represented by the operator (d/ds). The interim output is given by the expression $\{g'_e(y_s)(x * h'_s) * \omega_e\}$ which is interpreted as follows (Lyon and Shamma, 1996): the derivative of the threshold nonlinearity with respect to its argument $[g'_e(y_s)]$ produces a positive half-wave rectified square-wave which samples (or is multiplied by) the function $x * h'_s$. This latter expression can be thought of as an internal representation of the spectrogram of $x(t)$ that is more resolved than that of the cochlear outputs (y_s) as it is produced by convolving the input signal $x(t)$ with the "spatial derivative" of the cochlear filters (h'_s) which are much sharper filters ($Q_{3\,\mathrm{dB}} = 12$) because of the steep (high-frequency) slopes of the cochlear filters h_s [see Fig. 2 in Carlyon and Shamma (2003) for more details]. Rectifying and sampling this "sharpened spectrogram" [by $g'_e(y_s)$] effectively generates a baseband signal proportional to its power. The low-pass filter ω_e then smooths it further, followed by a final measurement of the LIN output power using a half-wave rectifier $L(\cdots)$ and a low-pass filter [$W(t)$]. The final E spectrogram is then approximately a smoothed version of the sharpened spectrogram $x * h'_s$.

Consequently, in this idealized E route, the acoustic spectrogram is reliably encoded if the inner hair cell has only the threshold nonlinearity [$g_e(\cdots)$]. It begins to deteriorate if saturation is added, and is completely lost when the dynamic range is reduced to one bit "clipping," in which case the spectrogram can only be conveyed by TFS cues, as we discuss next.

2. The TFS route

To explore the contributions of the "pure" TFS cues, we construct an idealized pathway in which the inner hair cell compressive nonlinearity has a one-bit dynamic range, and hence preserves only the zero-crossings of the cochlear filter outputs (saturating the envelope of the firing rate or E cues). Figure 1 depicts this by having the cochlear filter outputs ($y_s = x * h_s$) become fully compressed (into square-waves) by the nonlinear function $g(\cdots)$, thus flattening the envelopes of the cochlear filter outputs. Each resulting output $g(x * h_s)$ is subsequently filtered by the inner hair cell membrane low-pass filter $\omega(t)$ with a 0.16 ms time-constant that reflects the gradual loss of neural phase-locking for frequencies above 1 kHz [$c_s = g(x * h_s) * \omega$]. The zero-crossing rates of the square-waves here reflect the carrier of the cochlear filter output signal (approximately near the center frequency or CF of the filters), while the clipped amplitudes of the low-passed waveforms c_s exhibit a gradual fall-off with frequency due to the loss of phase-locking at higher frequencies. Applying the LIN derivative produces a pattern $\{g'(y_s)(x * h'_s) * \omega\}$ that strongly resembles the expression $\{g_e(y_s)(x * h'_e)\}$ in the E route. Once again, the sharply resolved spectrogram ($x * h'_s$) is half-wave rectified by the sampling function $g'(y_s)$ which here samples the spectrogram only near the zero-crossings of y_s and generates a baseband that subsequently (after LIN rectification and more smoothing) yields the final TFS spectrogram in Fig. 1. This TFS spectrogram, like the E spectrogram, essentially reflects a smoothed version of the sharpened internal spectrogram $x * h'_s$. The theoretical underpinnings of this claim have already been discussed in detail in Lyon and Shamma (1996).

Intuitively, the TFS spectrogram emerges because different spectral components in the input signal produce phase-locked responses in localized populations of auditory-nerve channels (each according to its frequency tuning). The key informative cues in these patterns are the *borders* between the different phase-locked responses, whose clarity and locations depend on the relative amplitudes and frequencies of the underlying spectral components. The LIN detects these borders and estimates the relative size of the neighboring spectral components that induce them. Note that the "absolute phase" of the phase-locked responses is immaterial for this estimate because it does not affect the location or salience of the borders, and hence randomizing the carrier of cochlear filter outputs relative to each other is inconsequential for the TFS spectrogram (Yang *et al.*, 1992; Lyon and Shamma, 1996; Carlyon and Shamma, 2003).

3. Auditory spectrogram: combining the E and TFS

The E and TFS routes are idealizations. In reality, the inner hair cell nonlinearity is neither infinitely compressive, nor is neural phase-locking absent in the spectral regions that are most important for speech. The two routes of information transfer simply coexist in the same channels, depicted in the top "E&TFS" route in Fig. 1. The LIN would then normally extract an amalgam spectrogram that we refer to as the "auditory spectrogram." This full spectrogram is generated by using a sigmoid nonlinearity, $g_c(y) = 1/(1 + e^{-y})$, with a finite dynamic range (20–30 dB) that allows partial E fluctuations to remain, and an inner hair cell time constant that preserves some neural phase-locking over much of the speech range (up to 6 kHz). Therefore, subsequent to the LIN, the intermediate expression prior to the final auditory spectrogram resembles that seen earlier in the E and TFS spectrograms, namely, $\{g'_c(y_s)(x * h'_s) * \omega\}$, where the sampling is now done by the derivative of the sigmoidal nonlinearity, $g'_c(\cdots)$, as explained in detail in Lyon and Shamma (1996).

While the neural E and TFS cues may normally contribute equally to the auditory spectrogram as we shall demonstrate by later simulations, their properties can nevertheless diverge under certain severe circumstances such as (1) extreme signal conditions that cause differential deterioration of neural E and TFS cues. For instance, very high sound levels may cause more auditory channels to saturate diminishing the quality of the effective E spectrogram. The TFS-cues in this case remain viable, thus contributing to the stability of the final auditory spectrogram. By contrast, neural TFS cues would be substantially diminished if the speech signal were to be transposed up in frequency to the non-phase-locked response region, leaving only the E cues to contribute to the final auditory spectrogram. (2) Another source of E and TFS imbalance is cochlear pathologies such as reduced dynamic range, loss of auditory-nerve fibers, or increased cochlear filter bandwidths, all of which could differentially affect the spectral representations attributed to the neural E or TFS cues.

4. Summary

The model depicted in Fig. 1 demonstrates that the final auditory, E, and TFS spectrograms are all theoretically comparable: they are patterns that are obtained by passing the input signal $x(t)$ through a bank of narrowband (cochlear) filters, h'_s, then sampled and smoothed to give the well-resolved auditory, E, and TFS spectrograms $L(g'_c(y_s)(x * h'_s) * \omega) * W(t)$, $L(g'_e(y_s)(x * h'_s) * \omega_e) * W(t)$ and $L(g'(y_s)(x * h'_s) * \omega) * W(t)$ in terms of the power in each channel. The fundamental difference between the three above expressions is the different time-widths of the nonlinear sampling functions $g'_c(y_s)$, $g'_e(y_s)$, and $g'(y_s)$ which cause them to differ in some details and noise robustness as we discuss later. Specifically, in the TFS spectrogram computations, $g'(y_s)$ is formally a *dirac δ-function* being the derivative a unit step function (Fig. 1), and hence it is very narrow and samples the value of the $(x * h'_s)$ precisely at the zero-crossings of the y_s. The E spectrogram samples are computed with $g'_e(y_s)$ which occur at the same zero-crossings of y_s except that they span the whole positive *half-period* between them. Therefore the resulting spectrogram appears smoother than the TFS spectrogram since it is sampled with wider pulses. The auditory spectrogram is intermediate between these two extremes. Thus, depending on the steepness and saturation level of the sigmoid nonlinearity g_c, its derivative $g'_c(y_s)$ can be made to look as narrow as that of the TFS route (by making the sigmoid a step) or as wide as that of the E route (by removing saturation and making it a pure threshold).

We should emphasize that the neural E and TFS cues exist in the auditory pathway because of the physiological properties of the auditory channels (neural phase-locking, inner hair-cell nonlinearities, and low-pass filtering) independently of any external manipulation of the acoustic input. In other words, the E and TFS spectrograms represent conceptually the spectrogram of the input $x(t)$ regardless of any prior processing it may have undergone.

C. Model assessment

In order to compare the spectrograms generated from the various cues, we shall adopt two methods. The first is a straightforward pattern-match (or the correlation coefficient) between the spectrograms, defined as $r = \langle S_1 S_2 \rangle$, where $\langle \cdots \rangle$ denotes the inner product, and S_1 and S_2 are the normalized zero-mean spectrograms (see more details below). Specifically, as shown in Fig. 2, we shall quantify several comparisons between the various spectrograms generated by the models (all measurements will be rounded to the nearest two significant decimals):

(1) The match between the E and TFS spectrograms, and their corresponding auditory spectrograms, denoted by the dotted red and blue lines and the symbols r_{E} and r_{TFS}, respectively. More importantly for this study, we shall report the ratio $r_{\mathrm{E}}/r_{\mathrm{TFS}}$ $(= R)$, a metric that captures succinctly the balance between the LIN-extracted spectrograms.

(2) In an exactly analogous manner, we shall also compute the match between the E and TFS spectrograms and the corresponding *clean* auditory spectrogram, denoted by the solid red and blue lines and the symbols $r_{\mathrm{E\_CLEAN}}$ and $r_{\mathrm{TFS\_CLEAN}}$, respectively. We will also report the match between the auditory spectrograms of the vocoded and clean signals, $r_{\mathrm{VOC\_CLEAN}}$ (shown by solid green line).

(3) Finally, the auditory, E, and TFS spectrograms will also be compared with their counterparts for the clean signal. These are represented in Fig. 2 by the dashed green, red, and blue lines and the symbols $r_{\mathrm{VOC\_CLEAN}}$, $r_{\mathrm{E\_E}}$, and $r_{\mathrm{TFS\_TFS}}$. The $r_{\mathrm{E\_E}}$ and $r_{\mathrm{TFS\_TFS}}$ metrics are analogous to the correlation-coefficient based metrics that have been used in previous modeling work (Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006; Sheft *et al.*, 2008), or in experimental and computational studies of auditory-nerve spike-train responses (Heinz and Swaminathan, 2009; Swaminathan, 2010). The key difference between these measurements and ours is that the $r_{\mathrm{E\_E}}$ and $r_{\mathrm{TFS\_TFS}}$ quantify neural E and TFS contributions by first transforming them (with the LIN) into a common centralspectrogram representation which facilitates their direct comparison.

