EURASIP Journal on Applied Signal Processing 2005:9, 1350–1364

Neuromimetic Sound Representation for Percept Detection and Manipulation

Dmitry N. Zotkin
Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD 20742, USA
Email: dz@umiacs.umd.edu

Taishih Chi
Neural Systems Laboratory, The Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
Email: tschi@isr.umd.edu

Shihab A. Shamma
Neural Systems Laboratory, The Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
Email: sas@eng.umd.edu

Ramani Duraiswami
Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD 20742, USA
Email: ramani@umiacs.umd.edu

Received 2 November 2003; Revised 4 August 2004

The acoustic wave received at the ears is processed by the human auditory system to separate different sounds along the intensity, pitch, and timbre dimensions. Conventional Fourier-based signal processing, while endowed with fast algorithms, is unable to easily represent a signal along these attributes. In this paper, we discuss the creation of maximally separable sounds in auditory user interfaces and use a recently proposed cortical sound representation, which performs a biomimetic decomposition of an acoustic signal, to represent and manipulate sound for this purpose. We briefly overview algorithms for obtaining, manipulating, and inverting a cortical representation of a sound and describe algorithms for manipulating signal pitch and timbre separately. The algorithms are also used to create sound of an instrument between a "guitar" and a "trumpet." Excellent sound quality can be achieved if processing time is not a concern, and intelligible signals can be reconstructed in reasonable processing time (about ten seconds of computational time for a one-second signal sampled at 8 kHz). Work on bringing the algorithms into the real-time processing domain is ongoing.

Keywords and phrases: anthropomorphic algorithms, pitch detection, human sound perception.

1. INTRODUCTION

When a natural sound source such as a human voice or a musical instrument produces a sound, the resulting acoustic wave is generated by a time-varying excitation pattern of a possibly time-varying acoustical system, and the sound characteristics depend both on the excitation signal and on the production system. The production system (e.g., human vocal tract, the guitar box, or the flute tube) has its own characteristic response. Varying the excitation parameters produces a sound signal that has different frequency components, but still retains perceptual characteristics that uniquely identify the production instrument (identity of the person, type of instrument—piano, violin, etc.), and even the specific type of piano on which it was produced. When one is asked to characterize this sound source using descriptions based on Fourier analysis, one discovers that concepts such as frequency and amplitude are insufficient to explain such perceptual characteristics of the sound source. Human linguistic descriptions that characterize the sound are expressed in terms of pitch and timbre. The goal of anthropomorphic algorithms is to reproduce these percepts quantitatively.

The perceived sound pitch is closely coupled with its harmonic structure and frequency of the first harmonic, or F_0. On the other hand, the timbre of the sound is defined broadly as everything other than the pitch, the loudness, and the spatial location of the sound. For example, two musical instruments might have the same pitch if they play the same note,

but it is their differing timbre that allows us to distinguish between them. Specifically, the spectral envelope and the spectral envelope variations in time that include, in particular, onset and offset properties of the sound are related to the timbre percept.

Most conventional techniques of sound manipulation result in simultaneous changes in both the pitch and the timbre and cannot be used to control or assess the effects in pitch and timbre dimensions independently. A goal of this paper is the development of controls for independent manipulation of pitch and timbre of a sound source using a *cortical sound representation* introduced in [2], where it was used for assessment of speech intelligibility and for prediction of the cortical response to an arbitrary stimulus, and later extended in [3] providing fuller mathematical details as well as addressing invertibility issues. We simulate the multiscale audio representation and processing believed to occur in the primate brain [4], and while our sound decomposition is partially similar to existing pitch and timbre separation and sound morphing algorithms (in particular, MFCC decomposition algorithm in [5], sinusoid-plus-noise model and effects generated with it in [6], and parametric source models using LPC and physics-based synthesis in [7]), the neuromorphic framework provides a view of processing from a different perspective, supplies supporting evidence to justify the procedure performed, tailors it to the way the human nervous system processes auditory information, and extends the approach to include decomposition in the time domain in addition to frequency. We anticipate our algorithms to be applicable in several areas, including musical synthesis, audio user interfaces, and sonification.

In Section 2, we discuss the potential applications for the developed framework. In Sections 3 and 4, we describe the processing of the audio information through the cortical model [3] in forward and backward directions, respectively, and in Section 5, we propose an alternative, faster implementation of the most time-consuming cortical processing stage. We discuss the quality of audio signal reconstruction in Section 6 and show examples of timbre-preserving pitch manipulation of speech and timbre interpolation of musical notes in Sections 7 and 8, respectively. Finally, Section 9 concludes the paper.

2. APPLICATIONS

The direct application that motivated us to undertake the research described (and the area it is currently being used in) is the development of advanced auditory user interfaces. Auditory user interfaces can be broadly divided into two groups, based on whether speech or nonspeech audio signals are used in the interface. The field of sonification [8] ("... use of nonspeech audio to convey information") presents multiple challenges to researchers in that they must both identify and manipulate different percepts of sound to represent different parameters in a data stream while at the same time creating efficient and intuitive mappings of the data from the numerical domain to the acoustical domain. An extensive resource describing sonification work is the International Community for Auditory Display (ICAD) web page (see http://www.icad.org/), which includes past conference proceedings. While there are some isolated examples of useful sonifications and attempts at creating multidimensional audio interfaces (e.g., the Geiger counter or the pulse oximeter [9]), the field of sonification, and as a consequence audio user interfaces, is still in the infancy due to the lack of a comprehensive theory of sonification [10].

What is needed for advancements in this area are identification of perceptually valid attributes ("dimensions") of sound that can be controlled; theory and algorithms for sound manipulation that allow control of these dimensions; psychophysical proof that these control dimensions convey information to a human observer; methods for easy-to-understand data mapping to auditory domain; technology to create user interfaces using these manipulations; and refinement of acoustic user interfaces to perform some specific example tasks. Our research addresses some of these issues and creates the basic technology for manipulation of existing sounds and synthesis of new sounds achieving specified attributes along the perceptual dimensions. We focus on neuromorphic-inspired processing of pitch and timbre percepts, having the location and ambience percepts described earlier in [11]. Our real-time pitch-timbre modification and scene rendering algorithms are capable of generating stable virtual acoustic objects whose attributes can be manipulated in these perceptual dimensions.

The same set of percepts may be modified in the case when speech signals are used in audio user interfaces. However, the purpose of percept modification in this case is not to convey information directly but rather to allow for maximally distinguishable and intelligible perception of (possibly several simultaneous) speech streams under stress conditions using the natural neural auditory dimensions. Applications in this area might include, for example, an audio user interface for a soldier where multiple sound streams are to be attended to simultaneously. To our knowledge, much research has been devoted to selective attention to one signal from a group [12, 13, 14, 15, 16] (the well-known "cocktail party effect" [17]), and there have only been a limited number of studies (e.g., [18, 19]) on how well a person can simultaneously perceive and understand multiple concurrent speech streams. The general results obtained in these two papers suggest that increasing separation along most of the perceptual characteristics leads to improvement in the recognition rate for several competing messages. The characteristic that provides the most improvement is the spatial separation of the sounds, which is beyond the scope of this paper; these spatialization techniques are well described in [11]. Pitch was a close second, and in Section 7 of this paper, we present a cortical-representation-based pitch manipulation algorithm that can be used to achieve the desired perceptual separation of the sounds. Timbre manipulations did not result in significant improvements in recognition rate in [18, 19], though.

Another area where we anticipate our algorithms to be applicable to is musical synthesis. Synthesizers often use sampled sound that has to be pitch shifted to produce different notes [7]. Simple resampling that was widely used in the

past in commercial-grade music synthesizers preserves neither the spectral nor the temporal envelope (onset and decay ratios) of an instrument. More recent wavetable synthesizers can impose the correct temporal envelope on the sound but may still distort the spectral envelope. The spectral and the temporal envelopes are parts of the timbre percept, and their incorrect manipulation can lead to poor perceptual quality of the resulting sound samples.

The timbre of the instrument usually depends on the size and the shape of the resonator; it is interesting that for some instruments (piano, guitar), the resonator shape (which determines the spectral envelope of the produced sound) does not change when different notes are played, and for others (flute, trumpet), the length of resonating air column changes as the player opens different holes in the tube to produce different notes. Timbre-preserving pitch modification algorithm described in Section 7 provides a physically correct pitch manipulation technique for instruments with the resonator shape independent of the note played. It is also possible to perform timbre interpolation between sound samples; in Section 8, we describe the synthesis of a new musical instrument with the perceptual timbre lying in between two known instruments—the guitar and the trumpet. The synthesis is performed in the timbre domain, and then a timbre-preserving pitch shift described in Section 7 is applied to form different notes of the new instrument. Both operations use the cortical representation, which turned out to be extremely useful for separate manipulations of percepts.

3. THE CORTICAL MODEL

In a complex acoustic environment, sources may simultaneously change their loudness, location, timbre, and pitch. Yet, humans are able to integrate effortlessly the multitude of cues arriving at their ears and derive coherent percepts and judgments about each source [20]. The cortical model is a computational model for how the brain is able to obtain these features from the acoustic input it receives. Physiological experiments have revealed the elegant multiscale strategy developed in the mammalian auditory system for coding of spectro-temporal characteristics of the sound [4, 21]. The primary auditory cortex (AI), which receives its input from the thalamus, employs a multiscale representation in which the dynamic spectrum is repeatedly represented in AI at various degrees of spectral and temporal resolutions. This is accomplished by cells whose responses are selective to a range of spectro-temporal parameters such as the local bandwidth, the symmetry, and onset and offset transition rates of the spectral peaks. Similarly, psychoacoustical investigations have shed considerable light on the way we form and label sound images based on relationships among their physical parameters [20]. A mathematical model of the early and the central stages of auditory processing in mammals was recently developed and described in [2, 3]. It is a basis for our work and is briefly summarized here; a full formulation of the model is available in [3] and analysis code in the form of a Matlab toolbox ("NSL toolbox") can be downloaded from http://www.isr.umd.edu/CAAR/pubs.html.

