Spectral shape analysis in the central auditory system

Article *in* IEEE Transactions on Speech and Audio Processing · October 1995

DOI: 10.1109/89.466657 · Source: IEEE Xplore

Kuansan Wang
Microsoft
128 PUBLICATIONS **2,886** CITATIONS

Shihab Shamma
University of Maryland, College Park
359 PUBLICATIONS **12,376** CITATIONS

Spectral Shape Analysis in the Central Auditory System

Kuansan Wang and Shihab A. Shamma

Abstract—**A model of spectral shape analysis in the central auditory system is developed based on neurophysiological mappings in the primary auditory cortex and on results from psychoacoustical experiments in human subjects. The model suggests that the auditory system analyzes an input spectral pattern along three independent dimensions: a logarithmic frequency axis, a local symmetry axis, and a local spectral bandwidth axis. It is shown that this representation is equivalent to performing an affine wavelet transform of the spectral pattern and preserving both the magnitude (a measure of the scale or local bandwidth of the spectrum) and phase (a measure of the local symmetry of the spectrum). Such an analysis is in the spirit of the cepstral analysis commonly used in speech recognition systems, the major difference being that the double Fourier-like transformation that the auditory system employs is carried out in a local fashion. Examples of such a representation for various speech and synthetic signals are discussed, together with its potential significance and applications for speech and audio processing.**

I. Introduction

SPECTRAL shape is the most important physical correlate of the percept of timbre [26]. In the early stages of auditory processing, a robust and enhanced representation of the acoustic spectrum is extracted in a series of reasonably well understood operations (described in detail in [44]). However, it is uncertain how this spectral pattern is further elaborated upon at higher auditory stages to separate the cues and features associated with different sound percepts such as pitch and timbre. In this paper, an attempt is made to formulate a model for spectral shape analysis in the central auditory system based on organizational principles recently discovered in the primary auditory cortex (A1).

Much like other primary sensory areas, neurons in A1 exhibit certain organizational characteristics that reflect systematic response selectivity to various stimulus features. For instance, each cell tends to be tuned to a specific range of tone frequencies and intensities that are known as the response area of the cell [22]. These response areas are organized topographically across the surface of A1, analogous to the topographic organization of the receptive fields in the primary visual and somatosensory cortecies [15]. Moreover, in the visual cortex, several complex stimulus features are also topographically organized, such as selectivity to edge orientations and to the direction of motion [14]. This discovery has led to major advances in our understanding of the functional organization of the visual cortex and of visual perception. Given the known similarity in the developmental, anatomical, and physiological properties of all primary sensory cortecies [25], it is natural to explore whether recently discovered response maps in A1 may be integrated within a functional framework similar to that of the visual cortex.

The topographical organization of A1 cells originates from the segregation of neural response selectivities, i.e., neurons exhibiting similar response patterns tend to be gathered in vicinity. As such, the problem of analyzing the feature mapping mechanism may be tackled by examining the organization of neural response selectivities. In this work, a perceptron-like single neuron model is adopted, namely, the response of a single neuron is assumed to be a weighted sum of its inputs followed by a compressive nonlinearity. Assuming that the compressive nonlinearity is a monotonic function, the neuron response is fully described by the weighted sum, which, mathematically, is an inner product between the input and the weighting function. Consequently, the neural response selectivities are determined by how the inputs "match" the weighting function associated with each neuron. Modeling the topographic organization of these weighting functions (analogous to the so-called receptive fields in vision) will be the focus of this study.

This paper is organized as follows. In Section II, relevant physiological results from cortical mapping experiments are briefly reviewed. They suggest that the auditory cortex bases its analysis and representation of the spectral shape on two factors: the local symmetry and the local bandwidth of the spectrum. Based on such viewpoint, a mathematical model of A1 is formulated in Section III. Specifically, it is shown that the local symmetry and bandwidth of a function can be closely linked to its Fourier domain properties. Consequently, the fundamental analysis in A1 can be interpreted as a multiscale estimation on the phase and modulus of the local Fourier transform of the spectral pattern of its input. Physiological and psychoacoustical findings in support of these hypotheses are reviewed in Section IV. In Section V, the model is tentatively

Manuscript received November 28, 1993; revised March 2, 1995. This work was supported by the Office of Naval Research, the Air Force Office of Scientific Research, and NSF Engineering Research Centers grant (NSFD CD-8803012). A portion of this paper was presented at NNSP-93, Baltimore, MD, September 1993. The associate editor coordinating the review of this paper and approving it for publication was James M. Kates.

K. Wang is with AT&T Bell Laboratories, Murray Hill, NJ 07974 USA.

S. Shamma is with the Department of Electrical Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA.

IEEE Log Number 9413734.