Previous studies have used a variety of models, ranging from simple gamma-tone or gamma-chirp filters (e.g., Zeng *et al.* 2004; Gilbert and Lorenzi, 2006; Sheft *et al.* 2008) to complex physiologically based auditory-nerve models (and data) that include details of the cochlear nonlinearities and synaptic adaptation (e.g., Heinz and Swaminathan, 2009; Swaminathan and Heinz, 2012; Ibrahim and Bruce, 2010). The model depicted in Fig. 1 is intermediate in complexity as it lacks certain features deemed uncritical for the goals of this study, such as adaptation and cochlear nonlinear filters. These and other details may influence the look of the final

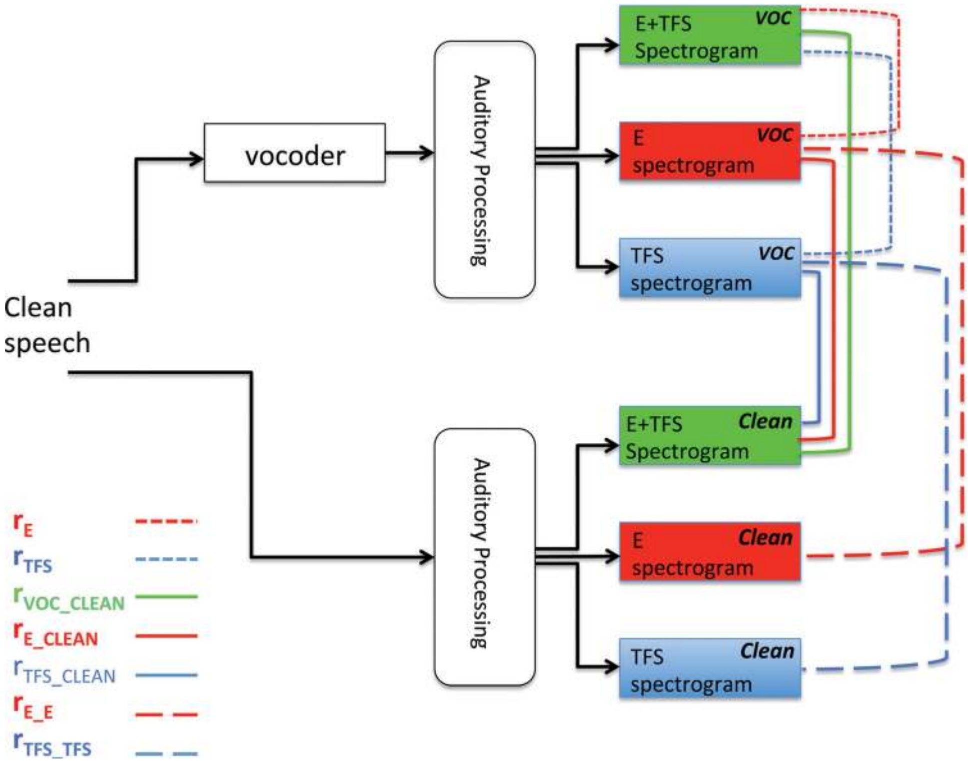

FIG. 2. Schematic of the various correlation measures computed from the model. Three correlation measures are shown: (Red) E spectrogram versus *clean* auditory spectrogram, $r_{\mathrm{E\_CLEAN}}$ (solid), versus the *vocoded* auditory spectrogram, r_{E} (dashed), and versus the clean E spectrogram, $r_{\mathrm{E\_E}}$ (dotted). (Blue) TFS spectrogram versus clean auditory spectrogram $r_{\mathrm{TFS\_CLEAN}}$ (solid), versus vocoded auditory spectrogram r_{TFS} (dashed), and versus clean TFS spectrogram $r_{\mathrm{TFS\_TFS}}$ (dotted). (Green) vocoded versus clean auditory spectrogram $r_{\mathrm{VOC\_CLEAN}}$ (solid)

LIN spectrograms, but should not change any of our major conclusions.

The correlation measures computed in this paper are slightly different from the definition above in two respects: (1) to compute the match between the patterns S_1 and S_2, the inner product $\langle S_1(t,f)S_2(t,f+\delta_f)\rangle$ is computed at a couple of shifts around zero ($\delta_f = 0 +/- 2$), and the maximum value is noted. This allows the measure to remain stable for a small amount of overall pattern shifts; (2) all correlations involving vocoded signals (Figs. 3–6) were computed over time intervals where the speech signal is strong, i.e., where the clean signal simply exceeded 5% of its maximum power [e.g., at the six syllabic segments of Fig. 3(A) clean signal]. The reason for avoiding the silent regions is the inevitable noise-bursts introduced by the infinite clipping in the TFS route (also typical of FM vocoders) that amplifies random fluctuations in the silent intervals [see TFS spectrograms of Figs. 4(A), 5(A), and 6(A)].

The second method we propose to assess the intelligibility of the auditory, E, and TFS spectrograms entails reconstructing the audio signal that produces the closest spectrogram to the target spectrogram in the mean square

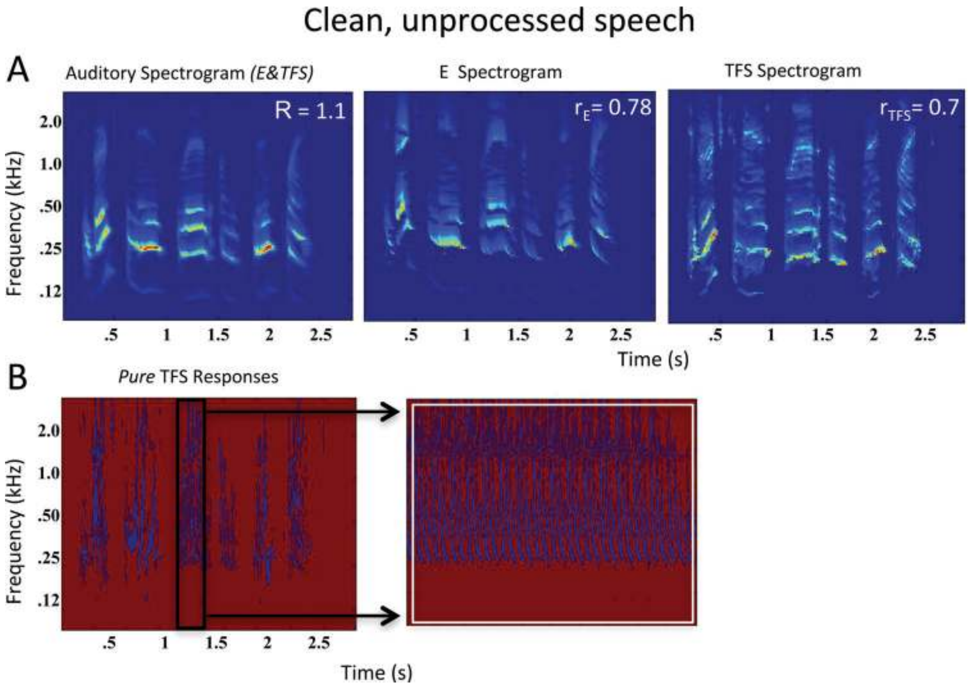

FIG. 3. Spectrograms of a speech signal derived from different response cues on the auditory nerve. (A) The "auditory spectrogram" utilizes both E and TFS cues. The E spectrogram is derived from only the envelope of the neural responses. The TFS spectrogram is extracted from the zero-crossings of the neural responses. All three representations capture the essential features of the speech signal including its harmonics and formant peaks, and temporal dynamics. (B) Cochlear responses after infinite clipping by the hair cell nonlinearity in the TFS route. The responses resemble 1-0 square-waves where all information is preserved in the patterns of zero-crossings. The responses in the small segment (between 1.2 and 1.4 s) are magnified to show more clearly the phase-locked patterns due to the different signal harmonics, and the borders between them that the LIN detects to generate the TFS spectrogram.