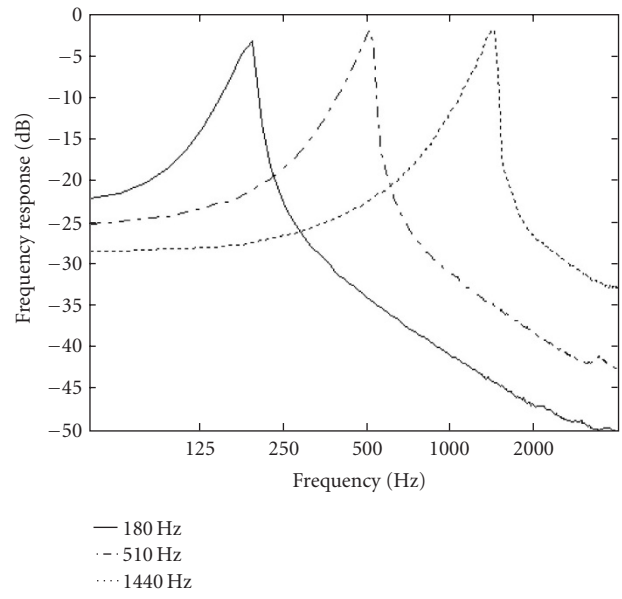


Figure 1: Tuning curves for cochlear filter bank filters tuned at 180 Hz, 510 Hz, and 1440 Hz (channels 24, 60, and 96), respectively.

The model consists of two basic stages. The first stage of the model is an early auditory stage, which models the transformation of an acoustic signal into an internal neural representation, called the "auditory spectrogram." The second is a central stage, which analyzes the spectrogram to estimate its spectro-temporal features, specifically its spectral and temporal modulations, using a bank of modulation selective filters mimicking those described in the mammalian primary auditory cortex.

The first processing stage converts the audio signal $s(t)$ into an auditory spectrogram representation $y(t, x)$ (where x is the frequency on a logarithmic frequency axis) and consists of a sequence of three steps described below.

(i) In the analysis step, the acoustic wave creates a complex pattern of mechanical vibrations on a basilar membrane in mammalian cochlea. For an acoustic tone of a given frequency, the amplitude of the traveling wave induced in the membrane slowly increases along it up to a certain point x and then sharply decreases. The position of the point x depends on the frequency, with different frequencies resonating at different points along the membrane. These maximum response points create a tonotopical frequency axis with frequencies approximately logarithmically decreasing from the base of the cochlea. This process is simulated by a *cochlear filter bank*—a bank of highly asymmetric constant Q bandpass filters (also called *channels*) spaced equally over the log-frequency axis; we denote the impulse response of each filter by $h(t; x)$. There are 128 channels with 24 channels per octave covering a total of 5(1/3) octaves with the lowest channel frequency of 90 Hz in the implementation of the model that we use, and equivalent rectangular bandwidth (ERB) filter quality $Q_{\text{ERB}} \approx 4$. Figure 1 shows the frequency-response curves of a few cochlear filters.

(ii) In the transduction step, the mechanical vibrations of the membrane are transduced into the intracellular potential of the inner hair cells. Membrane displacements cause the flow of liquid in the cochlea to bend the *cilia* (tiny hair-like formations) that are attached to the inner hair cells. This bending opens the cell channels and enables ionic current to flow into the cell and to change its electric potential, which is later transmitted by auditory nerve fibers to the cochlear nucleus. In the model, these steps are simulated by a highpass filter (equivalent to taking a time-derivative operation), nonlinear compression $g(z)$, and the lowpass filter $w(t)$ with cut-off frequency of 2 kHz, representing the fluid-cilia coupling, ionic channel current, and hair cell membrane leakage, respectively.

(iii) Finally, in the reduction step, the input to the anteroventral cochlear nucleus undergoes lateral inhibition operation followed by envelope detection. Lateral inhibition effectively enhances the frequency selectivity of the cochlear filters from $Q \approx 4$ to $Q \approx 12$ and is modeled by a spatial derivative across the channel array. Then, the nonnegative response of the lateral inhibitory network neurons is modeled by a half-wave rectifier, and an integration over a short window, $\mu(t;\tau) = e^{-t/\tau}$, with $\tau = 8$ milliseconds, is performed to model the slow adaptation of the central auditory neurons.

In mathematical form, the three steps described above can be expressed as

$$\begin{aligned} y_1(t,x) &= s(t) \oplus h(t;x), \\ y_2(t,x) &= g\big(\partial_t y_1(t,x)\big) \oplus w(t), \\ y(t,x) &= \max\big(\partial_x y_2(t,x), 0\big) \oplus \mu(t;\tau), \end{aligned} \tag{1}$$

where $\oplus$ denotes a convolution with respect to t.

The above sequence of operations essentially consists of a bank of constant Q filters with some additional operations and efficiently computes the time-frequency representation of the acoustic signal, which is called the auditory spectrogram (Figure 2). The auditory spectrogram is invertible through an iterative process (described in the next section); perceptually perfect inversion can be achieved, albeit at a very significant computational expense. A time slice of the spectrogram is called the auditory spectrum.

The second processing stage mimics the action of the higher central auditory stages (especially the primary auditory cortex). We provide a mathematical derivation (as presented in [3]) of the cortical representation below, as well as qualitatively describe the processing.

The existence of a wide variety of neuron spectro-temporal response fields (SRTF) covering a range of frequency and temporal characteristics [21] suggests that they may, as a population, perform a multiscale analysis of their input spectral profile. Specifically, the cortical stage estimates the spectral and temporal modulation content of the auditory spectrogram using a bank of modulation selective filters $h(t,x;\omega,\Omega,\varphi,\theta)$. Each filter is tuned ($Q = 1$) to a combination of a particular spectral modulation and a particular temporal modulation of the incoming signal, and filters are centered at different frequencies along the tonotopical axis.

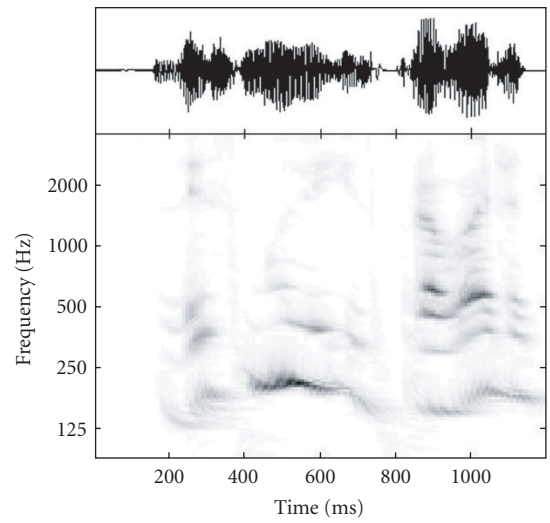


Figure 2: Example auditory spectrogram for the sentence "This movie is provided …".

These two types of modulations are defined as follows.

(i) Temporal modulation, which defines how fast the signal energy is increasing or decreasing along the time axis at a given time and frequency. It is characterized by the parameter ω, which is referred to as rate or velocity and is measured in Hz, and by characteristic temporal modulation phase φ.

(ii) Spectral modulation, which defines how fast the signal energy varies along the frequency axis at a given time and frequency. It is characterized by the parameter Ω, which is referred to as density or scale and is measured in cycles per octave (CPO), and by characteristic spectral modulation phase θ.

The filters are designed for a range of rates from 2 to 32 Hz and scales from 0.25 to 8 CPO, which corresponds to the ranges of neuron spectro-temporal response fields found in the primate brain. The impulse response function for the filter $h(t,x;\omega,\Omega,\varphi,\theta)$ can be factored into $h_s(x;\Omega,\theta)$-spectral and $h_t(t;\omega,\varphi)$-temporal parts, respectively. The spectral impulse response function $h_s(x;\Omega,\theta)$ is defined through a phase interpolation of the spectral filter seed function $u(x;\Omega)$ with its Hilbert transform $\bar{u}(x;\Omega)$, and the temporal impulse response function is similarly defined via the temporal filter seed function $v(t;\omega)$:

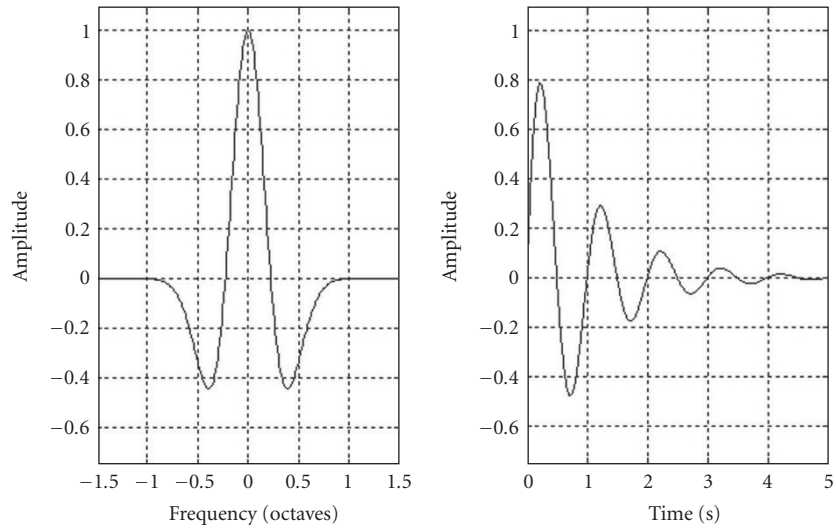
$$\begin{aligned} h_s(x;\Omega,\theta) &= u(x;\Omega)\cos\theta + \bar{u}(x;\Omega)\sin\theta, \\ h_t(t;\omega,\varphi) &= v(t;\omega)\cos\varphi + \bar{v}(t;\omega)\sin\varphi. \end{aligned} \tag{2}$$

The Hilbert transform is defined as

$$\bar{f}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{f(z)}{z-x}dz. \tag{3}$$

We choose

$$u(x) = \big(1-x^2\big)e^{-x^2/2}, \qquad v(t) = e^{-t}\sin(2\pi t) \tag{4}$$



(a) (b)

Figure 3: Tuning curves for the basis (seed) filter for the rate-scale decomposition. The seed filter is tuned to the rate of 1 Hz and the scale of 1 CPO. (a) Spectral response. (b) Temporal response.

as the functions that produce the basic seed filter tuned to a scale of 1 CPO and a rate of 1 Hz. Figure 3 shows its spectral and temporal responses generated by functions $u(x)$ and $v(t)$, respectively. Differently tuned filters are obtained by dilation or compression of the filter (4) along the spectral and temporal axes:

$$u(x;\Omega) = \Omega u(\Omega x), \qquad v(t;\omega) = \omega v(\omega t). \tag{5}$$

The response $r_c(t,x)$ of a cell c with parameters ω_c, Ω_c, φ_c, θ_c to the signal producing an auditory spectrogram $y(t,x)$ can therefore be obtained as

$$r_c(t,x;\omega_c,\Omega_c,\varphi_c,\theta_c) = y(t,x) \otimes h(t,x;\omega_c,\Omega_c,\varphi_c,\theta_c), \tag{6}$$

where $\otimes$ denotes a convolution both on x and on t.

