

Multiresolution Spectrotemporal Analysis of Complex Sounds

Taishih Chi, Powen Ru* and Shihab A. Shamma**

*Center for Auditory and Acoustics Research, Institute for Systems Research
Electrical and Computer Engineering Department, University of Maryland, College Park,
MD 20742, USA*

* Now at Cybernetics InfoTech Inc.

** Corresponding author. Tel.: +1 301 405 6842 E-mail: sas@isr.umd.edu (S.A. Shamma)

Abstract

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 of sound likely critical in the perception of timbre. Several types of complex stimuli are used to demonstrate the spectrotemporal information extracted and represented by the model. Also outlined are several reconstruction algorithms to resynthesize the sound so as to evaluate the fidelity of the representation and contribution of different features and cues to the sound percept. Simplified versions of this model representations have already been used in a variety of applications, as in the assessment of speech intelligibility [Elhilali et al., 2003, Chi et al., 1999] and in explaining the perception of monaural phase sensitivity [Carlyon and Shamma, 2002].

1. 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 [Meddis et al., 1990] 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 recently begun to be uncovered [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 still rare that one finds any ideas from central auditory processing being applied in the design of speech and audio processing systems (e.g., [Kleinschmidt et al., 2001]). 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].

To exploit the accumulating experimental findings from the central auditory system, it is essential that they be re-formulated as mathematical models and signal processing algorithms. In previous publications from our group, we have outlined the general principles of a model of central auditory processing, and demonstrated its 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, 2002]. Here we provide a more complete mathematical specification of the model, illustrating how signals are transformed through various stages of the model, and showing the sufficiency of this representation by proposing algorithms for signal reconstruction from the cortical representation. The model 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 physiological observations are particularly important in the model. 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) [Kowalski et al., 1996, Miller et al., 2002, Schreiner and Urbas, 1988a, Langner, 1992]¹. These latter cortical time-scales are commensurate with the dynamics of the vocal tract in speech, with the rate of change of pitch in musical melody, with the transient dynamics that differentiate a struck from a bowed string (e.g., a piano versus a violin), and with the

¹Cortical 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]

rhythms of percussion instruments. 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 we describe in this paper incorporates the two physiological findings above within a framework that consists of two major auditory transformations. An *early* stage that captures monaural processing from the cochlea to the midbrain. It transforms the acoustic stimulus to an auditory time-frequency spectrogram-like representation which combines relatively simple bandpass spectral selectivity with moderate temporal dynamics. The second is called *cortical* stage because it reflects the more complex spectrotemporal analysis presumed to take place in mammalian AI.

Our focus will be on specifying this model for the analysis of sound timbre - the most ill-defined of the rich and varied percepts of complex sounds such as music and speech. Unlike the other basic attributes of sound - loudness, pitch, and location - timbre is a multidimensional percept that is difficult to relate to a concise cue or to reduce to a simple ordered scale. Instead, it has been customary to propose several "descriptive" scales to quantify it using intuitive notions such as sharp-to-dull and continuant-to-transient [Plomp, 1976]. Nevertheless, psychoacoustical evidence is abundantly clear that the *shape and dynamics* of a sound spectrum strongly influence its timbre. Consequently, the proposed model will provide an integrated *spectro-temporal* representation of the monaural acoustic spectrum which potentially serves as an accurate *quantitative descriptor* of timbre. The representation we elaborate upon here has applications where timbre plays a key role as in speech recognition and music synthesis.

In the following section, we review the cortical physiological data and psychoacoustical results that motivated and justified the model's development. The mathematical formulation of the early and cortical stages of the model are summarized in sections 3 and 4, together with an illustration of the way in which a variety of complex sounds are represented. In section 5, we describe algorithms to *reconstruct* audible approximate versions of the original sounds from the model's representations. The intelligibility of the reconstructed signals based on various ranges of spectro-temporal modulations is assessed via the Spectro-Temporal Modulation Index (STMI) [Elhilali et al., 2003]. We end in section 6 with a summary and a brief assessment of the utility of the model in several psychoacoustical studies including measurements of speech intelligibility.

2. 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 Spectro-Temporal Response Fields (STRF) of AI cells. Figure 1(b) provides examples of a variety of measured STRFs with their excitatory and inhibitory fields and dynamics. 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 (Figure 1(a)). 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 unit to all ripple densities and velocities it is possible by an inverse Fourier transform to compute the two dimensional impulse response of the cell, or its STRF, and hence characterize simultaneously the cells' spectral and temporal response selectivity.

Figure 1

Measured STRFs in AI exhibit a variety of spectral widths, asymmetries, and dynamics as illustrated in Figure 1(b). Some units are broadly tuned (**i,ii**) and hence are most responsive to low density ripples; others are narrowly tuned (**iv**) and respond well to fine features of the profile and to high density ripples. STRFs also exhibit a variety of asymmetric inhibitory surrounds (e.g., contrast the symmetrically inhibited STRF in unit **iv** with the asymmetric STRF in **iii**). Finally, STRFs may be slow (**iii,v**) or fast (**iv**), or selective to upward-moving ripples(**v**). From a functional and psychoacoustical perspective, such rich variety implies that each STRF acts as a *selective filter* specific to a particular range of spectral resolutions (or *scales*) and tuned to a limited range of temporal modulations *rates*. The collection of all such STRFs then would constitute a filter bank spanning the broad range of psychoacoustically observed scale and rate sensitivity in humans and animals [Green, 1986, Viemeister, 1979, Chi et al., 1999, Amagai 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 [Lieberman et al., 1967, 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, Dau et al., 1996b, 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.

3. 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].

3.1. Mathematical formulation

The stages of the early auditory model are illustrated in Figure 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 ($Q_{ERB} = 5.88$) bandpass filters with center frequencies (CF) that are uniformly distributed along a logarithmic frequency axis (x), over 5.3 octaves (24 filters/octave). The impulse response of each filter is denoted by $h(t; x)$ (see Appendix I for details of filter implementations). 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 $\tau = 8$ msec mimicking the further loss of phase-locking observed in the midbrain. The mathematical formulation for this model can be summarized as followed:

$$y_{coch}(t, x) = s(t) *_t h(t; x) \quad (1)$$

$$y_{AN}(t, x) = g(\partial_t y_{coch}(t, x)) *_t w(t) \quad (2)$$

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

$$y_{final}(t, x) = y_{LIN}(t, x) *_t \mu(t; \tau) \quad (4)$$

where $*_t$ denotes convolution operation in the time domain.

Figure 2

3.1.1. Limitations and extensions of the early auditory stage

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. These properties are

critical in some applications. The lateral inhibition model is very schematic and lacks details of single neurons. We also have no explicit adaptive properties in our current model [Westerman and Smith, 1984, Meddis et al., 1990]. All of these details are likely to be important in certain circumstances, and should be added when needed.

3.2. 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 combinations, 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 next section.

3.2.1. Three tones: 250, 1000, 4000 Hz

Figure 3(a) illustrate the response patterns due to a low, medium, and high frequency tones. The low frequency tones (250, 1000 Hz) evoke the typical travelling-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 Figure 3(a), substantial phase-locking is only seen for the 250 Hz tone.

3.2.2. 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 octave) on a logarithmic frequency axis (135 - 7465 Hz). At this inter-tone spacing, 2 to 4 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 inter-tone 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 inter-tone 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].

3.2.3. Harmonic complexes

Unlike broadband noise, harmonic complexes have uniform inter-tone 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 CF's, the unresolved higher harmonics interact producing the strong 80 Hz periodic envelope modulations. When the harmonics are random-phase (Figure 3(b), right panel), the envelope modulations become irregular and less peaked, but still preserve their periodicity of 80 Hz. The key general observation to make about these envelope modulations is that they

relate to inter-component interactions, and hence are affected by the spacing, phase, and relative amplitudes of the components - factors reflecting the perceptual timbre 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.

Figure 3

3.2.4. Ripples: spectro-temporally modulated noise

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

Left panel of Figure 3(c) displays the y_{final} output for a downward sweeping ripple ($\omega = 16$ Hz, $\Omega = 1$ cyc/oct). 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 CF's, 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 Figure 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 in speech, or from one note or instrument to another in music (see next example).

3.2.5. Speech and music

Speech and music are an elaboration of harmonic or noise ripples in that they are conceptually constructed of a spectro-temporal 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 in the context of a full sentence. Figure 4(b) displays the responses to the B3 note played on a bowed violin. Both responses exhibit very similar patterns to those of the ripple. Specifically, the three kinds of modulations are clearly seen throughout the response. For example, in the voiced part of the word /drew/ (300-450 ms in Figure 4(a)), phase-locked responses to the fundamental (≈ 100 Hz) and its second and third harmonics are clearly evident. At higher CF (≈ 2000 Hz), the intermediate modulations due to unresolved harmonics (≈ 100 Hz) again reflect the fundamental frequency or source characteristics. The slowest modulations in speech, such as the divergence

of first and second formants, the convergence of the second and third formants, or the onset of the vocalic segment at 60 ms, are due to vocal tract motion and not source characteristics, and hence are the primary purveyors of intelligibility [Chi et al., 1999]. The same types of modulations are seen in the violin sound in Figure 4(b). Note especially the slow modulations encoding the gradual onset and offset of the note, and the regular modulations at ≈ 6 Hz seen in most channels responses. As in speech, these 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]).

Figure 4

4. 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 (STRF) of these filters are also centered at different frequencies along the tonotopic axis [Chi et al., 1999].

An example of the STRF of such filter in the bank is shown in Figure 5(a). Three features are of particular interest: (i) it is centered on a particular center frequency (CF). The location of the excitatory (red) and inhibitory (blue) stripes on the vertical axis indicates that it is sensitive to frequencies of about a 2 octave range around the CF (between about 0.5 CF and 2 CF); (ii) the modulation rate along the time axis is about 16 Hz; (iii) the excitatory portions are separated on the vertical axis by about 1 octave, giving rise to a spectral "scale" sensitivity to peaks separated by 1 octave, 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 spectro-temporal 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.

4.1. Mathematical formulation

We assume a bank of STRFs composed of the product of a spatial impulse response $h_{IRS}(x)$ and temporal impulse response $h_{IRT}(t)$, i.e., $\text{STRF} \equiv h_{IRT}(t) \cdot h_{IRS}(x)$. These impulse responses are defined by sinusoidally interpolating symmetric seed functions $h_s(\cdot), h_t(\cdot)$ and their Hilbert transform

$$h_{IRS}(x; \Omega, \phi) = h_s(x; \Omega) \cos \phi + \hat{h}_s(x; \Omega) \sin \phi \quad (5)$$

$$h_{IRT}(t; \omega, \theta) = h_t(t; \omega) \cos \theta + \hat{h}_t(t; \omega) \sin \theta \quad (6)$$

where Ω and ω are the spectral density and velocity parameters of the filters; ϕ and θ are characteristic phases; the Hilbert transform is defined as

$$\hat{h}_s(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{h_s(z)}{z - x} dz$$

We choose the second derivative of a Gaussian function and one gamma function to approximate $h_s(\cdot)$ and $h_t(\cdot)$ respectively (Figure 5(b))

$$h_s(x) = (1 - x^2)e^{-\frac{x^2}{2}}$$

$$h_t(t) = t^3 e^{-4t} \cos(2\pi t)$$

and the impulse responses for different scales and rates are given by dilation

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

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

Therefore, the spectrotemporal response of a cell c for an input spectrogram $y(t, s)$ is given by

$$r_c(t, x; \omega_c, \Omega_c, \theta_c, \phi_c) = y(t, x) *_{tx} [h_{IRT}(t; \omega_c, \theta_c) \cdot h_{IRS}(x; \Omega_c, \phi_c)] \quad (7)$$

where $*_{tx}$ denotes convolution with respect to both t and x . This multiscale multirate (or *multiresolution spectrotemporal*) response is called "cortical representation". Substituting Eqs.(5), (6) into Eq.(7), the cortical representation at cell c can be rewritten as

$$r_c(t, x; \omega_c, \Omega_c, \theta_c, \phi_c) = y(t, x) *_{tx} [h_t h_s \cos \theta_c \cos \phi_c + h_t \hat{h}_s \cos \theta_c \sin \phi_c + \hat{h}_t h_s \sin \theta_c \cos \phi_c + \hat{h}_t \hat{h}_s \sin \theta_c \sin \phi_c] \quad (8)$$

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 the different directional STRFs. Let

$$\begin{aligned} z_{\Downarrow}(t, x; \omega_c, \Omega_c) &= y(t, x) *_{tx} [h_{TW}(t; \omega_c) h_{SW}(x; \Omega_c)] \\ &= |z_{\Downarrow}(t, x; \omega_c, \Omega_c)| e^{j\psi_{\Downarrow}(t, x; \omega_c, \Omega_c)} \end{aligned} \quad (9)$$

$$\begin{aligned} z_{\Uparrow}(t, x; \omega_c, \Omega_c) &= y(t, x) *_{tx} [h_{TW}^*(t; \omega_c) h_{SW}(x; \Omega_c)] \\ &= |z_{\Uparrow}(t, x; \omega_c, \Omega_c)| e^{j\psi_{\Uparrow}(t, x; \omega_c, \Omega_c)} \end{aligned} \quad (10)$$

where $*$ denotes the complex conjugate; \Downarrow and \Uparrow denote downward and upward moving direction respectively; $h_{SW}(\cdot)$ and $h_{TW}(\cdot)$ are the analytical forms of the impulse responses $h_{IRS}(\cdot)$ and $h_{IRT}(\cdot)$

$$h_{SW}(x; \Omega_c) = h_s(x; \Omega_c) + j\hat{h}_s(x; \Omega_c) \quad (11)$$

$$h_{TW}(t; \omega_c) = h_t(t; \omega_c) + j\hat{h}_t(t; \omega_c) \quad (12)$$

Substituting Eqs.(11),(12) into Eqs.(9),(10) and comparing with Eq.(8), the cortical response at cell c can be simplified to

$$\begin{aligned} r_c(t, x; \omega_c, \Omega_c, \theta_c, \phi_c) &= \frac{1}{2} [\Re\{z_{\Downarrow}\} \cos(\theta_c + \phi_c) + \Im\{z_{\Downarrow}\} \sin(\theta_c + \phi_c) \\ &+ \Re\{z_{\Uparrow}\} \cos(\phi_c - \theta_c) + \Im\{z_{\Uparrow}\} \sin(\phi_c - \theta_c)] \end{aligned} \quad (13)$$

$$= \frac{1}{2} [|z_{\Downarrow}| \cos(\psi_{\Downarrow} - \theta_c - \phi_c) + |z_{\Uparrow}| \cos(\psi_{\Uparrow} + \theta_c - \phi_c)] \quad (14)$$

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; $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote the real part and imaginary part respectively.