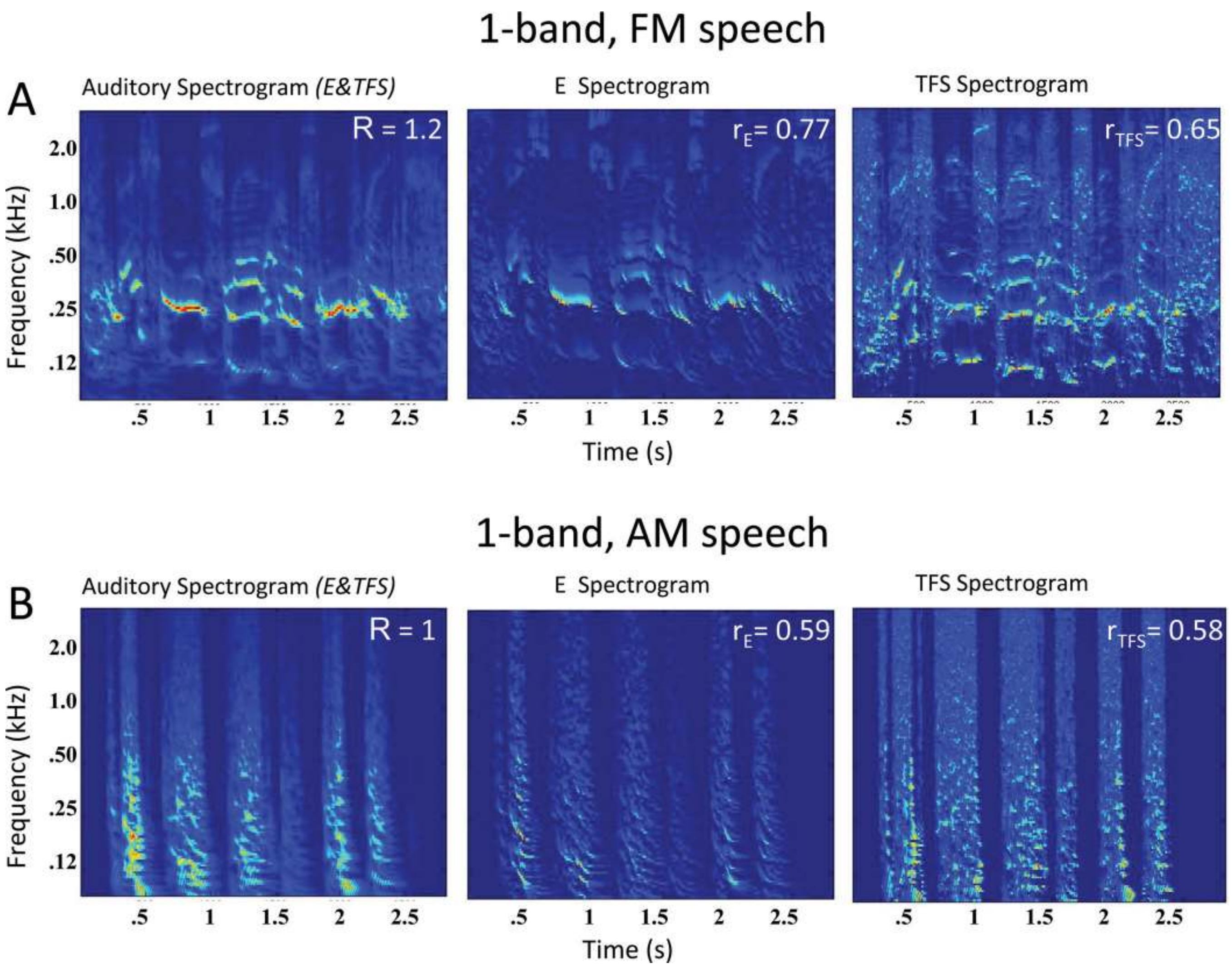

FIG. 4. Auditory, E, and TFS spectrograms for one-band vocoded speech. (A) One-band FM vocoded speech in which the AM of the signal is flattened while the FM carrier is untouched. Speech is highly intelligible, and all spectrograms depict well the harmonic and formant structure of the speech. (B) One-band AM vocoded speech in which the AM is maintained while the FM carrier is replaced by noise. The speech is unintelligible and, consistently, all spectrograms appear relatively flat and lacking the informative spectral details. The matches between each of the E- and TFS spectrograms and the corresponding auditory spectrogram are indicated as $r_{\rm E}$ and $r_{\rm TFS}$. Also indicated is the ratio between the two correlation indices, $R = r_{\rm E}/r_{\rm TFS}$.

error sense (for a detailed presentation of the audio-reconstruction technique, see Yang *et al.*, 1992). The reconstruction procedure is iterative and relies on convex projection methods as detailed in the above referenced study. The audio samples are available at /www.isr.umd.edu/∼sas/Audio_samples_JASA_2011 and are named according to the figure numbers in the article of Yang *et al.* (1992).

III. RESULTS

This section provides simulations of the E and TFS spectrograms described in Figs. 1(B) and Fig. 2 for different vocoded speech signals. We shall illustrate the key conclusion of this study: that these two neural representations remain comparable to each other for a wide range of vocoded input signals. Consequently, we shall argue that changes of intelligibility in vocoded speech are not due to a change in the balance of E and TFS cues. Since the conclusions are essentially identical for the various correlation metrics, we shall detail first the $r_{\rm E}$ and $r_{\rm TFS}$ measures and their ratios, and later summarize the main findings for the other correlations.

A. Clean, unprocessed speech

Figure 3(A) illustrates the auditory spectrogram of a clean speech signal and the two variations representing exclusively the neural E and TFS cues. Also shown in Fig. 3(B) is the phase-locked structure in the responses [or the c_s in the TFS route of Fig. 1(B)]. The auditory spectrogram of the sentence displays the typical features seen in the (logarithmic axis) spectrogram, including the harmonic structure, the formants and their transitions, and the phonetic/syllabic segments of the speech signal. This spectrogram effectively combines the neural E and TFS information at the outputs of the cochlear channels. Figure 3(A) (second panel) illustrates the E spectrogram which displays the same essential features seen in the auditory spectrogram. Figure 3(B) displays the neural TFS responses that persist at the cochlear filter outputs if their outputs are fully compressed leaving no

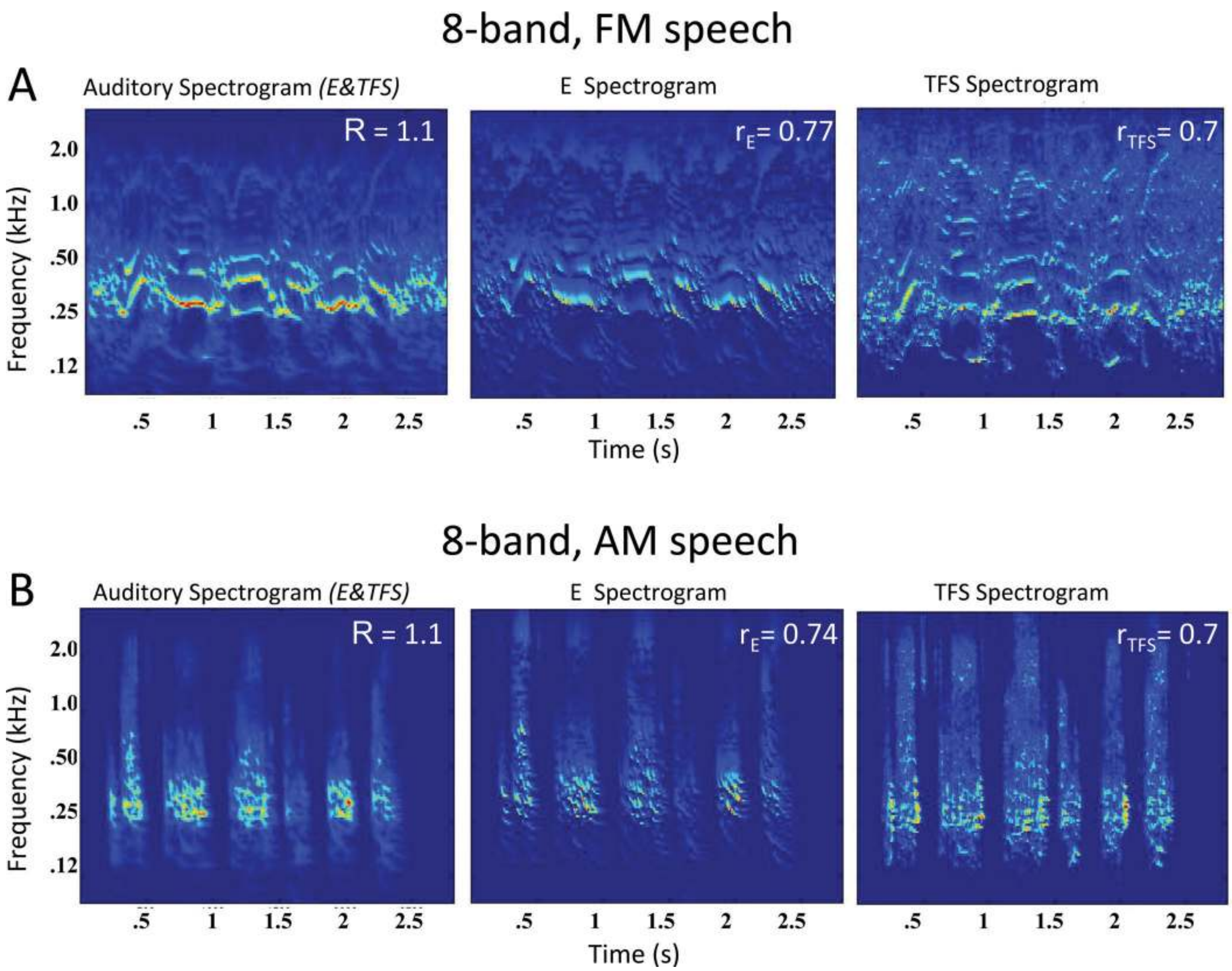

FIG. 5. Auditory, E, and TFS spectrograms for eight-band vocoded speech. (A) Eight-band FM vocoded speech in which the AM of the subband outputs is flattened while maintaining the carriers. Speech is less intelligible than the one-band case, and the E and TFS spectrograms resemble the auditory spectrogram in preserving the low harmonics but lacking other spectral details. (B) Eight-band AM vocoded speech in which the AM of the subbands is preserved while the FM carriers are randomized. Speech is intelligible, and all three spectrograms consistently depict the envelope of the speech spectrum. In both cases, the matches between each of the E and TFS spectrograms and the corresponding auditory spectrogram are indicated as r_E and r_{TFS}. Also indicated is the ratio $R = r_E/r_{TFS}$.

meaningful neural E information. What is quite evident in the enlarged segment of the responses is the precise phase-locked structure of the carriers across the array, and the discontinuities at locations that correspond to some of the harmonics of the speech spectrum. When the LIN (spatial derivative) is applied to this pattern, the discontinuities produce large final outputs giving rise to the TFS spectrogram [third panel in Fig. 3(A)] that resembles the auditory and E spectrograms. While these three representations are clearly not identical because they involve different idealizations of the nonlinearity (sigmoid, threshold, and one-bit compression), they nevertheless are broadly similar and it is useful to compute a measure of the match between them so as to assess the relative values for different input signals. The match between the E and TFS spectrograms on the one hand, and the full auditory spectrogram on the other are quantified in terms of the coefficients r_E and $r_{TFS} = 0.78$ and 0.7, respectively (see Sec. II). However, the absolute values of these coefficients are less important for the goals of this study than the way their ratio changes for different vocoded signals. The absolute values also depend on the specific parameters of the nonlinearities and hair cell model, such as the slopes of the sigmoid in the auditory spectrogram, the threshold-level of the nonlinearity in the E spectrogram, and the exact limits of the phase-locking for the TFS spectrogram. These model parameters are fixed throughout the study and hence play no further role in maintaining a stable ratio between the $\mathrm{r_E}$ and $\mathrm{r_{TFS}}$ across the different vocoder conditions.