An alternative representation of the filter can be derived in the complex domain. Denote

$$\begin{aligned} \tilde{h}_s(x;\Omega) &= u(x;\Omega) + j\bar{u}(x;\Omega), \\ \tilde{h}_t(t;\omega) &= v(t;\omega) + j\bar{v}(t;\omega), \end{aligned} \tag{7}$$

where $j = \sqrt{-1}$. Convolution of $y(t,x)$ with a downward-moving STRF obtained as $\tilde{h}_s(x;\Omega)\tilde{h}_t(t;\omega)$ and an upward-moving SRTF obtained as $\tilde{h}_s(x;\Omega)\tilde{h}_t^*(t;\omega)$ (where asterisk denotes complex conjugation) results in two complex response functions:

$$\begin{aligned} z_d(t,x;\omega_c,\Omega_c) &= y(t,x) \otimes [\tilde{h}_s(x;\Omega_c)\tilde{h}_t(t;\omega_c)] \\ &= |z_d(t,x;\omega_c,\Omega_c)|\, e^{j\psi_d(t,x;\omega_c,\Omega_c)}, \\ z_u(t,x;\omega_c,\Omega_c) &= y(t,x) \otimes [\tilde{h}_s(x;\Omega_c)\tilde{h}_t^*(t;\omega_c)] \\ &= |z_u(t,x;\omega_c,\Omega_c)|\, e^{j\psi_u(t,x;\omega_c,\Omega_c)}, \end{aligned} \tag{8}$$

and it can be shown [3] that

$$\begin{aligned} r_c(t,x;\omega_c,\Omega_c,\varphi_c,\theta_c) = \frac{1}{2}[\,&|z_d| \cos(\psi_d - \varphi_c - \theta_c) \\ &+ |z_u| \cos(\psi_u + \varphi_c - \theta_c)\,] \end{aligned} \tag{9}$$

(the arguments of z_d, z_u, ψ_d, and ψ_u are omitted here for clarity). Thus, the complex wavelet transform (8) uniquely determines the response of a cell with parameters ω_c, Ω_c, φ_c, θ_c to the stimulus, resulting in a dimensionality reduction effect in the cortical representation. In other words, knowledge of the complex-valued functions $z_d(t,x;\omega_c,\Omega_c)$ and $z_u(t,x;\omega_c,\Omega_c)$ fully specifies the six-dimensional cortical representation $r_c(t,x;\omega_c,\Omega_c,\varphi_c,\theta_c)$. The cortical representation thus can be obtained by performing (8) which results in a four-dimensional (time, frequency, rate, and scale) hypercube of (complex) filter coefficients that can be manipulated as desired and inverted back into the audio signal domain.

Essentially, the filter output is computed by a convolution of its spectro-temporal impulse response (STIR) with the input auditory spectrogram, producing a modified spectrogram. Since the spectral and temporal cross-sections of an STIR are typical of a bandpass impulse response in having alternating excitatory and inhibitory fields, the output at a given time-frequency position of the spectrogram is large only if the spectrogram modulations at that position are tuned to the rate, scale, and direction of the STIR. A map of the responses across the filter bank provides a unique characterization of the spectrogram that is sensitive to the spectral shape and dynamics over the entire stimulus.

To emphasize the features of the model that are important for the current work, note that every filter in the rate-scale analysis responds well to the auditory spectrogram features that have high correlation with the filter shape.

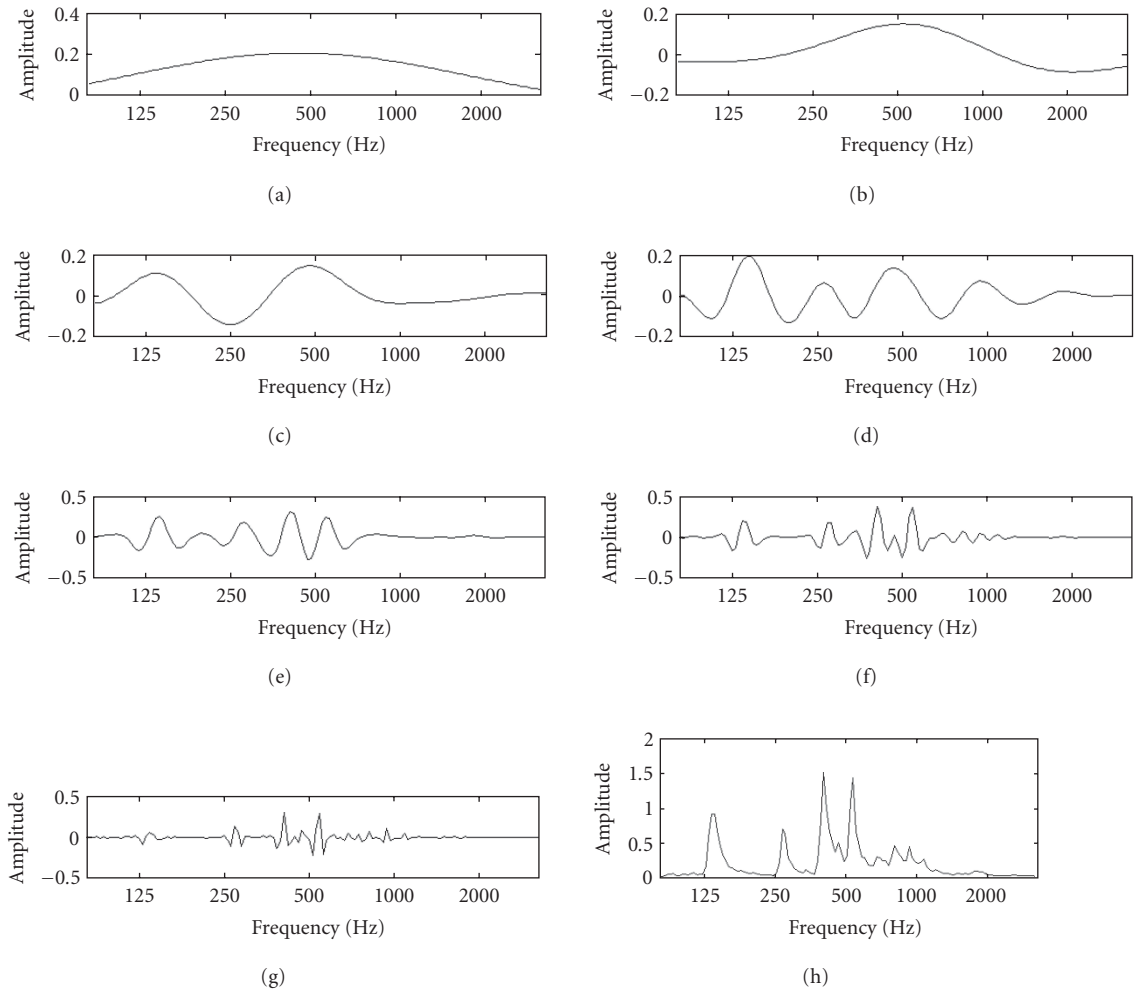


Figure 4: Sample scale decomposition of (h) the auditory spectrum using different scales: (a) DC, (b) 0.25, (c) 0.5, (d) 1.0, (e) 2.0, (f) 4.0, and (g) 8.0.

The filter shown in Figure 3 is tuned to the scale of 1 CPO and essentially extracts features that are of about this particular width on the log-frequency axis. A scale analysis performed with filters of different tuning (different width) will thus decompose the spectrogram into sets of decomposition coefficients for different scales, separating the "wide" features of the spectrogram from the "narrow" features. Some manipulations can then be performed on parts of the decomposed spectrogram, and a modified auditory spectrogram can be obtained by inverse filtering. Similarly, rate decomposition allows for segregation of "fast" and "slow" dynamic events along the temporal axis. A sample scale analysis of the auditory spectrogram is presented in Figure 4 (Figure 4h is the auditory spectrum, Figure 4a is the DC level of the signal which is necessary for the reconstruction, and the remaining 6 plots show the results of processing of the given auditory spectrum with filters of scales ranging from 0.25 to 8 CPO), and the rate analysis is similar.