The expressions above show that the cortical model response r_c can be re-expressed in terms of magnitude responses $|z_{\Downarrow}|, |z_{\Uparrow}|$ and phase responses $\psi_{\Downarrow}, \psi_{\Uparrow}$; which are obtained by complex wavelet transform (Eqs.(9),(10)). In other words, the six-dimensional response $r_c(t, x; \omega_c, \Omega_c, \theta_c, \phi_c)$ can be fully specified by four functions $|z_{\Downarrow}|, |z_{\Uparrow}|, \psi_{\Downarrow}$ and ψ_{\Uparrow} . 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) .

Figure 5

4.2. Examples of cortical representations

Because of the multi-dimensionality 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 section 3.

4.2.1. Three tones

Figure 6(a) shows three particularly useful summary views of the cortical responses to the three-tone auditory spectrogram in Figure 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 Figure 6(a), the

remaining third variable is integrated out over its domain. For example, in 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 ($\omega_c \leq 4$ 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 Figure 3(a)); a weaker output due to phase-locking is also seen at 1 kHz. Center panel displays the output in the scale-frequency plane, integrating all filter outputs along the rate axis. STRFs with fine resolution relative to the inter-tone 2 octave spacing (i.e., tuned to $\Omega_c > 0.5$ cyc/oct) respond to each tone separately. STRFs with broad bandwidths ($\Omega_c < 0.5$ cyc/oct) smear the representation of the tones 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 ≤ 4 Hz and at 0.5 cyc/oct (since the tones are separated by 2 octaves). 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.

4.2.2. Noise and harmonic spectra

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 Figure 6(b)-(c).

The noise evokes a rate-frequency response (Figure 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 Figure 3(b) earlier. By contrast, the response to the harmonic stimulus (top panel of Figure 6(c)) is dominated by the phase-locked responses to the resolved low-order harmonics, and all intermediate-rate modulations at high CF (≥ 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 "travelling waves").

The scale-frequency panels (bottom panels) of Figure 6(b)-(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.

4.2.3. 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 Figure 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.

Figure 6

4.2.4. Speech and music

Unlike other stimuli, speech and music are typically non-stationary, 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 .

Figure 7

5. Reconstruction

We derive in this section computational procedures to re-synthesize 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. Detailed mathematical analysis of the proposed projection algorithms are discussed in Appendix III. The ability to reconstruct the audio signal from the final representation is extremely useful in building the intuition of the role of different spectro-temporal 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.

5.1. Reconstruction from auditory spectrogram

The most important component of the forward analysis stage - the *linear* filter bank operation (Eq.(1)) - is invertible and the inverse operation can be derived as follows [Akansu and Haddad, 1992]. From Eq.(1),

$$\begin{aligned} Y_{coch}(\omega, x) &= S(\omega)H(\omega; x) \\ \Rightarrow \sum_x Y_{coch}(\omega, x)H^*(\omega; x) &= S(\omega) \sum_x H(\omega; x)H^*(\omega; x) \\ \Rightarrow S(\omega) &= \sum_x Y_{coch}(\omega, x)H^*(\omega; x) / \sum_x |H(\omega; x)|^2 \end{aligned} \quad (15)$$

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, $\sum_x |H(\omega; x)|^2$, is flat except at the lowest and highest frequency skirts where it drops precipitously, causing large noise and numerical errors in the inversion procedure. 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.(15))

$$\begin{aligned}\tilde{S}(\omega) &= \sum_x \tilde{Y}_{coch}(\omega, x) H_1^*(\omega; x) \\ \tilde{s}(t) &= \sum_x \tilde{y}_{coch}(t, x) *_t h_1^*(-t; x) \\ &= \sum_x \tilde{y}_{coch}(t, x) *_t h_1(-t; x)\end{aligned}\quad (16)$$

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 basic assumption of this method is that the auditory spectrogram $y_{final}(t, x)$ roughly represents 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 $\tilde{y}_{final}(t, x)$ from $\tilde{y}_{coch}(t, x)$. 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., $\tilde{y}_{coch}^{(k)}(t, x) \leftarrow r^{(k)}(t, x) \tilde{y}_{coch}^{(k)}(t, x)$.
5. Reconstruct time waveform $\tilde{s}^{(k+1)}(t)$ by inverse filtering (Eq.(16)), 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).

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 an 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 quantified the Mean Opinion Score (MOS) of the reconstructed signals using the "Perceptual Evaluation of Speech Quality" (PESQ) (available from <http://www.itu.int/> under "ITU Publications" [ITU-T, 2001]). The average PESQ score of 50 reconstructed sentences /Come home right away/ (Figure 9) with different initial patterns after 200 iterations is 4.122 (toll quality) with 0.075 standard deviation.

Figure 8

5.2. Reconstruction from the cortical representation

The cortical stage is modeled by a bank of spectro-temporal 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.(9), (10), (14)). This stage is formally identical

to the cochlear analysis stage (Eq.(1) versus Eq.(7)), and hence the one-dimensional inverse filtering technique (Eq.(15)) can be extended to solve the inverse problem of two-dimensional cortical filtering process.

The Fourier representations of Eqs.(9) and (10) can be written as

$$Z_{\downarrow}(\omega, \Omega; \omega_c, \Omega_c) = Y(\omega, \Omega)H_{TW}(\omega; \omega_c)H_{SW}(\Omega; \Omega_c) \quad (17)$$

$$Z_{\uparrow}(\omega, \Omega; \omega_c, \Omega_c) = Y(\omega, \Omega)H_{TW}^*(-\omega; \omega_c)H_{SW}(\Omega; \Omega_c) \quad (18)$$

and from Eqs.(11), (12)

$$H_{SW}(\Omega; \Omega_c) = H_s(\Omega; \Omega_c)[1 + \text{sgn}(\Omega)] \quad (19)$$

$$H_{TW}(\omega; \omega_c) = H_t(\omega; \omega_c)[1 + \text{sgn}(\omega)] \quad (20)$$

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

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

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

$$\tilde{Y}(\omega, \Omega) = \frac{\sum_{\omega_c, \Omega_c} Z_{\downarrow} H_{TW\downarrow}^* H_{SW}^* + \sum_{\omega_c, \Omega_c} Z_{\uparrow} H_{TW\uparrow}^* H_{SW}^*}{\sum_{\omega_c, \Omega_c} |H_{TW\downarrow} H_{SW}|^2 + \sum_{\omega_c, \Omega_c} |H_{TW\uparrow} H_{SW}|^2} \quad (21)$$

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

Figure 9

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). 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 Appendix II, two algorithms are proposed to reconstruct original speech from the modulation energy distributions ($|z_{\downarrow}|$ and $|z_{\uparrow}|$ in Eq.(14)) only. While the "quality" of the reconstructed signals is worse due to a smaller dynamic range or to propagation of errors in the reconstruction procedure (see Appendix II), they are completely intelligible as can be heard on the website <http://www.isr.umd.edu/CAAR/pubs.html>.

5.3. Intelligibility of the reconstructed signals

To demonstrate the quality and 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 Spectro-Temporal 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 Eq.(14), $\forall c$). Temporal modulations are then filtered out by nulling the outputs of the undesired filters (parameterized 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 section 5.2). 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 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 presented in through a speaker in a sound-insulating chamber (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 Figure 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 cyc/oct). Some intelligibility remains even with filters as broad as 0.5-1 cyc/oct (or about an octave) consistent with previous experimental findings [Shannon et al., 1995].

Figure 10

6. 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 (separable) spectrotemporal mother-wavelet (see Figure 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 wide array of percepts, especially those related to the percept of timbre. Examples include (as we elaborate below) the assessment of speech quality and intelligibility, discrimination of musical timbre, and more generally, quantifying the perception of any complex sound subjected to arbitrary spectral and temporal change (in magnitude or phase).

6.1. Representation and perception of temporal modulations

A fundamental aspect of the model is its extraction and processing of various temporal modulations that carry the signal information. Specifically, we noted the existence of two types of modulations that convey different percepts. The first are the intermediate-rate modulations (a few hundred Hertz) prevalent in the early auditory pathway, and which (together with the spectral pattern) reflect the timbre of the sound. These modulations essentially ride on top of the cochlear filter outputs. The second are the slow-modulations that ride on top of the intermediate ones and convey the percepts that render a speech signal intelligible, or a musical note expressive. It is, therefore, clear that the roles and contributions of these temporal modulations to the auditory percept are often distinguishable. For example, consider the effect of smoothing-out the response envelopes on each channel in the auditory spectrogram to remove all modulations above a certain rate (say 20 Hz) as was done by Drullman in order to examine the relationship between speech intelligibility and the slow envelope modulations. The reconstructed signal from such a spectrogram sounds slightly reverberant but mostly intelligible for all cut-off rates above 4 Hz [Drullman et al., 1994]. This finding has recently been criticized by Ghitza [Ghitza, 2001] on the grounds that the auditory spectrum of the reconstructed signal recreates temporal modulations faster than the intended cutoff rate. While this is true, these recreated rates are purely due to component interactions and hence are of the intermediate kind which only affect the timbre of the signal, and play only a small role in its intelligibility. Consequently, Drullman's original manipulations and interpretations are basically valid, and there is no need to adopt additional strategies to eliminate the re-created intermediate modulations if intelligibility is the focus of the investigation.

6.2. 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]) or purely temporal (e.g., [Drullman et al., 1994]) approaches. 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 wavelet transform). Consequently, all the constraints imposed during the iterations are completely different.

6.3. Applications of the multiresolution model

We have recently adapted and tested the auditory model in two very different contexts. In the first application, 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, 2002]. 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. 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, 2002]. In the second application, we used the model to analyze the effects of noise, reverberations, and other distortions on the joint spectro-temporal 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 [Drullman et al., 1994, Hermansky and Morgan, 1994, Shannon et al., 1995, Arai et al., 1996, Dau et al., 1996a, Greenberg et al., 1996]. Therefore, an intelligibility index which reflects the integrity of these modulations can be effective regardless of the source of the degradation. Such a Spectro-Temporal 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].

We believe that this approach can be successfully extended beyond these two applications

to explain the complex phenomena associated with "auditory scene analysis" and "informational masking". For example, it is well known that two sounds (A,B) are more readily streamed when they differ in timber, pitch, or other perceptual qualities [Darwin and Carlyon, 1995]. We hypothesize that this "perceptual distance" can be directly measured from the cortical multiresolution spectrotemporal representation of these sounds. And, therefore, it is possible to predict or test the effects of manipulating this distance on the perception of streaming in a variety of sequences. The same approach may also be fruitfully employed to address informational masking, a phenomenon which refers to masking of a signal that exceeds what can be accounted for by "energetic masking" or interference from the masker. For example, speech is a more effective masker than noise even when the latter is spectrally filtered to match the speech spectrum. This "unaccounted for" or additional masking beyond what would be expected from a spectral overlap between the masker and signal, is called informational masking. However, we propose that informational masking could simply reflect a failure of "energetic masking" models to account for all masking. For instance, most such models implicitly assume a simple spectral representation of the signals roughly as found at the cochlea or early auditory stages. However, "energetic masking" can take place between more elaborate representations (e.g., cortical-like) that are more subtly sensitive to various features of the signal beyond just the simple spectra. In this light, we predict that informational masking can be recast as simple energetic masking at a higher more sophisticated representational level. For instance, the spectrally-shaped noise masker discussed above is not effective as a competing speech signal because the noise lacks the temporal modulations that characterize speech signals, and hence would have little overlap with (and hence masking of) speech in the cortical representation. Reverse speech, by contrast, does share most of its spectrotemporal features with its normal counterpart, and hence would be a more effective masker of speech. In this case, any remaining differences between normal and reverse speech maskers would have to be referred to an even higher "semantic and syntactic" representation that is beyond the proposed cortical model.

6.4. 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 be modified to reflect future physiological and anatomical findings with no significant effect on the computations. For example, real cortical STRFs (Figure 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.

Certain other details of the cortical filters (STRFs) are critical and reflect as much as possible physiological properties and parameter ranges. One example is the *Quadrant-separability* we impose between the spectral and temporal dimensions of the STRFs which

constrains their shape and function (see [Depireux et al., 2001] for details). One consequence of such a construction 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 be inseparable and hence cannot be implemented with our formulation. In our physiological investigations, we have rarely come across cortical STRFs that violate this property [Depireux et al., 2001]. Another important aspect of the cortical STRFs is their relatively slow dynamics compared to the pre-cortical stages [Miller et al., 2002]. This arises from a combination of possible mechanisms including slow NMDA excitatory inputs [Krukowski and Miller, 2001], slow cortical inhibition [Thomson and Destexhe, 1999], synaptic depression and facilitation [Abbott et al., 1997]. Our linear filters incorporate this important property by having transfer functions tuned to a range of low rates (usually < 20 Hz).