For this clean signal, the ratio R $(= r_E/r_{TFS} = 1.1)$ can be thought of as the balance between the E and TFS spectrograms (under the specific conditions and parameters of the model). This ratio (R) as we show below remains stable between 1–1.2 regardless of the manipulations applied to the vocoded input. Note that this ratio is unconstrained and can in principle attain any value between 0 and infinity, and hence its confinement to this narrow range testifies to its stability. The audio signals reconstructed from the above E and

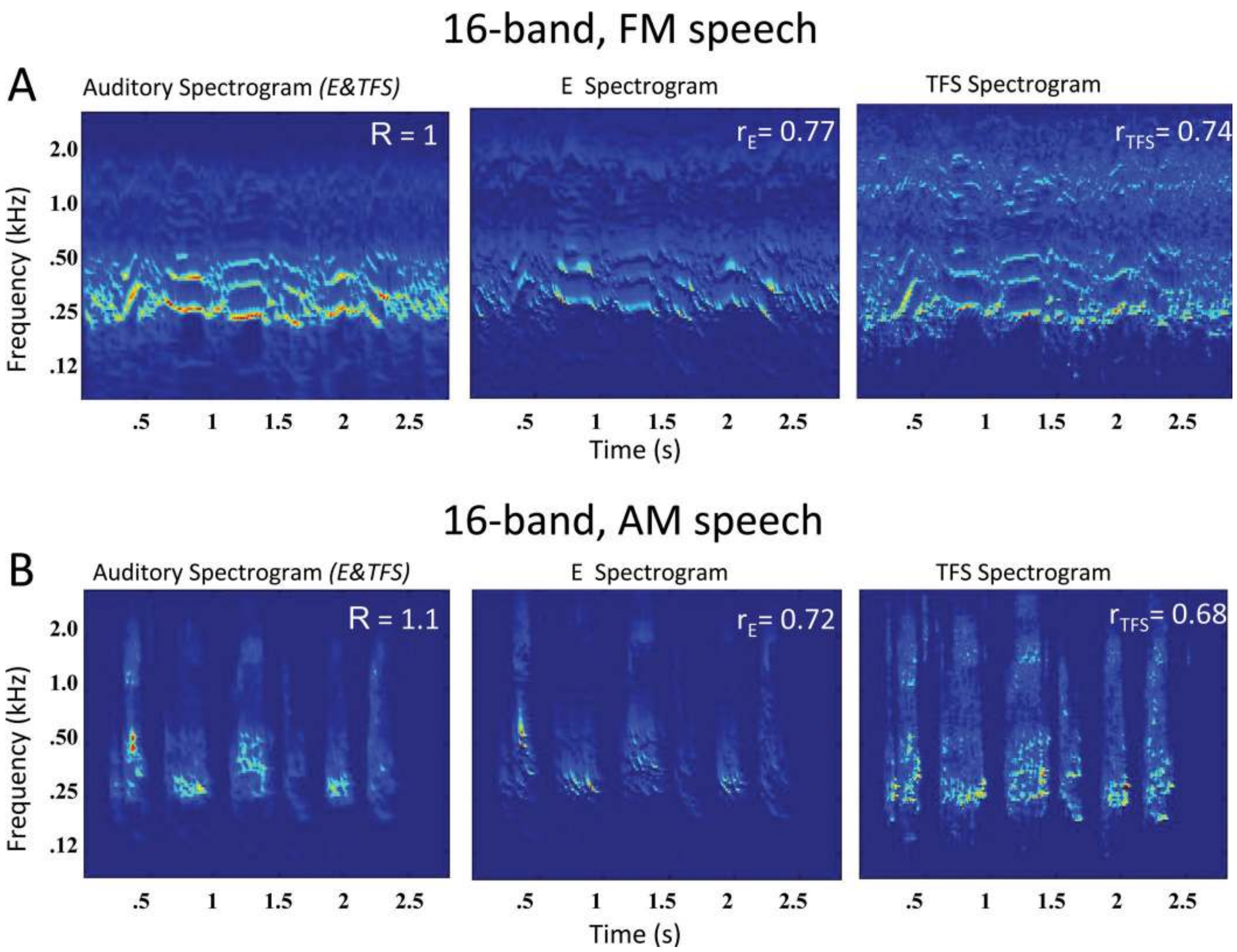

FIG. 6. Auditory, E, and TFS spectrograms for 16-band vocoded speech. (A) 16-band FM vocoded speech in which the AM of the subband outputs is flattened while maintaining the carriers. Speech is significantly less intelligible than the one-band case, but the E and TFS spectrograms still resemble the auditory spectrogram in preserving some of the low harmonics. (B) 16-band AM vocoded speech in which the AM of the subbands is preserved while the FM carriers are randomized. Speech is highly intelligible, and all three spectrograms equally depict the envelope of the speech spectrum. The matches between each of the E and TFS spectrograms and the corresponding auditory spectrogram are indicated as r_E and r_{TFS}. Also indicated is their ratio, $R = r_E/r_{TFS}$.

TFS spectrograms are both highly intelligible, demonstrating further the similarity between these two representations.

B. One-band AM and FM vocoded speech

When a speech signal is fully clipped while preserving its zero-crossings or FM information (referred to as the one-band FM condition), it remains highly intelligible (Smith *et al.*, 2002; Gilbert and Lorenzi, 2006). By contrast, when the carrier (FM) is replaced by noise while preserving the AM (one-band AM condition), intelligibility is barely above chance level (Shannon *et al.*, 1995; Smith *et al.*, 2002). The model outputs are consistent with these findings in that both E and TFS convey relatively intact spectrograms for the one-band FM speech [Fig. 4(A)], and largely flattened spectrograms for the one-band AM speech [Fig. 4(B)]. In Fig. 4(A), the E and TFS spectrograms ($r_E = 0.77$; $r_{TFS} = 0.65$) are similar except for one significant deviation, namely, the bursts of noise introduced by the one-bit "clipping" which amplifies random fluctuations (recording noise) within the quiet intervals between the syllabic segments of speech. Nevertheless, the balance between the two remains relatively stable as in the clean speech case before [Fig. 3(A)] at $R = 1.2$. Furthermore, a similar balance is maintained for the E and TFS spectrograms in Fig. 4(B) ($r_E = 0.59$; $r_{TFS} = 0.58$) where $R = 1$. That is, while flattening the AM or randomizing the FM produces two signals with substantially different spectrograms and intelligibility, the relative contributions of E and TFS to the auditory spectrogram remain comparable. Thus, one cannot conclude much regarding the relative importance of the E or TFS for these vocoder manipulations.

C. Multi-band vocoded speech

The multi-channel vocoder has been the primary tool for manipulating speech signals in order to explore the nature of the information encoded in the AM and FM. Multichannel vocoders are commonly conceived of as resembling cochlear frequency analysis and hence assumed to be useful in controlling the envelope and fine structure of the cochlear outputs. This has led to many unwarranted conclusions regarding the efficacy of the neural E and TFS information in the early auditory pathway. We shall illustrate this in the cases of the 8- and 16-band vocoders, which are intermediate and share many properties with the wideband filter of the 1-band vocoder and the narrowband filters of the 32-band vocoder (where the bandwidth of analysis filters approaches the width of normal cochlear filters). These vocoders have been used to simulate speech perception for cochlear implantees (e.g., Friesen *et al.*, 2001) and listeners with moderate sensorineural hearing loss (e.g., Baskent, 2006).

In one important manipulation (referred to as the eight-band AM speech), the speech signal is first filtered into eight bands, and then the carriers of the filter outputs are replaced by noise to destroy the FM information they carry. Since speech remains highly intelligible after combining all filter outputs, it has been argued that the AM (and implicitly the neural E), and not the FM (or implicitly the neural TFS cues), carry all the speech information (e.g., Drullman, 1995; Shannon *et al.*, 1995; Smith *et al.*, 2002). This interpretation is further reinforced by the finding that intelligibility is more degraded when the filter outputs are compressed while leaving the FM untouched (i.e., the eight-band FM condition) unless listeners are given extensive training or presented with isolated speech segments (e.g., bisyllables) or high-context speech material (Gilbert and Lorenzi, 2006; Gilbert *et al.*, 2007; Lorenzi *et al.*, 2006; Sheft *et al.*, 2008; Hopkins *et al.*, 2010).

Associating the AM and FM of the vocoder with the neural E and TFS of the auditory nerve has led to the conclusion that it is the neural E cues and not the neural TFS cues that are essential to convey the intelligibility of clean speech. This conclusion, however, is false because the TFS spectrogram does not reflect the carrier (i.e., FM) of the vocoder outputs in a simple way. For example, in the eight-band FM vocoded signal of Fig. 5(A), the AM envelopes are completely flattened within each band causing the E spectrogram to deteriorate. Likewise, the TFS spectrogram retains partially the harmonic structure of the speech signal seen in the E and auditory spectrograms ($r_{\mathrm{E}} = 0.77$; $r_{\mathrm{TFS}} = 0.7$). Similarly, in the eight-band AM vocoder of Fig. 5(B), the FM carriers are completely randomized causing the signal spectrogram to change significantly. Nevertheless, the LIN extracts the TFS spectrogram by detecting the discontinuities at the borders of the phase-locked responses between the eight adjacent bands as explained in Fig. 3(B), and the E and TFS spectrograms still resemble each other and the auditory spectrogram ($r_{\mathrm{E}} = 0.74$; $r_{\mathrm{TFS}} = 0.7$). More importantly, however, while these FM and AM vocoder manipulations affect substantially the spectrogram of the speech signal and its intelligibility [left panels of Figs. 5(A) and 5(B)], they do not alter the balance or relative contributions of the E and TFS information ($R = 1.1$). Hence, these manipulations cannot provide the presumed insights into the balance of neural E and TFS encoding of speech.

