Multiresolution spectrotemporal analysis of complex sounds

Taishih Chi,[a] Powen Ru,[b] and Shihab A. Shamma[c]
Center for Auditory and Acoustics Research, Institute for Systems Research Electrical and Computer Engineering Department, University of Maryland, College Park, Maryland 20742

(Received 22 June 2004; revised 2 May 2005; accepted 12 May 2005)

A computational model of auditory analysis is described that is inspired by psychoacoustical and neurophysiological findings in early and central stages of the auditory system. The model provides a unified multiresolution representation of the spectral and temporal features likely critical in the perception of sound. Simplified, more specifically tailored versions of this model have already been validated by successful application in the assessment of speech intelligibility [Elhilali *et al.*, Speech Commun. **41**(2-3), 331–348 (2003); Chi *et al.*, J. Acoust. Soc. Am. **106**, 2719–2732 (1999)] and in explaining the perception of monaural phase sensitivity [R. Carlyon and S. Shamma, J. Acoust. Soc. Am. **114**, 333–348 (2003)]. Here we provide a more complete mathematical formulation of the model, illustrating how complex signals are transformed through various stages of the model, and relating it to comparable existing models of auditory processing. Furthermore, we outline several reconstruction algorithms to resynthesize the sound from the model output so as to evaluate the fidelity of the representation and contribution of different features and cues to the sound percept. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1945807]

PACS number(s): 43.66.Ba, 43.71.An, 43.71.Gv [KWG] Pages: 887–906

I. INTRODUCTION

Cochlear frequency analysis has for decades influenced the development of algorithms and perceptual measures for the analysis and recognition of speech and audio. Examples include the formulation of the articulation index (Kryter, 1962) to estimate the effect of noise on speech intelligibility, and the exploitation of models of psychoacoustical masking for the efficient coding of speech and music (Pan, 1995). However, cochlear analysis of sound and the extraction of the acoustic spectrum in the cochlear nucleus are only the earliest stages in a sequence of substantial transformations of the neural representation of sound as it journeys up to the auditory cortex via the midbrain and thalamus. And, while much is known about the neural correlates of sound pitch, location, loudness, and the representation of the spectral profile in these early stages, the response properties and functional organization in the more central structures of the inferior colliculus, medial geniculate body, and the cortex have only begun to be uncovered relatively recently (deRibaupierre and Rouiller, 1981; Kowalski *et al.*, 1996; Schreiner and Urbas, 1988b; Miller *et al.*, 2002; Lu *et al.*, 2001; Eggermont, 2002; Ulanovsky *et al.* 2003). Consequently, it is less common that one finds ideas from central auditory processing being applied in psychoacoustics (Houtgast, 1989; Dau *et al.*, 1997a, b; Ewert and Dau, 2000; Grimault *et al.*, 2002) and in design of speech and audio processing systems (Arai *et al.*, 1996; Pitton *et al.*, 1996; Greenberg and Kingsbury, 1997; Tchorz and Kollmeier, 1999; Hansen and Kollmeier, 1999; Kleinschmidt *et al.*, 2001; Atlas and Shamma, 2003). Interestingly, the opposite has occurred, that is, numerous useful algorithms and representations that were developed decades ago, based only on engineering intuition, have turned out to be in hindsight grounded on solid auditory neural processing strategies (Hermansky and Morgan, 1994; Atal, 1974).

To exploit the accumulating physiological findings from the central auditory system and from psychoacoustic experiments, it is essential that they be reformulated as mathematical models and signal processing algorithms. To achieve this objective, this paper provides two specific contributions:

(1) It describes a detailed computational model of central auditory processing. Simplified, specifically tailored, versions of this model have already appeared in previous publications from our group where we demonstrated its successful applications in the objective evaluation of speech intelligibility (Elhilali *et al.*, 2003; Chi *et al.*, 1999) and the perception of phase of complex sounds (Carlyon and Shamma, 2003). Here we provide a more complete mathematical formulation of the model, illustrating how complex signals are transformed through various stages of the model, and relating it to comparable existing models of auditory processing. This expanded version of the model is completely consistent with the earlier versions and has been validated to account for the types of signals and distortions considered in earlier publications.

(2) It outlines algorithms for reconstructing the input acoustic signal from its final model outputs. These algorithms are important in that they demonstrate the sufficiency of this model representation by reconstructing faithful replica of the original inputs. They also enable the model to be used in assessing the perceptual significance of various output features, as well as in applications where a modified final acoustic waveform is necessary such as noise suppression for general audio and hearing aids.

[a]Present address: Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, Republic of China
[b]Present address: Cybernetics InfoTech Inc.
[c]Electronic mail: sas@eng.umd.edu

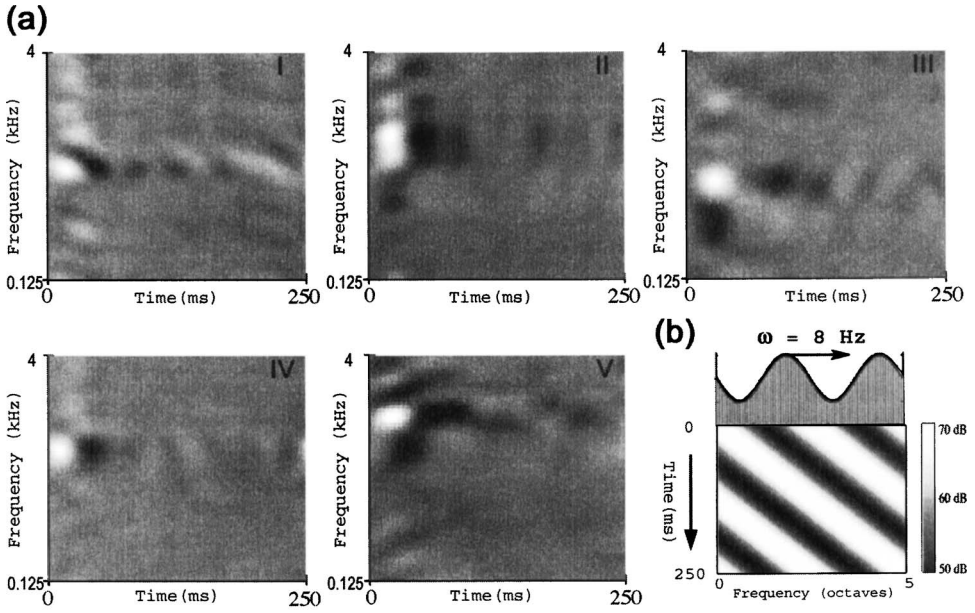

FIG. 1. Details of the dynamic ripple stimulus and examples of spectrotemporal response fields (STRFs) in primary auditory cortex (A1). (a) Example STRFs recorded from A1 of the ferret. White (black) color indicates regions of strongly excitatory (suppressed) responses. The STRFs display a wide range of properties from temporally fast (iv) to slow (iii, v), spectrally sharp (iv) to broad (i, ii), with symmetric (iv) or asymmetric (iii) inhibition. (b) The moving ripple spectral profile $[S(t,x)]$ is defined by the expression: $S(t,x)=1+A\cdot\sin(2\pi\cdot(\omega\cdot t+\Omega\cdot x)+\Phi)$, where A is the modulation depth, Φ is the phase of the profile, ω is called ripple velocity (in Hz), and Ω controls the spectral variation (or modulation)—also called ripple density (in cycles/octave). It usually consists of many simultaneously presented tones, depicted schematically by the vertical lines along the frequency axis. The tones are usually equally spaced along the logarithmic frequency axis and spanning 5 oct (e.g., 0.25–8 kHz or 0.5–16 kHz). The sinusoidal spectral profile $S(t,x)$ is depicted by the dashed curve. The spectrogram of one ripple profile is shown in the bottom panel ($\Omega=0.4$ cycles/octave, $\omega=8$ Hz).

The model we describe is not biophysical in spirit, but rather it abstracts from the physiological data an interpretation that we believe is likely to be relevant in the design of sound engineering systems. Two particularly important physiological observations are incorporated. The first is the apparent progressive loss of temporal dynamics from the periphery to the cortex. Thus, on the auditory nerve, rapid phase locking to individual spectral components of the stimulus survives up to 4–9 kHz. It diminishes to moderate rates of synchrony in the midbrain (under 1 kHz), and to the much lower rates of modulations in the cortex (less than 30 Hz)[1] (Kowalski *et al.*, 1996; Miller *et al.*, 2002; Schreiner and Urbas, 1988a; Langner, 1992). Another important change in the nature of the neural responses is the emergence of elaborate selectivity to combined spectral and temporal features, selectivity that is typically much more complex than the relatively simple tuning curves and dynamics of auditory-nerve fiber responses (Nelken and Versnel, 2000; Shamma *et al.*, 1993; Edamatsu *et al.*, 1989).

The computational model consists of two major auditory transformations. An *early* stage captures monaural processing from the cochlea to the midbrain. It transforms the acoustic stimulus to an auditory time-frequency spectrogramlike representation that combines relatively simple bandpass spectral selectivity with moderate temporal dynamics (<1000 Hz). The second is called the *cortical* stage because it reflects the more complex spectrotemporal analysis presumed to take place in mammalian AI. In the following section, we review the cortical physiological data and psychoacoustical results that motivated and justified this model's development. The mathematical formulation of the early and cortical stages are summarized in Secs. III and IV, together with an illustration of the way in which a variety of complex sounds are represented at each stage. In Sec. V, algorithms to *reconstruct* audible approximate versions of the original sounds from the model's representations are described. We also provide an example of how the reconstructed signals can be used to assess the contribution of different ranges of spectro-temporal modulations to the intelligibility of speech. Finally, we end in Sec. VI with a summary and a brief assessment of the utility of the model in a variety of potential applications.

II. AUDITORY CORTICAL PHYSIOLOGY

The *cortical* stage of the model is strongly inspired by extensive data and ideas gained from physiological and psychoacoustical experiments over the last decade. Specifically, much insight has been gained from measurements of the so-called spectro-temporal response fields (STRF) of AI cells. Examples of a variety of measured STRFs are shown in Fig. 1(a). A STRF summarizes the way a cell responds to the stimulus. Along its ordinate—"frequency axis"—the color white depicts acoustic frequencies that excite responses, black denotes frequencies that suppress (or inhibit) responses, while gray denotes frequency regions of no response. Thus, some STRFs are responsive (excited or suppressed) over a broad range of frequencies, exceeding an octave (**ii**), while others are quite narrowly tuned (**iv**). Along its abscissa—"Time axis"—the STRF depicts the response dynamics to an "impulse" of energy delivered at each frequency. In most STRFs in Fig. 1, the impulse response con-

sists of a damped wave of alternating excitatory (white) and inhibitory (black) responses. The response fades rapidly in some STRFs (**iv**), while it lasts twice as long in others (**v**). Finally, this combined time-frequency sensitivity can take more complex forms that are "inseparable" as in the *oriented* STRFs of **i** and **iii**.

Another way to understand the STRF of a cell is through the implied response selectivity to special test stimuli. STRFs have been measured in many ways (Calhoun and Schreiner, 1995; deCharms *et al.*, 1998), one of which is the "ripple analysis method" (Shamma *et al.*, 1995; Kowalski *et al.*, 1996; Klein *et al.*, 2000). Ripples are broadband noise with sinusoidally modulated spectrotemporal envelopes with different parameters [Fig. 1(b)]. They serve the same function as regular sinusoids in measuring the transfer function of linear filters, except that they are two dimensional (spectral and temporal). AI cells respond well to ripples and are usually selective to a narrow range of ripple parameters that reflect details of their *spectrotemporal transfer functions*. By compiling a complete description of the responses of a cell to all ripple densities and velocities it is possible by an inverse Fourier transform to compute the corresponding STRF.

Therefore, a cell's STRF and its ripple spectrotemporal transfer functions are uniquely related through the Fourier transform. For instance, broadly tuned cells are most responsive to low ripple densities, whereas the opposite is true for narrowly tuned cells. Similarly, STRFs with relatively sluggish dynamics respond poorly to fast ripple rates. Finally, oriented STRFs imply strong selectivity to correspondingly oriented ripples (i.e., of an appropriate rate-density combination). From a functional perspective, the rich variety of STRFs found in AI implies that each STRF acts as a *modulation selective filter* of its input spectrogram, specifically tuned to a particular range of spectral resolutions (also called *scales*) and a limited range of temporal modulations (or *rates*). The collection of all such STRFs then would constitute a filterbank spanning the broad range of psychoacoustically observed scale and rate sensitivity in humans and animals (Viemeister, 1979; Green, 1986; Dau *et al.*, 1997a; Amagai *et al.*, 1999; Chi *et al.*, 1999).

Evidence of the importance of spectrotemporal modulations in the perception of complex sounds has come from experiments in which systematic degradations of the speech signal were correlated with the gradual loss of intelligibility (Drullman *et al.*, 1994; Shannon *et al.*, 1995). All such experiments have consistently pointed to the importance of the slow temporal (<30 Hz) and broad spectral modulations in conveying a robust level of intelligibility (Drullman, 1995; Fu and Shannon, 2000). In fact, the relationship between the temporal modulations and speech intelligibility has long been codified in the formulation of the widely used speech transmission index (STI) (Houtgast *et al.*, 1980). In an extension of such ideas, and inspired by the neurophysiological data briefly reviewed here, we formulated and tested a spectro-temporal modulation index (STMI) (Chi *et al.*, 1999; Elhilali *et al.*, 2003), which assesses the integrity of *both* the spectral and temporal modulations in a signal as a measure of intelligibility. The STMI proved reliable in capturing the deleterious effects of noise and reverberations, as well as of previously difficult to characterize distortions such as nonlinear compression, phase jitter, and phase shifts (Elhilali *et al.*, 2003).

In summary, there is physiological and psychoacoustical evidence that the auditory system, particularly at the level of AI, analyzes the dynamic acoustic spectrum of the stimulus extracted at its earlier stages. It does so by explicitly representing its spectrotemporal modulations by employing arrays of spectrally and temporally selective STRFs. In the remainder of this paper, we elaborate on the mathematical formulation of these computations, and detail a method to invert the representations back to the acoustic stimulus so as to hear the effects of arbitrary manipulations.

III. THE EARLY STAGE: THE AUDITORY SPECTROGRAM

Sound signals undergo a series of transformations in the early auditory system and are converted from a one-dimensional pressure time waveform to a two-dimensional pattern of neural activity distributed along the tonotopic (roughly a logarithmic frequency) axis. This two-dimensional pattern, which we shall call the *auditory spectrogram*, represents an enhanced and noise-robust estimate of the Fourier-based spectrogram (Wang and Shamma, 1994). Details of the biophysical basis and anatomical structures involved are available (Shamma, 1985b; Shamma *et al.*, 1986; Yang *et al.*, 1992).