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 20-1000 Hz). The second stage would be slower filters (≤ 20 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 pre-cortically) 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].

7. 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 re-synthesize the acoustic signals from the full or just the envelope of the auditory spectrogram and the multiresolution representation. Such algorithms are shown to converge and can be thought as an implementation of the gradient descent search for solving nonlinear inverse problems. The perceptually tolerable re-synthesized sounds imply that these representations carry the crucial information about the timbre and intelligibility of the sound.

To validate our model, the output representations of the model have been adapted for several applications and shown promising results when used to measure the perceptual distance between two sounds [Carlyon and Shamma, 2002] 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 pre-processor 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 I

Cochlear Filter Bank

The cochlear filter bank is modeled by a set of constant-Q bandpass filters. The frequency response of each individual filter is obtained by dilating or compressing a seed filter response along the linear frequency axis, i.e., $H_m(f) = H_0(f/\alpha^m)$. The exact shape is of no importance as long as the measured cochlear filter has shape of moderate lower skirt (6to12 dB/oct) and steeper upper skirt (-40 to -500 dB/oct) [Allen, 1985]. The 3-dB bandwidth of this highly asymmetric response is roughly 0.2 octave.

The magnitude frequency response of one candidate is

$$|H(x)| = \begin{cases} (x_h - x)^\alpha e^{-\beta(x_h - x)}, & 0 \leq x \leq x_h \\ 0, & x > x_h \end{cases} \quad (22)$$

where x_h is the high frequency limit, $\alpha = .3$ and $\beta = 8$ are positive real numbers.

The center frequency (CF) of the filter at location x on the logarithmic frequency axis (in octaves) is defined as

$$f_x = f_0 \cdot 2^x \text{ (Hz)}$$

where f_0 is a reference frequency (1 kHz in this study).

The phase is chosen to have the corresponding impulse response, $h(t)$, a minimum-phase signal, whose magnitude and phase responses are related by [Oppenheim and Schaffer, 1989]

$$\angle H(f) = -\mathcal{H}\{\log |H(f)|\} \quad (23)$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. This choice will ensure the inverse filter ($H^{-1}(x)$) is stable such that the reconstruction from the spectrogram is feasible.

The frequency response of the other filters can be generated by simply dilating the $H(f)$.

$$H(f, x) = H(2^x f) \quad (24)$$

$$h(t, x) = 2^{-x} h(2^{-x} t) \quad (25)$$

In this study, x was discretized to 128 channels whose characteristic frequencies range roughly from 190 Hz to 7.2 kHz with 24 channels/octave resolution. The channel density was chosen to have a half semi-tone spacing (1/24 octave) between adjacent channels. As a consequence, the spectrum of C note is merely a translated version of the spectrum of E note. The filter-bank is designed for 16-kHz sampled input signal. If 8-kHz sampled signal is encountered, the range of CFs is then shifted down by one octave, i.e., 95 Hz \sim 3.6 kHz. Figure 11 shows the response of the cochlear filter with characteristic frequency around 1 kHz. As indicated in Figure 11(a), the left (right) slope at 3-dB edge marked by asterisks is roughly 20 dB/oct (-145 dB/oct) and the 3-dB bandwidth is about .23 octave. The equivalent rectangular bandwidth (ERB), which is defined as the bandwidth of the rectangular filter with the same peak response and passing the same total power when driven by white noise, of the reference filter (CF = 1 kHz) is 170 Hz. Therefore, the Q_{ERB} ($= CF/ERB(CF)$) of the cochlear filters used in this study is around 5.88. Figure 11(b) and (c) show the magnitude and phase responses on the linear frequency axis and the impulse response is given in Figure 11(d).

Figure 11

The computational load in the cochlear model (Eq.(1) ~ (4)) lies in the filtering process (Eq.(1)). It's well known that IIR filter bank is much more efficient in computation than the conventional overlap-and-add approach [Oppenheim and Schaffer, 1989]. Therefore, the cochlear filters in this work are implemented by IIR filters. The goal is to find a discrete-time transfer function $H(z) = B(z; N_b)/A(z; N_a)$ with a given complex frequency response $H(f)$ where $B(z; N_b) = \sum_{n=0}^{N_b} b_n z^{-n}$ and $A(z; N_a) = \sum_{n=0}^{N_a} a_n z^{-n}$. This is a discrete-time filter identification problem that may not have closed-form solutions but only optimal approximations. Let \mathbf{b} denotes moving average coefficients $[b_1..b_{N_b}]$ and \mathbf{a} denotes autoregressive coefficients $[a_1..a_{N_a}]$. The iterative algorithm for solving \mathbf{b} and \mathbf{a} is composed of two stages: estimation and optimization. First, it uses an equation error method to identify the best model from the available data. Based on Levi's algorithm [Levi, 1959], $\check{\mathbf{b}}$ and $\check{\mathbf{a}}$ are obtained by solving a system of linear equations:

$$(\check{\mathbf{b}}, \check{\mathbf{a}}) = \arg \min_{\mathbf{b}, \mathbf{a}} \sum_{m=1}^M |H(z_m)A(z_m; N_a) - B(z_m; N_b)|^2 \quad (26)$$

where $z_m = e^{2\pi f_m}$. Second, the damped Gauss-Newton method is applied for iterative search [Dennis and Schnabel, 1983] with the initial estimate $(\check{\mathbf{b}}, \check{\mathbf{a}})$. This step minimizes the sum of the squared error between the actual and the desired frequency response:

$$(\check{\mathbf{b}}, \check{\mathbf{a}})(\epsilon, K) = \arg \min_{\mathbf{b}, \mathbf{a}} \sum_{m=1}^M |H(z_m) - B(z_m; N_b)/A(z_m; N_a)|^2 \quad (27)$$

The algorithm stops when the mean squared error is less than tolerance ϵ or after K iterations, whichever comes first. This algorithm is available in Signal Processing Toolbox of MATLAB. The parameters were assigned as follows: $N_b = N_a$, $\epsilon = 10^{-10}$ and $K = \lceil 80000/f_c \rceil$ where f_c is the characteristic frequency of the desired frequency response $H(f)$ and $\lceil x \rceil$ denotes the smallest integer greater than x . For each channel (cochlear filter), the coefficients were searched over an order range of 4 ~ 24 and the optimal order was picked based on the SNR of the 128-ms truncated impulse response.

Appendix II

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 [Irimo 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.(9),(10)). 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.(11),(12)) 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 algorithms and convex projection algorithms.

II.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 *nonnegative* 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.(9),(10)).
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).
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.(21)).
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 Figure 12(a). In Appendix IV, this algorithm is shown to be equivalent to a gradient descent search method.

II.2. Algorithm II: filter-by-filter

Because of the highly overlapped bandwidths of the filters in both the ω and Ω domains, the magnitude constraints of adjacent cortical filters are highly redundant. Consequently, *Algorithm I* yields accurate reconstruction when it converges, but with very high computational cost. To overcome this drawback, certain properties of the filters, e.g., their analytical form could provide implicit additional constraints during the iterations.

Observed from Eqs.(17) ~ (20), $Z_{\downarrow}(\omega, \Omega; \omega_c, \Omega_c)$ 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, a second algorithm is proposed to reduce the computation complexity as follows:

1. Initialize a nonnegative 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.(21) from cortical responses up to filter i ($\tilde{Z}^{(N_i)}(1), \dots, \tilde{Z}^{(N_i)}(i)$) 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.(9),(10)) when $i < N_f$ (number of filters).
6. Go to step 3 and update filter indicator $i = i + 1$.

The block diagram of this filter-by-filter algorithm is depicted in Figure 12(b).

Figure 12

II.3. Comparing algorithms I and II

Algorithm II resolves constraints of one filter at a time (step 3), thus reduces dramatically computation time when compared to Algorithm I. The algorithm is recursive in that the initial phase of filter i ($\tilde{\psi}_{\downarrow}^{(1)}(i)$ and $\tilde{\psi}_{\uparrow}^{(1)}(i)$ in step 3) is estimated 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 from one filter to adjacent filter do not change rapidly. A drawback of Algorithm II is that the

overall performance of this filter-by-filter algorithm depends entirely on the reconstruction result from the first filter since the errors propagate and are magnified down the chain. The reconstructed spectrograms using the two algorithms are compared in the bottom two panels of Figure 9. The processing time of Algorithm I (the third panel from top; 100 iterations) is 150 times longer than that of Algorithm II (bottom panel; $N_i = 10$ for each filter). The plots illustrate that the filter-by-filter algorithm has a smaller dynamic range resulting in some distortion near onsets and offsets, and boosting the representation of the lower harmonics and other weak features in the original spectrogram.

To overcome drawbacks of both proposed algorithms - high computational load versus propagation of errors, a hybrid algorithm might be used, for example, by performing 10 iterations of the first direct-projection algorithm, thus providing a better starting point to initialize the second algorithm for all filters.

The STMI^Ts of the reconstructed speech (second to bottom panel) in Figure 9 are 0.95, 0.89 and 0.91, respectively. These scores indicate that all reconstruction algorithms preserve the slow temporal modulations very well, as can be seen in the auditory spectrograms in Figure 9. However, speech quality in these signals is not high due to the numerous sources of distortion and error in these reconstruction procedures as discussed above.

Appendix III

General Projection Algorithms

Levi and Stark studied the signal restoration by generalized projections [Levi and Stark, 1983, Levi and Stark, 1984]. For any closed set \mathcal{C} , the element $g \triangleq P_{\mathcal{C}}h$ is called the projection of h onto \mathcal{C} , if $g \in \mathcal{C}$ satisfies

$$\|g - h\| = \min_{y \in \mathcal{C}} \|y - h\|$$

where $\|\cdot\|$ is the norm, or length, and is usually taken as the L_2 -space distance. The projection g is unique for a *convex* set \mathcal{C} but is not necessarily unique (or may not exist) for a nonconvex set.

Assume two closed sets $\mathcal{C}_1, \mathcal{C}_2$ with projection operators P_1, P_2 and operator $T_i = 1 + \lambda_i(P_i - 1)$, $i = 1, 2$. The summed distance error is defined as

$$\begin{aligned} J(f_n) &= d(f_n, \mathcal{C}_1) + d(f_n, \mathcal{C}_2) \\ &= \|P_1 f_n - f_n\| + \|P_2 f_n - f_n\| \end{aligned} \quad (28)$$

where $d(h, \mathcal{C})$, the distance between a point h and a set \mathcal{C} , is defined as

$$d(h, \mathcal{C}) = \min_{y \in \mathcal{C}} \|y - h\|$$

Theorem (Levi-Stark, 1984) For a restricted set of λ_i ,

$$J(f_{n+1}) \leq J(T_2 f_n) \leq J(f_n) \quad (29)$$

where $f_{n+1} = T_1 T_2 f_n$. (See [Levi and Stark, 1984] for a proof.)

The restricted set includes $\lambda_i = 1$, hence, the iterative algorithm $f_{n+1} = T_1 T_2 f_n$ reduces to the Gerchberg-Saxton algorithm

$$f_{n+1} = P_1 P_2 f_n, \quad f_0 \text{ arbitrary}$$

with error-reduction property, $J(f_{n+1}) \leq J(f_n)$, which was described in [Fienup, 1982].

The restoration from magnitude cortical representation stated in Appendix II does not fall under the category of problems covered by projection onto convex sets. The proposed algorithms are generalized projections in the sense of Levi-Stark. For the first (direct-projection) algorithm, the set \mathcal{C}_1 of auditory spectrograms $y(t, x)$ where $y \geq 0$ for all t, x is convex, however, the set

$$\mathcal{C}_2 = \{y(t, x) \mid |y(t, x) * [h_{TW}(t; \omega) h_{SW}(x; \Omega)]| = M(t, x; \omega, \Omega) \text{ for all } \omega, \Omega\}$$

with the magnitude constraint $M(t, x; \omega, \Omega)$ is not convex. The projection operators are defined as

$$P_1 y(t, x) = \begin{cases} y(t, x), & y(t, x) \geq 0 \\ 0, & y(t, x) < 0 \end{cases}$$

and

$$P_2 y(t, x) = \mathcal{CF}^{-1}(M(t, x; \omega, \Omega) e^{j\psi(t, x; \omega, \Omega)}, \forall(\omega, \Omega))$$

where \mathcal{CF}^{-1} is the inverse cortical filtering (Eq.(21)) and $\psi(t, x; \omega, \Omega)$ is the phase of the cortical representation of $y(t, x)$ (Eqs.(9), (10)). Thus, the proposed reconstruction algorithm

$$y_{n+1} = P_2 P_1 y_n, \quad y_0 \text{ arbitrary}$$

has the non-increasing error property.

As for the second algorithm, the projections are done filter by filter. The sets \mathcal{C}_1 and \mathcal{C}_2 for each downward (**upward**) cortical filter (ω_i, Ω_j) are

$$\mathcal{C}_1 = \{z_{ij}(t, x) \mid |z_{ij}(t, x)| = M_{ij}(t, x)\}$$

$$\mathcal{C}_2 = \{z_{ij}(t, x) \mid Z_{ij}(\omega, \Omega) \text{ has non-zero elements only in the first (**second**) quadrant}\}$$

where $z_{ij}(t, x) \equiv z(t, x; \omega_i, \Omega_j)$ for short notation and $Z_{ij}(\omega, \Omega)$ is the Fourier transform of $z_{ij}(t, x)$. The projection operators defined for the nonconvex set \mathcal{C}_1 and convex set \mathcal{C}_2 , respectively, are

$$P_1 z_{ij}(t, x) = M_{ij}(t, x) e^{j\psi_{ij}(t, x)}$$

$$P_2 z_{ij}(t, x) = \mathcal{F}^{-1}\{Z_{ij}(\omega, \Omega) \cdot 1(\omega_i \geq (\leq) 0, \Omega_j \geq 0)\}$$

where \mathcal{F}^{-1} is inverse Fourier transform, $\psi_{ij}(t, x)$ is the phase of $z_{ij}(t, x)$ and the indication function is defined as

$$1(\omega_i \geq 0, \Omega_j \geq 0) = \begin{cases} 1, & \omega_i \geq 0 \text{ and } \Omega_j \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Therefore, the step 3 of the filter-by-filter algorithm can be written as

$$z_{n+1} = P_2 P_1 z_n$$

with the non-increasing error property. Unlike choosing an arbitrary starting point in the direct-projection algorithm, the initial $z_0(t, x; \omega_i, \Omega_j)$ for filter (ω_i, Ω_j) is estimated from reconstructed $z(t, x; \omega \leq \omega_i, \Omega \leq \Omega_j)$. For each filter, the location of the starting point (initial estimate z_0) shall strongly affect the fidelity of the reconstruction since the generalized projection algorithms do not guarantee a unique solution for nonconvex sets.