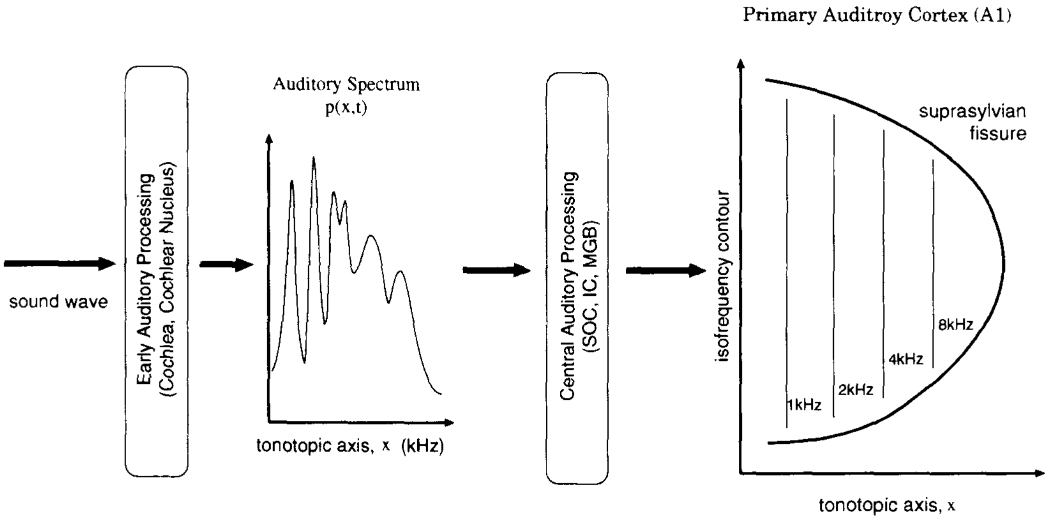


Fig. 1. Schematic description of the auditory model. The sound wave is transformed in the early auditory processing stages into an *auditory spectrum* ($p(x,t)$) profile distributed along the tonotopic axis. This profile is assumed to be the input to the cortical model. In the primary auditory cortex (A1), the tonotopic axis is expanded into two dimensions, with the iso-frequency contours, which are perpendicular to the tonotopic axis, potentially constituting an extra dimension for mapping other stimulus features.

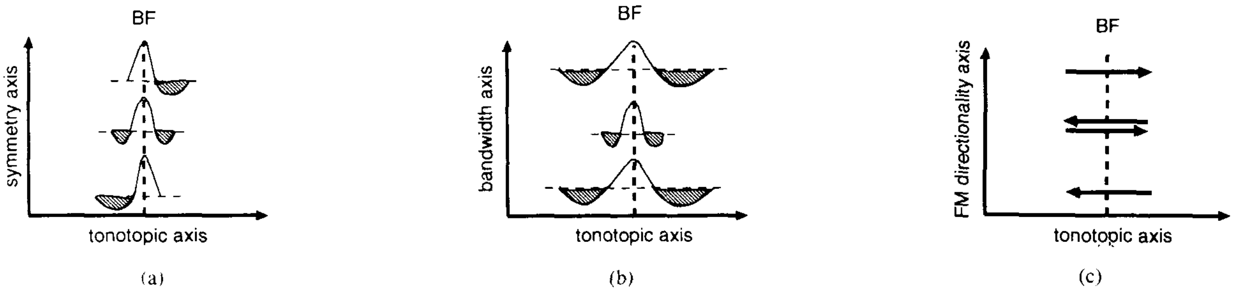


Fig. 2. Schematic summary of the functional organization of the A1: (a) Spatial distribution of response area asymmetry; (b) spatial distribution of response area bandwidths; (c) FM directional selectivity of the neural responses.

extended to deal with time-varying spectra, such as FM tone sweeps, and in Section VI, examples of the model outputs for various signals are discussed. Finally, the relevance of this type of representation to speech and audio processing systems are briefly explored in Section VII.

II. Organization of Physiological Response Properties in Primary Auditory Cortex

The cochlea of the inner ear analyzes a complex sound into a topographically organized array of channels that are tuned to different (best) frequencies (BF's). These BF's are roughly logarithmically ordered along the length of the cochlea, creating the tonotopic axis of the auditory system, which is preserved through several central processing stages all the way up to the auditory cortex [22]. However, unlike the 1-D nature of the tonotopic axis in the cochlea, in A1, the tonotopic axis becomes 2-D, with many cells tuned to the same frequency lined up along the so-called *iso-frequency planes* that run perpendicularly to the tonotopic axis (illustrated in Fig. 1).

A. Symmetry

There have been numerous studies on the question of what specific features of the stimulus might be represented along these iso-frequency planes. For example, it has recently been reported [37] that cortical cells exhibit a systematic change in the *symmetry* of their tuning curves, as illustrated in Fig. 2(a). Thus, in terms of excitatory and inhibitory responses within the tuning curve, the experimental data show that in the center of A1, the response area has a centered excitatory band that is symmetrically flanked by inhibitory side bands. Toward the edges, the response areas become more asymmetric with stronger inhibitory sidebands above BF in one direction and below the BF in the opposite direction. One possible implication of such a systematic change in symmetry is that the auditory cortex explicitly analyzes the symmetry of the stimulus spectrum. Thus, cells with asymmetric response areas are best excited by signals with spectral shapes of the opposite asymmetry [37].