Additional useful insights into the rate-scale analysis can be obtained if we consider it as a two-dimensional wavelet decomposition of an auditory spectrogram using a set of basis functions, which are called *sound ripples*. The sound ripple is simply a spectral ripple that drifts upwards or downwards in time at a constant velocity and is characterized by the same two parameters—scale (density of peaks per octave) and rate (number of peaks drifting past any fixed point on the log-frequency axis per 1-second time frame). Thus, an upward ripple with scale 1 CPO and rate 1 Hz has alternating peaks and valleys in its spectrum with 1 CPO periodicity, and the spectrum shifts up along the time axis, repeating itself with 1 Hz periodicity (Figure 5). If this ripple is used as an input audio signal for the cortical model, strong localized response is seen at the filter with the corresponding selectivity of $\omega = 1$ Hz, $\Omega = 1$ CPO. All other basis functions are obtained by dilation (compression) of the seed-sound ripple (Figure 5) in both time and frequency. (The difference between the ripples and the filters used in the cortical model is that the seed spectro-temporal response used in cortical model (4) and shown in Figure 3 is local; the seed-sound ripple can be obtained from it by reproducing the spatial

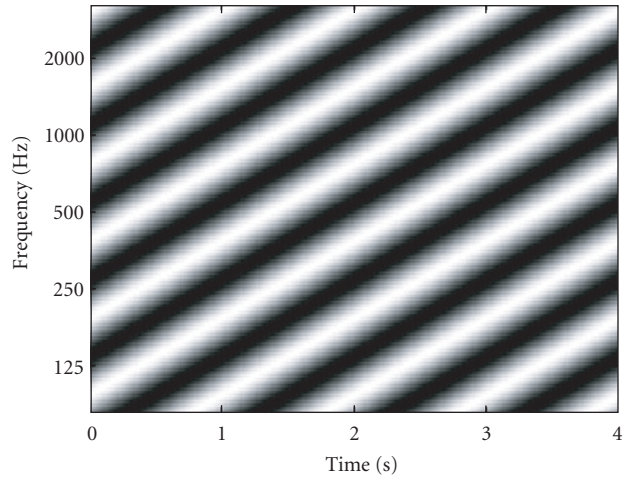


Figure 5: Sound ripple at the scale of 1 CPO and the rate of 1 Hz.

response at every octave and removing the time decay from the time response, and multiscale decomposition can then be viewed as overlapping the auditory spectrogram with different sound ripples and performing local cross-correlations at various places over the spectrogram.) In Figure 6, we show the result of filtering of the sample spectrogram showed earlier using two particular differently tuned filters, one with $\omega = 8$ Hz, $\Omega = 0.5$ CPO, and the other with $\omega = -2$ Hz, $\Omega = 2$ CPO. It can be seen that the filter output is the highest when the spectrogram features match the tuning of the filter both in rate and scale.

As such, to obtain a multiscale representation of the auditory spectrogram, complex filters having the "local" sound ripples (5) of different rates, scales, and central frequencies as their real parts and Hilbert transforms of these ripples as their imaginary parts are applied to the input audio signal as a wavelet transform (8). The result of this decomposition is a four-dimensional hypercube of complex filter coefficients that can be modified and inverted back to the acoustic signal. The phase of the coefficient shows the best-fitting direction of the filter over a particular location of the auditory spectrogram. This four-dimensional hypercube is called the cortical representation of the sound. It can be manipulated to produce desired effects on the sound, and in the following sections, we show some of the possible sound modifications.

In the cortical representation, two-dimensional rate-scale slices of the hypercube reveal the features of the signal that are most prominent at a given time. The rate-scale plot evolves in time to reflect changing ripple content of the spectrogram. Example of rate-scale plots are shown in Figure 7 where brightness of the pixel located at the intersection of particular rate and scale values corresponds to the magnitude of response of the filter tuned to these rate and scale values. For simplification of data presentation, these plots are obtained by integration of the response magnitude over the tonotopical axis. The first plot is a response of the cortical model to a single downward-moving sound ripple with $\omega = 3$ Hz, $\Omega = 2$ CPO; the best-matching filter (or, in other words, the "neuron" with the corresponding SRTF) responds best. The responses of 2 Hz and 4 Hz units are not equal here because of the cochlear filter bank asymmetry in the early processing stage. The other three plots show the evolution of the rate-scale response at different time instants of the sample auditory spectrogram shown in Figure 2 (at approximately 600, 720, and 1100 milliseconds, respectively); one can indeed trace the plot time stamps back to the spectrogram and see that the spectrogram has mostly sparse downward-moving and mostly dense upward-moving features appearing before the 720- and 1100-millisecond marks, respectively. The peaks in the test sentence plots are sharper in rate than in scale, which can be explained by the integration performed over the tonotopical axis in these plots (the speech signal is unlikely to elicit significantly different rate-scale maps at different frequencies anyway because it consists mostly of equispaced harmonics that can rise or fall only in unison, so the rate at which the highest response is seen is not likely to differ at different points on the tonotopical axis; the prevalent scale does change somewhat though due to higher number of harmonics per octave at higher frequencies).

4. RECONSTRUCTING THE AUDIO FROM THE MODEL

After altering the cortical representation, it is necessary to reconstruct the modified audio signal. Just as with the forward path, the reconstruction consists of two steps, corresponding to the central processing stage and the early processing stage. The first step is the inversion of the cortical multiscale representation back to a spectrogram. It is a one-step inverse wavelet transform operation because of the linear nature of the transform (8), which in the Fourier domain can be written as

$$\begin{aligned} Z_d(\omega,\Omega;\omega_c,\Omega_c) &= Y(\omega,\Omega)\tilde{H}_s(\Omega;\Omega_c)\tilde{H}_t(\omega;\omega_c), \\ Z_u(\omega,\Omega;\omega_c,\Omega_c) &= Y(\omega,\Omega)\tilde{H}_s(\Omega;\Omega_c)\tilde{H}_t^*(-\omega;\omega_c), \end{aligned} \tag{10}$$

where capital letters signify the Fourier transforms of the functions determined by the corresponding lowercase letters. From (10), similar to the usual Fourier transform case, one can write the formula for the Fourier transform of the reconstructed auditory spectrogram $y_r(t,x)$ from its decomposition coefficients Z_d, Z_u as

$$Y_r(\omega,\Omega) = \frac{\sum_{\omega_c,\Omega_c} Z_d(\omega,\Omega;\omega_c,\Omega_c)\tilde{H}_t^*(\omega;\omega_c)\tilde{H}_s^*(\Omega;\Omega_c) + \sum_{\omega_c,\Omega_c} Z_u(\omega,\Omega;\omega_c,\Omega_c)\tilde{H}_t(-\omega;\omega_c)\tilde{H}_s^*(\Omega;\Omega_c)}{\sum_{\omega_c,\Omega_c} |\tilde{H}_t(\omega;\omega_c)\tilde{H}_s(\Omega;\Omega_c)|^2 + \sum_{\omega_c,\Omega_c} |\tilde{H}_t^*(-\omega;\omega_c)\tilde{H}_s(\Omega;\Omega_c)|^2}. \tag{11}$$

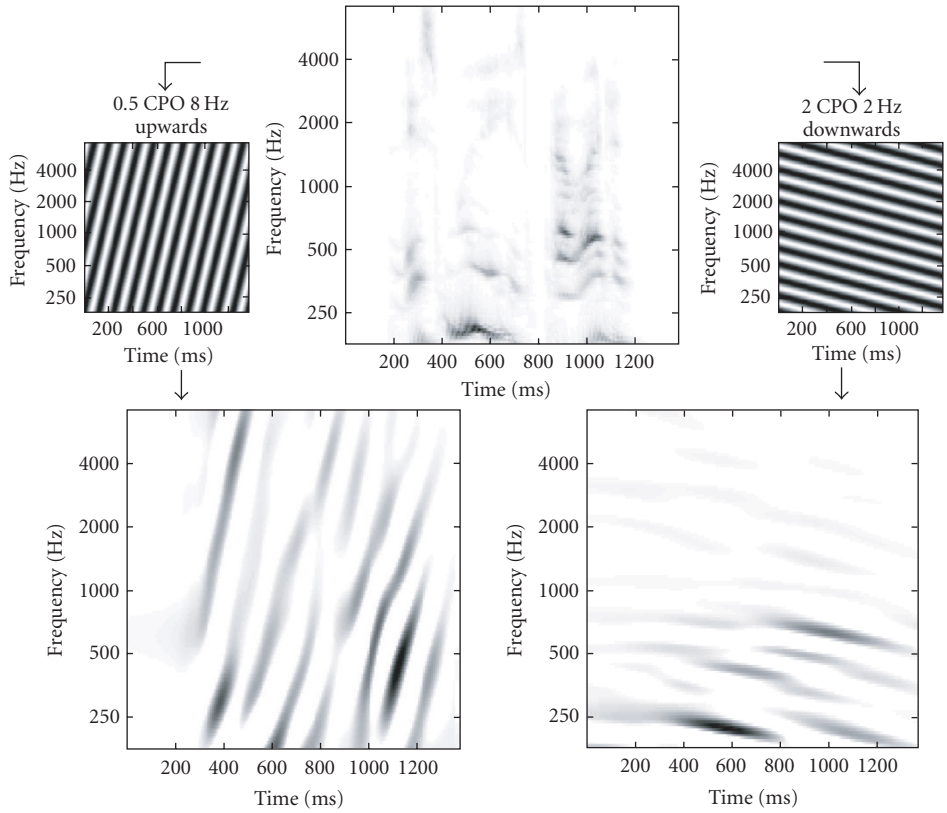


Figure 6: Wavelet transform of a sample auditory spectrogram (shown in Figure 2) using two sound ripples.

Then, $y_r(t,x)$ is obtained by inverse Fourier transform of $Y_r(\omega,\Omega)$ and is rectified to ensure that the resulting spectrogram is positive. The subscript r here and below refers to the reconstructed version of the signal. Excellent reconstruction quality is obtained within the effective band because of the linear nature of involved transformations.

The second step (going from the auditory spectrogram to the acoustic wave) is a complicated task due to the nonlinearity of the early auditory processing stage (nonlinear compression and half-wave rectification), which leads to the loss of the component phase information (because the auditory spectrogram contains only the magnitude of each frequency component), and a direct reconstruction cannot be performed. Therefore, the early auditory stage is inverted iteratively using a convex projection algorithm adapted from [22], which takes the spectrogram as an input and reconstructs the acoustic signal that produces the spectrogram closest to the given one.

Assume that an auditory spectrogram $y_r(t,x)$ is obtained using (11) after performing some manipulations in the cortical representation, and it is now necessary to invert it back to the acoustic signal $s_r(t)$. Observe that the analysis (first) step of the early auditory processing stage is linear and thus invertible. If an output of the analysis step $y_{1r}(t,x)$ is known, the acoustic signal $s_r(t)$ can be obtained as

$$s_r(t) = \sum_x y_{1r}(t,x) \oplus h(-t;x). \tag{12}$$

The challenge is to proceed back from $y_r(t,x)$ to $y_{1r}(t,x)$. In the convex projection method, an iterative adaptation of the estimate $\hat{y}_{1r}(t,x)$ is performed based on the difference between $y_r(t,x)$ and the result of the processing of $\hat{y}_{1r}(t,x)$ through the second and third steps of the early auditory processing stage. The processing steps are listed below.