Appendix IV

Gradient Descent Search Method

In this Appendix, we employ the gradient descent search method to solve the reconstruction problem and show that it is identical to the proposed direct-projection algorithm. To simplify notation, $z_{ij}(t, x) \equiv z(t, x; \omega_i, \Omega_j)$ is used through this Appendix.

The squared error in the cortical domain is defined as

$$\begin{aligned} E &= \sum_{t,x,i,j} [|z_{ij}(t, x)| - M_{ij}(t, x)]^2 \\ &= \sum_{t,x,i,j} [|y(t, x) * h_{ij}(t, x)| - M_{ij}(t, x)]^2 \end{aligned} \quad (30)$$

where $M_{ij}(t, x)$ denotes the desired magnitude cortical response, and $h_{ij}(t, x)$ is the impulse response of cortical cell tuned at (ω_i, Ω_j) . This error metric E is to be minimized by varying a set of parameters. In this work, the elements of $\tilde{y}(t, x)$, the estimate of $y(t, x)$, are treated as independent parameters. The Gradient descent method iteratively seeks the bottom point of the error surface of the parameters by applying successive adjustments to the parameters in the direction opposite to the gradient vector [Haykin, 1994].

The partial derivative of E with respect to a given point $y(t', x')$ is

$$\frac{\partial E}{\partial y(t', x')} = 2 \cdot \sum_{t,x,i,j} (|z_{ij}(t, x)| - M_{ij}(t, x)) \cdot \frac{\partial |z_{ij}(t, x)|}{\partial y(t', x')} \quad (31)$$

Since

$$\begin{aligned} \frac{\partial z_{ij}^{(*)}(t, x)}{\partial y(t', x')} &= \frac{\partial}{\partial y(t', x')} [y(t, x) *_{tx} h_{ij}^{(*)}(t, x)] \\ &= \frac{\partial}{\partial y(t', x')} \left[\sum_{\tau_1} \sum_{\tau_2} y(t - \tau_1, x - \tau_2) \cdot h_{ij}^{(*)}(\tau_1, \tau_2) d\tau_1 d\tau_2 \right] \\ &= h_{ij}^{(*)}(t - t', x - x') \end{aligned}$$

then

$$\begin{aligned} \frac{\partial |z_{ij}(t, x)|}{\partial y(t', x')} &= \frac{\partial}{\partial y(t', x')} [|z_{ij}(t, x)|^2]^{1/2} \\ &= \frac{1}{2 \cdot |z_{ij}(t, x)|} [z_{ij}^*(t, x) h_{ij}(t - t', x - x') + z_{ij}(t, x) h_{ij}^*(t - t', x - x')] \\ &= \frac{1}{2} [h_{ij}(t - t', x - x') \cdot e^{-j\psi_{ij}(t, x)} + h_{ij}^*(t - t', x - x') \cdot e^{j\psi_{ij}(t, x)}] \end{aligned}$$

where $z_{ij}(t, x) = |z_{ij}(t, x)| e^{j\psi_{ij}(t, x)}$. Thus, Eq.(31) can be rewritten as

$$\frac{\partial E}{\partial y(t', x')} = \sum_{t,x,i,j} [z_{ij}(t, x) - M_{ij}(t, x) e^{j\psi_{ij}(t, x)}] \cdot h_{ij}^*(t - t', x - x')$$

$$\begin{aligned}
& + \sum_{t,x,i,j} [z_{ij}^*(t,x) - M_{ij}(t,x)e^{-j\psi_{ij}(t,x)}] \cdot h_{ij}(t-t',x-x') \\
= & 2 \cdot \Re[\sum_{t,x,i,j} z_{ij}(t,x) \cdot h_{ij}^*(t-t',x-x') - \sum_{t,x,i,j} M_{ij}(t,x)e^{j\psi_{ij}(t,x)} \cdot h_{ij}^*(t-t',x-x')] \\
= & 2 \cdot \Re[\sum_{i,j} z_{ij}(t',x') * h_{ij}^*(-t',-x') - \sum_{i,j} M_{ij}(t',x')e^{j\psi_{ij}(t',x')} * h_{ij}^*(-t',-x')] \quad (32)
\end{aligned}$$

where the first term is equal to $y(t',x')$ (see Eq.(16), inverse filtering process).

According to the gradient steepest descent method, the adjustment applied to the parameter $\tilde{y}^{(k)}(t',x')$ at iteration k is defined by

$$\Delta \tilde{y}^{(k)}(t',x') = -\eta \frac{\partial \tilde{E}^{(k)}}{\partial \tilde{y}^{(k)}(t',x')}$$

Therefore, with the choice of the learning-rate parameter $\eta = 1/2$ and $\tilde{y}^{(k)}(t',x')$ being real $\forall k$ (object-domain constraint), the gradient descent algorithm can be stated as

$$\begin{aligned}
\tilde{y}^{(k+1)}(t',x') & = \tilde{y}^{(k)}(t',x') - \frac{1}{2} \cdot \frac{\partial \tilde{E}^{(k)}}{\partial \tilde{y}^{(k)}(t',x')} \\
& = \Re[\sum_{i,j} M_{ij}(t',x')e^{j\tilde{\psi}_{ij}^{(k)}(t',x')} * h_{ij}^*(-t',-x')] \quad (33)
\end{aligned}$$

From the above equation, $\tilde{y}^{(k+1)}$ is computed by inverse cortical filtering the modified cortical representations whose magnitude $|\tilde{z}_{ij}^{(k)}|$ have been replaced by cortical domain magnitude constraints M_{ij} . Evidently, this gradient steepest-descent method is identical to the proposed direct-projection algorithm for our reconstruction problem.

References

- [Abbott et al., 1997] Abbott, L. F., Varela, J. A., Sen, K., and Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, 275:220–224.
- [Akansu and Haddad, 1992] Akansu, A. N. and Haddad, R. A. (1992). *Multiresolution Signal Decomposition*. Academic Press.
- [Allen, 1985] Allen, J. B. (1985). Cochlear model. *IEEE Trans. Acoust. Speech and Signal Proc.*, pages 3–28.
- [Amagai et al., 1999] 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(3):2029–2035.
- [Arai et al., 1996] Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (1996). Intelligibility of speech with filtered time trajectories of spectral envelopes. *Proc. ICSLP*, pages 2490–2492.
- [Baer and Moore, 1993] Baer, T. and Moore, B. C. J. (1993). Effects of spectral smearing on the intelligibility of sentences in noise. *J. Acoust. Soc. Am.*, 94(3):1229–1241.
- [Bates, 1984] Bates, R. H. T. (1984). Uniqueness of solutions to two-dimensional fourier phase problems for localized and positive images. *Computer Vision, Graphics, and Image Processing*, 25:205–217.
- [Calhoun and Schreiner, 1995] Calhoun, B. and Schreiner, C. (1995). Spectral envelope coding in cat primary auditory cortex. *J. Auditory Neuroscience*, pages 39–61.
- [Carlyon and Shamma, 2002] Carlyon, R. and Shamma, S. (2002). An account of monaural phase sensitivity. *submitted to J. Acoust. Soc. Am.*
- [Chi et al., 1999] 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(5):2719–2732.
- [Darwin and Carlyon, 1995] Darwin, C. and Carlyon, R. (1995). Auditory grouping. In Moore, B. C. J., editor, *The Handbook of Perception and Cognition, Vol 6: Hearing*, pages 387–424. Academic Press.
- [Dau et al., 1996a] Dau, T., Puschel, D., and Kohlrausch, A. (1996a). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, 99(6):3615–3622.
- [Dau et al., 1996b] Dau, T., Puschel, D., and Kohlrausch, A. (1996b). A quantitative model of the "effective" signal processing in the auditory system. II. Simulation and measurements. *J. Acoust. Soc. Am.*, 99(6):3623–3631.
- [deCharms et al., 1998] deCharms, R. C., Blake, D. T., and Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443.

- [Dennis and Schnabel, 1983] Dennis, J. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall.
- [Depireux et al., 2001] 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 and Rouiller, 1981] 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, 1995] Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am.*, 97(1):585–592.
- [Drullman et al., 1994] Drullman, R., Festen, J., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064.
- [Edamatsu et al., 1989] 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, 2002] 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 et al., 2003] Elhilali, M., Chi, T., and Shamma, S. A. (2003). A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech Communication*, 41(2-3):331–348.
- [Elhilali et al., 2004] 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.
- [Fienup, 1982] Fienup, J. R. (1982). Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21:2758–2769.
- [Fienup and Wackerman, 1987] Fienup, J. R. and Wackerman, C. C. (1987). Phase-retrieval stagnation problems and solutions. *J. Opt. Soc. Am. A*, 3(11):1897–1907.
- [Fu and Shannon, 2000] 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(1):589–597.
- [Gerchberg and Saxton, 1972] Gerchberg, R. W. and Saxton, W. O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246.
- [Ghitza, 2001] 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(3):1628–1640.

- [Green, 1986] Green, D. M. (1986). Frequency and the detection of spectral shape change. In *Auditory Frequency Selectivity*, pages 351–359. Plenum Press.
- [Greenberg et al., 1998] Greenberg, S., Arai, T., and Silipo, R. (1998). Speech intelligibility derived from exceedingly sparse spectral information. In *International Conference on Spoken Language Processing*, pages 2803–2806, Sydney.
- [Grimault et al., 2002] Grimault, N., Bacon, S. P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *J. Acoust. Soc. Am.*, 111(3):1340–1348.
- [Hayes, 1982] Hayes, M. H. (1982). The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform. *IEEE Trans. Acoust. Speech and Signal Proc.*, ASSP-30(2):140–154.
- [Hayes, 1987] Hayes, M. H. (1987). The unique reconstruction of multidimensional sequences from fourier transform magnitude or phase. In Stark, H., editor, *Image Recovery: Theory and Application*, pages 195–230. Academic Press.
- [Hayes et al., 1980] Hayes, M. H., Lim, J. S., and Oppenheim, A. V. (1980). Signal reconstruction from phase or magnitude. *IEEE Trans. Acoust. Speech and Signal Proc.*, ASSP-28(6):672–680.
- [Haykin, 1994] Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Macmillan, New York.
- [Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). Rasta processing of speech. *IEEE Trans. Speech and Audio Proc.*, 2(4):578–589.
- [Houtgast et al., 1980] 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 and Kawahara, 1993] Irino, T. and Kawahara, H. (1993). Signal reconstruction from modified auditory wavelet transform. *IEEE Trans. Signal Proc.*, 41(12):3549–3554.
- [ITU-T, 2001] 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, Feb. 2001.
- [Joris and Yin, 1992] Joris, P. and Yin, T. C. (1992). Responses to amplitude-modulated tones in the auditory nerve of the cat. *J. Acoust. Soc. Am.*, 91(1):215–232.
- [Klein et al., 2000] 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. Neuroscience*, 9:85–111.
- [Kleinschmidt et al., 2001] Kleinschmidt, M., Tchorz, J., and Kollmeier, B. (2001). Combining speech enhancement and auditory feature extraction for robust speech recognition. *Speech Communication*, 34(1-2):75–91.

- [Kowalski et al., 1996] 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.
- [Krukowski and Miller, 2001] Krukowski, A. E. and Miller, K. D. (2001). Thalamocortical nmda conductances and intracortical inhibition can explain cortical temporal tuning. *Nature Neurosci.*, 4:424–430.
- [Kryter, 1962] Kryter, K. (1962). Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.*, 34(11):1689–2147.
- [Langner, 1992] Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research*, 60:115–142.
- [Langner and Schreiner, 1988] 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 and Stark, 1983] Levi, A. and Stark, H. (1983). Signal restoration from phase by projections onto convex sets. *J. Opt. Soc. Am.*, 73(6):810–822.
- [Levi and Stark, 1984] 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.
- [Levi, 1959] Levi, E. C. (1959). Complex-curve fitting. *IRE Trans on Automatic Control*, AC(4):37–44.
- [Liberman et al., 1967] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol Rev*, 74:431–461.
- [Lu et al., 2001] Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neurosci.*
- [Lyon and Shamma, 1996] Lyon, R. and Shamma, S. (1996). Auditory representation of timbre and pitch. In Hawkins, H., McMullen, E. T., Popper, A., and Fay, R., editors, *Auditory Computations*, pages 221–270. Springer Verlag.
- [Meddis et al., 1990] 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(4):1813–1816.
- [Miller et al., 2002] 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 and Unbehauen, 1997] Mou-yan, Z. and Unbehauen, R. (1997). Methods for reconstruction of 2-d sequences from fourier transform magnitude. *IEEE Trans. Image Proc.*, 6(2):222–233.