A. Mathematical formulation

The stages of the early auditory model are illustrated in Fig. 2. In brief, the first operation is an affine wavelet transform of the acoustic signal $s(t)$. It represents the spectral analysis performed by the cochlear filter bank. This analysis stage is implemented by a bank of 128 overlapping constant-Q ($Q_{10dB} \simeq 3$) bandpass filters with center frequencies (CFs) that are uniformly distributed along a logarithmic frequency axis (x), over 5.3 oct (24 filters/octave). The impulse response of each filter[2] is denoted by $h(t;x)$. These cochlear filter outputs $y_{coch}(t,x)$ are transduced into auditory-nerve patterns $y_{AN}(t,x)$ by a hair cell stage consisting of a high-pass filter, a nonlinear compression $g(\cdot)$, and a membrane leakage low-pass filter $w(t)$ accounting for decrease of phase-locking on the auditory nerve beyond 2 kHz. The final transformation simulates the action of a lateral inhibitory network (LIN) postulated to exist in the cochlear nucleus (Shamma, 1989), which effectively enhances the frequency selectivity of the cochlear filter bank (Lyon and Shamma, 1996; Shamma, 1985b). The LIN is simply approximated by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectifier to produce $y_{LIN}(t,x)$. The final output of this stage is obtained by integrating $y_{LIN}(t,x)$ over a short window, $\mu(t;\tau)=e^{-t/\tau}u(t)$, with time constant $\tau = 8$ ms mimicking the further loss of phase locking observed in the midbrain. The mathematical formulation for this model can be summarized as follows:

$$y_{coch}(t,x) = s(t) \otimes_t h(t;x), \tag{1}$$

$$y_{AN}(t,x) = g(\partial_t y_{coch}(t,x)) \otimes_t w(t), \tag{2}$$

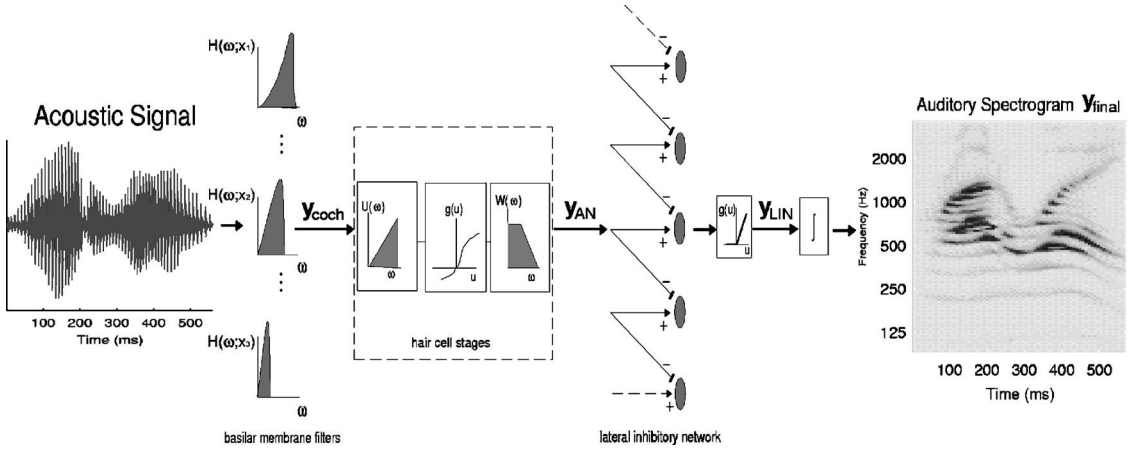

FIG. 2. Schematic of early auditory stages. The acoustic signal is analyzed by a bank of constant-Q cochlear-like filters. The output of each filter (y_{coch}) is processed by a hair cell model (y_{AN}) followed by a lateral inhibitory network, and is finally rectified (y_{LIN}) and integrated to produce the auditory spectrogram (y_{final}).

$$y_{LIN}(t,x) = \max(\partial_x y_{AN}(t,x), 0), \tag{3}$$

$$y_{final}(t,x) = y_{LIN}(t,x) \otimes_t \mu(t;\tau), \tag{4}$$

where $\otimes_t$ denotes convolution operation in the time domain.

The model described above attempts to capture many of the important properties of auditory processing that are critical for our objectives and further detailed in the following sections. In creating such a computational model, one has to balance many conflicting requirements and hence make compromises on what simplifications to apply and what details to include. For instance, our cochlear filtering is essentially linear, lacking such phenomena as two-tone suppression and level-dependent tuning, which are critical in some applications (Carney, 1993). The lateral inhibition model is very schematic and lacks details of single neurons (Shamma, 1989). We also have no explicit adaptive properties in our current model (Westerman and Smith, 1984; Meddis *et al.*, 1990; Dau *et al.*, 1996). All of these details are likely to be important in certain circumstances and should be added when needed (Cohen, 1989).

B. Examples of the auditory spectrogram

Examples of the information preserved at the LIN output $[y_{LIN}(t,x)]$ and midbrain levels $[y_{final}(t,x)]$ of the model are described for five types of progressively more complex stimuli; a three-tone combination, noise, a harmonic complex, ripples, and speech and music segments. Understanding details of the auditory spectrogram $y_{final}(t,x)$ is important since it serves as the input to the cortical analysis stage as we discuss in the next section.

1. Three tones: 250, 1000, and 4000 Hz

Figure 3(a) illustrate the response patterns due to a low-, medium-, and high-frequency tones. The low-frequency tones (250 and 1000 Hz) evoke the typical traveling-wave phase-locked patterns observed experimentally in the auditory nerve (Pfeiffer and Kim, 1975; Shamma, 1985a). For the high-frequency tone, phase locking is lost and only the envelope is preserved. These patterns remain the same at the midbrain stage except that the upper limit of phase locking decreases to below 1000 Hz. Thus, in the right panel of Fig. 3(a), substantial phase locking is only seen for the 250-Hz tone.

2. Noise

Figure 3(b) (left panel) depicts the $y_{final}(t,x)$ generated by a broadband noise constructed with 59 random-phase tones that are equally spaced (0.1 oct) on a logarithmic frequency axis (135–7465 Hz). At this intertone spacing, two to four tones interact within each constant-Q cochlear filter, producing a modulated carrier at the CF of each filter. The envelope modulations at each filter reflect its bandwidth and the intertone spacing in the stimulus. In the low frequency regions (<1000 Hz), the output $[y_{final}(t,x)]$ captures both the carrier and envelope. At higher CF regions, the predominant representation is that of the envelopes as carrier phase-locking diminishes. Note that the modulation rates of the envelope increase (in Hz) with CF as filter bandwidths and stimulus intertone spacing become wider. Maximum rates are limited by maximum filter bandwidths, and hence do not exceed a few hundred Hertz in most mammals (Joris and Yin, 1992).

3. Harmonic complexes

Unlike broadband noise, harmonic complexes have uniform intertone spacing equal to the fundamental frequency of the harmonic series. Consequently, the fundamental component and low-order harmonics remain well resolved by the filters, whereas many high-order harmonics fall within the bandwidth of a cochlear filter at high CFs. Figure 3(b) (middle panel) illustrates the responses to *in-phase* harmonic series stimulus with the fundamental at 80 Hz. Low-order harmonics (<8th) are well resolved (as indicated by the arrows), each dominating the response within one filter, and hence there are little envelope modulations. At high CFs, the unresolved higher harmonics interact, producing the strong 80-Hz periodic envelope modulations. When the harmonics are random phase [Fig. 3(b), right panel], the envelope modulations become irregular and less peaked, but still pre-

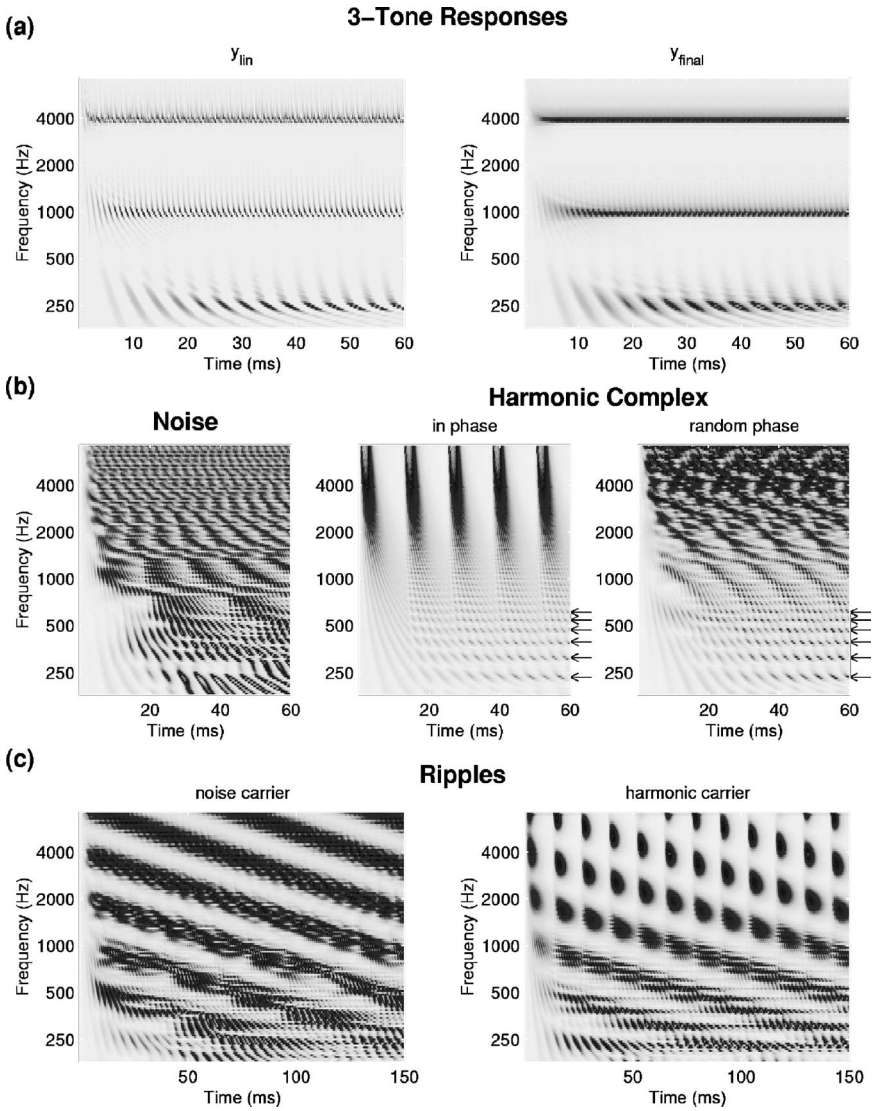

FIG. 3. Examples of early auditory responses for progressively more complex stimuli. (a) A three-tone (250, 1000, 4000 Hz) combination; left panel shows the response at the LIN output $[y_{LIN}(t,x)]$ and right panel shows the response at midbrain level of the model $[y_{final}(t,x)]$. (b) The midbrain output $y_{final}(t,x)$ to a broadband noise (left), broadband in-phase harmonic complex (middle), and a broadband random-phase harmonic complex (right). (c) The $y_{final}(t,x)$ output to a spectro-temporally modulated noise (left) and spectro-temporally modulated in-phase harmonic series (right). All stimuli are sampled at 16 kHz.

serve their periodicity of 80 Hz. The key general observation to make about these envelope modulations is that they relate to intercomponent interactions, and hence are affected by the spacing, phase, and relative amplitudes of the components—factors reflecting the perceptual timber of the sound. In the next two example stimuli, we distinguish these intermediate rate modulations from *slow modulations* created by production mechanisms which, in speech and music, strongly determine the intelligibility of speech and identity of an instrument.

4. Ripples: Spectrotemporally modulated noise

The model's outputs for a spectro-temporal modulated broadband noise—also called a *ripple*—are shown in Fig. 3(c) (left panel). The stimulus is generated by amplitude modulating each of the 59 components of the noise described earlier in Fig. 3(b) (left panel) so as to produce a spectrotemporal profile as depicted in Fig. 1(b). Detailed definition and description of these stimuli can be found in Chi *et al.* (1999) and Kowalski *et al.* (1996).

The left panel of Fig. 3(c) displays the y_{final} output for a downward sweeping ripple (ω=16 Hz, Ω=1 cycle/octave). At low CFs ($\ll$1000 Hz), the responses exhibit temporal modulations at three different time scales simultaneously. The *slow* modulations (16 Hz) reflect the spectrotemporal sinusoidal envelope of the ripple. They ride on top of the *intermediate* modulations due to component interactions (30–400 Hz). These in turn ride on one top of the *fast* responses phase locked to the tones of the stimulus. At high CFs, only the slow and intermediate modulations survive. At very low CFs ($<$250 Hz), slow and intermediate modulation rates may become comparable due to the narrower bandwidths of the filters, and hence the distinct view of the ripple modulations deteriorates.

Figure 3(c) (right panel) illustrates the responses to the *same* ripple spectrotemporal envelope, but this time carried by the harmonic series of Fig. 3(b) (middle panel). The slow modulations are again well represented in the responses, but this time riding on a totally different pattern of intermediate modulations that reflect the 80-Hz periodicity of the fundamental. It is in this sense that we distinguish between these two types of envelope modulations: the intermediate are strictly due to component interactions whereas the slow modulations are superimposed on top and are related to the evolution of the spectrum, e.g., from one syllable to another

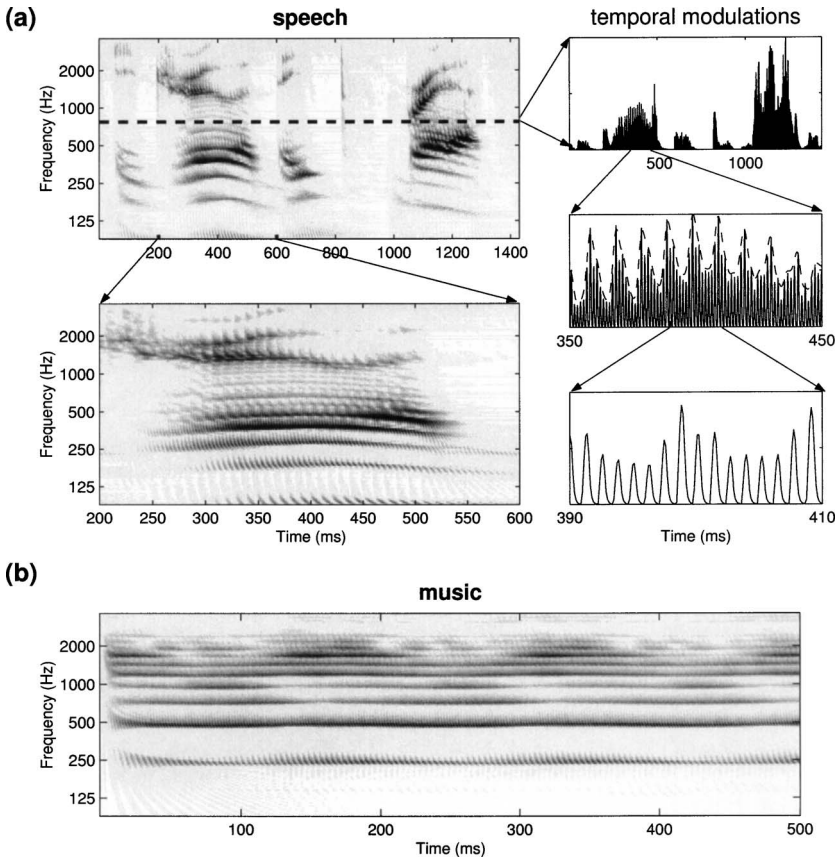

FIG. 4. Examples of auditory spectrograms $[y_{final}(t,x)]$ for speech and music stimuli. (a) The auditory spectrogram of the utterance /He drew a deep breath/ spoken by a male with a pitch of approximately 100 Hz. The dashed line marks the auditory channel at 750 Hz whose temporal modulations are depicted to the right at different time scales. At the coarsest scale (top panel), the slow modulations (few Hz) roughly correlate with the different syllabic segments of the utterance. At an intermediate scale (middle panel), modulations due to interharmonic interactions occur at a rate that reflects the fundamental (100 Hz) of the signal. This is clearly shown by the dashed line envelope of the response. At the finest scale (bottom panel), the fast temporal modulations are due to the frequency component driving this channel best (around 750 Hz). (b) The auditory spectrogram of the note (B3) played on a violin. Again, note the modulations of the energy in time, especially at the higher CF channels (>1500 Hz).

in speech, or from one note or instrument to another in music (see next example).