(i) Initialize the reconstructed signal $\hat{s}_r^{(1)}(t)$ by a Gaussian-distributed white noise with zero mean and unit variance. Set iteration counter $k = 1$.

(ii) Compute $\hat{y}_{1r}^{(k)}(t,x)$, $\hat{y}_{2r}^{(k)}(t,x)$, and $\hat{y}_r^{(k)}(t,x)$ from $\hat{s}_r^{(k)}(t)$ using (1).

(iii) Compute the ratio $r^{(k)}(t,x) = y_r(t,x)/\hat{y}_r^{(k)}(t,x)$.

(iv) Adjust $\hat{y}_{1r}^{(k+1)}(t,x) = r^{(k)}(t,x)\hat{y}_{1r}^{(k)}(t,x)$.

(v) Compute $\hat{s}_r^{(k+1)}(t)$ using (12). Increase k by 1.

(vi) Repeat from step 2 unless the preset number of iterations is reached or a certain quality criterion is met (e.g., the ratio $r^{(k)}(t,x)$ is sufficiently close to unity everywhere).

Sample auditory spectrograms of the original and the reconstructed signals are shown later, and the reconstruction quality for the speech signal after a sufficient number of iterations is very good.

5. ALTERNATIVE IMPLEMENTATION OF THE EARLY AUDITORY PROCESSING STAGE

An alternative, much faster implementation of the early auditory processing stage was developed and can best be used for a fixed-pitch signal (e.g., a musical instrument tone).

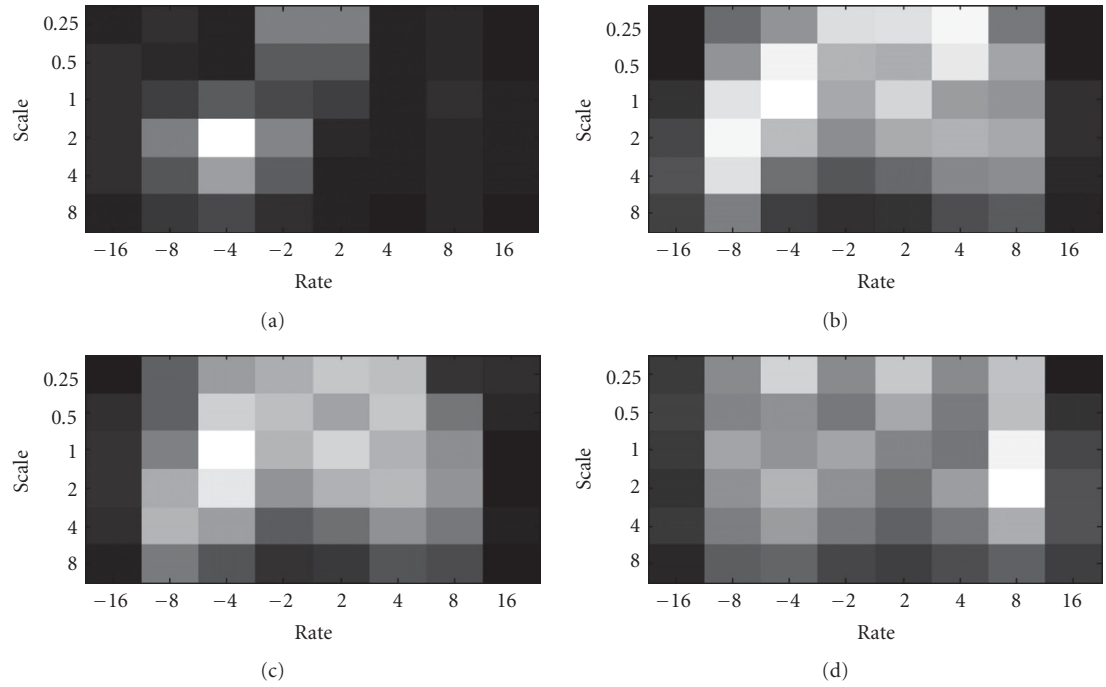


Figure 7: Rate-scale plots of response of cortical model to different stimuli. (a) Response to 2 CPO 3 Hz downward sound ripple. (b)–(d) Response at different temporal positions within the sample auditory spectrogram presented in Figure 2 (at 600, 720, and 1100 milliseconds, respectively).

In this implementation, which we will refer to as a log-Fourier transform early stage, a simple Fourier transform is used in place of the processing described by (1). We take a short segment of the waveform $s(t)$ at some time $t^{(j)}$ and perform a Fourier transform of it to obtain $S(f)$. The $S(f)$ is obviously discrete with the total of $L/2$ points on the linear frequency axis, where L is the length of the Fourier transform buffer. Some mapping must be established from the linear frequency axis f to the logarithmically growing tonotopical axis x. We divide a tonotopical axis into segments corresponding to channels. Assume that the cochlear filter bank has N channels per octave and that the lowest frequency of interest is f_0. Then, the lower $x_l^{(i)}$ and the upper $x_h^{(i)}$ frequency boundaries of the ith segment are set to be

$$x_l^{(i)} = f_0 2^{i/N}, \qquad x_h^{(i)} = f_0 2^{(i+1)/N}. \tag{13}$$

$S(f)$ is then remapped onto the tonotopical axis. A point f on a linear frequency axis is said to fall into the ith segment on the tonotopical frequency axis if $x_l^{(i)} < f \le x_h^{(i)}$. The number of points that falls into a segment obviously depends on the segment length, which becomes bigger for higher frequencies (therefore the Fourier transform of $s(t)$ must be performed with very high resolution and $s(t)$ padded appropriately to ensure that at least a few points on the f-axis fall into the shortest segment on the x-axis). Spectral magnitudes are then averaged for all points on the f-axis that fall into the same segment i:

$$y_{\text{alt}}(t^{(j)}, x^{(i)}) = \frac{1}{B^{(i)}} \sum_{x_l^{(i)} < f \le x_h^{(i)}} |S(f)|, \tag{14}$$

where $B^{(i)}$ is the total number of points on f-axis that falls into the ith segment on x-axis (the number of terms in the summation), and the averaging is performed for all i, generating a time slice $y_{\text{alt}}(t^{(j)}, x)$. The process is carried out for all time segments of $s(t)$, producing $y_{\text{alt}}(t, x)$, which can be substituted for the $y(t, x)$ computed using (1) for all further processing.

The reconstruction proceeds in an inverse manner. At every time slice $t^{(j)}$, a set of $y(t^{(j)}, x)$ is remapped to the magnitude spectrum $S(f)$ on a linear frequency axis f so that

$$S(f) = \begin{cases} y(t^{(j)}, x^{(i)}) & \text{if for some } i,\ x_l^{(i)} < f \le x_h^{(i)}, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

At this point, the magnitude information is set correctly in $S(f)$ to perform inverse Fourier transform but the phase information is lost. Direct one-step reconstruction from $S(f)$ is much faster than the iterative convex projection method described above but produces unacceptable results with clicks and strong interfering noise at the frequency corresponding to the processing window length. Processing the signal in heavily overlapping segments with gradual fade-in and fade-out windowing functions somewhat improves the results but the reconstruction quality is still significantly below the quality achieved using the iterative projection algorithm described in Section 4.

One way to recover the phase information and to use one-step reconstruction of $s(t)$ from magnitude spectrum $S(f)$ is to save the bin phases of the forward-pass Fourier transform and later impose them on $S(f)$ after it is recon-

structed from the (altered) cortical representation. Significantly better continuity of the signal is obtained in this manner. However, it seems that the saved phases carry the imprint of the original pitch of the signal, which produces undesirable effects if the processing goal is to perform a pitch shift.

However, the negative effect of the phase set carrying the pitch imprint can be reversed and used for good simply by generating the set of bin phases that corresponds to a desired pitch and imposing them on $S(f)$. Of course the knowledge of the signal pitch is required in this case, which is not always easy to obtain. We have used this technique in performing timbre-preserving pitch shift of musical instrument notes where the exact original pitch F_0 (and therefore the exact shifted pitch F_0') is known. To obtain the set of phases corresponding to the pitch F_0', we generate, in the time domain, a pulse train of frequency F_0' and take its Fourier transform with the same window length as used in the processing of $S(f)$. The bin phases of the Fourier transform of the pulse train are then imposed on the magnitude spectrum $S(f)$ obtained in (15). In this manner, very good results are obtained at reconstructing musical tones of a fixed frequency; it should be noted that such reconstruction is not handled well by the iterative convex projection method described above—the reconstructed signal is not a pure tone but rather constantly jitters up and down, preventing any musical perception, presumably because the time slices of $s(t)$ are treated independently by convex projection algorithm, which does not attempt to match signal features from adjacent time frames.

Nevertheless, speech reconstruction is handled better by the significantly slower convex projection algorithm, because it is not clear how to select F_0' to generate the phase set. If the log-Fourier transform early stage can be applied to the speech signals, significant processing speed-up can be achieved. A promising idea is to employ a pitch detection mechanism at each frame of $s(t)$ to detect F_0, to compute F_0', and to impose F_0'-consistent phases on $S(f)$ to enable one-step recovery of $s(t)$, which is the subject of ongoing work.