The 16-band vocoder has also been extensively used to investigate the role of E and TFS cues in speech perception for normal-hearing listeners and listeners with mild to moderate hearing loss (e.g., Lorenzi *et al.*, 2006, 2009; Heinz and Swaminathan, 2009; Ardoint *et al.*, 2010; Swaminathan, 2010). Our model simulations are shown in Figs. 6(A) and 6(B), where we arrive at exactly the same conclusions as in the one-band and eight-band vocoders [$R = 1$ and 1.1 for the spectrograms in Figs. 6(A) and 6(B), respectively], confirming the generality of our findings.

Finally, we reiterate that while the absolute values of correlation coefficients depend on the specifics of the model parameters and of the definition of the measures, their trends nevertheless are broadly consistent with well-known intelligibility assessments for the different vocoder signals discussed earlier (e.g., Smith *et al.*, 2002; Lorenzi *et al.*, 2006). For example, the match (correlation coefficients) between the auditory spectrograms of each vocoded signal (left panels of Figs. 4–6) versus that of the unprocessed sentence [left panel of Fig. 3(A)], fall-off gradually in the same order as the reported intelligibility: 0.77 for one-band FM condition [Fig. 4(A)], and 0.29 in the one-band AM condition [Fig. 4(B)]. The matches for the multiband vocoded signals are shown in Figs. 7(A) and 7(B) and are discussed later.[1]

D. Comparing the E and TFS spectrograms to the clean spectra

We have argued thus far that different vocoder conditions similarly affect the E and TFS spectrograms leaving the balance between them unchanged, and hence rendering the vocoder ineffective in investigating the relative importance of these representations to intelligibility. To illustrate this assertion in a different way, we computed the correlations $\mathrm{r}_{\mathrm{VOC\_CLEAN}}$, $\mathrm{r}_{\mathrm{E\_CLEAN}}$, and $\mathrm{r}_{\mathrm{TFS\_CLEAN}}$ (see Fig. 2) between the auditory (E + TFS), E, and TFS spectrograms and the clean auditory spectrogram of the speech signal (Fig. 3). The results for the different vocoder conditions are shown in Figs. 7(A) and 7(B) for the AM and FM vocoders.

First, we consider the correlations for the AM vocoder conditions. As expected, the correlation $r_{\mathrm{E\_CLEAN}}$ increases with the number of vocoder channels. However, what is crucial here is that $r_{\mathrm{TFS\_CLEAN}}$ also increases similarly confirming the earlier observation that the TFS spectrogram (extracted from purely the TFS cues) follows closely the E spectrogram. These similar trends are captured by the ratio R_{CLEAN} between the two correlation coefficients ($R_{\mathrm{CLEAN}} = r_{\mathrm{E\_CLEAN}}/r_{\mathrm{TFS\_CLEAN}}$) which remains roughly constant (∼1.09 +/− 0.06) *regardless* of the vocoder conditions, and just as was the case for the ratio R (∼1.06 +/− 0.05) discussed earlier.

The increasing $r_{\mathrm{E\_CLEAN}}$ and $r_{\mathrm{TFS\_CLEAN}}$ as a function of channel number also agree well with trends in $r_{\mathrm{VOC\_CLEAN}}$ and with measured intelligibility (e.g., Shannon *et al.*, 1995; Smith *et al.*, 2002). Nevertheless, as discussed earlier, both E and TFS cues contribute similarly to this increase in intelligibility, and hence one *cannot* deduce that E cues are more effective than TFS cues in conveying intelligibility. Instead, it is clear that vocoder manipulations alter *both* E and TFS spectrograms in parallel, roughly maintaining the balance of their effective contributions.

Figure 7(A) (right panel) displays the corresponding results for the FM-vocoder conditions. This vocoder distorts the auditory, E, and TFS spectrograms differently resulting in $r_{\mathrm{VOC\_CLEAN}}$, $r_{\mathrm{E\_CLEAN}}$ correlations that are flat or show a

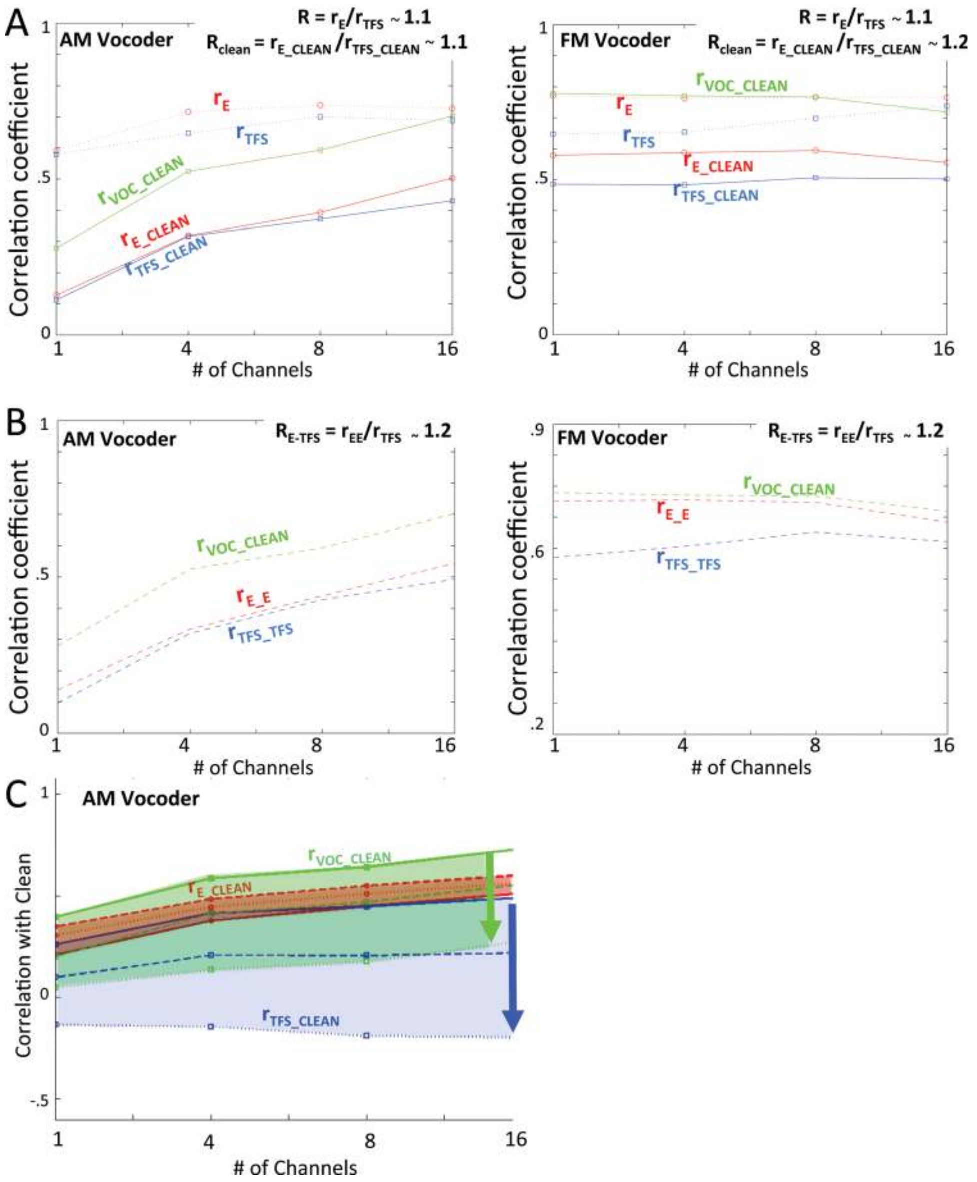

FIG. 7. Changes of various correlation coefficients (from Fig. 2) for various vocoder conditions. (A) Changes of $r_{\mathrm{VOC\_CLEAN}}$, $r_{\mathrm{E\_CLEAN}}$, and $r_{\mathrm{TFS\_CLEAN}}$ as a function of the number of channels in AM vocoder (left panel) and FM vocoder (right panel). Also displayed are the ratios R and R_{CLEAN} for each condition. (B) Changes of $r_{\mathrm{VOC\_CLEAN}}$, $r_{\mathrm{E\_E}}$, and $r_{\mathrm{TFS\_TFS}}$ as a function of the number of channels for AM vocoders (left panel) and FM vocoders (right panel). (C) The effects of weaker LIN on the balance of E and TFS cues. Correlations $r_{\mathrm{E\_CLEAN}}$, and $r_{\mathrm{TFS\_CLEAN}}$ are shown as a function of channel numbers for AM vocoders. As the LIN derivative is weakened from $\alpha = 1$, 0.9, and 0.7 (solid, dashed, and dotted lines, respectively; see text for details), $r_{\mathrm{E\_CLEAN}}$ remains relatively unchanged (red-shaded region). By contrast, the TFS correlation $r_{\mathrm{TFS\_CLEAN}}$ decreases substantially (blue-shaded region). The correlations of the auditory spectrogram are intermediate between the E and TFS (green-shaded region).

slight *decrease* as a function of increasing channel numbers, a trend that is consistent with published intelligibility results, at least when measured with isolated VCV syllables (e.g., Fig. 4 in Gilbert and Lorenzi, 2006). The $r_{\mathrm{TFS\_CLEAN}}$ remains roughly constant or slightly increases reflecting the stable shape of the TFS spectrogram. Again, despite these small changes, the ratio of the E and TFS correlations maintain roughly a stable value ($R_{\mathrm{CLEAN}} \sim 1.17$ +/− 0.06) just as in the case of the R (∼1.1 +/− 0.05) computed earlier from r_{E} and r_{TFS}. Combining the measurements from all vocoder conditions (AM, FM, and 1, 4, 8, and 16 channels), both ratios exhibit a small range of variability of about 5% around 1.1. Finally, Fig. 7(B) displays for comparison the correlations between the auditory, E, and TFS spectrograms in the vocoded conditions versus their clean counterparts ($r_{\mathrm{VOC\_CLEAN}}$, $r_{\mathrm{E\_E}}$, and $r_{\mathrm{TFS\_TFS}}$). The trends closely resemble those of earlier correlations with an average ratio $R_{\mathrm{E\text{-}TFS}} = 1.17$ +/− 0.1.