5. Speech and music

Speech and music are an elaboration of harmonic or noise ripples in that they are conceptually constructed of a spectrotemporal envelope superimposed on a broadband noise or harmonic complex. Figure 4(a) shows the $y_{final}(t,x)$ responses in detail to the utterance /He drew a deep breath/ spoken by a male speaker. Figure 4(b) displays the responses to the B3 note played on a bowed violin. Both responses exhibit similar features to those of the ripple. For example, it is possible to see in Fig. 4(a) the three kinds of *temporal* modulations, as highlighted for one channel (at 750 Hz) in the three right panels. Here the slow modulations that reflect the syllabic rates of speech (top panel) are superimposed upon the intermediate rate modulations due to unresolved harmonics ($\approx$100 Hz) of the fundamental pitch (middle panel), which in turn are riding upon the phase-locked responses to the acoustic energy near 750 Hz (bottom panel). Also evident in the spectrograms are the *spectral* modulations created by the resolved harmonics (<500 Hz), and the second and third formants (>750 Hz). The same types of modulations are seen in the violin sound in Fig. 4(b). Note especially the slow modulations encoding the gradual onset of the note, and the *periodic* modulations at $\approx$6 Hz seen in most channels responses. As in speech, these slow features reflect primarily motor production mechanisms due to the fingering (vibrato) and bowing characteristics.

The distinction between these three types of temporal scales (fast, intermediate, and slow) is essentially identical to one already proposed by Stuart Rosen (Rosen, 1992). In an incisive article, he dissected the acoustic speech waveform into these three time scales and related them to the various auditory and production aspects just as described above. The one point to emphasize here is that the temporal scales defined here are made with respect to the channel responses *after* the early auditory analysis rather than the original acoustic waveform [or as Rosen calls it, the normal hearing case (Rosen, 1992)].

IV. THE CORTICAL STAGE: SPECTROTEMPORAL ANALYSIS

The second analysis stage mimics aspects of the responses of higher central auditory stages (especially the primary auditory cortex). Functionally, this stage estimates the spectral and temporal modulation content of the auditory spectrogram. It does so computationally via a bank of filters that are selective to different spectrotemporal modulation parameters that range from slow to fast *rates* temporally, and from narrow to broad *scales* spectrally. The spectrotemporal receptive fields (STRFs) of these filters are also centered at different frequencies along the tonotopic axis (Chi *et al.*, 1999).

An example of the STRF of such a filter in the bank is shown in Fig. 5(a). Three features are of particular interest: (*i*) it is centered on a particular center frequency (CF). The location of the excitatory (white) and inhibitory (black) stripes on the vertical axis indicates that it is sensitive to

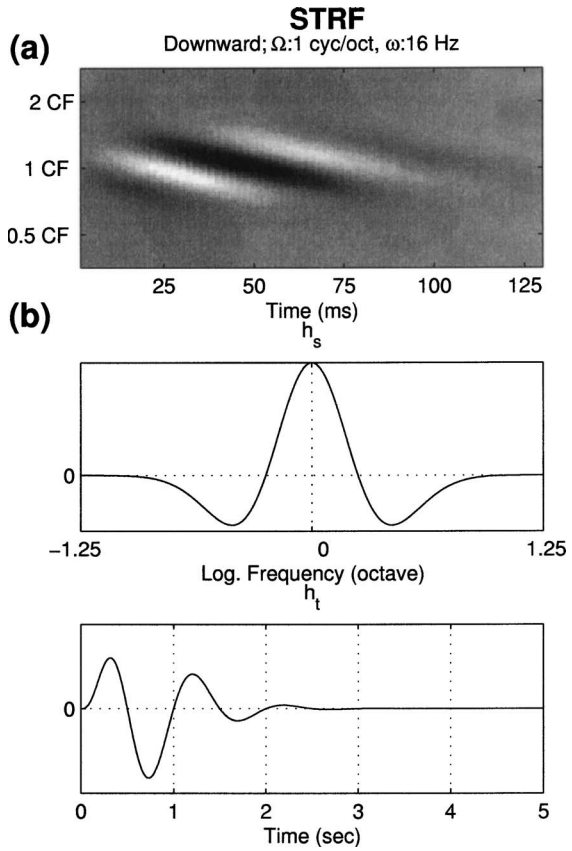

FIG. 5. A representative STRF and the seed functions of the spectrotemporal multiresolution cortical processing model. (a) An example of a STRF in the model. It is upward selective and tuned to (1 cycle/octave, 16 Hz). (b) Seed functions (noncausal h_s and causal h_t) used to generate all STRFs of the model. The abscissa of each figure is normalized to correspond to the tuning scale of 1 cycle/octave or rate of 1 Hz.

frequencies of about a 2-oct range around the CF (between about 0.5 CF and 2 CF); (*ii*) the modulation rate along the time axis is about 16 Hz; and (*iii*) the excitatory portions are separated on the vertical axis by about 1 oct, giving rise to a spectral "scale" sensitivity to peaks separated by 1 oct, or a scale of 1 cycle/octave. Finally, the bars sweep downwards diagonally from the top left, which is denoted in the model by assigning a positive sign to the rate parameter; bars sweeping up from bottom left to top right are designated by negative rate values. This distinction reflects the differential sensitivity of neurons in the auditory cortex to the direction in which spectral peaks move (Depireux *et al.*, 2001).

The filter output is computed by a convolution of its STRF with the input auditory spectrogram $[y_{final}(t,x)]$, i.e., it is a modified spectrogram. Note that the spectral and temporal cross sections of a filter's STRF are typical of a bandpass impulse response in having alternating excitatory (positive) and inhibitory (negative) fields. Consequently, the filter output is large only if the spectrotemporal modulations are commensurate with the rate, scale, and direction of the STRF. That is, each filter will respond best to a narrow range of these modulations. The output of the model consists of a map of the responses across the filter bank, with different stimuli being differentiated by which filters they activate best. The response map provides a unique characterization of the spectrogram, one that is sensitive to the spectral shape and dynamics of the entire stimulus. We now provide a mathematical formulation of the STRFs and procedures to compute, display, and interpret the model outputs.

A. Mathematical formulation

We assume a bank of "idealized" STRFs as depicted in Fig. 5(a). Each STRF is selective to a narrow range of temporal and spectral modulations and is also directionally sensitive to either upward or downward drifting modulations. A complete set of such STRFs with a range of temporal and spectral selectivity (e.g., 1–300 Hz, and 0.25–8 peaks or cycles/octave) would be sufficient to decompose and characterize the modulations in the auditory spectrogram. More realistic complex STRFs can be readily formed by superposition of these basic STRFs.

We define the STRF as a real function that is formed by combining two *complex* functions in a manner consistent with extensive physiological data. Specifically, experimental STRFs are not necessarily time-frequency separable. Instead, we have found that they are almost always so-called "quadrant separable."[3] This requires that the STRF be represented as the real of the product of a complex temporal and a complex spectral "impulse response" function, $h_{IRT}(t)$ and $h_{IRS}(x)$, as follows: $\mathrm{STRF} \equiv \mathcal{R}\{h_{IRT}(t) \cdot h_{IRS}(x)\}$, where

$$h_{IRS}(x;\Omega,\phi) = h_{irs}(x;\Omega,\phi) + j\hat{h}_{irs}(x;\Omega,\phi), \tag{5}$$

$$h_{IRT}(t;\omega,\theta) = h_{irt}(t;\omega,\theta) + j\hat{h}_{irt}(t;\omega,\theta). \tag{6}$$

$\mathcal{R}\{\cdot\}$ denotes the real part, and $h(\cdot)$ and $\hat{h}(\cdot)$ denote Hilbert transform pairs. The real functions $h_{irs}(\cdot)$ and $h_{irt}(\cdot)$ are defined by sinusoidally interpolating seed functions $h_s(\cdot)$, $h_t(\cdot)$ and their Hilbert transforms (Wang and Shamma, 1995),

$$h_{irs}(x;\Omega,\phi) = h_s(x;\Omega)\cos\phi + \hat{h}_s(x;\Omega)\sin\phi, \tag{7}$$

$$h_{irt}(t;\omega,\theta) = h_t(t;\omega)\cos\theta + \hat{h}_t(t;\omega)\sin\theta, \tag{8}$$

where Ω and ω are the spectral density and velocity parameters of the filters; ϕ and θ are characteristic phases; $h_s(\cdot)$ and $h_t(\cdot)$ are the spectral and temporal functions that determine the modulation selectivity of the STRF, and $\hat{h}_s(\cdot)$ and $\hat{h}_t(\cdot)$ are their Hilbert transforms. In addition, the directional sensitivity of the STRF is modeled as

$$\mathrm{STRF}_{\Downarrow} = \mathcal{R}\{h_{IRT}(t) \cdot h_{IRS}(x)\},$$

$$\mathrm{STRF}_{\Uparrow} = \mathcal{R}\{h^{*}_{IRT}(t) \cdot h_{IRS}(x)\},$$

where $*$ denotes the complex conjugate; $\Downarrow$ and $\Uparrow$ denote downward and upward moving direction respectively. Note, the downward STRF shown in Fig. 5(a) is a special case of $\theta = \phi = 0$.

We choose $h_s(\cdot)$ to be a Gabor-like function [commonly used in the vision literature to describe the analogous spatial aspect of a receptive field (Jones and Palmer, 1987)]. It is defined as the second derivative of a Gaussian function; $h_t(\cdot)$ is assumed to be a gamma function [e.g., as in Slaney (1998)]. Both are depicted in Fig. 5(b),

$$h_s(x) = (1 - x^2)e^{-x^2/2},$$

$$h_t(t) = t^2 e^{-3.5t}\sin(2\pi t),$$

and for different scales and rates,

$$h_s(x;\Omega) = \Omega h_s(\Omega x),$$

$$h_t(t;\omega) = \omega h_t(\omega t).$$

Therefore, the STRF in general is an inseparable spectrotemporal function of $h_s(\cdot)$ and $h_t(\cdot)$, with a specific highly constrained spectrotemporal structure known as "quadrant separable."

The spectrotemporal response of a downward (upward) cell c for an input spectrogram $y(t,s)$ is then given by

$$r_{c\Downarrow(\Uparrow)}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x)\otimes_{tx}\mathcal{R}\{[h_{IRT}^{(*)}(t;\omega_c,\theta_c)\cdot h_{IRS}(x;\Omega_c,\phi_c)]\}, \quad (9)$$

where $\otimes_{tx}$ denotes convolution with respect to both t and x. This multiscale multirate (or *multiresolution spectrotemporal*) response is called "cortical representation." Substituting Eqs. (5)–(8) into Eq. (9), the cortical representation at downward or upward cell c can be rewritten as

$$r_{c\Downarrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x)\otimes_{tx}[(h_t h_s - \hat{h}_t\hat{h}_s)\cos(\theta_c+\phi_c) + (\hat{h}_t h_s + h_t\hat{h}_s)\sin(\theta_c+\phi_c)] \quad (10)$$

and

$$r_{c\Uparrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x)\otimes_{tx}[(h_t h_s + \hat{h}_t\hat{h}_s)\cos(\theta_c-\phi_c) + (\hat{h}_t h_s - h_t\hat{h}_s)\sin(\theta_c-\phi_c)], \quad (11)$$

where $h_t \equiv h_t(t;\omega_c)$ and $h_s \equiv h_s(x;\Omega_c)$ to simplify notation.

A useful reformulation of the response r_c is in terms of the output *magnitude* and *phase* of a two-dimensional complex wavelet transform as follows. Let

$$z_{\Downarrow}(t,x;\omega_c,\Omega_c) = y(t,x)\otimes_{tx}[h_{TW}(t;\omega_c)h_{SW}(x;\Omega_c)] = |z_{\Downarrow}(t,x;\omega_c,\Omega_c)|e^{j\psi_{\Downarrow}(t,x;\omega_c,\Omega_c)}, \quad (12)$$

$$z_{\Uparrow}(t,x;\omega_c,\Omega_c) = y(t,x)\otimes_{tx}[h_{TW}^{*}(t;\omega_c)h_{SW}(x;\Omega_c)] = |z_{\Uparrow}(t,x;\omega_c,\Omega_c)|e^{j\psi_{\Uparrow}(t,x;\omega_c,\Omega_c)}, \quad (13)$$

with $h_{SW}(\cdot)$ and $h_{TW}(\cdot)$ defined as

$$h_{SW}(x;\Omega_c) = h_s(x;\Omega_c) + j\hat{h}_s(x;\Omega_c), \quad (14)$$

$$h_{TW}(t;\omega_c) = h_t(t;\omega_c) + j\hat{h}_t(t;\omega_c). \quad (15)$$

Substituting Eqs. (14) and (15) into Eqs. (12) and (13) and comparing with Eqs. (10) and (11), the cortical response at cell c can be simplified to

$$r_{c\Downarrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) = \mathcal{R}\{z_{\Downarrow}\}\cos(\theta_c+\phi_c) + \mathcal{I}\{z_{\Downarrow}\}\sin(\theta_c+\phi_c) = |z_{\Downarrow}|\cos(\psi_{\Downarrow}-\theta_c-\phi_c) \quad (16)$$

and

$$r_{c\Uparrow}(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) = \mathcal{R}\{z_{\Uparrow}\}\cos(\theta_c-\phi_c) - \mathcal{I}\{z_{\Uparrow}\}\sin(\theta_c-\phi_c) = |z_{\Uparrow}|\cos(\psi_{\Uparrow}+\theta_c-\phi_c) \quad (17)$$

where $z_{\Downarrow} \equiv z_{\Downarrow}(t,x;\omega_c,\Omega_c)$, $z_{\Uparrow} \equiv z_{\Uparrow}(t,x;\omega_c,\Omega_c)$, $\psi_{\Downarrow} \equiv \psi_{\Downarrow}(t,x;\omega_c,\Omega_c)$, and $\psi_{\Uparrow} \equiv \psi_{\Uparrow}(t,x;\omega_c,\Omega_c)$ for short notation; $\mathcal{R}\{\cdot\}$ and $\mathcal{I}\{\cdot\}$ denote the real part and imaginary part, respectively.