6. RECONSTRUCTION QUALITY

It is important to do an objective evaluation of the reconstructed sound quality. The second (central) stage of the cortical model processing is perfectly invertible because of the linear nature of the wavelet transformations involved, and it is the first (early) stage that presents difficulties for the inversion because of the phase information loss in the processing. Given the modified auditory spectrogram $y_r(t,x)$, the convex projection algorithm described above tries to synthesize the intermediate result $\hat{y}_{1r}(t,x)$ that, when processed through the two remaining steps of the early auditory stage, yields $\hat{y}_r(t,x)$ that is as close as possible to $y_r(t,x)$. The waveform $\hat{s}_r(t)$ can then be directly reconstructed from $\hat{y}_{1r}(t,x)$. The reconstruction error measure E is defined as the average relative magnitude difference between the target $y_r(t,x)$ and the candidate $\hat{y}_r(t,x)$:

$$E = \frac{1}{B}\sum_{i,j}\frac{\left|\hat{y}_r\left(t^{(j)},x^{(i)}\right) - y_r\left(t^{(j)},x^{(i)}\right)\right|}{y_r\left(t^{(j)},x^{(i)}\right)}, \qquad (16)$$

where B is the total number of terms in the summation. During the iterative update of $\hat{y}_{1r}(t,x)$, the error E does not drop monotonically; instead, the lower the error, the higher the chance that the next iteration actually increases the error, in which case the newly computed $\hat{y}_{1r}(t,x)$ should be discarded and a new iteration should be started from the best previously found $\hat{y}_{1r}(t,x)$.

In practical tests, it was found that the error E drops quickly to units of percents, and any further improvement requires very significant computational expense. For the purposes of illustration, we took the 1200-milliseconds auditory spectrogram (Figure 2) and ran the convex projection algorithm on it. It takes about 2 seconds to execute one algorithm iteration on a 1.7 GHz Pentium computer. In this sample run, the error after 20, 200, and 2000 iterations was found to be 4.73%, 1.60%, and 1.08%, respectively, which is representative of the general behavior observed in many experiments.

In Figure 8a, the original waveform $s(t)$ and its corresponding auditory spectrogram $y(t,x)$ from Figure 2 are plotted. The auditory spectrogram $y(t,x)$ is used as an input $y_r(t,x)$ to the convex projection algorithm, which, in 200 iterations, reconstructs the waveform $\hat{s}_r(t)$ shown in the top plot of Figure 8b. The spectrogram $\hat{y}_r(t,x)$ corresponding to the reconstructed waveform is also shown in the bottom plot of Figure 8b. Because the reconstruction algorithm attempts to synthesize a waveform $\hat{s}_r(t)$ such that its spectrogram $\hat{y}_r(t,x)$ is equal to $y_r(t,x)$, it can be expected that the spectrograms of the original and of the reconstructed waveforms would match. This is indeed the case in Figure 8, but the fine waveform structure is different in the original (Figure 8a) and in the reconstruction (Figure 8b), with noticeably less periodicity in some segments. However, it can be argued that because the original and the reconstructed waveforms produce the same results when processed through the early auditory processing stage, the perception of these should be nearly identical, which is indeed the case when the sounds are played to the human ear. Slight distortions are heard in the reconstructed waveform, but the sound is clear and intelligible. Increasing the number of iterations further decreases distortions; when the error drops to about 0.5% (tens of thousands of iterations), the signal is almost indistinguishable from the original.

We also compared the quality of the reconstructed signal with the quality of sound produced by existing pitch modification and sound morphing techniques. In [5], spectrogram modeling with MFCC coefficients plus residue spectrogram and iterative reconstruction process are used for sound morphing, and short morphing examples for voiced sounds are available for listening in the online version of the same paper. Book [7] also contains (among many other examples) some audio samples derived using algorithms that are relevant to our work and are targeted for the same application areas as we are considering, in particular, samples of cross-synthesis between the musical tone and the voice using channel vocoder and resynthesis of speech and musical tones using LPC with residual as an excitation signal and LPC with pulse train as an excitation signal. In our opinion, the signal quality we achieve is comparable to the quality of the relevant

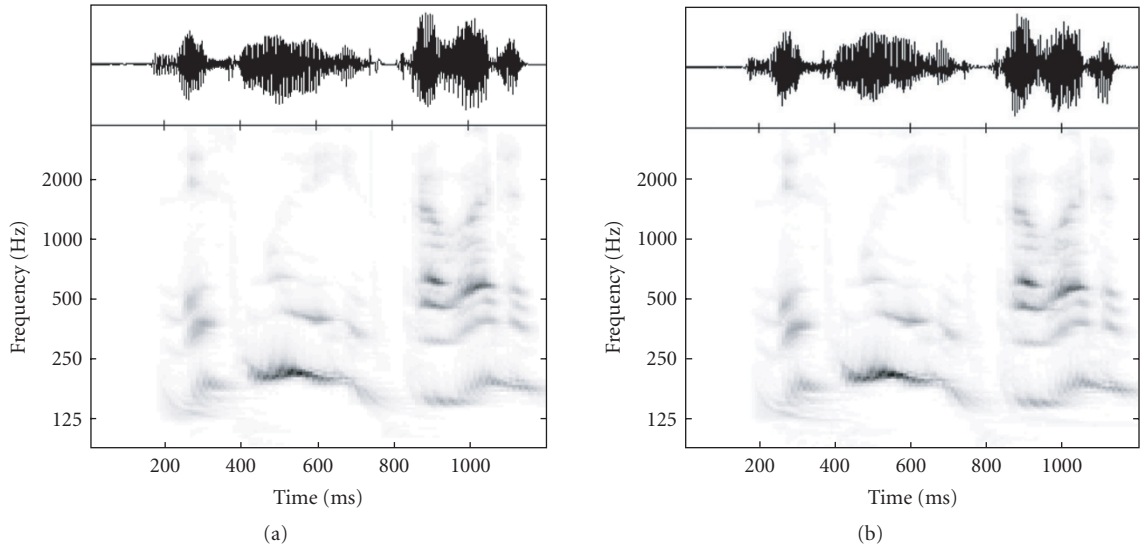


FIGURE 8: (a) Original waveform and corresponding spectrogram. (b) Reconstructed waveform and corresponding spectrogram after 200 iterations.

samples presented in these references, although the sound processing through a cortical representation is significantly slower than the algorithms presented in [5, 6, 7].

In summary, it can be concluded that reasonable quality of the reconstructed signal can be achieved in reasonable time, such as ten seconds or so of computational time per one second of a signal sampled at 8 kHz (although the iterative algorithm is not suitable for the real-time processing). If unlimited time (few hours) is allowed for processing, very good signal quality is achieved. The possibility of iterative signal reconstruction in real time is an open question and work in this area is continuing.

7. TIMBRE-PRESERVING PITCH MANIPULATIONS

For speech and musical instruments, timbre is conveyed by the spectral envelope, whereas pitch is mostly conveyed by the harmonic structure, or harmonic peaks. This biologically based analysis is in the spirit of the cepstral analysis used in speech [23], except that the Fourier-like transformation in the auditory system is carried out in a local fashion using kernels of different scales. The cortical decomposition is expressed in the complex domain, with the coefficient magnitude being the measure of the local bandwidth of the spectrum and the coefficient phase being the measure of the local symmetry at each bandwidth. Finally, just as it is the case with cepstral coefficients, the spectral envelope varies slowly. In contrast, the harmonic peaks are only visible at high resolution. Consequently, timbre and pitch occupy different regions in the multiscale representation. If X is the auditory spectrum of a given data frame, with the length N equal to the number of filters in the cochlear filter bank, and the decomposition is performed over M scales, then the matrix S of the scale decomposition of X has M rows, one per scale value, and N columns. If the first (top) row of S contains the decomposition over the finest scale and the Mth (bottom) row is the coarsest one, then the components of S in the upper left triangle can be associated with pitch, whereas the rest of the components can be associated with timbre information [24]. In Figure 9, sample plot of the scale decomposition of the auditory spectrum is shown. (Please note that this is a scale versus tonotopical frequency plot rather than scale-rate plot; all rate decomposition coefficients carry timbre information.) The brightness of a pixel corresponds to the magnitude of the decomposition coefficient, whereas the relative length and the direction of the arrow at the pixel show the coefficient phase. The white solid diagonal line shown in Figure 9 roughly separates timbre and pitch information in the cortical representation. The coefficients that lie above this line carry primarily pitch information, and the rest can be associated with timbre.

To control pitch and timbre separately, we apply modifications at appropriate locations in the cortical representation matrix and invert the cortical representation back to the spectrogram. Thus, to shift the pitch while holding the timbre fixed, we compute the cortical multiscale representation of the entire sound, shift (along the frequency axis) the triangular part of every time slice of the hypercube that holds the pitch information while keeping the timbre information intact, and invert the result. To modify the timbre keeping the pitch intact, we do the opposite. It is also possible to splice the pitch and the timbre information from two speakers, or from a speaker and a musical instrument. The result after inversion back to the sound is a "musical" voice that sings the utterance (or a "talking" musical instrument).

We express the timbre-preserving pitch shift algorithm in mathematical terms. The cortical representation consists of a set of complex coefficients $z_u(t, x; \omega_c, \Omega_c)$ and $z_d(t, x; \omega_c, \Omega_c)$. In the actual decomposition, the values of t, x, ω_c, and Ω_c are discrete, and the cortical representation of a sound is just a four-dimensional hypercube of complex coefficients $Z_{i,j,k,l}$. We agree that the first index i corresponds to the time axis,

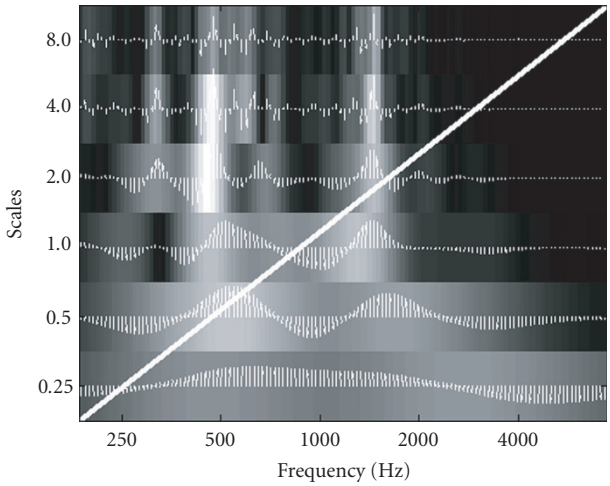


Figure 9: Plot of the sample auditory spectrum scale decomposition matrix. The brightness of the pixel corresponds to the magnitude of the decomposition coefficient, whereas the relative length and the direction of the arrow at the pixel show the coefficient phase. Upper triangle of the matrix of coefficients (above the solid white line) contains information about the pitch of the signal. The lower triangle contains information about the timbre.