In summary, regardless of all these correlation measures, the main message of this work is intuitive and can be readily seen by inspection of the spectra computed in Figs. 3–6. Namely, in any one condition, *all three extracted spectra* (E&TFS, E, and TFS spectrograms) have a fair amount of resemblance to each other, although they change *together* dramatically across different vocoder conditions. This resemblance between these three spectra is the key finding of this paper. It suggests that no matter what one does to change the spectra using one vocoder or another, the E and TFS cues nevertheless produce similar spectra within one condition, and therefore they are balanced relative to each other all the time.

E. The role of the LIN in extracting E and TFS cues

The LIN is a critical part of the model in that it extracts the available E and TFS cues from the phase-locked responses on the nerve and constructs a spectrogram representation for both so that they can be compared directly against each other and the auditory spectrogram. It is clear that significant changes in the LIN will likely distort the E and TFS representations and alter the conclusions we made. For instance, if the LIN is removed altogether, one has no

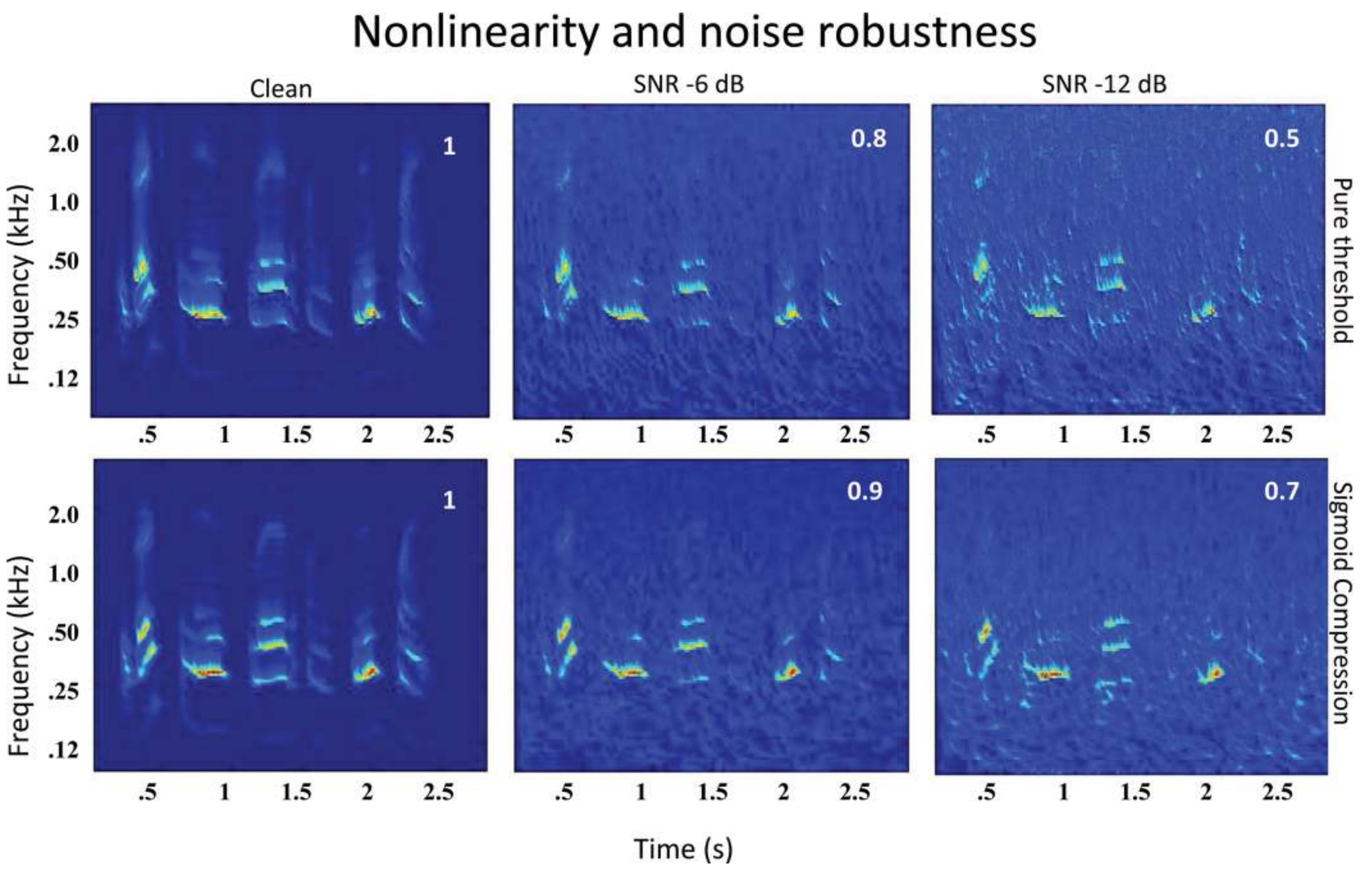

FIG. 8. The role of compression and TFS in enhancing noise robustness. The auditory spectrograms are computed for speech with increasing noise levels (from left to right: clean, −6 dB, and −12 dB SNRs). Two nonlinearities are used in computing the auditory spectrograms: In the upper panels, a purely threshold (or no saturation) nonlinearity is used [just as in the E-route in Fig. 1(B)]. In the bottom panels, the nonlinearity has compression (or saturation) just as in the E&TFS route in Fig. 1(B). Compression improves the robustness of the auditory spectrogram as reflected by the superior preservation of the spectral details of the clean spectrogram in the −6 and −12 dB SNR spectrograms (lower panels). This is quantified by the better matches between each of the noisy spectrograms and the corresponding clean spectrogram (correlation coefficients of 0.9 and 0.7 versus 0.8 and 0.5).

access anymore to the spectrograms and instead one has to resort to different measures that, for example, estimate the match between clean and vocoded phase-locked or averaged response patterns on the nerve (e.g., Swaminathan and Heinz, 2012). Clearly, one may derive different conclusions regarding the efficacy of the E and TFS cues depending on how they are computed.

To demonstrate the importance of the LIN, we introduced a gradual smoothing of the LIN's derivative of the auditory nerve responses described earlier in Methods and in Fig. 1. Specifically, we replaced the implementation of the LIN derivative as a *first-difference* between neighboring auditory channels by $(x_i - \alpha x_{i-1})$ where α gradually decreased from 1 (regular LIN) to 0.7 (weak LIN derivative). Figure 7(C) depicts how the correlation measures relative to the clean spectrogram, $r_{\mathrm{VOC\_CLEAN}}$, $r_{\mathrm{E\_CLEAN}}$, and $r_{\mathrm{TFS\_CLEAN}}$, decrease as smaller α's gradually weaken the ability of the LIN to extract the TFS cues from the phase-locked responses on the model nerve. Clearly, the balance of the extracted cues (reflected by $R_{\mathrm{CLEAN}} = r_{\mathrm{E\_CLEAN}}/r_{\mathrm{TFS\_CLEAN}}$) now varies with the changing mismatch between the E and TFS spectrograms. This simulation demonstrates how a deficient LIN would fail to extract the TFS cues from the same phase-locked responses on the nerve and leads (erroneously) to the conclusion that the TFS cues are unavailable to contribute to intelligibility.

F. Compression and robustness of the TFS spectrogram

While the E and TFS spectrograms are roughly similar for clean speech signals, they nevertheless differ in some important details because of the different sources of information and nonlinearities involved in their generation. For example, the TFS spectrogram of clean unprocessed speech discussed earlier (Fig. 3) displays a sharper harmonic structure in the mid-frequency range than the E spectrogram.

Another important difference between E and TFS cues concerns the robustness of the TFS spectrogram in extremely noisy conditions. Unlike the E spectrogram, the TFS involves a strong compressive nonlinearity converting the input waveform into one represented by its zero-crossings. Earlier analyses of the effects of this compression (Shamma and Morrish, 1987) have shown that compressing a waveform composed of multiple components leads to a relative enhancement of the stronger components and an increased robustness and resistance to added masking noise. Figure 8 illustrates this robustness and the role of the compressive nonlinearity. The clean, unprocessed speech signal shown earlier in Fig. 3 is now contaminated by a pink-noise masker. To quantify and highlight the role of the compressive nonlinearity in endowing this robustness, the auditory spectrograms are computed in Fig. 8 with a pure threshold nonlinearity (top row) and a compressive nonlinearity (bottom row), under increasing levels of noise [SNR = infinite (clean speech), −6, −12 dB from left to right]. The plots illustrate that under extreme noise conditions (rightmost spectrograms) the compressive nonlinearity (bottom spectrograms) preserve more of the original structure of the clean speech spectrograms, as reflected by the higher matches (correlation coefficients) between the clean and noisy spectrograms in the lower plots.

IV. DISCUSSION

A. Peripheral versus central reconstruction of the stimulus spectrogram

This study explored how E and TFS cues are expressed in the responses of the auditory nerve, and demonstrated using a simplified model of auditory processing how both cues covaried in a similar manner regardless of any vocoder manipulations. The model also illustrated the feasibility of a central decoding mechanism (the LIN) that could recover the speech spectrogram from the TFS patterns on the auditory-nerve. From a theoretical perspective, it has been known (Voelcker, 1966; Rice, 1973; Logan, 1977) that zero-crossings of a signal can provide nearly equivalent information to the full signal if they are processed appropriately (e.g., through narrowband filtering). Consistent with this principle, the work of Ghitza (2001), Zeng *et al.* (2004), Gilbert and Lorenzi (2006), and Heinz and Swaminathan (2009) illustrated how peripheral (i.e., cochlear) filtering could recover the speech spectrogram (referred to as the "envelope" spectrogram) from the broadband acoustic FM speech cues. The current study illustrates the same signal-processing principle, but applies it more centrally to the auditory-nerve responses rather than the acoustic stimulus.