The expressions above show that the cortical model response r_c can be reexpressed in terms of magnitude responses $|z_{\Downarrow}|, |z_{\Uparrow}|$ and phase responses $\psi_{\Downarrow}, \psi_{\Uparrow}$, which are obtained by complex wavelet transform [Eqs. (12) and (13)]. Clearly, the magnitude responses $|z_{\Downarrow}(t,x;\omega_c,\Omega_c)|$ and $|z_{\Uparrow}(t,x;\omega_c,\Omega_c)|$ represent the maximal downward ($\psi_{\Downarrow} = \theta_c + \phi_c$) and upward ($\psi_{\Uparrow} = -\theta_c + \phi_c$) cortical responses at location $(t,x;\omega_c,\Omega_c)$.

B. Examples of cortical representations

Because of the multidimensionality of the cortical response r_c, displaying it in an intuitive manner is not trivial, requiring user judgment as to which dimensional views provide the best insights. We illustrate next a variety of such views for the stimuli discussed earlier in Sec. III.

1. Three tones

Figure 6(a) shows three particularly useful summary views of the cortical responses to the three-tone auditory spectrogram in Fig. 3(a). These three displays are generated by first integrating $|z_{\Downarrow}|, |z_{\Uparrow}|$ over their duration, i.e., removing their dependence on t and becoming three dimensional. Next, to generate each of the 2-D panels in Fig. 6(a), the remaining third variable is integrated out over its domain. For example, in the left panel, the dependence on scale (Ω_c) is removed by integrating all STRF outputs along this dimension, hence emphasizing the representation of temporal modulations (rate) at each CF. Since this stimulus is stationary (sustained tones), it evokes only very low rate outputs ($\omega_c \leqslant 4$ Hz) at each of the three tone frequencies. There is, however, a strong output at x and ω_c of 250 Hz due to the phase-locked responses of this tone [seen in the auditory spectrogram of the stimulus in Fig. 3(a)]; a weaker output due to phase locking is also seen at 1 kHz. Note also that both phase-locked responses are much stronger in the "Downward" panel of the display because of the traveling wave delay evident in the spectrograms of Fig. 3(a).

The center panel of Fig. 6(a) displays the output in the scale-frequency plane, integrating all filter outputs along the rate axis. STRFs with fine resolution relative to the intertone 2-oct spacing (i.e., tuned to $\Omega_c > 0.5$ cycle/octave) respond to each tone separately. STRFs with broad bandwidths ($\Omega_c < 0.5$ cycle/octave) smear the representation of the tones

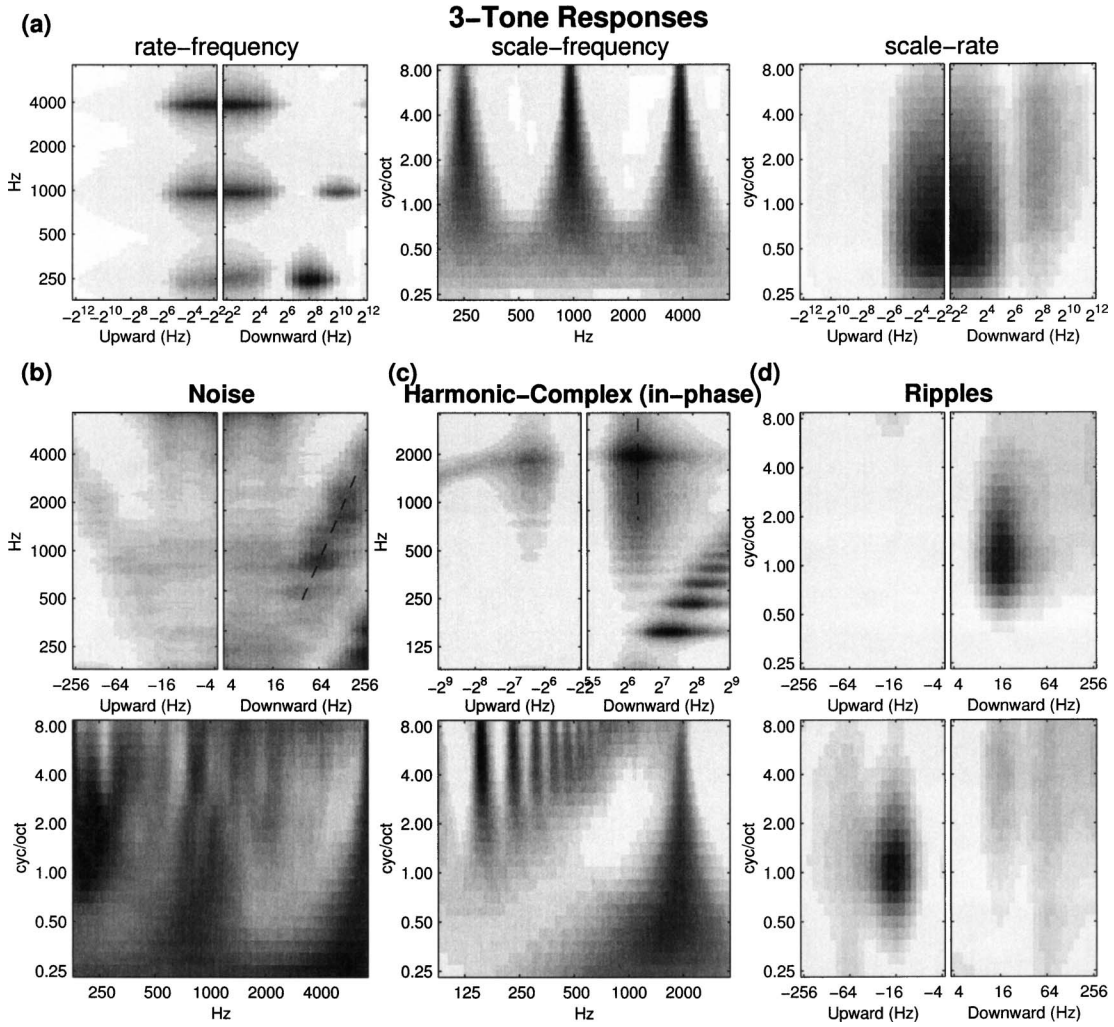

FIG. 6. Examples of cortical representations for stimuli as in Fig. 3: (a) a three-tone (250, 1000, 4000 Hz) combination, (b) a broadband noise, (c) broadband in-phase harmonic complex, and (d) ripples. For each of these stationary stimuli, the four-dimensional representation $|z_{\Downarrow}|$, $|z_{\Uparrow}|$ is first integrated over time to generate a three-dimensional representation. For the three-tone combinations, each of the remaining three variables (scale, rate, frequency) is integrated out over its domain to display these 2-D representations at left, center, and right panels of (a), respectively. For the broadband noise [in (b)] and in-phase harmonic complexes [in (c)], the top and bottom panels demonstrate the rate-frequency and scale-frequency cortical representations. The top and bottom panels of (d) show the scale-rate representations of a downward noise ripple (top) and an upward harmonic ripple (bottom), both modulated at 16 Hz, 1 cycle/octave. In each plot, the negative (positive) rate denotes upward (downward) moving direction.

into one broad peak. A "bifurcation" point emerges around the scale at which the peaks become resolved ($\Omega_c \approx 0.5$).

The right panel is particularly useful in summarizing the conjunction between the temporal and spectral modulations in a spectrogram. As expected, strong response can be seen at very low rate $\leqslant 4$ Hz and at 0.5 cycle/octave (since the tones are separated by 2 oct). A strong 250-Hz phase-locked response is also seen here but has been smeared out along the scale axis. Note, the frequency axis is integrated out, and hence the display is insensitive to pure translations of the spectrum along the x axis.

2. Noise and harmonic complexes

Like the tones, both stimuli here are stationary. However, the drastically different nature of their envelope modulations and underlying spectra creates distinctive cortical outputs as shown in Figs. 6(b) and 6(c).

The noise evokes a rate-frequency response [Fig. 6(b), top panel] which captures the increase in intermediate-rate temporal modulations with increasing CF (marked by the dashed line) due to the increasing bandwidth of the cochlear filters as discussed in Fig. 3(b) earlier. By contrast, the response to the harmonic stimulus [top panel of Fig. 6(c)] is dominated by the phase-locked responses to the resolved low-order harmonics, and all intermediate-rate modulations at high CF ($\geqslant 1000$ Hz) occur at a rate=80 Hz (marked by the dashed line). Finally, both panels exhibit larger energy in the "downward"-half of the plot due the accumulating phase lag of the cochlear filters (the well-known "traveling waves").

The scale-frequency panels (bottom panels) of Figs. 6(b) and 6(c) illustrate the contrast between the irregular versus regular nature of the two stimulus spectra. Note especially the distinctive and typical pattern associated with harmonic spectra in which "bifurcation" points shift systematically upwards, indicating the increasing crowding of the higher harmonics along the x axis.

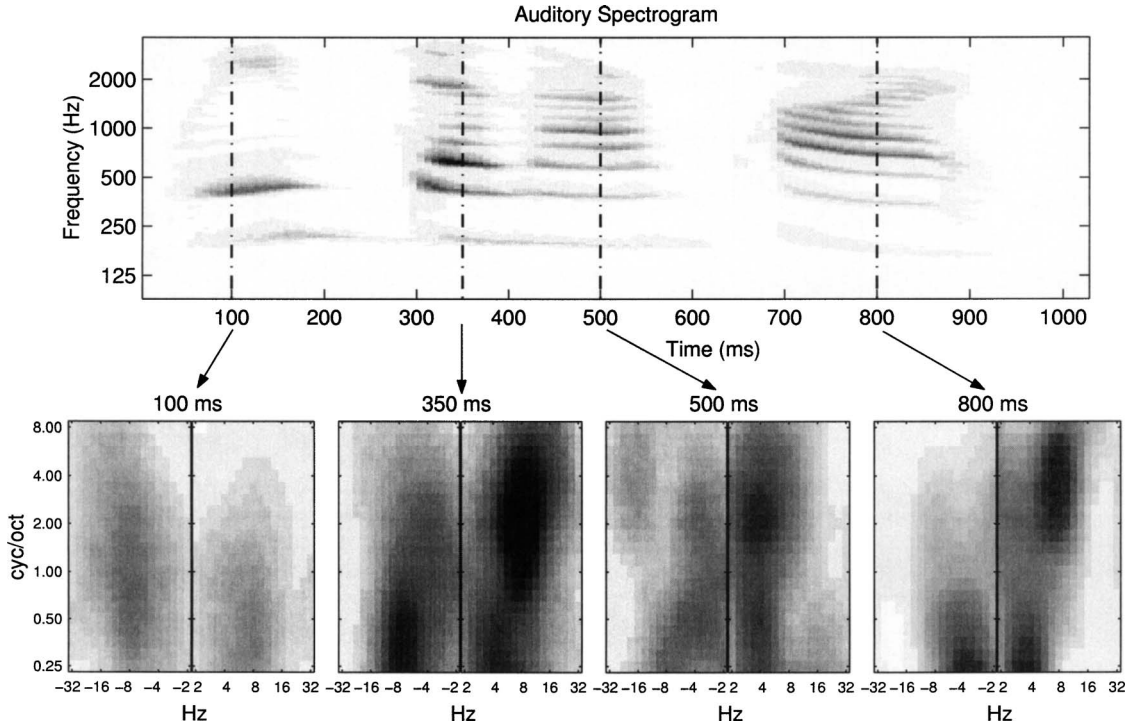

FIG. 7. The cortical multiresolution spectrotemporal representation of speech. The auditory spectrogram of the speech utterance /We've done our part/ spoken by a female speaker. The four bottom panels display scale-rate representation of the model output at the time instants marked by the vertical dashed lines in the auditory spectrogram. Each panel displays the spectrotemporal distribution of responses over the recent past (several 100 ms). For instance, the asymmetric responses at 350 ms reflect the downward shift in the pitch or frequency of all harmonics near the onset of the syllable (300 ms). They peak near 6–10 Hz because of the intersyllable time interval of about 120–180 ms (between the first and second syllables—/we've/ and /done/). They also peak at 2 cycle/octave because most of the spectral energy occurs near the second and third harmonics (which are separated by about 0.5 oct).

3. Ripples

Ripples with a single sinusoidal spectrotemporal modulation activate mostly STRFs with the corresponding selectivity. This is best illustrated by the localized response pattern in the scale-rate views of Fig. 6(d) due to a downward noise ripple (top panel) and an upward harmonic ripple (bottom panel). Regardless of the carrier, both ripples activate a localized response that captures the rate and scale of the slow modulations in the stimulus. Details of other views, however, would distinguish the two ripples from each other.

4. Speech and music

Speech and music are typically nonstationary, with spectrotemporal modulations that change their parameters. Consequently, it is often important to view the time evolution of the response patterns. Figure 7 illustrates one possible representation of the model outputs as a distribution of activity in the scale-rate plane as different phonemes and syllables are analyzed by the model. As before, these panels are computed by first integrating $|z_{\Downarrow}|, |z_{\Uparrow}|$ over frequency x, and then plotting the scale rate as a function of the third axis t.

These plots can uniquely summarize the salient features of the underlying spectrogram, and hence may potentially serve as efficient descriptors of the underlying speech segments. For instance, the downward-sweeping harmonic peaks near 350 and 800 ms generate strongly asymmetric patterns in the second and forth panels. The opposite symmetry is seen near 100 ms where the formant is sweeping upwards. Along the spectral dimensions, the main concentration of energy in the spectrogram shifts upwards from near 500 Hz at 100 ms (second harmonic) to 700 Hz at 350 ms (third harmonic), to near 1000 Hz at 800 ms (fourth and fifth harmonics). Consequently, the concentration of energy along the scale axis (illustrated in the lower series of panels) shifts upwards from near 1 cycle/octave (at 100 ms), to 2 cycles/octave (at 350 ms), to about 4 cycles/octave (at 800 ms). For further examples of such an analysis, please refer to Shamma (2003).

V. RECONSTRUCTION

We derive in this section computational procedures to resynthesize the original input stimulus from the output of early auditory and cortical stages. While the nonlinear operations in the early stage make it impossible to have perfect reconstruction, perceptually acceptable renditions are still feasible as we shall demonstrate. The ability to reconstruct the audio signal from the final representation is extremely useful in building the intuition of the role of different spectrotemporal cues in shaping the timbre percept as we shall elaborate in this section. Furthermore, it provides indirect measure of the fidelity and completeness of the representation as well as a potential means for manipulating timbre of musical instruments, morphing speech, and changing voice quality.