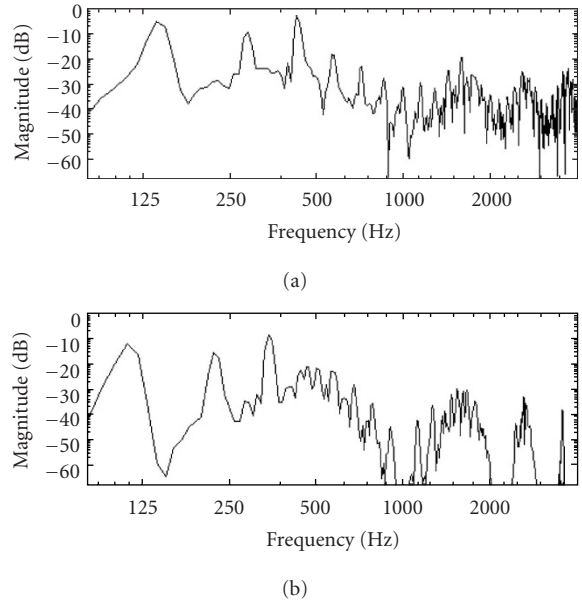


Figure 10: Spectrum of a speech signal (a) before and (b) after pitch shift. Note that the spectral envelope is filled with the new set of harmonics.

the second index j corresponds to the frequency axis, the third index k corresponds to the scale axis, and the fourth index l corresponds to the rate axis. Index j varies from 1 to N, where N is the number of filters in the cochlear filter bank, index k varies from 1 to M (in order of scale increase), where M is the number of scales, and, finally, index l varies from 1 to $2L$, where L is the number of rates (z_d and z_u are juxtaposed in $Z_{i,j,k,l}$ matrix as pictured on the horizontal axis in Figure 7: $l = 1$ corresponds to z_d with the highest rate, $l = 2$ to z_d with the next lower rate, $l = L$ to z_d with the lowest rate, $l = L+1$ to z_u with the lowest rate, $l = L+2$ to z_u with the next higher rate, and $l = 2L$ to z_u with the highest rate; this particular order is not critical for the pitch modifications described below as they do not depend on it). Then, the coefficient is assumed to carry pitch information if it lies above the diagonal shown in Figure 9 (i.e., if $(M-k)/j > (M-1)/N$), and to shift the pitch up by q channels, we fill the matrix $Z^*_{i,j,k,l}$ with the coefficients of matrix $Z_{i,j,k,l}$ as follows:

$$\begin{aligned} Z^*_{i,j,k,l} &= Z_{i,j,k,l}, \quad j < j_b, \\ Z^*_{i,j,k,l} &= Z_{j,j_b,k,l}, \quad j_b \le j < j_b + q, \\ Z^*_{i,j,k,l} &= Z_{i,j-q,k,l}, \quad j_b + q \le j, \end{aligned} \tag{17}$$

where $j_b = (M-k)N/(M-1)$ rounded to the nearest positive integer (note that j_b depends on k and therefore is different in different hyperslices of the matrix that have different values of k). The similar procedure shifts the pitch down by q channels:

$$\begin{aligned} Z^*_{i,j,k,l} &= Z_{i,j,k,l}, \quad j < j_b, \\ Z^*_{i,j,k,l} &= Z_{i,j+q,k,l}, \quad j_b \le j < N - q, \\ Z^*_{i,j,k,l} &= Z_{i,N,k,l}, \quad j_b \le j,\ N - q \le j. \end{aligned} \tag{18}$$

Finally, to splice the pitch of the signal S_1 with the timbre of the signal S_2, we compose Z^* from two corresponding cortical decompositions Z_1 and Z_2, taking the elements for which $(M-k)/j > (M-1)/N$ from Z_1 and all other ones from Z_2. Inversion of Z^* back to the waveform gives us the desired result.

We show one-pitch shift example here and refer the interested reader to http://www.isr.umd.edu/CAAR/ and http://www.umiacs.umd.edu/labs/pirl/NPDM/ for the actual sounds used in this example, and for more samples. We use the above-described algorithm to perform a timbre-preserving pitch shift of a speech signal. The cochlear model has 128 filters with 24 filters per octave, covering 5(1/3) octaves along the frequency axis. The cortical representation is modified using (18) to achieve the desired pitch modification and then inverted using the reconstruction procedure described in Section 4, resulting in a pitch-scaled version of the original signal. In Figure 10, we show plots of the spectrum of the original signal and of the signal having the pitch shifted down by 8 channels (one third of an octave) at a fixed point in time. The pitches of the original and of the modified signals are 140 Hz and 111 Hz, respectively. It can be seen from the plots that the signal spectral envelope is preserved and that the speech formants are kept at their original locations, but a new set of harmonics is introduced.

The algorithm is sufficiently fast to be used in real time if a log-Fourier transform early stage (described in Section 5) is substituted for a cochlear filter bank to eliminate the need for an iterative inversion process. Additionally, it is not necessary to compute the full cortical representation of the sound to do timbre-preserving pitch shifts. It is enough to perform only scale decomposition for every time frame of the auditory spectrogram because shifts are done along the frequency

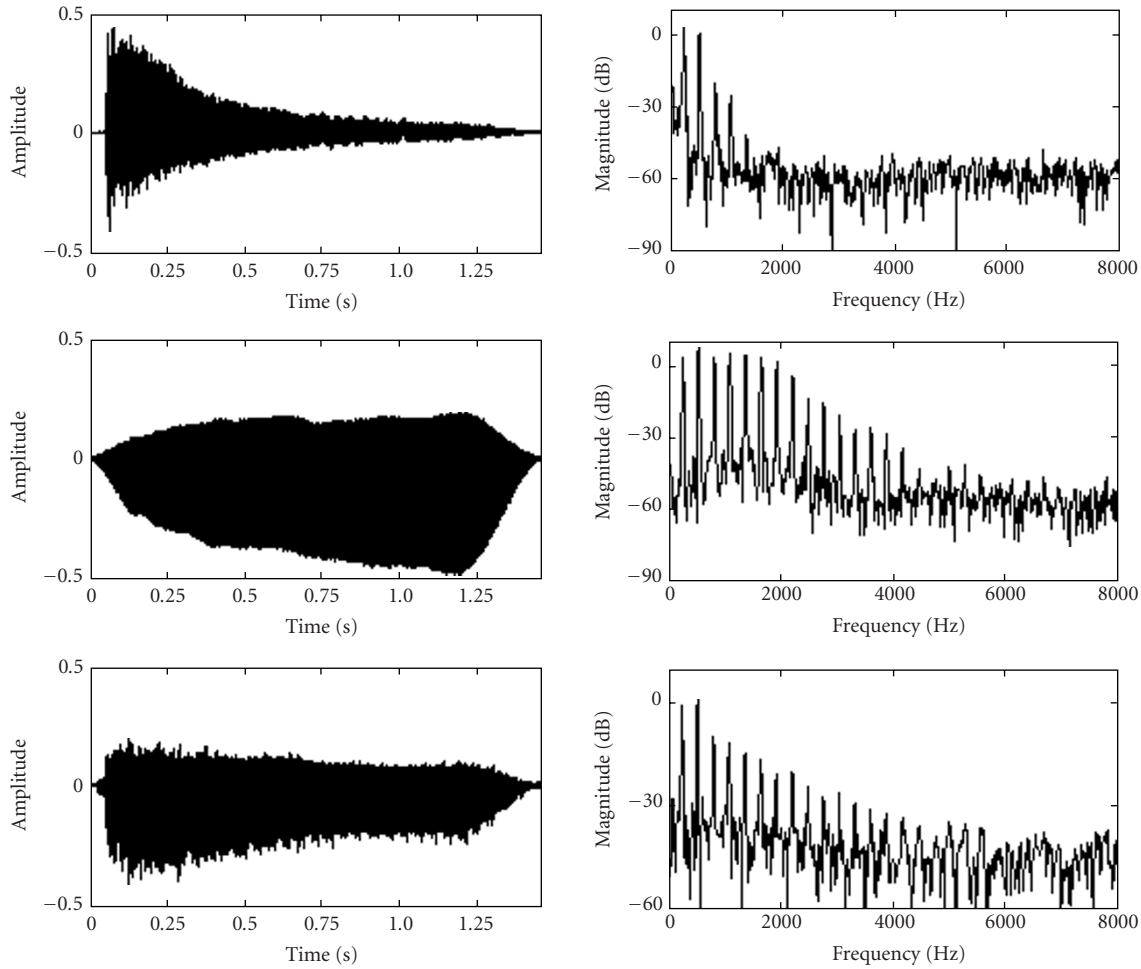


Figure 11: (Left column) Waveform plots and (right column) spectrum plots for guitar (top plots), trumpet (middle plots), and new instrument (bottom plots).

axis and can be performed in each time slice of the hypercube independently; thus, the rate decomposition is unnecessary. We have used the pitch-shift algorithm in a small-scale study in an attempt to generate maximally separable sounds to improve simultaneous eligibility of multiple competing messages [19]; it was found that the pitch separation does improve the perceptual separability of sounds and the recognition rate. Also, we have used the algorithm to generate, from one note of a given frequency, other notes of a newly created musical instrument that has the timbre characteristics of two existing instruments. This application is described in more details in the following section.