B. Bandwidth

Since the notion of spectral asymmetry is only meaningful within the bandwidth of the neuron's response area, its evaluation must be regarded as a local and possibly multiscale operation. Physiological support of this argument comes from the finding that cells along the iso-frequency planes vary considerably and systematically in the bandwidth

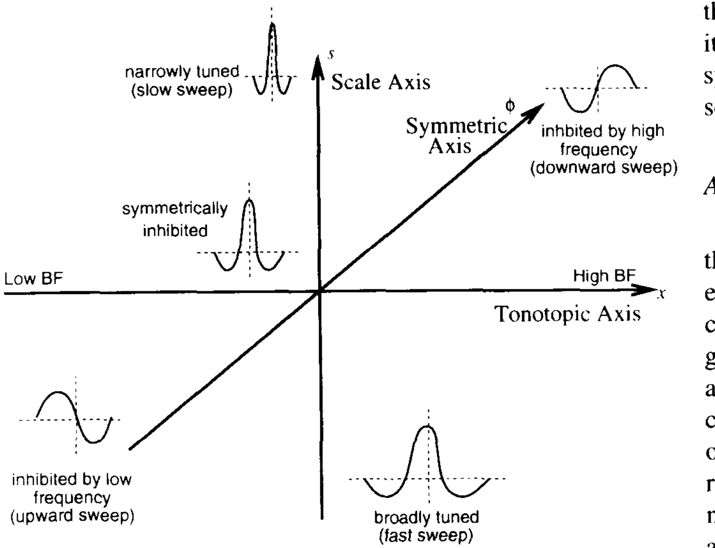


Fig. 3. Schematic of the cortical model. The response fields are assumed to be organized along three dimensions: the best frequency (tonotopic axis), the bandwidth (the scale axis), and the asymmetry (the phase axis).

of their tuning. Specifically, neurons in the center of A1 are more narrowly tuned compared with those near the edges of A1 [31] (Fig. 2(b)). From a functional point of view, this implies that the auditory system may employ a multiscale mechanism to analyze the spectral representation manifested by the precortical stages, and each scale resolves and extracts information encoded in a specific bandwidth.

C. Frequency Modulation (FM) Selectivity

Finally, it has also been observed that many neurons in A1 respond selectively to the direction and rate of a frequency modulated (FM) tone. Specifically, FM directional selectivity correlates well with the asymmetry of the response area [37]. Thus, asymmetric cells with strong below-BF inhibitory sidebands respond best to downward FM sweeps, and vice versa (Fig. 2(c)). Other experiments also indicate that the FM rate sensitivity seems to be correlated to the bandwidth of response areas [10], [21]. Therefore, as we shall discuss in detail later, it is likely that neural selectivity of transient stimuli, such as FM tones, is functionally linked to the neural selectivity of spectral shape.

In summary, properties of response areas in A1 are conceptually organized along three organizational axes as illustrated in Fig. 3:

1) the tonotopic axis (as reflected by the systematic change in BF's)
2) the symmetry axis (as reflected by the change in response area asymmetries along the iso-frequency planes)
3) the scale axis (as reflected by the systematic change in bandwidths along the iso-frequency planes).

These three axes will form the backbone of the mathematical model presented next, which is used later to investigate the analysis and representation of the acoustic spectrum in A1.

III. Mathematical Model for Cortical Processing

Based on experimental results reviewed earlier and the conceptual framework outlined above, we first introduce in this section a quantitative definition of symmetry and use it to elaborate the information represented on the tonotopic-symmetry plane. The discussion is then extended to a multiscale framework to account for the variation in bandwidth.

A. Preliminaries

Becuase we are dealing with monaural spectral analysis in the central auditory system, we shall approximately view several intermediate precortical stages (e.g., the superior olivary complex (SOC), the inferior colliculi (IC), and the medial geniculate body as in Fig. 1) as signal relay stations and assume here that the input to the cortical model is the so-called auditory spectrum $p(x,t)$ as it appears at the output of the cochlear nucleus. It is basically a spectro-temporal representation of the sound signal similar to a short-time, self-normalized power spectrum [44]. The index t refers to time, and x denotes the spatial location along the tonotopic axis, which is roughly close to a the logarithmic frequency axis. For notational simplicity, the t index will be dropped in the following text whenever no confusion arises. In [44], it is shown that the dynamic range of the auditory spectrum is bounded, i.e., the possible inputs to the model are from the signal space that is a subset of $L^2(\mathcal{R})$. In the following, we denote the inner product and the induced norm on $L^2(\mathcal{R})$ as $\langle f,g\rangle_x = \int_{\mathcal{R}} f(x)g(x)dx$ and $\|p\| = \sqrt{\langle p,p\rangle_x}$ respectively.