A. Reconstruction from auditory spectrogram

The most important component of the forward analysis stage—the *linear* filter bank operation [Eq. (1)]—is invert-

ible and the inverse operation can be derived as follows (Akansu and Haddad, 1992). From Eq. (1),

$$\begin{aligned} Y_{coch}(\omega,x) &= S(\omega)H(\omega;x) \\ &\Rightarrow \sum_x Y_{coch}(\omega,x)H^*(\omega;x) \\ &= S(\omega)\sum_x H(\omega;x)H^*(\omega;x) \Rightarrow S(\omega) \\ &= \sum_x Y_{coch}(\omega,x)H^*(\omega;x) \Big/ \sum_x |H(\omega;x)|^2, \end{aligned} \tag{18}$$

where $Y_{coch}(\omega,x)$, $S(\omega)$, and $H(\omega;x)$ are the Fourier transforms of $y_{coch}(t,x)$, $s(t)$, and $h(t;x)$ respectively. The overall response of the filter bank, $\Sigma_x|H(\omega;x)|^2$, is flat except at the lowest and highest frequency skirts where it drops precipitously, causing large noise and numerical errors in the inversion procedures. To avoid this problem, we shall simply ignore the response at these extreme frequencies and make the overall response unitary within the remaining band by introducing a real-valued weighting function $W(x)$:

$$H_1(\omega;x) = W(x)H(\omega;x)$$

such that

$$\sum_x |H(\omega;x)|^2 W(x) \simeq 1$$

within the effective band. Therefore, the time waveform $\tilde{s}(t)$ can be computed from the projected filter bank response $\tilde{y}_{coch}(t,x)$ [Eq. (18)]:

$$\begin{aligned} \tilde{S}(\omega) &= \sum_x \tilde{Y}_{coch}(\omega,x)H_1^*(\omega;x), \\ \tilde{s}(t) &= \sum_x \tilde{y}_{coch}(t,x) \otimes_t h_1^*(-t;x) = \sum_x \tilde{y}_{coch}(t,x) \otimes_t h_1(-t;x). \end{aligned} \tag{19}$$

The reconstruction from the envelope $y_{final}(t,x)$ back to $y_{coch}(t,x)$ is difficult to derive directly through the two nonlinear functions $g(\cdot)$ and $\max(\cdot,0)$. Instead, an iterative method based on the *convex projection* algorithm proposed in Yang *et al.* (1992) is used to reconstruct $s(t)$. The method is summarized in the following steps:

(1) Initialize a Gaussian distributed white noise with zero-mean and unit variance, i.e., $\tilde{s}^{(k)}(t) \sim \mathcal{N}(0,1)$, and set the iteration counter $k=1$.
(2) Compute $\tilde{y}_{coch}^{(k)}(t,x)$ and all the way to $\tilde{y}_{final}^{(k)}(t,x)$ with respect to $\tilde{s}^{(k)}(t)$.
(3) Find the ratio $r^{(k)}(t,x)$ between the target $y_{final}(t,x)$ and $\tilde{y}_{final}^{(k)}(t,x)$.
(4) Scale the filter-bank response, i.e., $\tilde{y}_{coch}^{(k)}(t,x) \leftarrow r^{(k)}(t,x) \times \tilde{y}_{coch}^{(k)}(t,x)$.
(5) Reconstruct time waveform $\tilde{s}^{(k+1)}(t)$ by inverse filtering [Eq. (19)], and update counter $k=k+1$.
(6) Go to step 2 unless certain criteria are met [e.g., the distortion rate of $\tilde{y}_{final}^{(k)}(t,x)$ or the number of iteration].

Note, the auditory spectrogram $y_{final}(t,x)$ is assumed roughly representing a local time-frequency (TF) energy distribution, and hence the estimated $\tilde{y}_{coch}(t,x)$ can be adjusted by the ratio of the target $y_{final}(t,x)$ divided by the computed spectrogram $\tilde{y}_{final}(t,x)$ pertaining to $\tilde{y}_{coch}(t,x)$. Figure 8 illustrates the similarity between original and reconstructed auditory spectrograms of two speech utterances after 100 iterations. Note that although this iterative algorithm does not give a unique reconstructed waveform because of the loss of the phase of the original components, the quality of reconstructed sounds using different initial conditions is very close and is reasonably similar to the original signal as can be heard at http://www.isr.umd.edu/CAAR/pubs.html. We shall discuss later in this section an objective assessment of the quality of this reconstructed speech using the mean opinion score (MOS) as quantified by the "perceptual evaluation of speech quality" (PESQ) index available from http://www.itu.int/ under "ITU Publications" (ITU-T, 2001).

B. Reconstruction from the cortical representation

The cortical stage is modeled by a bank of spectrotemporal filters which produce multiscale, multirate (or multiresolution) time-frequency cortical representations from an auditory spectrogram. This linear spectro-temporal filtering process is implemented by a two-dimensional complex wavelet transform [Eqs. (12), (13), (16), and (17)]. This stage is formally identical to the cochlear analysis stage [Eq. (1) versus Eq. (9)], and hence the one-dimensional inverse filtering technique [Eq. (18)] can be extended to solve the inverse problem of two-dimensional cortical filtering process.

The Fourier representations of Eqs. (12) and (13) can be written as

$$Z_{\Downarrow}(\omega,\Omega;\omega_c,\Omega_c) = Y(\omega,\Omega)H_{TW}(\omega;\omega_c)H_{SW}(\Omega;\Omega_c), \tag{20}$$

$$Z_{\Uparrow}(\omega,\Omega;\omega_c,\Omega_c) = Y(\omega,\Omega)H_{TW}^*(-\omega;\omega_c)H_{SW}(\Omega;\Omega_c), \tag{21}$$

and from Eqs. (14) and (15)

$$H_{SW}(\Omega;\Omega_c) = H_s(\Omega;\Omega_c)[1+\mathrm{sgn}(\Omega)], \tag{22}$$

$$H_{TW}(\omega;\omega_c) = H_t(\omega;\omega_c)[1+\mathrm{sgn}(\omega)], \tag{23}$$

where $H_s(\Omega;\Omega_c)$ and $H_t(\omega;\omega_c)$ are the Fourier transform of $h_s(x;\Omega_c)$ and $h_t(t;\omega_c)$, respectively, and

$$\mathrm{sgn}(A) = \begin{cases} 1, & A>0, \\ 0, & A=0, \\ -1, & A<0. \end{cases}$$

Therefore, reconstructing from the cortical representations back to auditory spectrogram is given by

$$\tilde{Y}(\omega,\Omega) = \frac{\Sigma_{\omega_c,\Omega_c} Z_{\Downarrow} H_{TW\Downarrow}^* H_{SW}^* + \Sigma_{\omega_c,\Omega_c} Z_{\Uparrow} H_{TW\Uparrow}^* H_{SW}^*}{\Sigma_{\omega_c,\Omega_c}|H_{TW\Downarrow}H_{SW}|^2 + \Sigma_{\omega_c,\Omega_c}|H_{TW\Uparrow}H_{SW}|^2}, \tag{24}$$

where $Z_{\Downarrow} \equiv Z_{\Downarrow}(\omega,\Omega;\omega_c,\Omega_c)$, $Z_{\Uparrow} \equiv Z_{\Uparrow}(\omega,\Omega;\omega_c,\Omega_c)$, $H_{TW\Downarrow} \equiv H_{TW}(\omega;\omega_c)$, $H_{TW\Uparrow} \equiv H_{TW}^*(-\omega;\omega_c)$, and $H_{SW} \equiv H_{SW}(\Omega;\Omega_c)$ for short notation. With similar considerations given to the lowest and highest frequencies of the overall two-dimensional transfer function, an excellent reconstruction

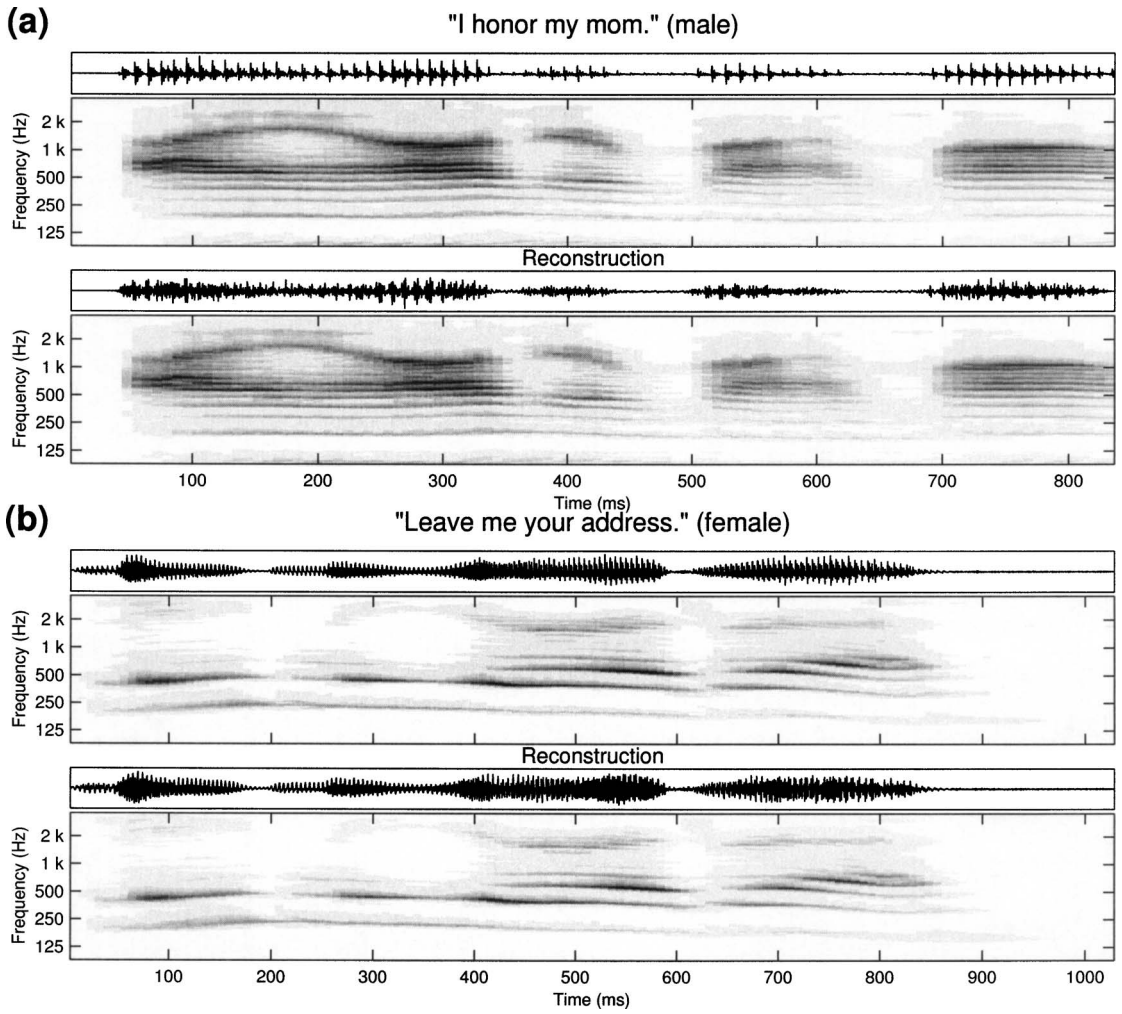

FIG. 8. Two examples of reconstructed acoustic waves from auditory spectrograms: (a) sentence /I honor my mom/ spoken by a male speaker and (b) sentence /Leave me your address/ spoken by a female speaker. The original speech signals are extracted from TIMIT corpus. In each example, the original time waveform $[s(t)]$, the target auditory spectrogram $[y_{final}(t,x)]$, the reconstructed time waveform $[\tilde{s}(t)]$, and the corresponding auditory spectrogram $[\tilde{y}_{final}(t,x)]$ are plotted from top to bottom panels.

within the effective band can be obtained. One example is shown in Fig. 9(b) with the rates up to 32 Hz and scales up to 8 cycles/octave used in the reconstruction. The reconstructed signals can be heard at http://www.isr.umd.edu/CAAR/pubs.html.

It is likely that temporal modulations faster than 20–40 Hz are encoded in the auditory cortex only by their energy distribution or envelope rather than by their actual phase-locked waveforms (Kowalski *et al.*, 1996; Lu *et al.*, 2001). Psychoacoustic experiments and previous models of temporal modulation sensitivity also support this conclusion (Dau *et al.*, 1997b; Sheft and Yost, 1990). Furthermore, in certain applications of the cortical model (Chi *et al.*, 1999), the output magnitude turns out to be an efficient and excellent indicator of the information and percepts of the stimulus. It is therefore useful to demonstrate that the "magnitude" of the response carries sufficient information about the stimulus that generated it. In the Appendix, two algorithms are proposed to reconstruct original speech from the modulation-energy-distributions $[|z_{\Downarrow}|$ and $|z_{\Uparrow}|$ in Eqs. (16) and (17)] only. While the "quality" of the reconstructed signals is worse due to a smaller dynamic range or to propagation of errors in the reconstruction procedures (see the Appendix), they are completely intelligible as can be heard on the website http://www.isr.umd.edu/CAAR/pubs.html.