In a broader perceptive, this is consistent with the long-articulated hypothesis that TFS cues on the auditory nerve can encode the acoustic spectrogram, especially given their accuracy and robustness (e.g., Young and Sachs, 1979). Lateral inhibition (or the LIN) is one possible way to recover the acoustic spectrogram that, although physiologically plausible, remains unconfirmed largely because of the absence of targeted studies. The LIN principle is fundamentally equivalent to a coincidence detection operation (Lyon and Shamma, 1996) and can be implemented by a variety of networks with neural receptive fields that may exhibit one-sided inhibition (as in this study) or two-sided inhibition (Shamma, 1985b), or even multiplicative coincidences (Loeb *et al.*, 1983; Deng and Geisler, 1987; Carney *et al.*, 2002; Yin *et al.*, 1987; Carney, 1990; Joris *et al.*, 1994; Shamma and Klein, 2000; Deshmukh *et al.* 2007; Howard and Rubel, 2010). The available physiological data from animal AVCN (especially regarding the T-chopper cell types) are consistent with these LIN notions (Blackburn and Sachs, 1990; Cedolin and Delgutte, 2007) but further evidence for the existence of the LIN has not been pursued in detail.

B. Estimation of neural E and TFS transmission at the peripheral and central levels

As stated earlier, the $r_{\mathrm{E\_E}}$ and $r_{\mathrm{TFS\_TFS}}$ metrics are not substantially different from the correlation-coefficient based metrics that have been used extensively in previous modeling work (Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006; Sheft *et al.*, 2008; Heinz and Swaminathan, 2009; Swaminathan, 2010). However, there are some important differences, The modeling and electrophysiological work conducted by Heinz and Swaminathan (2009) and Swaminathan and Heinz (2012) used shuffled auto- and cross-correlogram analyses to compute separate correlations for responses to E and TFS based on both model and recorded spike trains from auditory-nerve fibers. Specifically, their measures of TFS and E cues on the auditory nerve relied on a comparison between how these cues are expressed for the clean speech versus the vocoded signals. Consequently, when the expression of these cues is changed substantially by the vocoder, e.g., for TFS cues in an AM vocoder, the measure indicates extremely weak saliency [Figs. 3(B) and 3(C) in Swaminathan and Heinz (2012) show no TFS cues in PHENV and PDENV vocoded-speech conditions]. Furthermore, it is not possible with these measures to compare directly the E and TFS cues because they have different representations, and consequently they could only be compared using intelligibility assessments across different vocoder conditions. By contrast, our approach transformed the E and TFS cues into similar spectrogram representations that could subsequently be compared directly against a common auditory spectrogram or relative to each other.

In summary, our results are in conflict with some previous studies in that we predict balanced E and TFS contributions to the auditory spectrogram *regardless* of the vocoder, or of the intelligibility of the speech across different vocoder conditions. We arrive at this conclusion by addressing the expression of the E and TFS cues *relative* to each other, and show that it remains roughly constant regardless of whether the input speech (generated by different vocoder conditions) was intelligible or not. A logical follow up on this study in the future is to measure (or predict) the intelligibility of the TFS and E spectrograms for each vocoder condition and demonstrate that the two are comparably intelligible *regardless* of the vocoder condition, which is clearly a different conclusion from that of Swaminathan and Heinz (2012).

C. Effects of vocoder manipulations on the contribution of neural E and TFS cues

The simulations here demonstrate that the degradation of the FM component of speech by multi-channel (noise-excited) AM vocoders, and conversely of the AM component of speech by FM vocoders, leave the balance of the E and TFS information on the auditory nerve substantially unaltered. Thus, this study concludes that for normal-hearing listeners, the relative contribution of neural E and TFS cues across the speech-processing conditions tested in previous vocoder studies is largely unaffected and is comparable. The debate as to whether one or the other is more effective in transmitting speech information is thus based on a mistaken assumption—that vocoder manipulations of the signal can be used to change significantly the relative balance of neural E and TFS integrity with respect to the information they carry. It is important to note that these conclusions drawn from simulations based on speech stimuli presented in quiet also apply to other studies with vocoded speech in steady or fluctuating background maskers (e.g., Nelson *et al.*, 2003; Zeng *et al.*, 2005; Gnansia *et al.*, 2009; Hopkins *et al.*, 2008).

D. Cochlear pathologies

The auditory, E, and TFS spectrograms involve various physiological mechanisms postulated in normal auditory processing, such as cochlear filtering, hair cell filtering, neural phase-locking, level dependence, and threshold and saturation nonlinearities. Therefore, different pathologies, and assistive and prosthetic interventions that alter these mechanisms could significantly change the representation of neural E and TFS cues, and through them the auditory spectrogram.

One common change in hearing impairment is the broadening of the cochlear filters which has been assumed to cause significant loss of speech intelligibility, especially for speech presented against background noise (for a review, see Moore, 2007). It is evident from the formulations and model filters in the Methods section that the broadening of the cochlear filters [$h_s(t)$] does not necessarily cause a critical loss if the filters maintain their sharp edges and phase roll-off. Instead, since the effective filter for producing the auditory, E, and TFS spectrograms is the spatial-derivative (h'_s), the loss of the steep high-frequency edges of the filters and their accompanying rapid phase-shifts near the best-frequency are more likely to cause significant deterioration in the spectrograms, as suggested by recent physiological findings in animals with noise-damaged cochlear responses (Heinz *et al.*, 2010).

Another potential pathology is the loss of phase locking in the auditory nerve which would degrade or completely abolish the neural TFS cues while leaving the E spectrogram relatively intact. While several psychoacoustic studies conducted with listeners showing cochlear damage are consistent with this notion (e.g., Buss *et al.*, 2004), such a loss of phase locking has not been demonstrated in neurophysiological studies with noise- or drug-induced hearing loss in animals (e.g., Miller *et al.*, 1997; however, see Kale and Heinz, 2010 for a demonstration of a relative deficit in phase-locking to TFS cues compared to phase-locking to E cues). On the other hand, recent work demonstrated acute loss of afferent nerve terminals and delayed degeneration of the cochlear nerve in animals with noise-induced damage to the ear (Kujawa and Liberman, 2009). Surprisingly, these animals showed intact inner and outer hair cells (and thus, normal frequency selectivity). Irrespective of the true nature of the central mechanism proposed here (LIN, cross-correlation, or coincidence detection), it is most likely that a loss of auditory nerve fibers should strongly alter the conversion of neural TFS information to a TFS spectrogram.

The poorer-than-normal speech intelligibility in steady and fluctuating noise typically observed for hearing-impaired listeners with or without hearing aids (e.g., Lorenzi *et al.*, 2006; Hopkins *et al.*, 2008) suggests that the TFS cues conveyed by the reduced number of surviving auditory nerve fibers (showing either normal or degraded frequency selectivity) are not sufficient to restore the *robust* central TFS spectrograms demonstrated above (cf. Sec. III D).

E. Cochlear implants

With cochlear implants, the spatial spread of nerve stimulation is usually quite large resulting effectively in a very broadband filtering of acoustic information and poor frequency resolution (e.g., Wilson and Dorman, 2008). Nevertheless, simulations using broadband vocoders with normal-hearing listeners have demonstrated that some intelligibility is preserved both in quiet and in moderate levels of noise. This has been implicitly attributed to the "survival" of the E spectrogram (Friesen *et al.*, 2001; Shannon, 2007; Moore and Shannon, 2009), when in fact both TFS and E convey the same information in such normal-hearing listeners, and hence both cues contribute roughly comparable information to intelligibility (Figs. 3–5). The limitations for cochlear implant listeners, however, are severe because of the dissociation of phase-locked responses and their envelopes. Cochlear-implant stimulation may reproduce a broadly analyzed spectrogram, but it is usually coupled with arbitrary phase-locked responses that do not reflect what normal-hearing listeners receive. Instead, the TFS on the auditory nerve of implantees is completely uninformative, consisting of constant or simply variable pulse rates whose purpose is simply to deliver the charge and stimulate the nerve, albeit with the wrong temporal structure. Therefore, the key difference between cochlear implantees and normal-hearing subjects listening to eight-band AM vocoded speech is the contribution of the TFS spectrogram in normal-hearing listeners, which may explain the latter's much better listening competence.

It should be noted that inducing the "correct" pattern of phase-locked responses in cochlear prosthesis stimulation is not trivial. The neural fine-structure has to be consistent with the band-limited structure of the filters. Effectively, it needs to be equivalent to the carriers from such filter responses. That aside, to extract the TFS spectrogram, the LIN detects the relative phase of responses across the auditory-nerve fiber array. Consequently, to reconstruct the spectrogram from a few channels of a cochlear implant, each has to induce phase-locking patterns that are sufficiently distinct from its neighbors so as to create clear discontinuities or edges between them. These in turn then recreate the TFS spectrogram of the stimulus or contribute to the auditory spectrogram as explained earlier.

V. CONCLUSIONS

(1) For normal-hearing listeners, the relative contribution of neural E and TFS cues across the speech-processing

conditions tested in previous vocoder studies is comparable. The debate as to whether one or the other is more effective in transmitting phonetic information is based on a mistaken assumption—that vocoder manipulations of the signal can be used to change significantly the relative balance of neural E and TFS integrity with respect to the information they carry. Instead, we argue that the neural E and TFS of an input speech signal are approximately equivalent and their relative proportion remains roughly constant regardless of such manipulations (e.g., number of bands and distortions introduced) as well as signal demodulation techniques, e.g., coherent (Clark and Atlas, 2009) or the more common but noisier incoherent demodulation based on the Hilbert transform. Thus, we conclude that the neural TFS in normal-hearing listeners induces a spectrogram-like representation that resembles the E and auditory spectrograms, and which is readily extracted by a post-cochlear mechanism such as lateral inhibition, coincidence detection, or cross correlation. Consequently, both neural TFS and E may contribute approximately equally to the intelligibility of the input speech signal under normal circumstances.

(2) Simulations and theoretical analyses demonstrate that, over the frequency range where neural phase locking is present, the TFS spectrogram is more robust than the E spectrogram in extremely noisy conditions and under certain distortions such as highly compressed responses. Furthermore, peripheral mechanisms such as the fast-acting compression produced by outer hair cells and the threshold and saturation nonlinearities of inner hair cells limit specifically the transmission of neural E cues. Consequently, it is likely that in many challenging environmental conditions, the neural TFS cues are critical for maintaining sufficient levels of intelligibility.