8. TIMBRE MANIPULATIONS

Timbre of the audio signal is conveyed both by the spectral envelope and by the signal dynamics. Spectral envelope is represented in the cortical representation by the lower right triangle of the scale decomposition coefficients and can be manipulated by modifying these. Sound dynamics is captured by the rate decomposition. Selective modifications to enhance or diminish the contributions of components of a certain rate can change the dynamic properties of the sound. As an illustration, and as an example of information separation across the cells of different rates, we synthesize a few sound samples using simple modifications to make the sound either abrupt or slurred. One such simple modification is to zero out the cortical representation decomposition coefficients that correspond to the "fast" cells, creating the impression of a low-intelligibility sound in an extremely reverberant environment; the other one is to remove "slow" cells, obtaining an abrupt sound in an anechoic environment (see http://www.isr.umd.edu/CAAR/ and http://www.umiacs.umd.edu/labs/pirl/NPDM/ for the actual sound samples where the decomposition was performed over the rates of 2, 4, 8, and 16 Hz; from these, "slow" rates are 2 and 4 Hz and "fast" rates are 8 and 16 Hz). It might be possible to use such modifications in sonification (e.g., by mapping some physical parameter to the amount of simulated reverberation and by manipulating the perceived reverberation time by gradual decrease or increase of contribution of "slow" components) or in audio user interfaces in general. Similarly, in the musical synthesis, playback rate and onset and decay ratio can be modified with shifts along the rate axis while preserving the pitch.

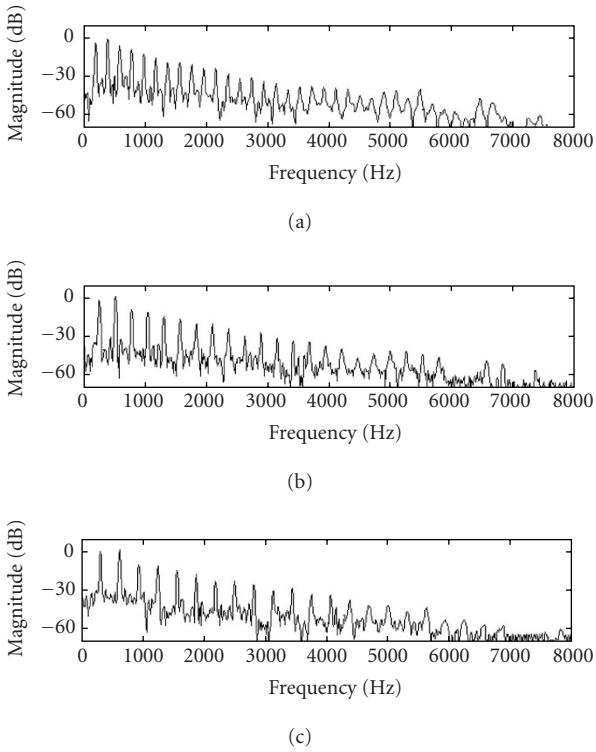


Figure 12: Spectrum of the new instrument playing (a) D#3, (b) C3, and (c) G2.

To show the ease with which timbre manipulation can be done using the cortical representation, we performed a timbre interpolation between two musical instruments to obtain a new in between synthetic instrument, which has both the spectral shape and the temporal spectral modulations (onset and decay ratio) that lie between the two original instruments. The two instruments selected were the guitar W_gC#3, and the trumpet, W_tC#3, playing the same note (C#3). The rate-scale decomposition of a short (1.5 seconds) instrument sample was performed and the geometric average of the complex coefficients in the cortical representations of these two instrument samples was computed and was converted back to the sound wave to obtain the new instrument sound sample W_nC#3. The behavior of the new instrument along the time axis is intermediate between two original ones, and the spectrum shape is also an average between two original instruments (Figure 11).

After the timbre interpolation, the synthesized instrument can only play the same note as the original ones. To synthesize other notes, we use the timbre-preserving pitch shift algorithm (Section 7) on the waveform W_nC#3 obtained by the timbre interpolation (third waveform in Figure 11) as an input. Figure 12 shows the spectrum of the new instrument for three different newly generated notes—D#3, C3, and G2. It can be seen that the spectral envelope is the same in all three plots (and is the same as the spectral envelope of the W_nC#3), but this envelope is filled with different sets of harmonics for different notes. For this synthesis, a log-Fourier transform early stage with pulse-train phase imprinting (Section 5) was used, as it is ideally suited for the task. A few samples of music made with the new instrument are available at http://www.umiacs.umd.edu/labs/pirl/NPDM/.

9. SUMMARY AND CONCLUSIONS

We developed and tested simple yet powerful algorithms for performing independent modifications of the pitch and the timbre of an audio signal and for performing interpolation between sound samples. These algorithms constitute a new application of the cortical representation of the sound [3], which extracts the perceptually important audio features simulating the processing believed to occur in auditory pathways in primates and thus can be used for making sound modifications tuned for and targeted to the ways the human nervous system processes information. We obtained promising results and are using these algorithms in ongoing development of auditory user interfaces.

ACKNOWLEDGMENTS

Partial support of ONR Grant N000140110571, NSF Grant IBN-0097975, and NSF Award IIS-0205271 is gratefully acknowledged. This paper is an extended version of paper [1].

REFERENCES

[1] D. N. Zotkin, S. A. Shamma, P. Ru, R. Duraiswami, and L. S. Davis, "Pitch and timbre manipulations using cortical representation of sound," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 5, pp. 517–520, Hong Kong, China, April 2003, reprinted in Proc. ICME '03, Baltimore, Md, USA, July 2003, vol. 3, pp. 381–384, because of the cancellation of ICASSP '03 conference meeting.

[2] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2, pp. 331–348, 2003.

[3] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," to appear in *Journal of the Acoustical Society of America*.

[4] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. A. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[5] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)*, vol. 2, pp. 1001–1004, Atlanta, Ga, USA, May 1996.

[6] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, G. D. Poli, A. Picialli, S. T. Pope, and C. Roads, Eds., Swets & Zeitlinger Publishers, Lisse, The Netherlands, 1997.

[7] P. R. Cook, *Real Sound Synthesis for Interactive Applications*, A. K. Peters, Natick, Mass, USA, 2002.

[8] S. Barrass, *Sculpting a Sound Space with Information Properties: Organized Sound*, Cambridge University Press, Cambridge, UK, 1996.

[9] G. Kramer, B. Walker, T. Bonebright, et al. "Sonification report: Status of the field and research agenda," prepared for NSF by members of the ICAD, 1997, http://www.icad.org/websiteV2.0/References/nsf.html.

[10] S. Bly, "Multivariate data mapping," in *Proc. Auditory Display: Sonification, Audification, and Auditory Interfaces*, G. Kramer, Ed., vol. 18 of Santa Fe Institute Studies in the Sciences of Complexity, pp. 405–416, Addison Wesley, Reading, Mass, USA, 1994.
[11] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, 2004.
[12] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, 2001.
[13] C. J. Darwin and R. W. Hukin, "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," *Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 335–342, 2000.
[14] C. J. Darwin and R. W. Hukin, "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 970–977, 2000.
[15] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *Journal of the Acoustical Society of America*, vol. 105, no. 6, pp. 3436–3448, 1999.
[16] W. A. Yost, R. H. Dye Jr., and S. Sheft, "A simulated cocktail party with up to three sound sources," *Perception and Psychophysics*, vol. 58, no. 7, pp. 1026–1036, 1996.
[17] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, pp. 35–50, 1992.
[18] P. F. Assmann, "Fundamental frequency and the intelligibility of competing voices," in *Proc. 14th International Congress of Phonetic Sciences*, pp. 179–182, San Francisco, Calif, USA, August 1999.
[19] N. Mesgarani, S. A. Shamma, K. W. Grant, and R. Duraiswami, "Augmented intelligibility in simultaneous multitalker environments," in *Proc. International Conference on Auditory Display (ICAD '03)*, pp. 71–74, Boston, Mass, USA, July 2003.
[20] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, Mass, USA, 1991.
[21] N. Kowalski, D. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.
[22] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824–839, 1992.
[23] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Mass, USA, 1998.
[24] R. Lyon and S. A. Shamma, "Auditory representations of timbre and pitch," in *Auditory Computations*, vol. 6 of *Springer Handbook of Auditory Research*, pp. 221–270, Springer-Verlag, New York, NY, USA, 1996.

Dmitry N. Zotkin was born in Moscow, Russia, in 1973. He received a combined B.S./M.S. degree in applied mathematics and physics from the Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia, in 1996, and received the M.S. and Ph.D. degrees in computer science from the University of Maryland, College Park, USA, in 1999 and 2002, respectively. Dr. Zotkin is currently an Assistant Research Scientist at the Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park. His current research interests are in multichannel signal processing for tracking and multimedia. He is also working in the general area of spatial audio, including virtual auditory scene synthesis, customizable virtual auditory displays, perceptual processing interfaces, and associated problems.

Taishih Chi received the B.S. degree from National Taiwan University, Taiwan, in 1992, and the M.S. and Ph.D. degrees from the University of Maryland, College Park, in 1997 and 2003, respectively, all in electrical engineering. From 1994 to 1996, he was a Graduate School Fellow at the University of Maryland. From 1996 to 2003, he was a Research Assistant at the Institute for Systems Research, University of Maryland. Since June 2003, he has been a Research Associate at the University of Maryland. His research interests are in neuromorphic auditory modeling, soft computing, and speech analysis.

Shihab A. Shamma obtained his Ph.D. degree in electrical engineering from Stanford University in 1980. He joined the Department of Electrical Engineering, the University of Maryland, in 1984, where his research has dealt with issues in computational neuroscience and the development of microsensor systems for experimental research and neural prostheses. His primary focus has been on uncovering the computational principles underlying the processing and recognition of complex sounds (speech and music) in the auditory system, and the relationship between auditory and visual processing. Other research interests include the development of photolithographic microelectrode arrays for recording and stimulation of neural signals, VLSI implementations of auditory processing algorithms, and development of algorithms for the detection, classification, and analysis of neural activity from multiple simultaneous sources.

Ramani Duraiswami is a member of the faculty in the Department of Computer Science and in the Institute for Advanced Computer Studies (UMIACS), the University of Maryland, College Park. He is the Director of the Perceptual Interfaces and Reality Laboratory there. Dr. Duraiswami obtained the B.Tech. degree in mechanical engineering from IIT Bombay in 1985, and a Ph.D. degree in mechanical engineering and applied mathematics from the Johns Hopkins University in 1991. His research interests are broad and currently include spatial audio, virtual environments, microphone arrays, computer vision, statistical machine learning, fast multipole methods, and integral equations.