C. Quality of the reconstructed speech signals

The multiscale auditory model (together with its reconstruction algorithms) can be essentially considered a "coding-decoding" system, and as such we can derive an objective assessment of the "quality" of the reconstructed speech by comparing it to the original clean samples using the standard perceptual evaluation of speech quality (PESQ) metric recommended by ITU (ITU-T, 2001). In this model-based method, we compare samples of clean speech signals to samples reconstructed from the auditory spectrogram and the full cortical representation (magnitude and phase included). The typical PESQ score obtained for the reconstruction from the auditory spectrogram is 4+ (toll quality). For instance, the average of 50 reconstructions of the sentence /Come home right away/ (Fig. 9) starting from different initial conditions and after 200 iterations is 4.04 with $\sigma = 0.075$. The average PESQ score for the reconstruction from

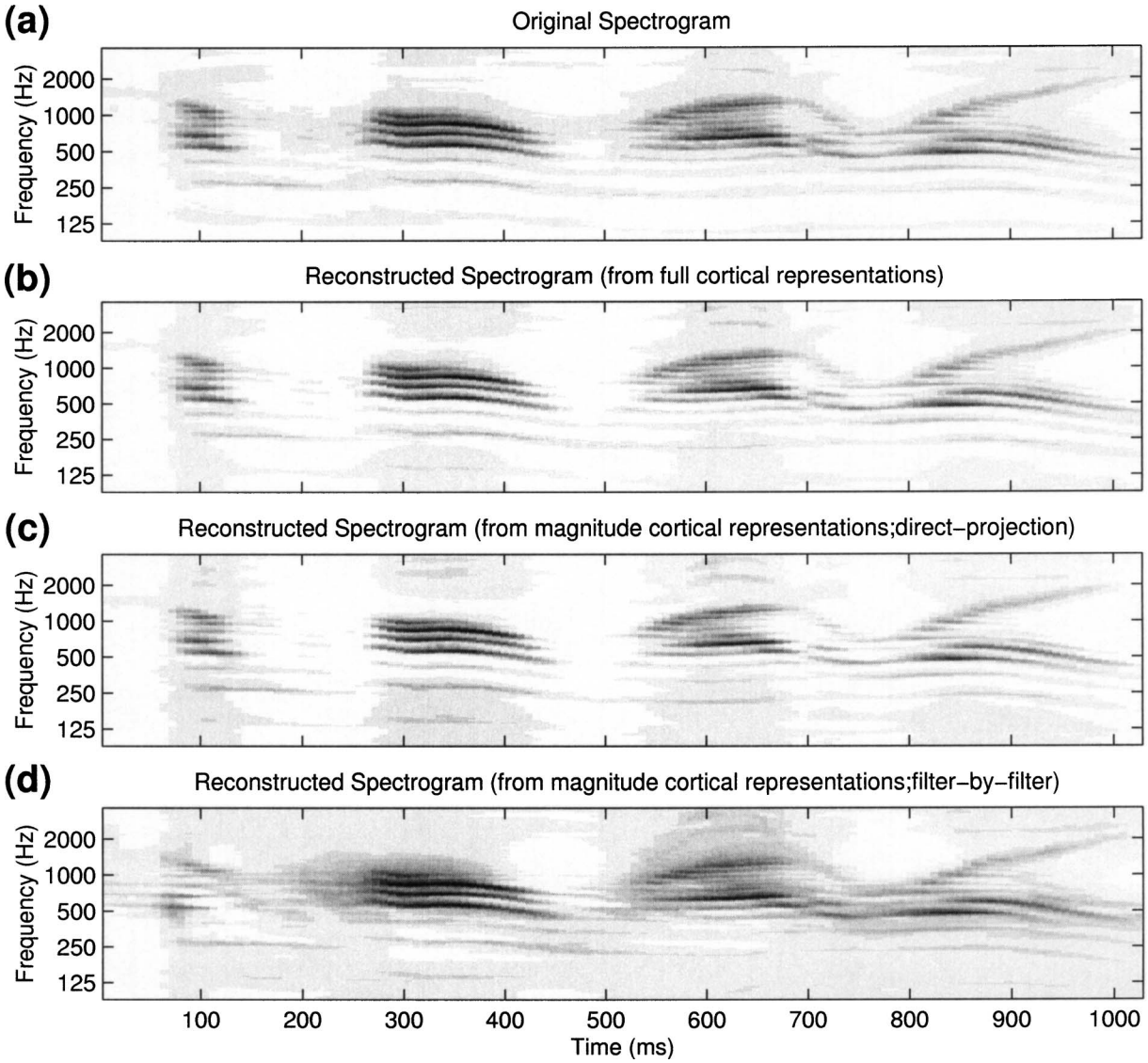

FIG. 9. Examples of reconstructed spectrograms. The top panel shows the original spectrogram of sentence /Come home right away/ spoken by a male speaker. The reconstructed spectrograms from *full* cortical representations and *magnitude* cortical representations (direct-projection and filter-by-filter algorithms) are demonstrated on the second to bottom panel, respectively. All spectrograms are reconstructed from those cortical representations which only include modulation rates up to 32 Hz.

the full cortical representation (with rates up to 128 Hz and scales up to 8 cycles/octave) is 4.02 (toll quality) with $\sigma = 0.069$.

D. Intelligibility of the reconstructed signals

To demonstrate the utility of the reconstructed speech signals from the model, we explore the assertions we made earlier in the Introduction regarding the critical role played by the slow spectrotemporal envelope modulations in preserving intelligibility of the speech signal. Specifically, we use the model to reconstruct a speech sentence after removing from its original version progressively more of its temporal and spectral modulations. We assess in psychoacoustic tests the perceptual effect of such manipulations, and compare the results to the spectrotemporal modulation index (STMI), a measure that was previously demonstrated to be a reliable correlate of human perception of speech intelligibility under a wide variety of interference signals and conditions (Elhilali *et al.*, 2003). We shall specifically employ a particular version of the STMI denoted by STMI^{T} (Elhilali *et al.*, 2003), where the superscript "T" refers to the use of a clean speech signal as the "template" to be compared to each of the "modulation reduced" (or distorted) versions reconstructed from the model.

We first compute the multiscale representation of the clean speech signal through the model [as in Eqs. (16) and (17), $\forall c$]. Temporal modulations are then filtered out by nulling the outputs of the undesired filters (parametrized by their center modulation rates ω_c and Ω_c). This "filtered" representation is then inverted to reconstruct the corresponding "modulation reduced" acoustic signal (as explained in Sec. V B). Figure 10(a) shows the STMI^{T} of the reconstructed speech as a function of the upper limit of *temporal* modulation rates (dashed line). Rates along the abscissa refer to the ω_c's of the cortical filters that are nulled in the STMI^{T} computations. Since the filters are fairly broad, these rates are

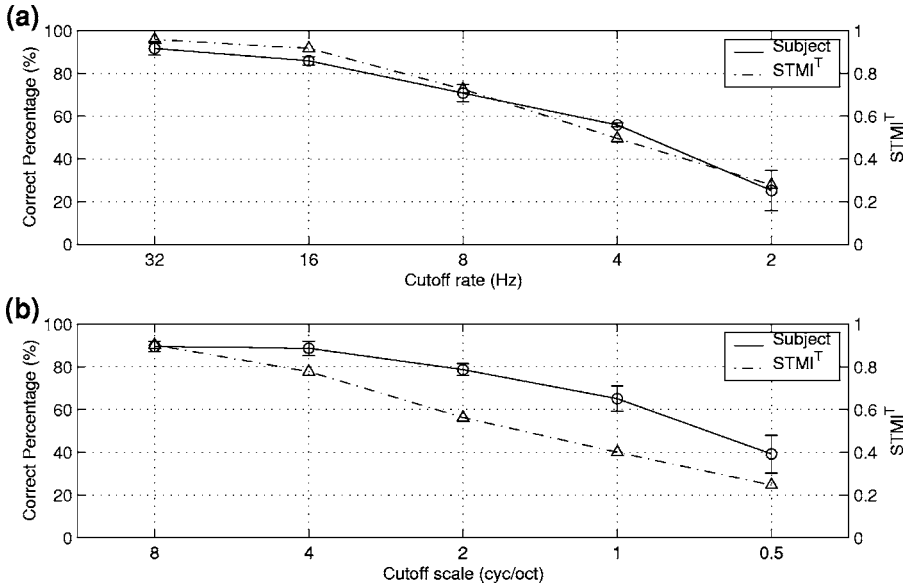

FIG. 10. The spectro-temporal modulation index ($STMI^T$) (Elhilali *et al.*, 2003) of reconstructed speech as a function of the range of spectral and temporal modulations preserved in the signal. (a) The $STMI^T$ (dashed line) and the experimental measurements of the correct phoneme recognition percentage of human subjects (solid line) as a function of the range of temporal modulations preserved. (b) The $STMI^T$ (dashed line) and the human performance (solid line) as function of the scales preserved.

gradual. Each value of the $STMI^T$ shown in the plot is the average of 20 sentences (a mix of males and females) extracted from the TIMIT corpus (the training portion of the New England dialect region). It is evident that intelligibility becomes marginal when temporal modulations around 4 Hz are filtered out, consistent with numerous previous experimental results (Elhilali *et al.*, 2003; Drullman *et al.*, 1994). These results are consistent with the average intelligibility scores measured with four native speakers. In these tests, each subject was to identify 300 reconstructed CVC word samples [see Elhilali *et al.* (2003) for experiment details]. The average percentage "correct phonemes" and the error bars with one standard deviation ranges are plotted in Fig. 10(a).

Figure 10(b) illustrates the $STMI^T$ and intelligibility scores obtained when the spectral profiles of the speech sentence are smoothed by removing progressively higher scales. While the $STMI^T$ and subjects' performance deviate from each other, the overall results confirm that the loss of spectrally sharp features diminishes intelligibility gradually beginning when the filters are effectively wider than about the critical bandwidth (3 cycles/octave). Some intelligibility remains even with filters as broad as 0.5–1 cycles/octave (or about an octave), consistent with previous experimental findings (Shannon *et al.*, 1995).

VI. DISCUSSION

We presented a model of auditory processing that transforms an acoustic signal into a multiresolution spectrotemporal representation inspired by experimental findings from the auditory cortex. The model consists of two major transformations of the acoustic signal:

(1) A frequency analysis stage associated with the cochlea, cochlear nucleus, and response features observed in the midbrain: This stage effectively computes an affine wavelet transform of the acoustic signal with a spectral resolution of about 10% (Lyon and Shamma, 1996).

(2) A spectrotemporal multiresolution analysis stage postulated to conclude in the primary auditory cortex: This stage effectively computes a two-dimensional affine wavelet transform with a Gabor-like spectrotemporal mother-wavelet [see Fig. 5(b)].

The model is intended to be a computational realization of the most basic aspects of auditory processing and not a biophysical description of its stages. Hence, there is only a loose correspondence between any specific structure and model parameters. However, we hypothesize that the model final representation of the acoustic signal captures explicitly and quantitatively the spectral and dynamic aspects that are directly perceived by a listener. Consequently, this representation may be utilized to account for a variety of phenomena, especially those related to the perception of timbre, such as in the assessment of speech quality and intelligibility (Elhilali *et al.*, 2003; Chi *et al.*, 1999), discrimination of musical timbre (Ru and Shamma, 1997), and, more generally, quantifying the perception of complex sounds subjected to arbitrary spectral and temporal changes (Carlyon and Shamma, 2003).

The spirit of this model shares much with others that have been proposed to quantify the perceptual relevance of temporal modulations in acoustic signals (Dau *et al.*, 1997a; Sheft and Yost, 1990; Houtgast, 1989; Bacon and Grantham, 1989; Viemeister, 1979). Dau and colleagues developed the most detailed of these models, consisting of a bank of purely temporal modulation selective filters. They also established its parameters and perceptual relevance in a series of extensive psychoacoustic experiments (Dau *et al.*, 1997a b). Our model is consistent with Dau's model in the details of its analysis of temporal modulations, e.g., possessing similar filter bandwidths in the modulation filterbank ($Q_{3dB}=1.8$ versus $Q_{3dB}=2$ in Dau's model). The two models fundamentally differ in the way temporal modulations from different spectral channels are integrated at the end. Dau's model is *fully separable*, integrating spectral information subsequent to an independent temporal analysis. By contrast the multiscale

cortical model is inseparable (but see footnote 3), postulating a "spectral" modulation filterbank that is fully integrated with the temporal modulation analysis. Under circumstances where *both* temporal and spectral features of the input spectrograms are manipulated [e.g., as in phase jitter or phase shift distortions described in Elhilali *et al.* (2003)], the two models respond differently.

A. Variations on the cortical model

As with the early auditory stage, the multiresolution cortical model is highly schematic and lacks realistic biophysical mechanisms and parameters. Nevertheless, the model aims to capture perceptually significant features in the auditory spectrogram, and hence justify its relevance through its successful application in accounting for a variety of perceptual thresholds and tasks as we have described above.

Many details of the model are somewhat arbitrary and can be probably modified to reflect future physiological and anatomical findings with no significant effect on the computations. For example, real cortical STRFs (Fig. 1) are far more complex than the simple Gabor-like shapes we have employed in the model. They are often tuned to multiple frequencies and are rarely purely selective to upward or downward frequency sweeps but rather are simply more responsive to one direction or the other. In many situations, these differences are not crucial as long as important spectrogram features (e.g., FM sweeps and AM modulations) are still encoded explicitly albeit in a different form.

One potentially interesting variation on our model is to split the spectrotemporal modulation analysis into two stages. The first would be a relatively fast bank of filters mimicking the temporal analysis hypothesized to exist in the inferior colliculus (Langner and Schreiner, 1988) (rates of 30–1000 Hz). The second stage would be slower filters ($\leqslant$30 Hz) operating on *each* output from the earlier stage. This latter stage would then capture all the important slow modulations of the spectrogram explicitly, whereas the earlier stage extracts the intermediate and fast modulations of the auditory spectrogram. The natural split between the dynamic factors involved in intelligibility (the slow rates found in the cortex) from those involved in sound quality (intermediate rates found precortically) becomes particularly advantageous when considering phenomena that contrast these two rate domains such as the streaming of two sounds based purely on their modulation rates (Roberts *et al.*, 2002; Grimault *et al.*, 2002).

B. Relation to previous reconstruction algorithms

The multiresolution representation and associated reconstruction algorithms presented here differ from previous methods for processing spectral and temporal envelopes in two ways. First, its formulation *combines* the spectral and temporal dimensions compared to the purely spectral (e.g., ter Keurs *et al.*, 1992; Baer and Moore, 1993), purely temporal (e.g., Drullman *et al.*, 1994), or a separable cascade of the two (e.g., Dau *et al.*, 1997b). Second, our reconstruction algorithm starts from a random noise signal without any prior information about the original speech. By contrast, previous experiments usually retained the carrier waveform of the speech in each frequency band (Drullman *et al.*, 1994) or the harmonic structure of the speech in each frame (ter Keurs *et al.*, 1992; Baer and Moore, 1993) and used them to resynthesize the filtered speech by superimposing the newly processed envelopes upon them. These carriers improve the quality of the reconstructed speech, but may contain residual intelligible information (Ghitza, 2001; Smith *et al.*, 2002).

Our algorithms are similar in spirit to Slaney's inversion algorithm (Slaney *et al.*, 1994), which also employs the iterative projection method and disposes of the fine structure in reconstructing the stimulus. The algorithm, however, differs fundamentally in all of its details in that it uses for its two-stage representation the cochleagram from a simpler Gammatone filter bank cochlear model (as opposed to the *early stage*) and the correlogram (as opposed to the *cortical multiscale* representation). Consequently, all the constraints imposed during the iterations are completely different.

C. Applications of the multiscale auditory model

The validity of the auditory model stems from its ability to account for psychoacoustic findings and from its successful application in a variety of perceptual tasks. To this end, we have recently adapted and tested the auditory model in several very different contexts. In the first, the auditory model was used to account for the detection of phase of complex sounds such as phase differences between the envelopes of sounds occupying remote frequency regions, and between the fine structures of partials that interact within a single auditory filter (Carlyon and Shamma, 2003). The approach was simply to interpret the discrimination between two stimuli as being proportional to the distance (or difference) measured between their cortical representation in the model (Tchorz and Kollmeier, 1999). Discriminations successfully accounted for phase differences between pairs of bandpass filtered harmonic complexes, and between pairs of sinusoidally amplitude modulated tones, discrimination between amplitude and frequency modulation, and discrimination of transient signals differing only in their phase spectra ("Huffman sequences") (Carlyon and Shamma, 2003).