- [Nelken and Versnel, 2000] 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.
- [Oppenheim and Schaffer, 1989] Oppenheim, A. V. and Schaffer, R. W. (1989). *Discrete-Time Signal Processing*. Prentice-Hall.
- [Pan, 1995] Pan, D. (1995). A tutorial on mpeg audio compression. *IEEE MultiMedia*, 2(2):60–74.
- [Papoulis, 1975] Papoulis, A. (1975). A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Trans. Circuits Syst.*, CAS-22(9):735–742.
- [Pfeiffer and Kim, 1975] Pfeiffer, R. R. and Kim, D. O. (1975). Cochlear nerve fiber responses: distributing along the cochlear partition. *J. Acoust. Soc. Am.*, 58(5):867–869.
- [Plomp, 1976] Plomp, R. (1976). *Aspects of Tone Sensation: A Psychophysical Study*, chapter 6, pages 85–110. Academic Press.
- [Roberts et al., 2002] 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, 1992] Rosen, S. (1992). Temporal information in speech: acoustic, auditory, and linguistic aspects. *Phil. Trans. Royal Soc. London (B)*, 336(10):367–373.
- [Schreiner and Urbas, 1988a] 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 and Urbas, 1988b] 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 and Fienup, 1990] 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 et al., 1986] 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(1):133–145.
- [Shamma, 1985a] 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(5):1612–1621.
- [Shamma, 1985b] 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(5):1622–1632.

- [Shamma, 1989] Shamma, S. A. (1989). Spatial and temporal processing in central auditory networks. In Koch, C. and Segev, I., editors, *Methods in Neural Modeling*, pages 247–289. MIT Press.
- [Shamma et al., 1993] Shamma, S. A., Fleshman, J. W., Wiser, P. R., and Versnel, H. (1993). Organization of the response areas in ferret primary auditory cortex. *J. Neurophysiol.*, 69(2):367–383.
- [Shamma et al., 1995] 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. Auditory Neuroscience*, 1(2):233–254.
- [Shannon et al., 1995] Shannon, R. V., Zeng, F.-G., Wygonski, J., Kamath, V., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270:303–304.
- [Slaney et al., 1994] Slaney, M., Naar, D., and Lyon, R. F. (1994). Auditory model inversion for sound separation. In *Proc. of IEEE ICASSP*, volume II, pages 77–80.
- [Smith et al., 2002] Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90.
- [ter Keurs et al., 1992] ter Keurs, M., Festen, J. M., and Plomp, R. (1992). Effect of spectral envelope smearing on speech reception. I. *J. Acoust. Soc. Am.*, 91(5):2872–2880.
- [Thomson and Destexhe, 1999] Thomson, A. M. and Destexhe, A. (1999). Dual intracellular recordings and computational models of slow inhibitory postsynaptic potentials in rat neocortical and hippocampal slices. *Neurosci.*, 92:1192–1215.
- [Ulanovsky et al., 2003] Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neurosci.*
- [Viemeister, 1979] Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am.*, 66(5):1364–1380.
- [Wang and Shamma, 1994] Wang, K. and Shamma, S. A. (1994). Self-normalization and noise-robustness in early auditory representations. *IEEE Trans. Speech and Audio Proc.*, 2(3):421–435.
- [Westerman and Smith, 1984] Westerman, L. A. and Smith, R. L. (1984). Rapid and short term adaptation in auditory nerve responses. *Hearing Research*, 15:249–260.
- [Yang et al., 1992] Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Trans. Inform. Theory*, 38(2):824–839.

Figure Captions

Figure 1

Details of the dynamic ripple stimulus and examples of spectrotemporal response fields (STRFs) in primary auditory cortex (A1). (a) 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 octaves (e.g., .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). (b) Example STRFs recorded from A1 of the ferret. Red (Blue) 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, vi) inhibition, and direction-selectivity; upward in (iii), downward in (vi).

Figure 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}).

Figure 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 pink 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 pink noise (left) and spectro-temporally modulated in-phase harmonic series (right). All stimuli are sampled at 16 kHz.

Figure 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 inter-harmonic interactions occur at a rate that reflects the fundamental (100 Hz) of the

signal. This is clearly shown by the red 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 ($> 1500\text{Hz}$).

Figure 5

A representative STRF and the seed functions of the spectrotemporal multiresolution cortical processing model. **(a)** An example of an STRF in the model. It is upward selective and tuned to (1 cyc/oct, 16 Hz). **(b)** Seed functions (non-causal h_s and causal h_t) of the model. The abscissa of each figure is normalized to correspond to the tuning scale of 1 cyc/oct or rate of 1 Hz.

Figure 6

Examples of cortical representations for stimuli as in Figure 3. (a) A three-tone (250, 1000, 4000 Hz) combination; (b) A broadband pink noise; (c) Broadband in-phase harmonic complex; (d) Ripples. For each of these stationary stimuli, the 4 dimensional representation $|z_{\downarrow}|, |z_{\uparrow}|$ is first integrated over time to generate a 3 dimensional representation. For the three-tone combinations, each of the remaining three variable (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 pink 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 cyc/oct. In each plot, the negative (positive) rate denotes upward (downward) moving direction.

Figure 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 inter-syllable time interval of about 120-180 ms (between the first and second syllables - /we've/ and /done/). They also peak at 2 cyc/octave because most of the spectral energy occurs near the 2nd and 3rd harmonics (which are separated by about .5 octave).

Figure 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.

Figure 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.

Figure 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.

Figure 11

The response of the representative cochlear filter with characteristic frequency around 1 kHz. (a) Magnitude response in dB on the tonotopic axis. (b) Linear frequency response: magnitude (dashed), real (solid) and imaginary (dotted) part. (c) Phase response. (d) Impulse response.

Figure 12

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

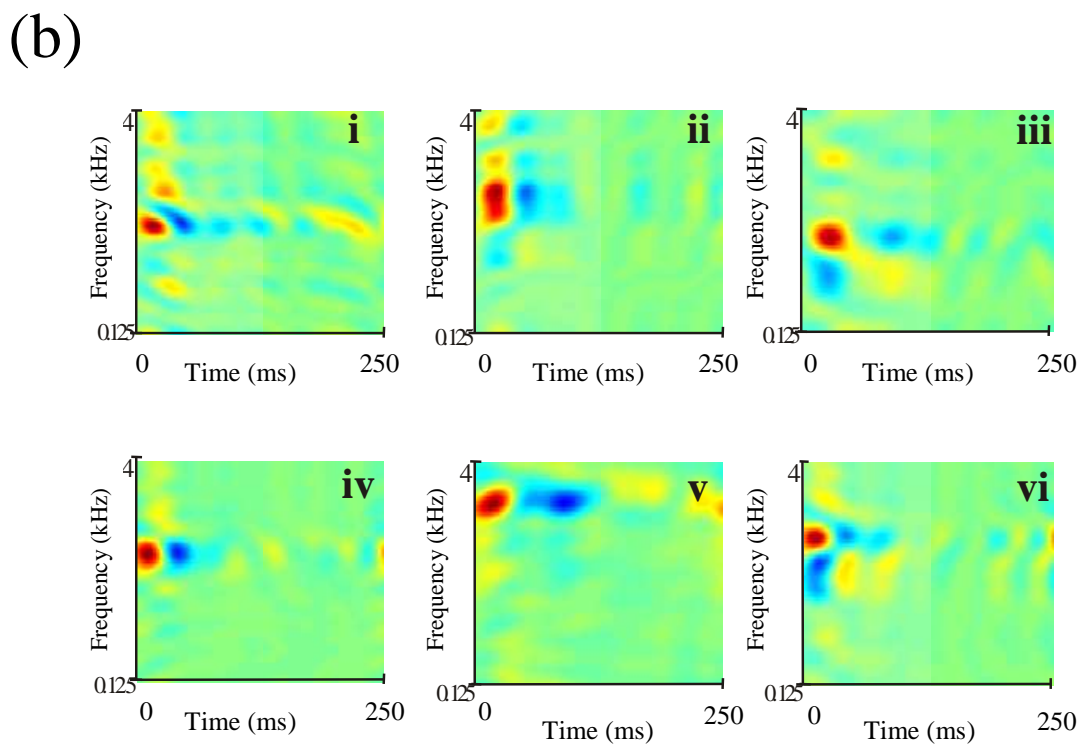
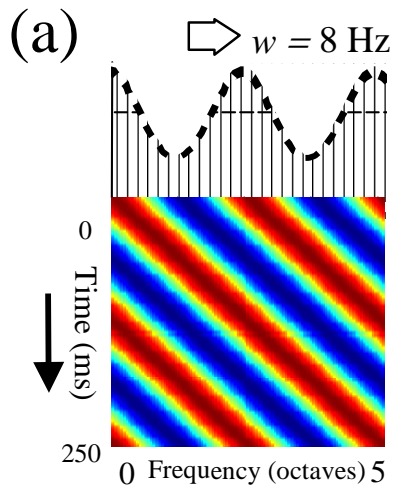


Figure 1

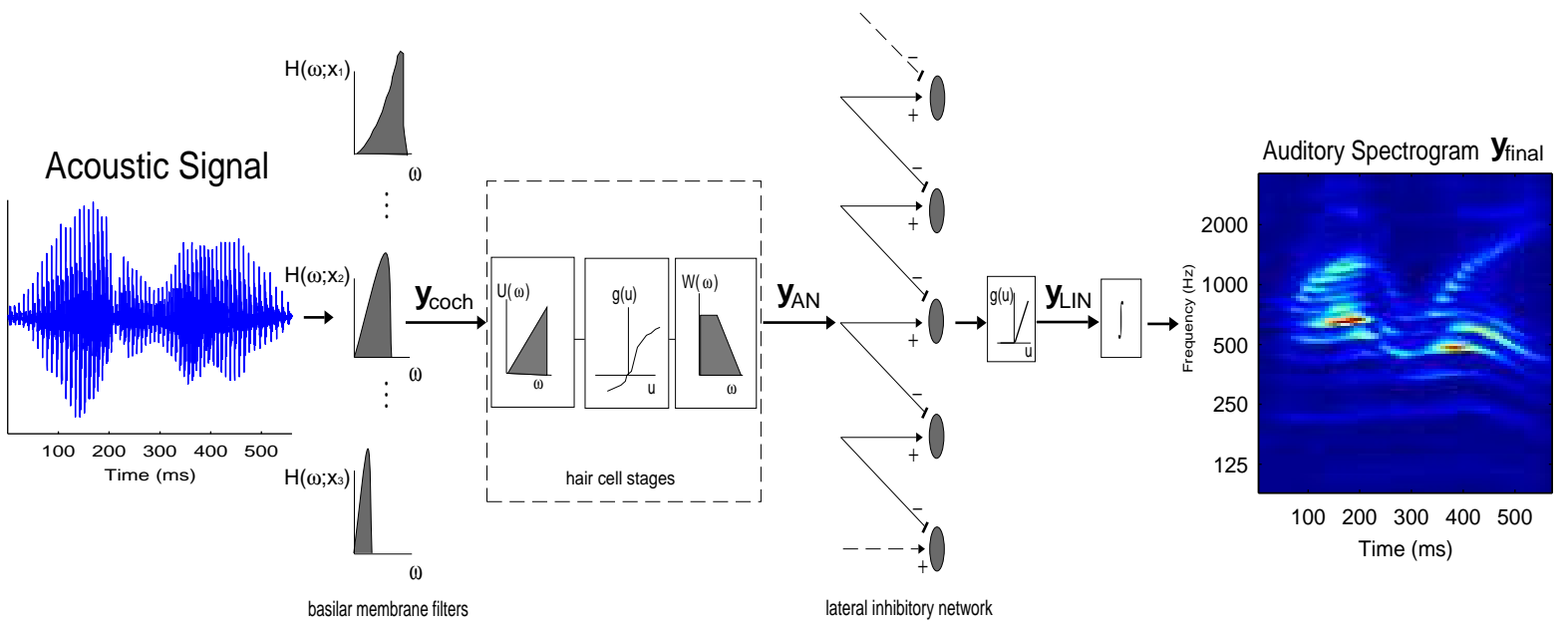
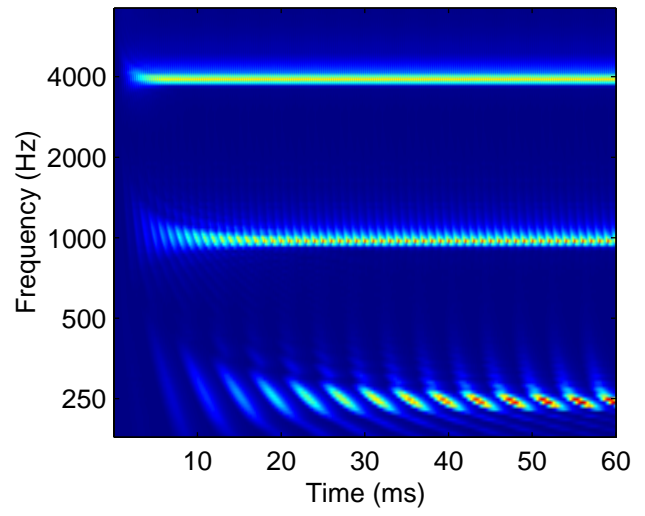
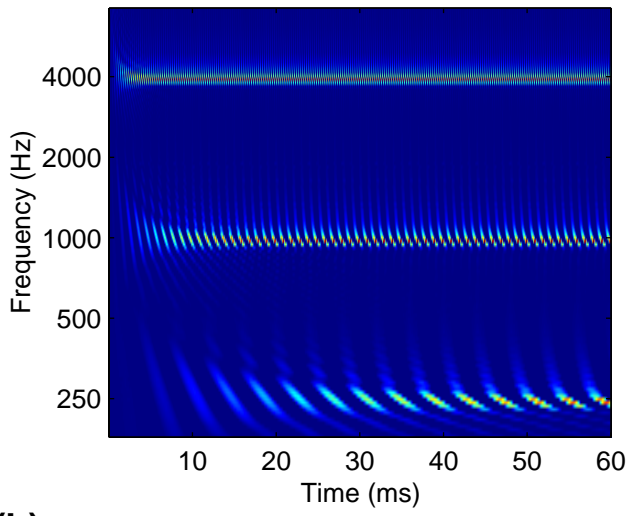
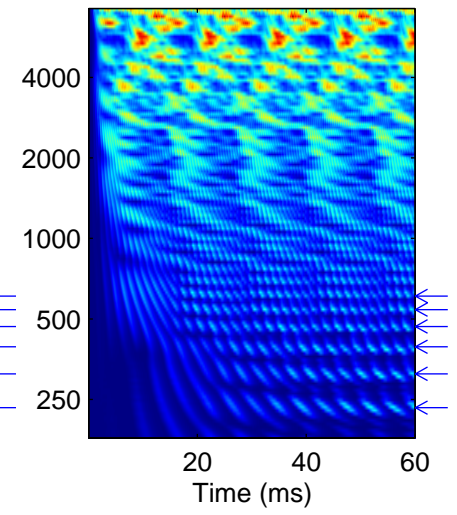
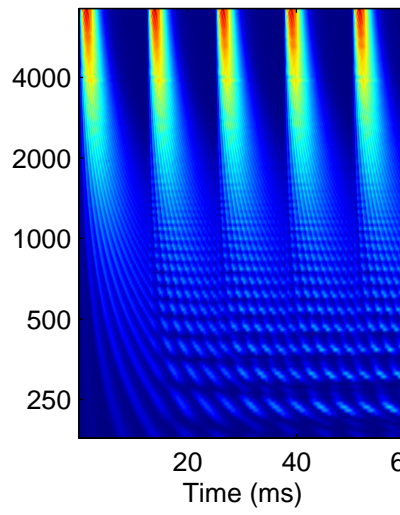
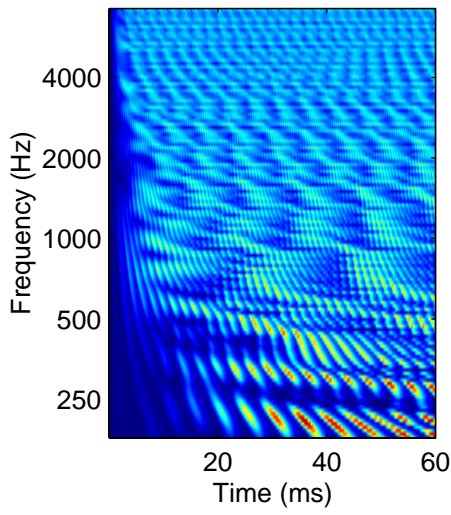