(3) Hearing impairment may cause severe disruptions of auditory-nerve responses resulting in degraded E and TFS spectrograms. Each pathology has its unique effects that can be simulated by manipulations in various modules of the model. These include shallower phase functions for auditory filters, broadening of auditory filters' bandwidth, reduced dynamic range, and partial loss of nerve fibers. Each of these deficits could disrupt the extraction of the E or TFS spectrograms according to its own unique role in the overall model of early auditory processing.

(4) Cochlear implants deliver mostly degraded representations of the E spectrogram. But more detrimental to perception is their inability to deliver adequate TFS information. In not doing so they give up cues that can serve as a source of robustness and a significant contributor to better intelligibility. The technological challenge in overcoming this limitation requires that current spread be minimized, and that nerve stimulation be able to induce phase-locked spiking patterns with sharp discontinuities similar to those in normal listeners.

(5) Our analysis emphasizes the notion that neural TFS cues in auditory-nerve responses may be converted centrally into spectrograms that complement the E cues via neural mechanisms such as lateral inhibition, coincidence detection, or cross-correlation. This form of reconstruction from TFS cues must be distinguished from the peripheral envelope reconstruction occurring at the output of cochlear filters in response to the acoustic FM cues in speech, as it requires the operation of more central mechanisms. Our analysis suggests that this form of central reconstruction process may be a crucial determinant of robust speech coding and the building of invariant speech representations.

ACKNOWLEDGMENTS

This research began with an unpublished, modeling study conducted by G. Gilbert in collaboration with S.S. and C.L. G. Gilbert unfortunately passed away in 2011, but must be acknowledged for his early contribution to the core ideas developed in the present study. The authors also wish to thank B. C. J. Moore for valuable comments on a preliminary version of this manuscript. S.S. was supported by a Blaise-Pascal Chair and NIH grants (Grant No. R01DC005779). C.L. was supported by a grant from ANR (HEARFIN project). The authors wish to thank two anonymous reviewers for valuable comments on previous versions of this manuscript.

[1] The small difference predicted here between intelligibility of AM and FM multichannel vocoders corresponds well only to reported measurements with isolated speech segments (VCV syllables) as in Gilbert and Lorenzi (2006), and not those measured with full sentences as in Smith *et al.* (2002). A possible reason for the low intelligibility measured with multi-band FM-vocoded sentences is the detrimental effects of the artificially boosted-noise in the gaps that normally serve to segment normal speech. Clearly, our global spectrogram measure is insensitive to these issues.

Ardoint, M., and Lorenzi, C. (**2010**). "Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine-structure or envelope cues," Hear. Res. **260**, 89–95.

Ardoint, M., Sheft, S., Fleuriot, P., Garnier, S., and Lorenzi, C. (**2010**). "Perception of temporal fine structure cues in speech with minimal envelope cues for listeners with mild- to-moderate hearing loss," Int. J. Audiol. **49**, 823–831.

Baskent, D. (**2006**). "Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels," J. Acoust. Soc. Am. **120**, 2908–2925.

Blackburn, C., and Sachs, M. (**1990**). "The representations of the steady-state vowel sound /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons," J. Neurophysiol. **63**(5), 1191–1212.

Buss, E., Hall, J. W., 3rd, and Grose, J. H. (**2004**). "Temporal fine structure cues to speech recognition and pure tone modulation in observers with sensorineural hearing loss," Ear. Hear. **25**, 242–250.

Carlyon, R. P., and Shamma, S. A. (**2003**). "An account of monaural phase sensitivity," J. Acoust. Soc. Am. **114**, 333–348.

Carney, L. H. (**1990**). "Sensitivities of cells in anteroventral cochlear nucleus of cat to spatiotemporal discharge patterns across primary afferents," J. Neurophysiol. **64**, 437–456.

Carney, L. H., Heinz, M. G., Evilsizer, M. E., Gilkey, R. H., and Colburn, H. S. (**2002**). "Auditory phase opponency: A temporal model for masked detection at low frequencies," Acta Acust. Acust. **88**, 334–347.

Cedolin, L., and Delgutte, B (**2007**). "Spatio-temporal representation of the pitch of complex tones in the auditory nerve," in *Hearing—From Sensory Processing to Perception*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Springer-Verlag, Berlin), pp. 61–70.

Clark, P., and Atlas, L. (**2009**). "Time-frequency coherent modulation filtering of non-stationary signals," IEEE Trans. Signal Process. **57**(11), 4323–4332.

Deng, L., and Geisler, C. (**1987**). "Responses of auditory-nerve fibers to nasal consonant-vowel syllables," J. Acoust. Soc. Am. **82**, 1977–1988.

Deshmukh, O., Espy-Wilson, C., and Carney, L. (**2007**). "Speech enhancement using the modified phase-opponency model," J. Acoust. Soc. Am. **121**, 3886–3898.

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

Drullman, R., Festen, J. M., and Plomp, R. (**1994**)."Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

Dudley, H., Riesz, R. R., and Watkins, S. S. A. (**1939**). "A synthetic speaker," J. Franklin Inst. **227**, 739–764.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

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 interruption on the intelligibility of speech based on temporal fine-structure or envelope cues," J. Acoust. Soc. Am. **122**, 1336–1339.

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

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

Heinz, M. G., and Swaminathan, J. (**2009**). "Quantifying Envelope and Fine-Structure Coding in Auditory Nerve Responses to Chimaeric Speech," J. Assoc. Res. Otolaryngol. **10**, 407–423.

Heinz, M. G., Swaminathan, J., Boley, J. D., and Kale, S. (**2010**). "Across-fiber coding of temporal fine structure: Effects of noise-induced hearing loss on auditory-nerve responses," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez- Poveda, A. R. Palmer, and R. Meddis (Springer, New York), pp. 621–630.

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.

Howard, M. A., and Rubel, E. W. (**2010**). "dynamic spike thresholds during synaptic integration preserve and enhance temporal response properties in the avian cochlear nucleus," J. Neurosci. **30**, 12063–12074.

Ibrahim, R. A., and Bruce, I. C. (**2010**). "Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), Chap. 40, pp. 429–438.

Johnson, D. H. (**1980**). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," J. Acoust. Soc. Am. **68**, 1115–1122.

Joris, P. X., and Yin, T. C. (**1992**). "Responses to amplitude-modulated tones in the auditory nerve of the cat," J. Acoust. Soc. Am. **91**, 215–232.

Joris, P. X., Carney, L. H., Smith, P. H., and Yin, T. C. (**1994**). "Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency," J. Neurophysiol. **71**, 1022–1036.

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.

Kujawa, S. G., and Liberman, M. C. (**2009**). "Adding insult to injury: Cochlear nerve degeneration after 'temporary' noise-induced hearing loss," J. Neurosci. **29**, 14077–14085.

Loeb, G. E., White, M. W., and Merzenich, M. M. (**1983**). "Spatial cross-correlation, A proposed mechanism for acoustic pitch perception," Biol. Cybern. **47**, 149–163.

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. USA **103**, 18866–18869.

Lyon, R., and Shamma, S. (**1996**). "Auditory representations of timbre and pitch," *Auditory Computation*, edited by Hawkins, H. L., McMullen, T. A., Popper, A. N., and Fay, R. R. (Springer, New York), pp. 221–270.

Miller, R. L., Schilling, J. R., Franck, K. R., and Young, E. D. (**1997**). "Effects of acoustic trauma on the representation of the vowel ⟨⟨eh⟩⟩ in cat auditory nerve fibers," J. Acoust. Soc. Am. **101**, 3602–3616.

Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological, and Technical Issues*, 2nd ed. (Wiley, Chichester), pp. 1–327.

Moore, B. C. J. (**2008**). "The choice of compression speed in hearing aids: Theoretical and practical considerations and the role of individual differences," Trends Amplif. **12**, 103–112.

Moore, D. R., and Shannon, R. V. (**2009**). "Beyond cochlear implants: awakening the deafened brain," Nat. Neurosci. **12**, 686–691.

Nelson, P. B., Jin, S. H., 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.

Rice, S. O. (**1973**). "Distortion produced by band limitation of an FM wave," Bell Syst. Tech. J. **52**, 605–626.

Sachs, M. B., and Young, E. D. (**1979**). "Encoding of steady state vowels in the auditory nerve: Representation in terms of discharge rate," J. Acoust. Soc. Am. **66**, 470–479.

Shamma, S. A. (**1985a**). "Speech processing in the auditory system: I. The representation of speech sounds in the responses of the auditory nerve," J. Acoust. Soc. Am. **78**, 1612–1621.

Shamma, S. A. (**1985b**). "Speech processing in the auditory system: II. Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," J. Acoust. Soc. Am. **78**, 1622–1632.

Shamma, S. A., and Klein, D. J. (**2000**). "The case of the missing pitch templates: How harmonic templates may form in the early auditory system," J. Acoust. Soc. Am. **107**, 2631–2644.

Shamma, S. A., and Morrish, K. A. (**1987**). "Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea," J. Acoust. Soc. Am. **81**, 1486–1498.

Shannon, R. V. (**2007**). "Understanding hearing through deafness," Proc. Natl. Acad. Sci. USA **104**, 6883–6884.

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.

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

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

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.

Voelcker, H. B. (**1966**). "Towards a unified theory of modulation. I. phase-envelope relationships," Proc. IEEE **54**, 340–354.

Wilson, B. S., and Dorman, M. F. (**2008**). "Cochlear implants: Current designs and future possibilities," J. Rehabil. Res. Dev. **45**, 695–730.

Yang, X., Wang, K., and Shamma, S. (**1992**). "Auditory representations of acoustic signals," IEEE Trans. Info. Theory **38**, 824–839.

Yin, T. C. T., Chan, J. C. K., and Carney, L. H. (**1987**). "Effects of interaural time delays of noise stimuli on low-frequency cells in the cats inferior colliculus III. Evidence for cross correlation," J. Neurophysiol. **58**, 562–583.

Young, E., and Sachs, M. (**1979**) "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403.

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. USA **102**, 2293–2298.