In a second application, we used the model to analyze the effects of noise, reverberations, and other distortions on the joint spectrotemporal modulations present in speech, and on the ability of a channel to transmit these modulations (Chi *et al.*, 1999; Elhilali *et al.*, 2003). The rationale behind this approach is that the perception of speech is critically dependent on the faithful representation of spectral and temporal modulations in the auditory spectrogram (Hermansky and Morgan, 1994; Drullman *et al.*, 1994; Shannon *et al.*, 1995; Arai *et al.*, 1996; Dau *et al.*, 1996; Greenberg *et al.*, 1998). Therefore, an intelligibility index which reflects the integrity of these modulations can be effective regardless of the source of the degradation. Such a spectrotemporal modulation index (STMI) was derived using the model representation of speech modulations and was validated by comparing its predictions of intelligibility to those of the classical *speech transmission index (STI)* and to error rates reported by human subjects listening to speech contaminated with

combined noise and reverberation. We further demonstrated that the STMI can handle difficult and nonlinear distortions such as phase jitter and shifts, to which the STI is not sensitive (Elhilali *et al.*, 2003).

In another application, the auditory model was used to discriminate speech from nonspeech signals (Mesgarani *et al.*, 2004), a relatively easy task for humans but one that has been very difficult to reliably automate. The proposed algorithm was largely based on learning a template of the unique representation of speech spectrotemporal modulations, a strategy that proved quite effective when compared to state-of-art alternatives. In a further recent extension of this application, it was possible to use the auditory model as a "filter" to remove "noise" modulations that lie outside of the range typical of speech (Mesgarani and Shamma, 2005). Subsequent reconstruction of the filtered signal demonstrated significant enhancement in sound quality.

VII. SUMMARY AND CONCLUSIONS

An auditory model inspired by existing psychophysical and physiological evidence is described. The first module mimics early auditory processing; it consists of a bank of constant-Q bandpass filters, followed by nonlinear compression and derivative across scale (frequency resolution sharpening) mechanisms, and ending with an envelope detector at each frequency band. The resulting output is an estimate of the spectrogram of the input stimulus with noise-robust and feature-enhanced properties (Wang and Shamma, 1994). The second module further analyzes the auditory spectrogram by a bank of linear spectro-temporal modulation filters, which effectively perform a two-dimensional complex wavelet transform. The result is a multiresolution representation which combines information about the temporal and spectral modulations and their distribution in time and frequency.

Several reconstruction algorithms adapted from convex projection methods are proposed to resynthesize the acoustic signals from the full or just the envelope of the auditory spectrogram and the multiresolution representation. The resynthesized sounds imply that these representations carry information critical to the perception of the timbre and the intelligibility of the sound.

To validate our model, the output representations of the model have been adapted for several applications and show promising results when used to measure the perceptual distance between two sounds (Carlyon and Shamma, 2003) or to assess the intelligibility of speech with various types of linear and nonlinear distortions (Elhilali *et al.*, 2003). In addition, we believe this model can be served as a preprocessor to segregate different auditory cues for sound grouping or streaming applications associated with the field of auditory scene analysis.

The proposed model has been implemented in a MATLAB environment, with a variety of computational and graphical modules to allow the user the flexibility of constructing any appropriate sequence of operations. The package also contains demos and help files for users, together with default parameter settings, making it easy learn for the new user. This software is available for download through our website at http://www.isr.umd.edu/CAAR/under "Publications."

APPENDIX: RECONSTRUCTION FROM MAGNITUDE CORTICAL REPRESENTATION

This restoration-from-magnitude problem (also called the phase retrieval problem) is encountered in many fields (Hayes, 1982; Fienup and Wackerman, 1987). Several approaches have been proposed in the past, including a generalized iterative projection algorithm to solve two-dimensional image restoration problems (Levi and Stark, 1984), reconstructing speech from auditory wavelet transform (Irino and Kawahara, 1993), and the error-reduction and extrapolation algorithms (Gerchberg and Saxton, 1972; Fienup, 1982; Papoulis, 1975). All these algorithms essentially perform iterative Fourier and inverse Fourier transforms between the object and Fourier domain, applying specific constraints in each domain. Mathematical convergence of these iterations is not generally guaranteed (Bates, 1984; Hayes, 1987; Seldin and Fienup, 1990). However, combining different algorithms improves the probability of convergence (Fienup, 1982; Mou-yan and Unbehauen, 1997).

In our case, there are no prescribed magnitude constraints in the Fourier domain (ω-Ω domain). Instead, the input and output (envelope) constraints are in the same time-frequency domain [see Eqs. (12) and (13)]. In general, complex signals (such as $z_{\Downarrow}$ and $z_{\Uparrow}$) cannot be uniquely determined from their modulus ($|z_{\Downarrow}|$ and $|z_{\Uparrow}|$) without additional information. Although the *analytical* form of the cortical filters [Eqs. (14) and (15)] narrows down the range of possible phases to be assigned to a given modulus, the lack of additional constraints about the locations of the poles or zeros of the cortical filters precludes a unique solution to our phase retrieval problem (Hayes *et al.*, 1980). The two algorithms proposed below are iterative and are inspired by traditional phase-retrieval and convex projection algorithms. Although the set of magnitude constraints is not convex, the proposed projections are generalized in the sense of Levi-Stark (Levi and Stark, 1983, 1984) and equivalent to the Gerchberg-Saxton algorithm with error-reduction property (Fienup, 1982). Detailed mathematical descriptions of the proposed projection operators can be found in Chi (2003).

1. Algorithm I: Direct projection

The first algorithm considers magnitude constraints of all filters ($|z_{\Downarrow}(t,x;\omega_c,\Omega_c)|$ and $|z_{\Uparrow}(t,x;\omega_c,\Omega_c)|$, $\forall c$) at the same time. It can be summarized as

(1) Initialize a *non-negative* auditory spectrogram $\tilde{y}^{(k)}(t,x)$ randomly and set the iteration counter $k=1$.
(2) Compute magnitude and phase cortical representations $|\tilde{z}_{\Downarrow}^{(k)}|$, $|\tilde{z}_{\Uparrow}^{(k)}|$, $\tilde{\psi}_{\Downarrow}^{(k)}$ and $\tilde{\psi}_{\Uparrow}^{(k)}$ associated with $\tilde{y}^{(k)}(t,x)$ by cortical filtering process [Eqs. (12) and (13)].
(3) Modify cortical representations by keeping phase $\tilde{\psi}_{\Downarrow}^{(k)}$ and $\tilde{\psi}_{\Uparrow}^{(k)}$ intact but replacing magnitude $|\tilde{z}_{\Downarrow}^{(k)}|$ and $|\tilde{z}_{\Uparrow}^{(k)}|$ with the prescribed magnitude responses $|z_{\Downarrow}|$ and $|z_{\Uparrow}|$ (constraints on the cortical output).

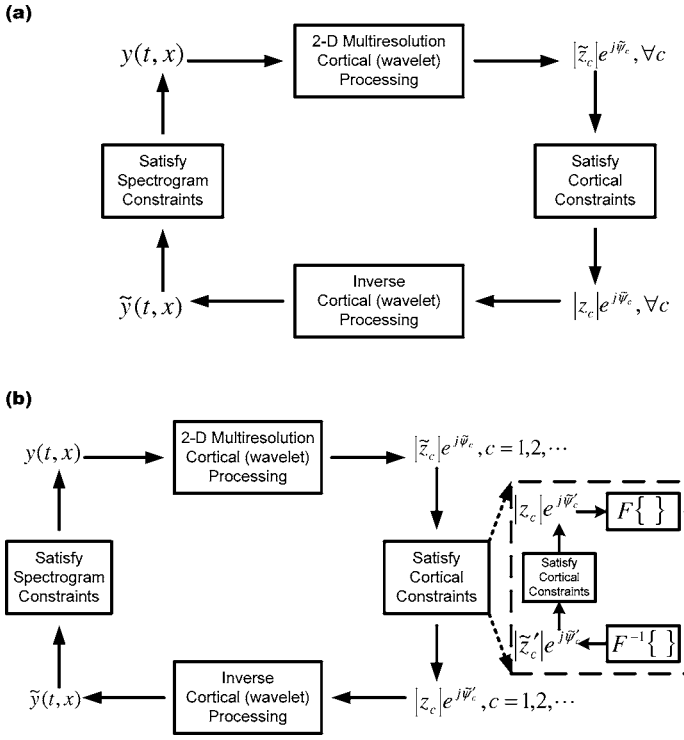

FIG. 11. Block diagrams of two proposed algorithms that reconstruct the spectrograms from magnitude cortical representation. (a) Direct projection algorithm; (b) Filter-by-filter algorithm.

(4) Synthesize $\tilde{y}^{(k+1)}(t,x)$ from modified cortical representations ($|z_{\Downarrow}|$, $|z_{\Uparrow}|$, $\tilde{\psi}_{\Downarrow}^{(k)}$, and $\tilde{\psi}_{\Uparrow}^{(k)}$) by inverse cortical filtering [Eq. (24)].
(5) Half-wave rectify $\tilde{y}^{(k+1)}(t,x)$ (constraints on the cortical input) and update counter $k=k+1$.

Repetitive application of step 2 to step 5 defines the iteration which is depicted in Fig. 11(a). This algorithm has been shown an implementation of the gradient descent search method in solving the nonlinear reconstruction problem (Chi, 2003).

2. Algorithm II: Filter-by-filter

The cortical filters are highly overlapped in both ω and Ω domains, therefore, the magnitude constraints of adjacent filters are redundant. Consequently, Algorithm I yields accurate reconstruction when it converges, but with very high computational cost. Here, taking the analytical form implementation of the cortical filters into account shall reduce the computational load dramatically.

Observed from Eqs. (20)–(23), $Z_{\Downarrow}(\omega,\Omega;\omega_c,\Omega_c)$ and $Z_{\Uparrow}(\omega,\Omega;\omega_c,\Omega_c)$ only have nonzero elements in the first and second quadrants of the (ω,Ω) space, respectively. With these additional implicit constraints and the fact that the frequency responses of the adjacent cortical filters are highly overlapped, the second algorithm is proposed as follows:

(1) Initialize a non-negative auditory spectrogram $\tilde{y}_{(i)}(t,x)$ randomly and set the filter indicator $i=1$.
(2) Compute cortical representations $|\tilde{z}_{\Downarrow}^{(1)}(i)|$, $|\tilde{z}_{\Uparrow}^{(1)}(i)|$, $\tilde{\psi}_{\Downarrow}^{(1)}(i)$ and $\tilde{\psi}_{\Uparrow}^{(1)}(i)$ of filter i, which has the lowest characteristic BF (ω_i,Ω_i) with coverage of DC response $(i=1)$. Here, $|\tilde{z}^{(1)}(i)|$ and $\tilde{\psi}^{(1)}(i)$ are short notations for $|\tilde{z}^{(1)}(t,x;\omega_i,\Omega_i)|$ and $\tilde{\psi}^{(1)}(t,x;\omega_i,\Omega_i)$.
(3) Set iteration counter $k=1$.
   (a) Replace $|\tilde{z}_{\Downarrow}^{(k)}(i)|$, $|\tilde{z}_{\Uparrow}^{(k)}(i)|$ with prescribed $|z_{\Downarrow}(i)|$, $|z_{\Uparrow}(i)|$ and compute $\tilde{Z}_{\Downarrow}^{(k)}(i)$, $\tilde{Z}_{\Uparrow}^{(k)}(i)$ by two-dimensional Fourier transforming $|z_{\Downarrow}(i)|$, $|z_{\Uparrow}(i)|$, $\tilde{\psi}_{\Downarrow}^{(k)}(i)$, and $\tilde{\psi}_{\Uparrow}^{(k)}(i)$.
   (b) Modify $\tilde{Z}_{\Downarrow}^{(k)}(i)$ and $\tilde{Z}_{\Uparrow}^{(k)}(i)$ by keeping the first- and second-quadrant components intact, respectively, and resetting all components in the other quadrants to zero.
   (c) Compute $|\tilde{z}_{\Downarrow}^{(k+1)}(i)|$, $|\tilde{z}_{\Uparrow}^{(k+1)}(i)|$, $\tilde{\psi}_{\Downarrow}^{(k+1)}(i)$, and $\tilde{\psi}_{\Uparrow}^{(k+1)}(i)$ by two-dimensional inverse Fourier transforming modified $\tilde{Z}_{\Downarrow}^{(k)}(i)$ and $\tilde{Z}_{\Uparrow}^{(k)}(i)$.
   (d) Update counter $k=k+1$; go to step 3 (a) when $k<N_i$ (predetermined number of iterations).
(4) Compute $\tilde{y}_{(i+1)}(t,x)$ by Eq. (24) from cortical responses up to filter $i(\tilde{Z}^{(N_i)}(1),\ldots,\tilde{Z}^{(N_i)}(i))$ and half-rectify it (constraint on the cortical input).
(5) Estimate cortical representations ($|\tilde{z}_{\Downarrow}^{(1)}(i+1)|$, $|\tilde{z}_{\Uparrow}^{(1)}(i+1)|$, $\tilde{\psi}_{\Downarrow}^{(1)}(i+1)$ and $\tilde{\psi}_{\Uparrow}^{(1)}(i+1)$) for adjacent filter $i+1$ by cortical forward filtering process [Eqs. (12) and (13)] when $i<N_f$ (number of filters).
(6) Go to step 3 and update filter indicator $i=i+1$.

Note, for each filter i, the starting pattern [initial estimate $\tilde{z}^{(1)}(i)$] shall strongly affect the fidelity of the reconstruction since the generalized projection algorithms do not guarantee a unique solution for nonconvex sets. The block diagram of this filter-by-filter algorithm is depicted in Fig. 11(b).

3. Comparing Algorithms I and II

Algorithm II resolves constraints of one filter at a time (step 3) and thus consumes much less computational time

than Algorithm I. The initial phase for filter i [$\tilde{\psi}_{\Downarrow}^{(1)}(i)$ and $\tilde{\psi}_{\Uparrow}^{(1)}(i)$ in step 3] is estimated recursively from the reconstruction result of previous $i-1$ filters (step 5). This is justified by the assumption that cortical filters have highly overlapped frequency responses, and hence the output phases of one filter and adjacent filters do not change rapidly. However, the overall performance primarily depends on the reconstruction result from the first filter because the errors propagate and are magnified through the iterations. The reconstructed spectrograms from both algorithms are plotted in the bottom two panels of Fig. 9. The processing time of Algorithm I (the third panel from top; 100 iterations) is 150 times longer than the processing time of Algorithm II (bottom panel; $N_i=10$ for each filter). Note, the reconstructed spectrogram at bottom panel shows a smaller dynamic range with apparent distortions near onsets, offsets, lower harmonics, and other weak features in the original spectrogram.

A hybrid algorithm can be used to balance the disadvantages of proposed algorithms, i.e., high computational load and propagation of errors. For example, the output after several iterations of the first direct-projection algorithm is a much better starting pattern than the random pattern to initialize the second algorithm for all filters.