Figure 2

(a)



(b)



(c)

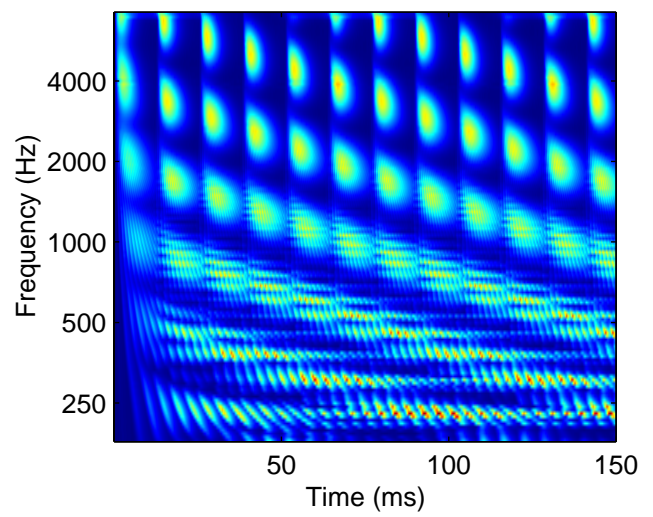
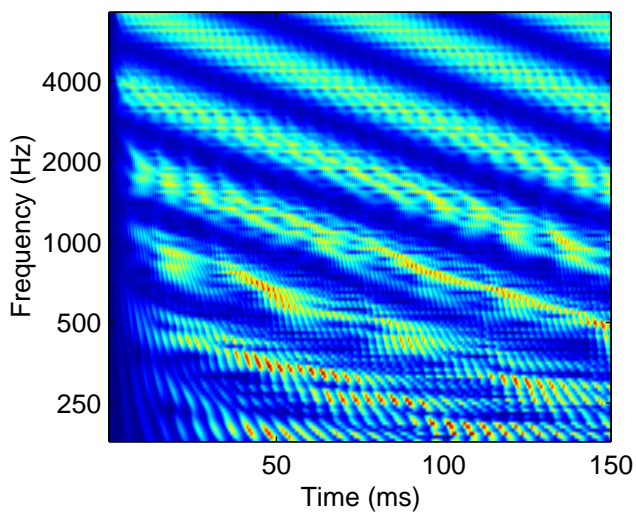


Figure 3

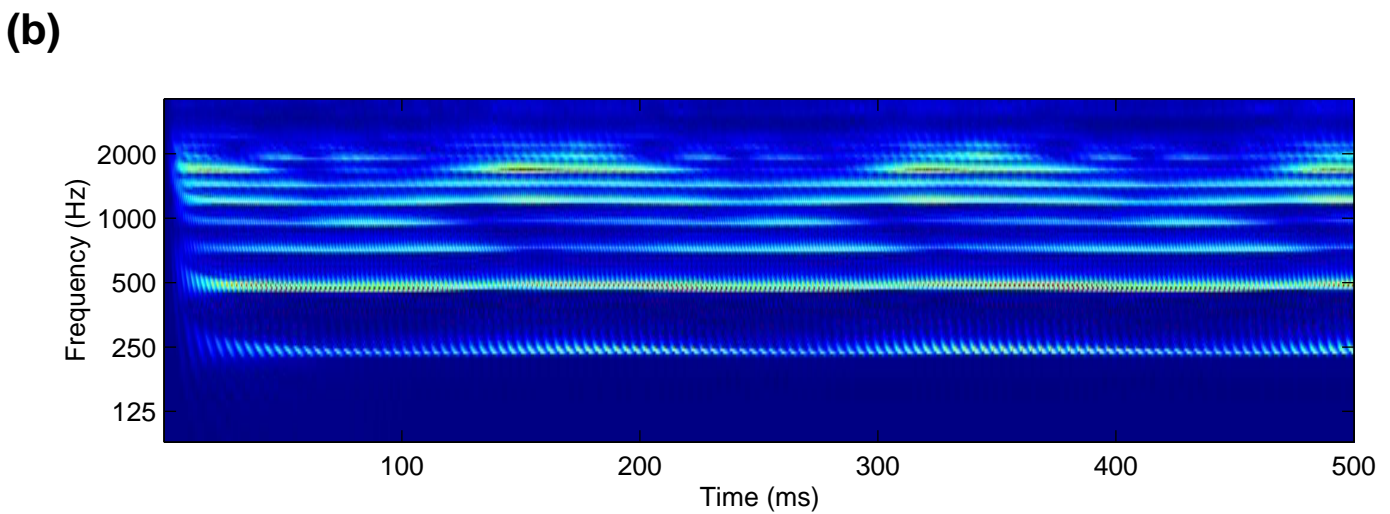
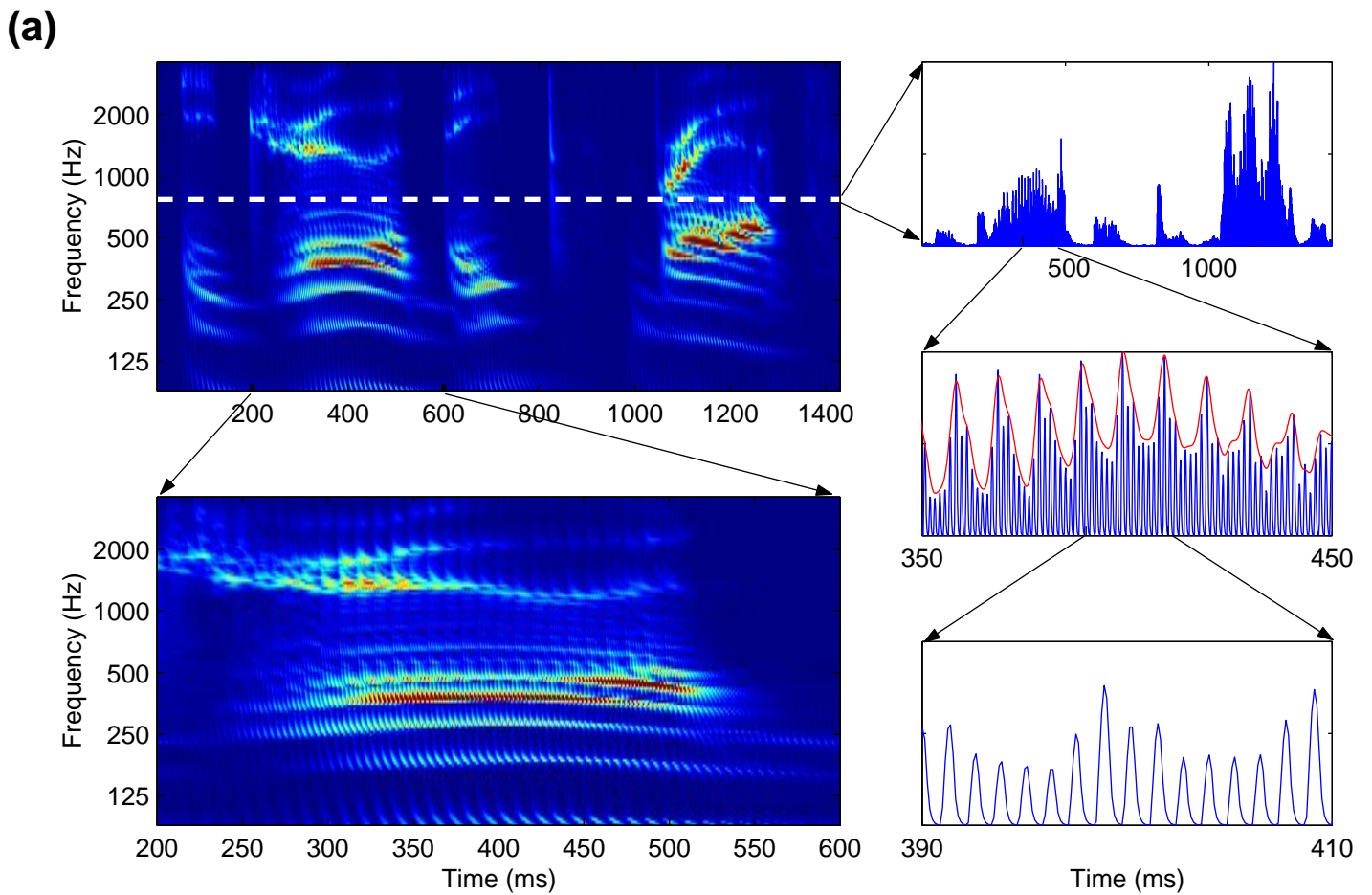