The Fourier transform of $p(x)$, which is defined as

$$P(k) = \int_{\mathcal{R}} p(x)e^{-jkx}dx$$

$$p(x) = \frac{1}{2\pi}\int_{\mathcal{R}} P(k)e^{jkx}dk$$

is an isometry and inner-product invariant mapping on $L^2(\mathcal{R})$, provided that the inner product in the Fourier domain is defined as

$$\langle P,Q\rangle_k = \frac{1}{2\pi}\int_{\mathcal{R}} P(k)Q^*(k)dk$$

where Q^* denotes the complex conjugate of Q, and the norm in the Fourier domain is defined as $\|P\| = \sqrt{\langle P,P\rangle_k}$. Note that the k domain here corresponds to the Fourier counterpart of the spatial x axis, and hence, k can be termed "spatial frequency." A spectral pattern that has a sinusoidal shape along the logarithmic frequency axis is also known in the psychoacoustical literature as a "ripple." Therefore, we also refer to k as the *ripple frequency* and $P(k)$ as the *ripple spectrum* in the following. Note also that the functions in the signal space and the Fourier domain are denoted in small and capital letters, respectively.

B. Encoding Spectral Asymmetry in the Cortical Model

With respect to a predefined origin x_c on the x axis, a signal $p(x)$ is said to be symmetric around x_c if $p(x-x_c)$ is an even function. In contrast, it is *antisymmetric* if $p(x-x_c)$ is odd. For any function $h(x)$ and any given reference x_c, one can always define a symmetric function

$$h_e(x) = \frac{h(x - x_c) + h(-x + x_c)}{2}$$

and an antisymmetric function

$$h_o(x) = \frac{h(x - x_c) - h(-x + x_c)}{2}$$

so that

$$h = h_e + h_o$$

Since the above decomposition is unique, the whole signal space is a direct sum of two subspaces that contain exclusively symmetric or antisymmetric functions. The so-called Hilbert transform

$$\hat{h}(x) = \frac{1}{\pi}\int_{\mathcal{R}} \frac{h(y)}{x - y} dy$$

is a one-to-one mapping between these two subspaces, i.e., if h is symmetric, $\hat{h}$ is antisymmetric, and vice versa. Note also that $||h|| = ||\hat{h}||$.

Now, consider the neural response functions for a tonotopic location x_c. It is evident from the above discussion that a family of functions $(w_s(x; x_c, \phi_c))$ representing the response functions (RF's) varying gradually in symmetry (along the symmetry axis in Figs. 2 and 3), can be generated by sinusoidally interpolating a symmetric (with respect to x_c) seed function $h_s(x)$ and its Hilbert transform

$$w_s(x; x_c, \phi_c) = h_s(x - x_c)\cos\phi_c - \hat{h}_s(x - x_c)\sin\phi_c \quad (1)$$

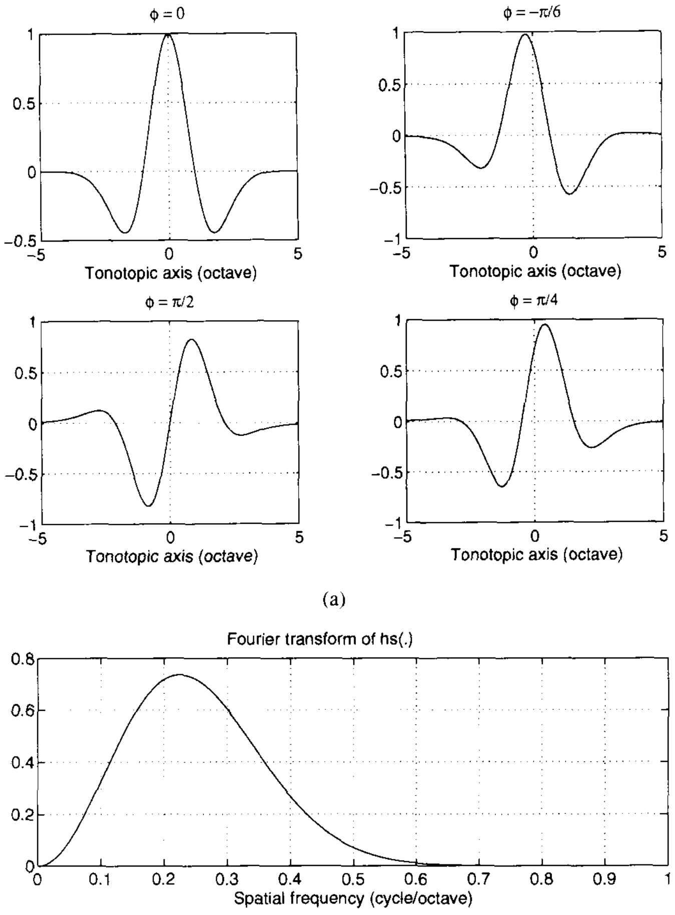
Therefore, the systematic change in RF's symmetry can be parameterized by ϕ_c, which will be referred to as the *symmetry index* or, for reasons that will become clear later, the *characteristic phase*. The cortical symmetry axis will be referred to as the ϕ axis, where each point ($\phi = \phi_c$) denotes a physical location along the iso-frequency contour $x = x_c$ for which the neural response functions have a symmetry index ϕ_c. For instance, $w_s(x; x_c, \pm\pi/2)$ are antisymmetric, and the symmetry gradually changes toward the center, where $w_s(x; x_c, 0)$ is symmetric (Fig. 4).