The STMI$^{\mathrm{T}}$s of the reconstructed speech (second to bottom panel) in Fig. 9 are 0.97, 0.97, and 0.91, respectively. These scores indicate that all reconstruction algorithms preserve the slow temporal modulations very well, as can be seen in the figure. However, the quality of the reconstructed sounds is not very good due to distortions as discussed above.

[1]Cortical cells may respond to transient stimuli with high precision (<1 ms), and at times phase lock to high rates exceeding 200 Hz for short intervals. These response patterns reflect the influence of complex mechanisms such as synaptic depression and feedforward inhibition that give rise to the cortical "slow down" in the first place. For details of these phenomena in the auditory cortex, see Elhilali *et al.* (2004).

[2]The cochlear filter is implemented by a minimum-phase signal $h(t)$ with magnitude frequency response

$$|H(x)| = \begin{cases} (x_h - x)^{\alpha} e^{-\beta(x_h - x)}, & 0 \leq x \leq x_h, \\ 0, & x > x_h, \end{cases}$$

where x_h is the cutoff frequency, $\alpha=0.3$, and $\beta=8$. Details of cochlear filter implementations can be found in Ru (2000).

[3]Quadrant-separability implies that in the 2D Fourier transform of the STRF, the temporal and spectral transfer functions are required to be separable only *within each quadrant* (not necessarily across quadrants) (Watson and Ahumada, 1985). This property implies that the STRF temporal cross sections (at different spectral locations) are all composed of the same essential temporal function except for an arbitrary (Hilbert) rotation. In our physiological investigations, we have rarely come across cortical STRFs that violate this property (Depireux *et al.*, 2001). An example of the consequence of such a constraint is that the STRFs cannot be strictly velocity-selective, i.e., respond to any arbitrary spectrum only when it sweeps past at a specific velocity because such STRFs would not be quadrant separable.

Akansu, A. N., and Haddad, R. A. (**1992**). *Multiresolution Signal Decomposition* Academic, Boston.

Amagai, S., Dooling, R., Shamma, S., Kidd, T., and Lohr, B. (**1999**). "Detection of modulation in spectral envelopes and linear-rippled noises by budgerigars," J. Acoust. Soc. Am. **105**, 2029–2035.

Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (**1996**). "Intelligibility of speech with filtered time trajectories of spectral envelopes," *Proc. ICSLP*, pp. 2490–2492.

Atal, B. S. (**1974**). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identifucation and verification," J. Acoust. Soc. Am. **55**, 1304–1312.

Atlas, L., and Shamma, S. (**2003**). "Joint acoustic and modulation frequency," EURASIP J. Appl. Signal Process. **7**, 668–675.

Bacon, S. P., and Grantham, D. W. (**1989**). "Modulation masking: Effects of modulation frequency, depth, and phase," J. Acoust. Soc. Am. **85**, 2575–2580.

Baer, T., and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. **94**, 1229–1241.

Bates, R. H. T. (**1984**). "Uniqueness of solutions to two-dimensional fourier phase problems for localized and positive images," Comput. Vis. Graph. Image Process. **25**, 205–217.

Calhoun, B., and Schreiner, C. (**1995**). "Spectral envelope coding in cat primary auditory cortex," J. Aud. Neuroscie. **1**, 39–61.

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

Carney, L. H. (**1993**). "A model for the responses of low-frequency auditory-nerve fibers in cat," J. Acoust. Soc. Am. **93**, 401–417.

Chi, T. (**2003**). "Computational Spectro-temporal Auditory Model with Applications to Acoustical Information Processing," Ph.D. thesis, University of Maryland, College Park, MD.

Chi, T., Gao, Y., Guyton, C. G., Ru, P., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**, 2719–2732.

Cohen, J. R. (**1989**). "Application of an auditory model to speech recognition," J. Acoust. Soc. Am. **85**, 2623–2633.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration," J. Acoust. Soc. Am. **102**, 2906–2919.

Dau, T., Puschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the effective signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**, 3615–3622.

deCharms, R. C., Blake, D. T., and Merzenich, M. M. (**1998**). "Optimizing sound features for cortical neurons," Science **280**(5368), 1439–1443.

Depireux, D., Simon, J., Klein, D., and Shamma, S. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**(3), 1220–1234.

deRibaupierre, F., and Rouiller, E. (**1981**). "Temporal coding of repetitive clicks: presence of rate selective units in the cat's medial geniculate body (mgb)," J. Physiol. (London) **318**, 23–24.

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

Drullman, R., Festen, J., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Edamatsu, H., Kawasaki, M., and Suga, N. (**1989**). "Distribution of combination-sensitive neurons in the ventral fringe area of the auditory cortex of the mustached bat," J. Neurophysiol. **61**(1), 202–207.

Eggermont, J. J. (**2002**). "Temporal modulation transfer functions in cat primary auditory cortex: Separating stimulus effects from neural mechanisms," J. Neurophysiol. **87**, 305–321.

Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," Speech Commun. **41**(2–3), 331–348.

Elhilali, M., Fritz, J. B., Klein, D. J., Simon, J. Z., and Shamma, S. A. (**2004**). "Dynamics of precise spike timing in primary auditory cortex," J. Neurosci. **24**(5), 1159–1172.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**, 1181–1196.

Fienup, J. R. (**1982**). "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769.

Fienup, J. R., and Wackerman, C. C. (**1987**). "Phase-retrieval stagnation problems and solutions," J. Opt. Soc. Am. A **3**(11), 1897–1907.

Fu, Q.-J., and Shannon, R. V. (**2000**). "Effect of stimulation rate on phoneme recognition by nucleus-22 cochlear implant listeners," J. Acoust. Soc. Am. **107**, 589–597.

Gerchberg, R. W., and Saxton, W. O. (**1972**). "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik (Jena) **35**, 237–246.

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.
Green, D. M. (**1986**). "Frequency and the detection of spectral shape change," in *Auditory Frequency Selectivity* (Plenum, New York), pp. 351–359.
Greenberg, S., and Kingsbury, B. (**1997**). "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. ICASSP*, pp. 1647–1650.
Greenberg, S., Arai, T., and Silipo, R. (**1998**). "Speech intelligibility derived from exceedingly sparse spectral information," in *Proc. of the Intl. Conf. on Spoken Language Processing*, Sydney, pp. 2803–2806.
Grimault, N., Bacon, S. P., and Micheyl, C. (**2002**). "Auditory stream segregation on the basis of amplitude-modulation rate," J. Acoust. Soc. Am. **111**, 1340–1348.
Hansen, M., and Kollmeier, B. (**1999**). "Continuous assessment of time-varying speech quality," J. Acoust. Soc. Am. **106**, 2888–2899.
Hayes, M. H. (**1982**). "The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-30**(2), 140–154.
Hayes, M. H. (**1987**). "The unique reconstruction of multidimensional sequences from fourier transform magnitude or phase," in *Image Recovery: Theory and Application*, edited by H. Stark (Academic, San Diego), pp. 195–230.
Hayes, M. H., Lim, J. S., and Oppenheim, A. V. (**1980**). "Signal reconstruction from phase or magnitude," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-28**(6), 672–680.
Hermansky, H., and Morgan, N. (**1994**). "Rasta processing of speech," IEEE Trans. Speech Audio Process. **2**(4), 578–589.
Houtgast, T. (**1989**). "Frequency selectivity in amplitude-modulation detection," J. Acoust. Soc. Am. **85**(4), 1676–1680.
Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics," Acustica **46**, 60–72.
Irino, T., and Kawahara, H. (**1993**). "Signal reconstruction from modified auditory wavelet transform," IEEE Trans. Signal Process. **41**(12), 3549–3554.
ITU-T (**2001**). "Perceptual evaluation of speech quality (pesq): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, February.
Jones, J. P., and Palmer, L. A. (**1987**). "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," J. Neurophysiol. **58**(6), 1233–1258.
Joris, P., and Yin, T. C. (**1992**). "Responses to amplitude-modulated tones in the auditory nerve of the cat," J. Acoust. Soc. Am. **91**, 215–232.
Klein, D. J., Depireux, D. A., Simon, J. Z., and Shamma, S. A. (**2000**). "Robust spectro temporal reverse correlation for the auditory system: Optimizing stimulus design," J. Comput. Neurosci. **9**, 85–111.
Kleinschmidt, M., Tchorz, J., and Kollmeier, B. (**2001**). "Combining speech enhancement and auditory feature extraction for robust speech recognition," Speech Commun. **34**(1–2), 75–91.
Kowalski, N., Depireux, D., and Shamma, S. A. (**1996**). "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra," J. Neurophysiol. **76**(5), 3503–3523.
Kryter, K. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–2147.
Langner, G. (**1992**). "Periodicity coding in the auditory system," Hear. Res. **60**, 115–142.
Langner, G., and Schreiner, C. E. (**1988**). "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," J. Neurophysiol. **60**(6), 1799–1822.
Levi, A., and Stark, H. (**1983**). "Signal restoration from phase by projections onto convex sets," J. Opt. Soc. Am. **73**(6), 810–822.
Levi, A., and Stark, H. (**1984**). "Image restoration by the method of generalized projections with application to restoration from magnitude," J. Opt. Soc. Am. A **1**(9), 932–943.
Lu, T., Liang, L., and Wang, X. (**2001**). "Temporal and rate representations of time-varying signals in the auditory cortex of awake primates," Nat. Neurosci. **11**, 1131–1138.
Lyon, R., and Shamma, S. (**1996**). "Auditory representations of timbre and pitch," in *Auditory Computation*, edited by H. Hawkins, E. T. McMullen, A. Popper, and R. Fay (Springer Verlag, New York), pp. 221–270.
Meddis, R., Hewitt, M. J., and Shackleton, T. M. (**1990**). "Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse," J. Acoust. Soc. Am. **87**, 1813–1816.
Mesgarani, N., and Shamma, S. (**2005**). "Speech enhancement based on filtering the spectrotemporal modulations," in *Proc. ICASSP*. Vol. 1, pp. 1105–1108.
Mesgarani, N., Slaney, M., and Shamma, S. (**2004**). "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," IEEE Trans. Speech Audio Process. (accepted for publication).
Miller, L. M., Escabi, M. A., Read, H. L., and Schreiner, C. E. (**2002**). "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," J. Neurophysiol. **87**(1), 516–527.
Mou-yan, Z., and Unbehauen, R. (**1997**). "Methods for reconstruction of 2-d sequences from fourier transform magnitude," IEEE Trans. Image Process. **6**(2), 222–233.
Nelken, I., and Versnel, H. (**2000**). "Responses to linear and logarithmic frequency-modulated sweeps in ferret primary auditory cortex," Eur. J. Neurosci. **12**(2), 549–562.
Pan, D. (**1995**). "A tutorial on mpeg audio compression," IEEE Multimedia **2**(2), 60–74.
Papoulis, A. (**1975**). "A new algorithm in spectral analysis and band-limited extrapolation," IEEE Trans. Circuits Syst. **CAS-22**(9), 735–742.
Pfeiffer, R. R., and Kim, D. O. (**1975**). "Cochlear nerve fiber responses: distributing along the cochlear partition," J. Acoust. Soc. Am. **58**, 867–869.
Pitton, J. W., Wang, K., and Juang, B.-H. (**1996**). "Time-frequency analysis and auditory modeling for automatic recognition of speech," Proc. IEEE **84**(9), 1199–1215.
Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. **112**(5), 2074–2085.
Rosen, S. (**1992**). "Temporal information in speech: acoustic, auditory, and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**(10), 367–373.
Ru, P. (**2000**). "Perception-Based Multi-resolution Auditory Processing of Acoustic Signal," Ph.D. thesis, University of Maryland, College Park, MD.
Ru, P., and Shamma, S. A. (**1997**). "Presentation of musical timbre in the auditory cortex," J. New Music Res. **26**(2), 154–169.
Schreiner, C. E., and Urbas, J. V. (**1988a**). "Representation of amplitude modulation in the auditory cortex of the cat. i: The anterior field," Hear. Res. **21**, 227–241.
Schreiner, C. E., and Urbas, J. V. (**1988b**). "Representation of amplitude modulation in the auditory cortex of the cat. ii: Comparison between cortical fields," Hear. Res. **32**, 49–63.
Seldin, J. H., and Fienup, J. R. (**1990**). "Numerical investigation of the uniqueness of phase retrieval," J. Opt. Soc. Am. A **7**(3), 412–427.
Shamma, S. (**2003**). "Physiological foundations of temporal integration in the perception of speech," J. Phonetics **31**, 495–501.
Shamma, S., Chadwick, R., Wilbur, J., Morrish, K., and Rinzel, J. (**1986**). "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," J. Acoust. Soc. Am. **80**, 133–145.
Shamma, S. A. (**1985a**). "Speech processing in the auditory system I: The representation of speech in the response 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. (**1989**). "Spatial and temporal processing in central auditory networks," in *Methods in Neuronal Modeling*, edited by C. Koch and I. Segev (MIT, Cambridge, MA), pp. 247–289.
Shamma, S. A., Versnel, H., and Kowalski, N. (**1995**). "Ripple analysis in the ferret auditory cortex: I. Response characteristics of single units to sinusoidally rippled spectra," J. Aud. Neurosci. **1**(2), 233–254.
Shamma, S. A., Fleshman, J. W., Wiser, P. R., and Versnal, H. (**1993**). "Organization of the response areas in ferret primary auditory cortex," J. Neurophysiol. **69**(2), 367–383.
Shannon, R. V., Zeng, F.-G., Wygonski, J., Kamath, V., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Sheft, S., and Yost, W. (**1990**). "Temporal integration in amplitude modulation detection," J. Acoust. Soc. Am. **88**, 796–805.
Slaney, M. (**1998**). "Auditory toolbox: Version 2," Technical Report 1998-010, Interval Research Corporation.
Slaney, M., Naar, D., and Lyon, R. F. (**1994**). "Auditory model inversion for sound separation," in *Proc. ICASSP*, Vol. **II**, pp. 77–80.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**(6876), 87–90.
Tchorz, J., and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition," J. Acoust. Soc. Am. **106**, 2040–2050.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am. **91**, 2872–2880.
Ulanovsky, N., Las, L., and Nelken, I. (**2003**). "Processing of low-probability sounds by cortical neurons," Nat. Neurosci. **6**, 391–398.
Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.
Wang, K., and Shamma, S. A. (**1994**). "Self-normalization and noise-robustness in early auditory representations," IEEE Trans. Speech Audio Process. **2**(3), 421–435.
Wang, K., and Shamma, S. A. (**1995**). "Representation of spectral profiles in primary auditory cortex," IEEE Trans. Speech Audio Process. **3**(5), 382–395.
Watson, A. B., and Ahumada, A. J. (**1985**). "Model of human visual-motion sensing," J. Opt. Soc. Am. A **2**(2), 322–342.
Westerman, L. A., and Smith, R. L. (**1984**). "Rapid and short term adaptation in auditory nerve responses," Hear. Res. **15**, 249–260.
Yang, X., Wang, K., and Shamma, S. A. (**1992**). "Auditory representations of acoustic signals," IEEE Trans. Inf. Theory **38**(2), 824–839.