Figure 4

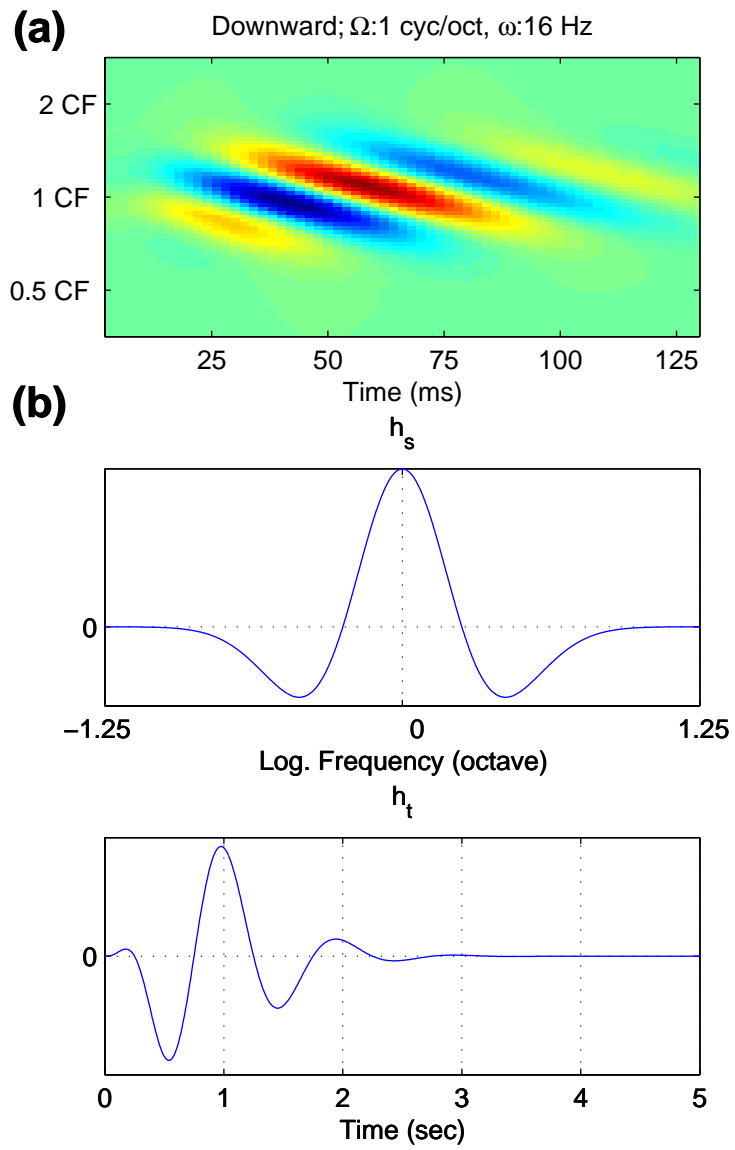


Figure 5

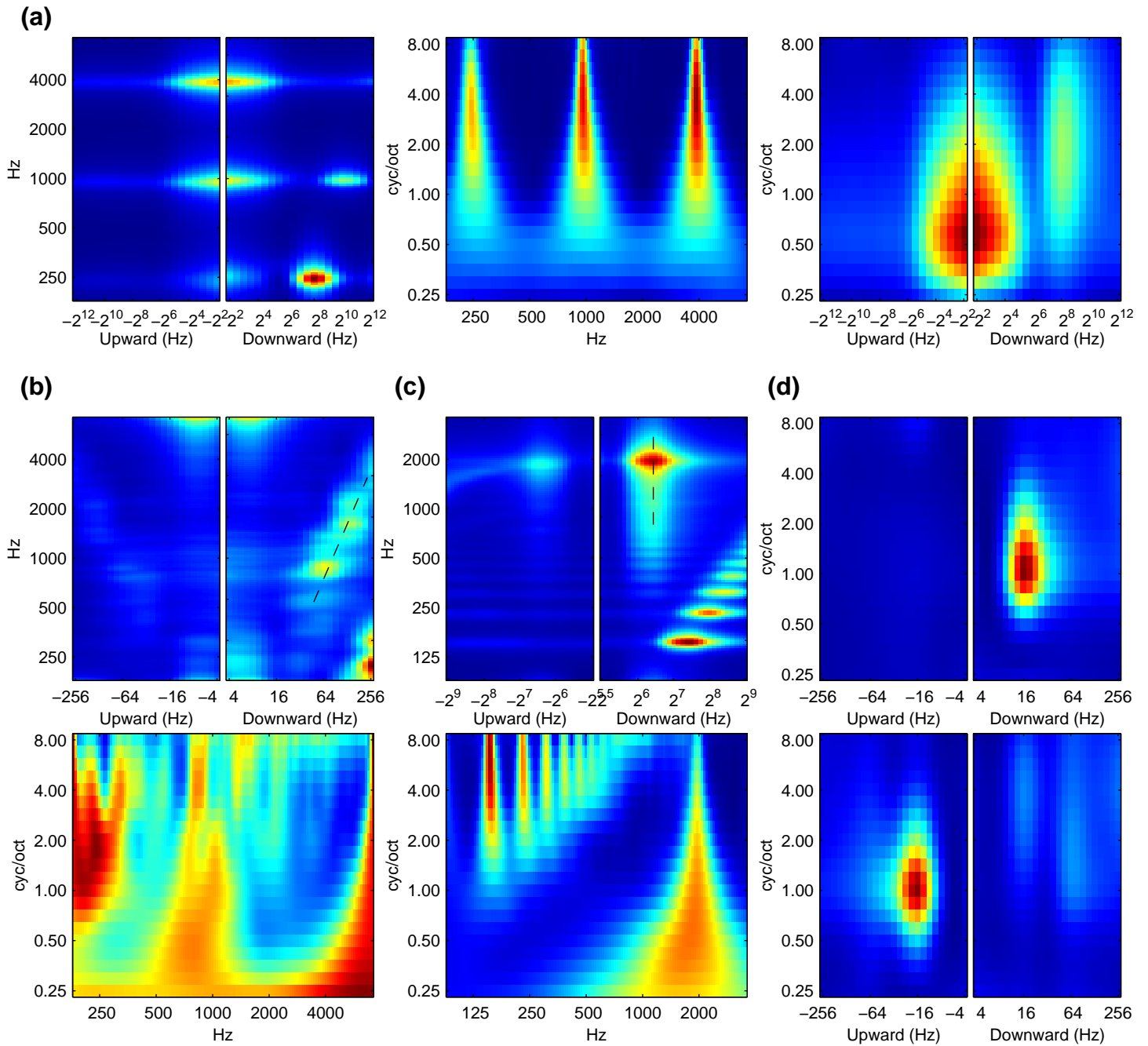


Figure 6

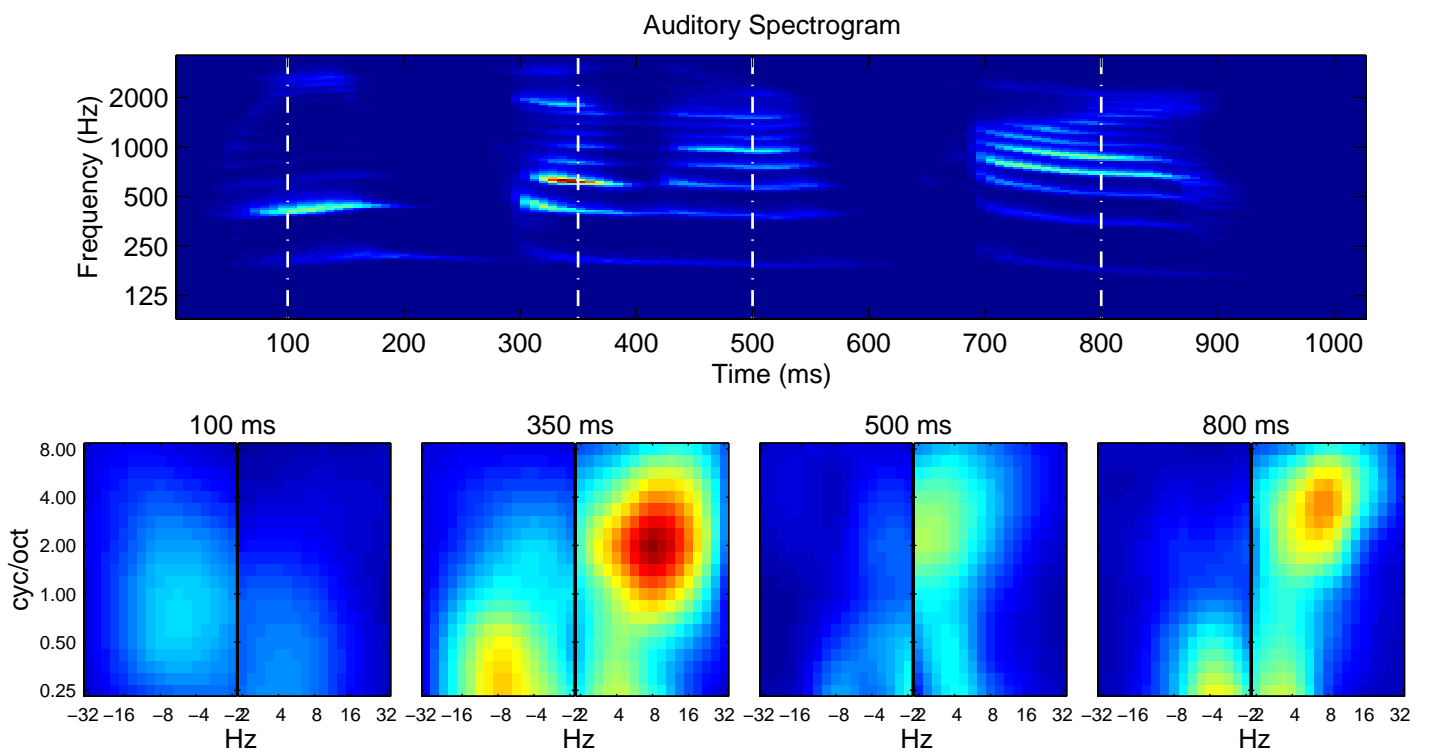


Figure 7

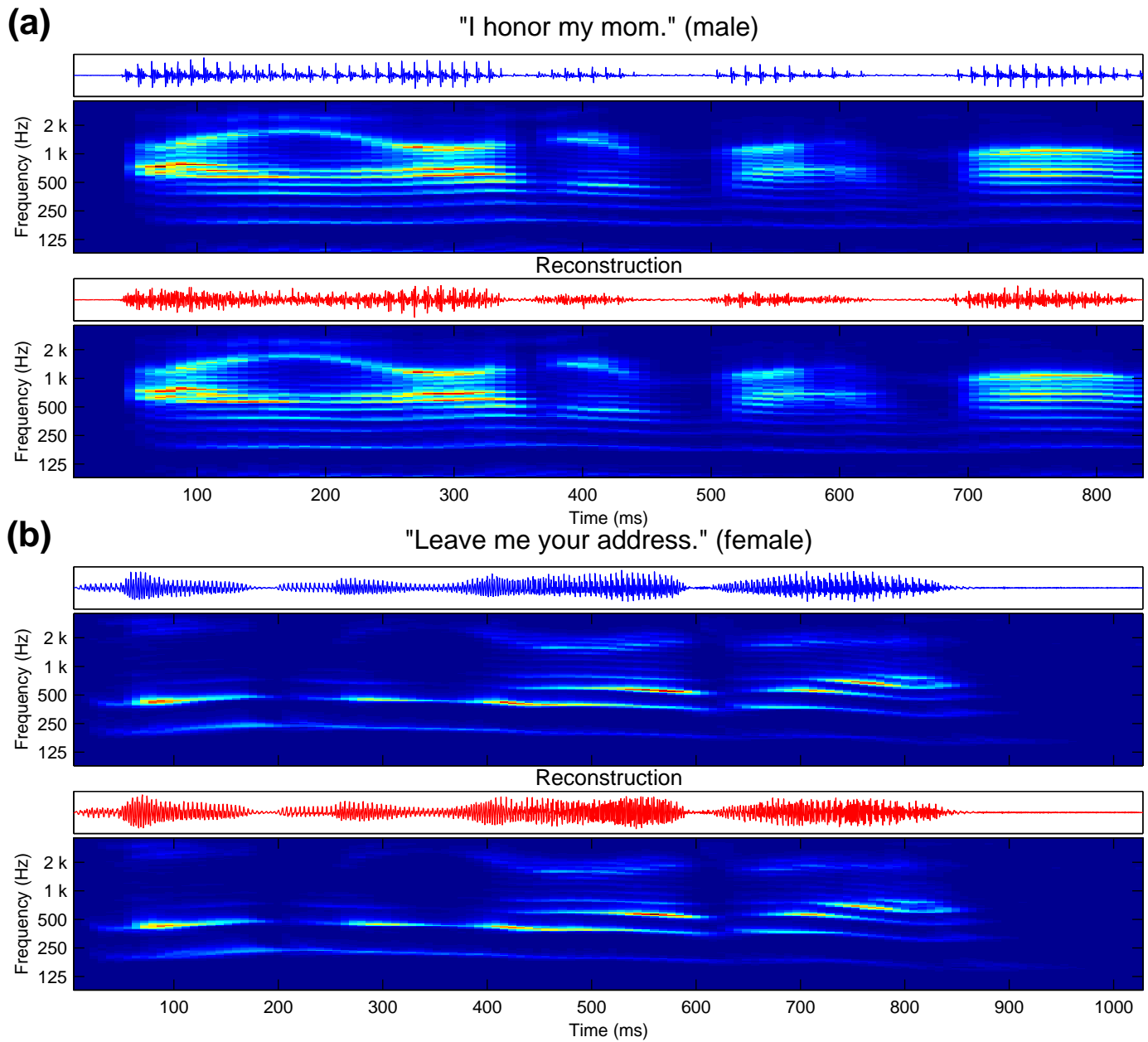


Figure 8

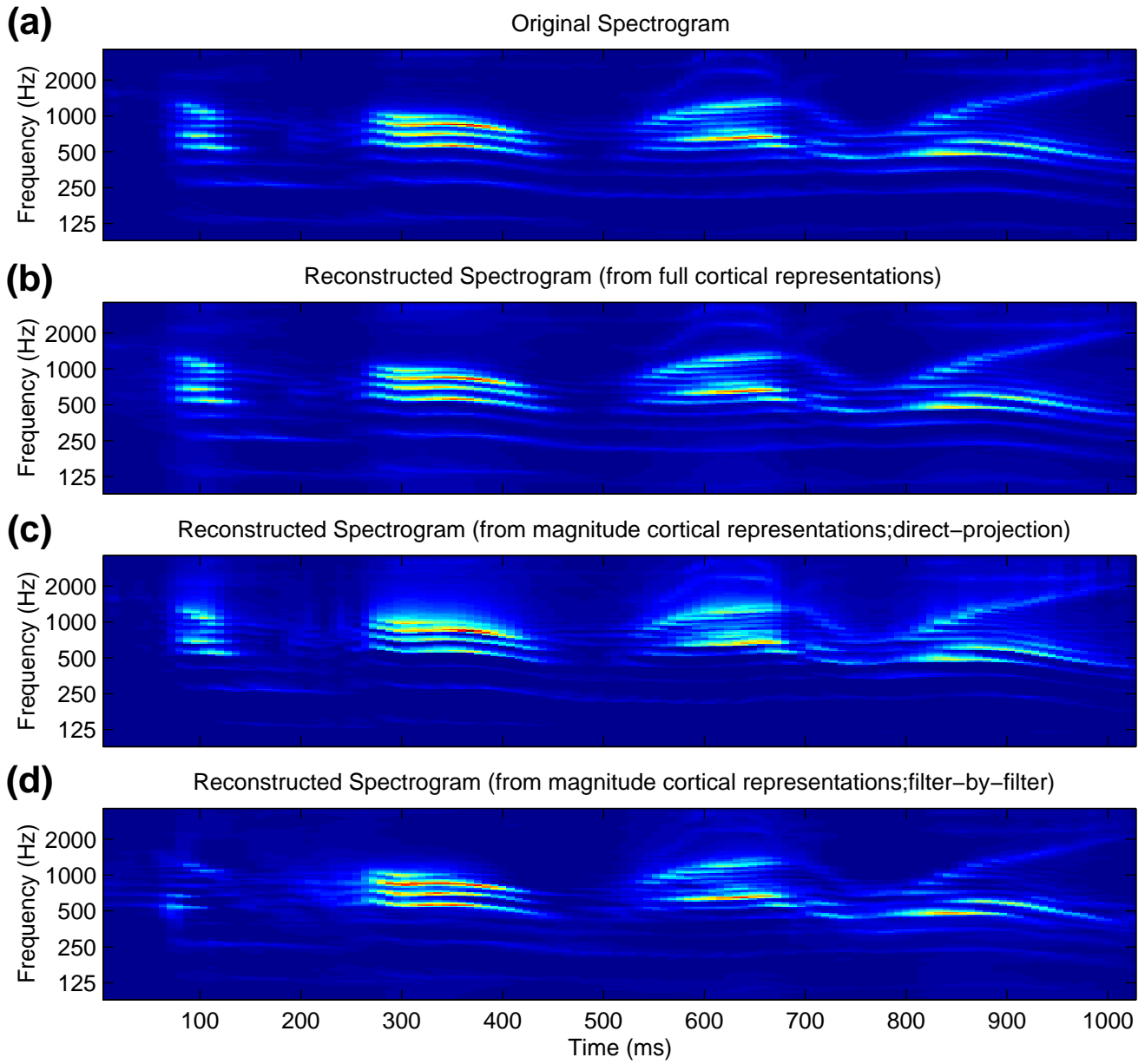