One implication of employing a sinusoidal interpolation is that the Fourier transform of the RF $W_s(k; x_c, \phi_c)$ has a constant antisymmetric phase $-\phi_c \,\mathrm{sgn}(k)$ (with respect to x_c). In addition, note that the family of functions thus generated have the same magnitude since

$$|W_s(k; x_c, \phi_c)| = |H_s(k)e^{j(kx_c + \phi_c)}| = H_s(k) \quad (2)$$

for all k and ϕ_c. An illustration of the RF's and their Fourier magnitude are shown in Fig. 4 in which x_c is arbitrarily set to 0.

As outlined before, it is assumed in this work that the cortical symmetry selectivity may be described by an inner product of the input auditory spectrum $p(x)$ and a response function associated with each neuron, denoted by $w_s(x; x_c, \phi_c)$ described above. Namely, the symmetry selectivity to a spectral pattern $p(x)$ on the tonotopic-symmetry (x-ϕ) plane is



(a)

(b)

Fig. 4. Examples of sinusoidal interpolation for generating cortical RF's $w_s(x, \phi)$ with various symmetry indices φ_c: (a) The seed function $h_s(x)$ used here is the negative second derivative of a Gaussian function, which is obtained by curve fitting physiological data published in [37]. In this figure, each unit length in the abscisa is normalized to correspond to an acoustic octave, namely, $h_s(x) = (1 - x^2)e^{-x^2/2}$ (top left panel). Its Fourier transform $H_s(k) = k^2 e^{-k^2/2}$ is shown in (b).

$$r_s(x, \phi) = \langle p(y), w_s(y; x, \phi)\rangle_y = \int_{\mathcal{R}} p(y) w_s(y; x, \phi) dy.$$

Since in the Fourier domain

$$R_s(k; x, \phi) = P(k)W_s(k; x, \phi) = P(k)H_s(k)e^{j\phi\,\mathrm{sgn}(k)}$$

we then have

$$\begin{aligned} r_s(x, \phi) &= \langle P(k)e^{jkx}, W_s^*(k; x, \phi)\rangle_k \\ &= \langle \Re\{P(k)e^{jkx}\}, H_s(k)\rangle_k \cos\phi \\ &\quad - \langle \Im\{P(k)e^{jkx}\}, H_s(k)\,\mathrm{sgn}(k)\rangle_k \sin\phi \\ &= a_s(x)\cos(\phi - \psi_s(x)) \end{aligned} \quad (3)$$

where $\Re\{P\}$ and $\Im\{P\}$ indicate the real and imaginary part of P, respectively, and

$$a_s(x) = |\langle P(k)e^{jkx}, H_s(k)(1 - j\,\mathrm{sgn}(k))\rangle_k|. \quad (4)$$

Equation (3) indicates that at each crosss section on the tonotopic axis, the cortical response is always a sinusoid in ϕ whose amplitude $a_s(x)$ is given by (4) and phase by

$$\psi_s(x) = \tan^{-1} \frac{\langle \Im\{Pe^{jkx}\}, H_s \operatorname{sgn}(k)\rangle_k}{\langle \Re\{Pe^{jkx}\}, H_s\rangle_k}. \tag{5}$$

The implication of the above two equations is that due to the redundancy, the 2-D response $r_s(x, \phi)$ can be fully specified by two 1-D functions $a_s(x)$ and $\psi_s(x)$, which are, respectively, the "windowed" estimates of the magnitude and phase of $P(k)e^{jkx}$. If $h_s(\cdot)$ has limited support[1] in both the x and k domains, $a_s(x)$ and $\psi_s(x)$ are two local measures. The (single scale) cortical selectivity can therefore be interpreted as evaluating the windowed Fourier magnitude and phase of the input auditory spectrum. For this reason, $a_s(\cdot)$ and $\psi_s(\cdot)$ will be referred to as the *local magnitude response* and the *local phase response*, respectively.

C. Multiscales in the Cortical Model

As indicated in (4) and (5), the "locality" of such analysis is determined by the support of the seed function $H_s(k)$, which is inversely proportional to the support of $h_s(x)$. Therefore, the changing bandwidth of the RF's can be modeled by a systematic dilation of the seed function $h_s(\cdot)$ as

$$H(k, s) \equiv H_s(k) = H_m(k/\alpha^s) \tag{6}$$

for some *mother function* $h_m(\cdot)$ and dilation factor α. The multiscale RF can therefore be modeled as $w(x; x_c, \phi_c, s_c) \equiv w_{s_c}(x; x_c, \phi_c)$. Following the previous derivation, the multiscale cortical selectivity $r(x, \phi, s) \equiv r_s(x, \phi)$ can therefore be described by the local magnitude response $a(x, s) = a_s(x)$ and local phase response $\psi(x, s) = \psi_s(x)$, respectively.

There are several implications of this dilation assumption:

1) Since $w(x; x_c, \phi_c, s_c)$ are dilated from a seed function, their Fourier transform $W(k; x_c, \phi_c, s_c)$ can be described as constant-Q (ripple) bandpass filters, where each is tuned around a *characteristic ripple frequency* $k_c(s_c)$. Consequently, the width of a given RF is inversely proportional to its characteristic ripple. For instance, the characteristic ripple of the RF shown in Fig. 4 is $k_c = \sqrt{2}$ rad/octave = 1/4.44 cycles/octave. The zero crossings occur at 1 and -1, i.e., the excitatory bandwidth is two octaves, which is slightly less than half of the reciprocal of k_c.
2) The characteristic ripple $k_c(s) = \alpha^s k_c(0)$ is mapped in a logarithmic fashion along the scale axis s, i.e.

$$\log k_c(s) = s \cdot \log \alpha + \log k_c(0) \tag{7}$$

where $k_c(0)$ denotes the characteristic ripple of the mother function.

The above mapping of spectral ripples onto a scale axis is functionally analogous to the logarithmic mapping of an acoustic frequency onto the spatial tonotopic axis of the cochlea. In fact, the analogy between the cortical and cochlear processing is quite accurate if one views the spectral profile as an acoustic signal (i.e., tonotopic axis as the time axis) and the cortical constant Q filters as the cochlear filters. This suggests that the sequence of cochlear and cortical analyses of the acoustic signal is conceptually a form of a double Fourier transform (more accurately, a double affine wavelet transform) on the acoustic waveform, which is similar *in spirit* to the cepstral analysis. These two types of analyses will be briefly compared in Section VII.

IV. Functional Interpretation of the Cortical Model

The cortical model can be described as a *ripple analysis model* in that it analyzes the auditory spectrum via a bank of ripple analysis filters $W(k; x, \phi, s)$, each tuned around a characteristic ripple frequency and phase. What is the direct physiological evidence for such a tuning in cortical cells, and what is the resolution and perceptual implications of such an analysis?

A. Characteristics of Cortical Responses to Rippled Spectra

Recent experiments in the A1 of ferrets and cats have confirmed that cortical cells exhibit the response properties demanded by the ripple analysis model [3], [38]. For instance, almost every cell in A1 is observed to tune to a characteristic ripple frequency (k_c) and phase (ϕ_c). Furthermore, supporting the linearity of the analysis, k_c's and ϕ_c's (and their topographic distributions across A1) are found to correspond well to the sharpness and asymmetry of the response area tuning curves (and to their previously elaborated mappings [32], [37]). The joint distribution of these two response parameters (at each tonotopic location x) cover a reasonable range of values to represent the input spectrum well.

B. Perceptual Implications of the Ripple Analysis Model

The cortical model, which is viewed as a ripple analysis of the auditory spectrum, designates spectral ripples with added perceptual significance. Therefore, it is important to assess the range and sensitivity of human subjects to various ripple frequencies and phases.

1) Ripple Contrast: Such data have only recently become available [8], [9], [13], [42]. Detection thresholds for different ripples (which is also known as the contrast sensitivity) reveal that humans are most sensitive in the range 1–4 cycles/octaves, but can still detect ripples over 10 cycles/octave fairly well. These ripple rates are substantially higher than those measured physiologically in the ferret and cat [3], [38] and likely reflect species-specific specializations analogous to those found with the contrast-sensitivity-functions in vision (e.g., [4, Fig. 5.2]). The origin of this contrast sensitivity curve may lie in the early stages of the auditory system and may not be related to any specific cortical mechanisms (see [44] for a detailed discussion).

2) Ripple Frequency: Another important perceptual measure is the sensitivity to a change in ripple frequency, which is also known as ripple frequency-difference-limens (FDL's). In

[1] A support of a function f is defined as $\{x: f(x) \neq 0\}$.

the context of the cortical model, this measure corresponds to the sensitivity to dilations in the input ripple, or equivalently, to translations in model output $r(x, \phi, s)$ along the scale s axis. The data suggest that thresholds for dilations are approximately constant at 20% for a broad range of input ripple frequencies (0.7–6 cycles/octave) [42], rising at both lower and higher ripples. The constancy of this threshold for the important intermediate ripple frequencies permits the assumption of a simple uniform organization along the scale axis of the model.