Figure 9

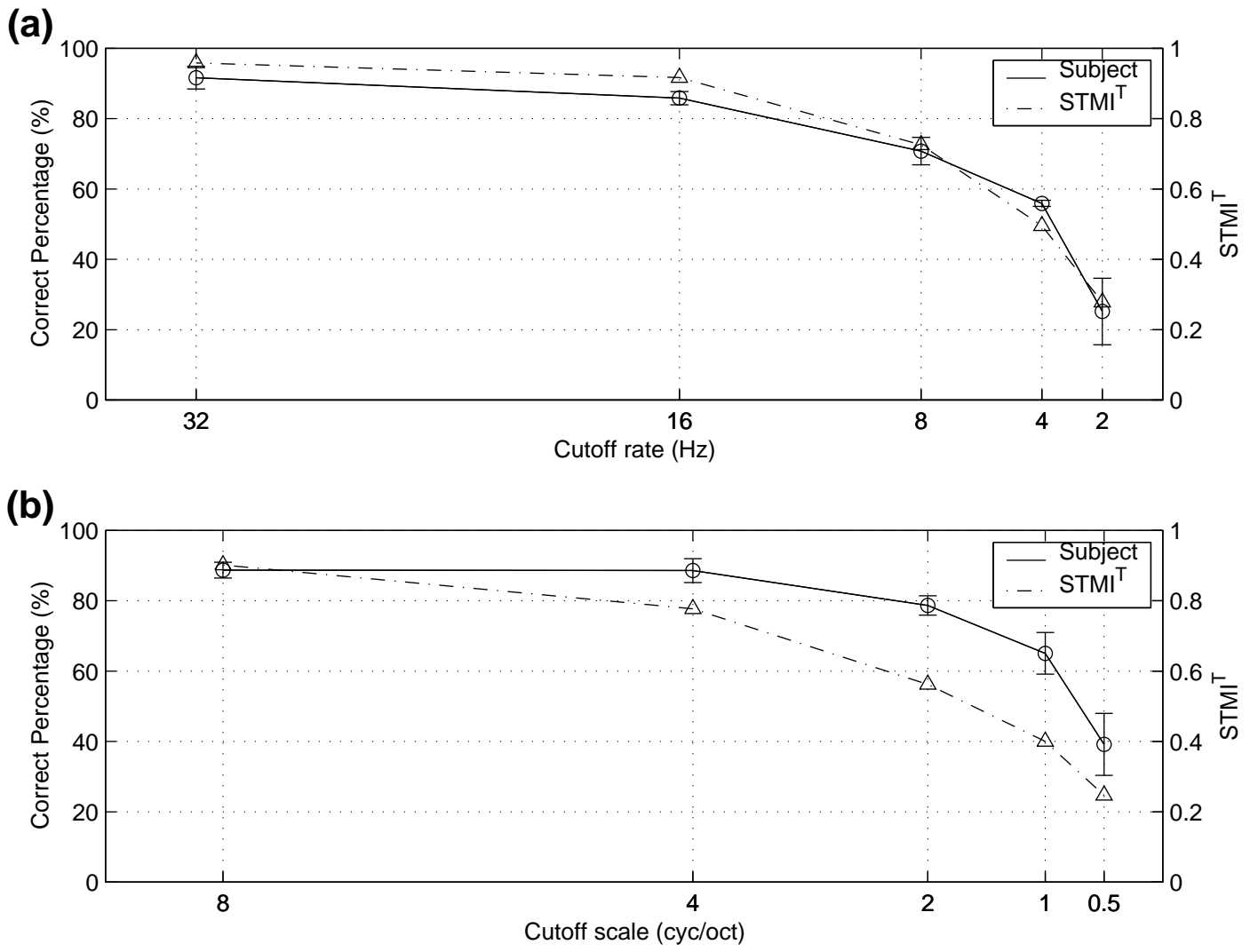


Figure 10

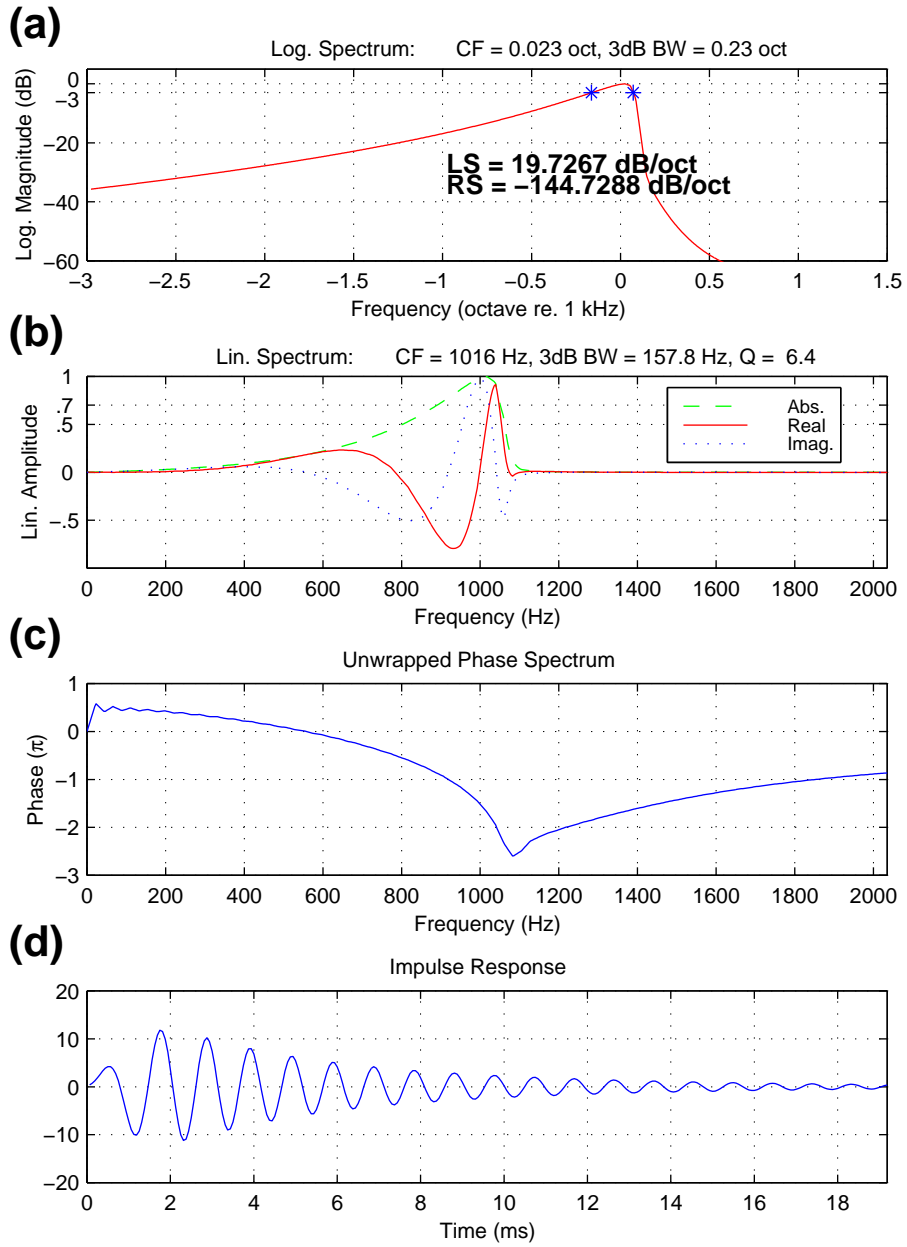
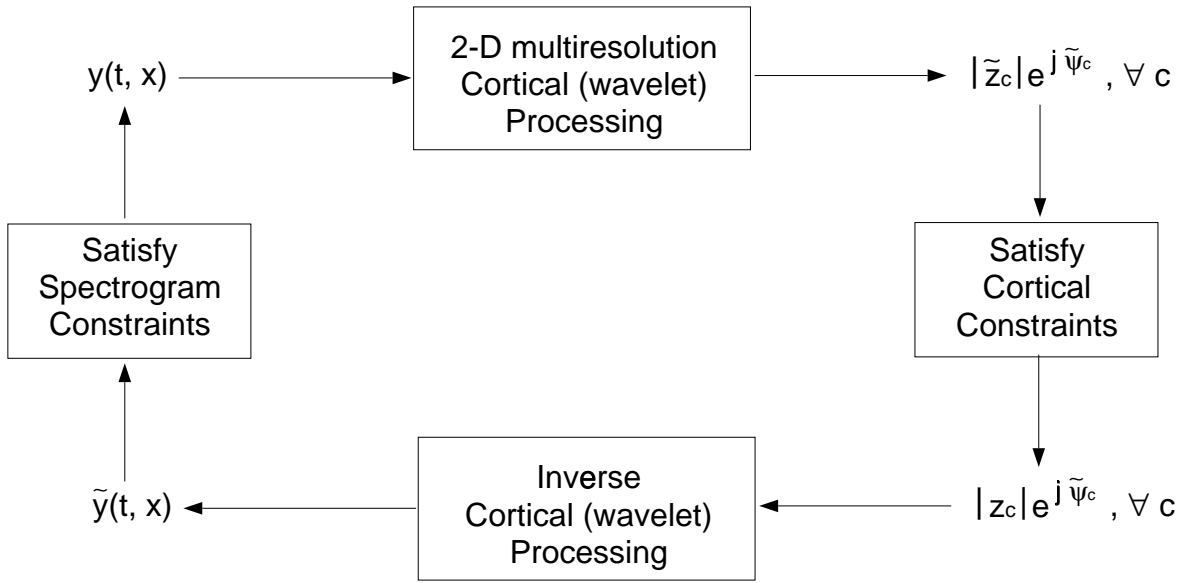


Figure 11

(a)



(b)

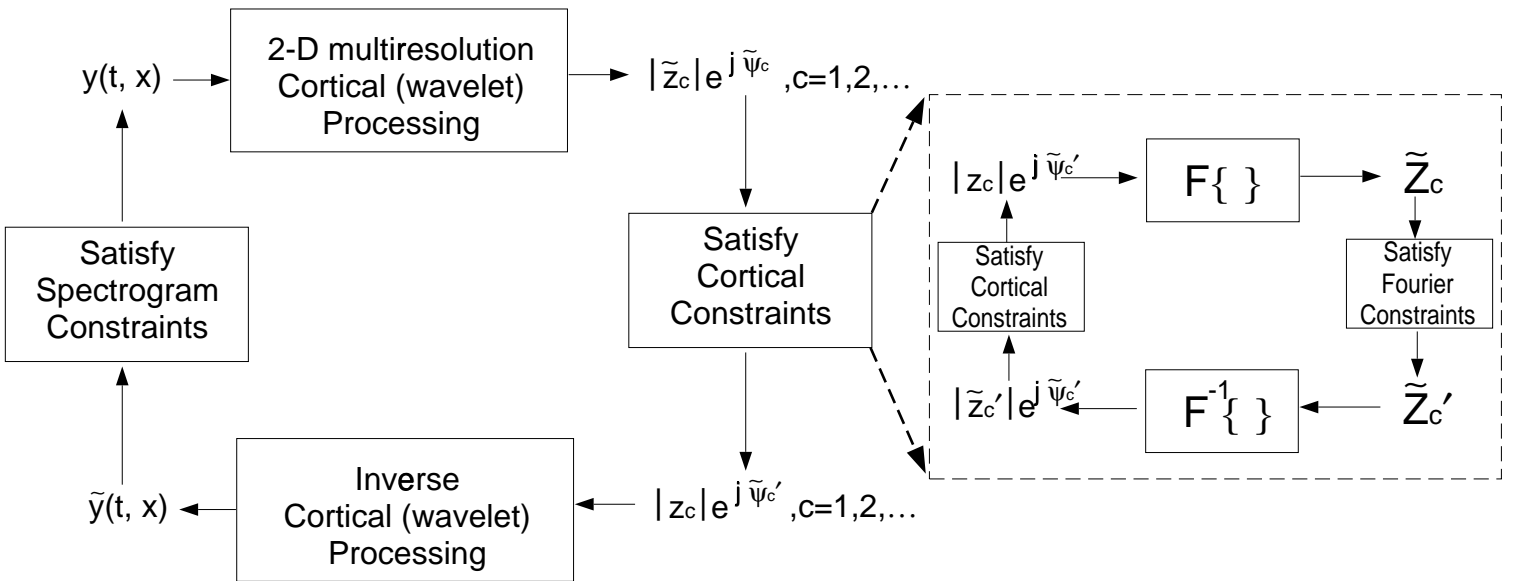


Figure 12