3) Ripple Phase: A third perceptual measure is the sensitivity to ripple phase changes or ripple phase-difference-limens (PDL's). These are also found to be constant at approximately 6° for lower ripple frequencies (< 2 cycles/octave) [42]. For higher ripples, PDL's increase roughly linearly with ripple frequency. This functional form of the phase sensitivity can be explained in a straightforward manner. First, note that the spatial position on the tonotopic axis is closely related to its spatial phase. Specifically, the pattern variation in terms of location and phase at spatial frequency k is given by

$$|\Delta\phi| = k|\Delta x|. \tag{8}$$

In the 3-D space of the cortical model, these two factors $\Delta\phi$ and Δx both account for the measurement of the displacement Δd of the pattern, i.e., Δd is a combination of Δx and $\Delta\phi$. Therefore, a change in the input profile can be detected if either $\Delta\phi$ or Δx of the model output exceeds the resolution of the corresponding axis. From (8), it can be concluded that at a low spatial frequency, Δd will be dominated mostly by $\Delta\phi$, and the trend reverses at high spatial frequencies. Therefore, the constant threshold at a low spatial frequency may simply reflect a constant resolution of the ϕ axis, whereas the linearly increasing threshold for high ripples may imply a similarly constant resolution on the tonotopic axis. From the experimental results, it can be estimated that the tonotopic resolution is about 0.016 octave, or 0.5%. This estimate is close to the results obtained from the frequency resolution experiments reported in [25] which are based on totally different paradigms.

In summary, considering the constancy of the phase sensitivity and the aforementioned arguments regarding the combined resolution along the tonotopic and phase axes, it is reasonable to assume a uniform organization along the ϕ axis of the cortical model similar to that along the other two axes (s and x).

C. Parameters of the Cortical Model

The basic operation of the cortical model is a local Fourier, or wavelet, analysis of the spectral profile. The analysis is performed at each point along the tonotopic axis by an ordered bank of filters, where each is tuned to a particular ripple (or spatial) frequency (k_c) and phase (ϕ_c). The filters are assumed to be of constant Q and with approximately 1 octave bandwidths (measured at the 3-dB points). These filter properties are derived from psychophysical and physiological estimates both from the auditory [13], [42] and the visual systems [4].

Many other details of the model remain uncertain because of the lack of experimental data. For instance, should the filters be considered of constant gain, constant energy, or some other intermediate form? In this paper, we adopt the simplest approach and assume the filters to be equally weighted (in the sense of (6)). This choice is justified for several reasons. First, as mentioned earlier, the ripple contrast sensitivity may well be accounted for by the early auditory system, and hence, no further cortical weighting is needed [44]. Second, even if a relatively gradual weighting on the filters is imposed, it can be readily shown that as in the auditory periphery [44], a compressive output nonlinearity (e.g., due to saturation and threshold in neural firing rates) coupled with subsequent reduction stages (discussed below) can effectively compensate for a nonuniform filter weighting.

D. Relation to Visual Processing

An appealing aspect of the local Fourier analysis model is its conceptual similarity to *spatial frequency analysis* that has long been prevalent in visual processing [4]. This approach is supported by substantial anatomical, neurophysiological, and psychophysical data, which is elegantly detailed in [4]. Ironically, in the vision community, the idea that the brain performs a local Fourier transformation of the input profile is motivated by its similarity to the cochlear transformations of the auditory system.

The correspondence between auditory ripple analysis and visual spatial frequency analysis is deeper than a mere analogy. As evidence to this claim, consider the closely matched values of the filter parameters and detection thresholds measured in the visual system, e.g., roughly constant Q and one-octave-wide filters, with constant 6° phase sensitivity increasing at higher ripple frequencies [4, Table 6.1, and Figs. 6.11 and 8.3]. These remarkable equivalences likely reflect modality-independent limitations imposed by identically structured sensory areas in the central nervous system [25]. For instance, the resolution of the analysis filters may simply be dictated by developmental rules limiting the minimum divergence or convergence of dendritic fields along the sensory epithelium (be it auditory or visual). Clearly, exploring further equivalences (and differences) between similarly defined psychophysical measures (e.g., FDL's for ripples *versus* gratings, which apparently have not been reported in the literature) would shed considerable light on the underlying functional organization of both systems.

V. Modeling the Dynamic Response Properties

Spectra of speech and other environmental sounds are rarely stationary for longer than a few milliseconds. Thus, it is important to include in the cortex model dynamic elements that can account for the basic response properties observed in physiological experiments with dynamic stimuli. We shall first discuss qualitative properties of the model arising from considerations of two such response features: 1) the correlations between FM directional and rate selectivity on the one hand and the symmetry and bandwidth of the response area on the other and the adaptive character of cortical responses, i.e., cells respond best near the onset of the stimulus and gradually weaken if it remains stationary.

A. Correlations of FM and Spectral Shape Selectivities

Numerous physiological experiments have demonstrated a consistent correlation between the FM and spectral shape selectivities for neurons in A1, and several hypotheses have been proposed to account for the data. One prevailing qualitative theory, which is based on the combination of temporal summation and lateral inhibition in the auditory system [10], [12], [14], suggests that FM responses may be explained by examining the temporal sequence activating the excitatory and inhibitory portions of the RF. For example, if an FM sweep first traverses the excitatory response area, discharges will be evoked and cannot be influenced by the inhibition later activated by the ongoing sweep. Conversely, if an FM sweep first traverses the inhibitory area, the inhibition may still be effective at the time the tone sweeps through the excitatory area. The response, assuming a temporal summation of the instantaneous triggers, will therefore be smaller in this direction of modulation. This theory may also explain why the response area bandwidth is correlated with the FM rate preference and why FM directional selectivity decreases with the FM rate [11].

B. The Adaptive Nature of Cortical Responses

When presented with a stationary stimulus like a continuous vowel or a single tone at BF, cortical cells usually respond strongly only near the onset of the stimulus. Furthermore, two successive stimuli can only be equally effective if a period of the order of tens of milliseconds between them has elapsed. The origins of these adaptation effects are uncertain but are likely to be present at many precortical locations along the auditory pathway. Their perceptual correlates, however, are quite significant. For instance, it has been demonstrated that in order for human subjects to correctly perceive speech with salient spectral transitions, the sound segments must contain the portions during which large temporal derivatives on the spectrum occur [7], [40]. These derivatives, in effect, may be accounted for by the adaptive influences discussed above.

In our dynamic model, we shall adopt a strategy similar to that in [7] by using a leaky derivative whose step response is the exponentially decaying function $e^{-t/\tau_o}u(t)$. In the following, we shall denote the adapted cortical selectivity as $\partial_t r(x, \phi, s, t)$, where ∂_t represents a leaky derivative operator. This leaky derivative, together with the temporal summation described in the previous subsection, will be included in the following to explain and predict the neural response selectivities to dynamic signals.

C. Dynamic Cortical Model

Aside from the leaky temporal derivative of the output, the basic new element added to the cortical model is a (leaky) temporal summation or smoothing stage at the input. Its effect is simply to increase the spatial spread of a dynamically varying auditory spectrum. Thus, given a nonstationary auditory spectrum $p_{\rm stim}(x, t)$, the input to the cortical model $p(x, t)$ is computed as

$$p(x,t) = p_{\rm stim}(x,t) *_t e^{-t/\tau}u(t)$$

where $e^{-t/\tau}u(t)$ is the impulse response of the leaky integrator. The cortical model output $r(\cdot)$ is then computed as before (3). Finally, the effects of adaptation are added at the end to obtain the final cortical output $\partial_t r(x, \phi, s, t)$.

D. Example of Model Responses to FM Tones

The above simple dynamic operations can account for the directional and rate selectivity to FM tones observed in cortical responses. To see this, consider the idealized spectrum of a single FM tone, which appears as a Dirac-delta pulse sweeping across the tonotopic axis

$$p_{\rm fm}(x,t) = \delta(x - vt)$$

where v represents the rate and direction of the FM sweep. Applying the leaky integrator (with τ time constant), the spectrum becomes

$$\begin{aligned} p(x,t) &= p_{\rm fm}(x,t) *_t e^{-t/\tau}u(t) \\ &= \frac{1}{v}e^{(x-vt)/v\tau}u(vt - x). \end{aligned}$$

The instantaneous spectral profile for the FM sweep above is an exponentially decaying pattern along the tonotopic dimension, as shown in the top panels of Fig. 5. It is evident from the above equations that the faster the sweep, the "broader" the resulting profile $p(x, t)$. Since a cortical neuron responds best to the stimulus whose spectral shape best matches its response area function, a broadly tuned neuron will prefer fast sweeps, and vice versa. Similarly, the tail of the exponential stretches to the low- (high-) frequency region for upward (downward) sweeps, corresponding to a negative (positive) characteristic phase. Therefore, a neuron asymmetrically inhibited by low frequency will prefer a downward sweep, and vice versa.

To be more specific, note that the FM response of a neuron is usually characterized by the spike counts within a short period of time $[t_1, t_1 + T]$ during which the tone traverses the response area of the neuron. Accordingly, the response of such a neuron with RF $w(x_c, \phi_c, s_c)$ can be described by

$$\begin{aligned} \gamma(x_c, \phi_c, s_c) &= \int_{t_1}^{t_1+T} \partial_t r(x_c, \phi_c, s_c, t)dt \\ &= \int_{t_1}^{t_1+T} \partial_t[p(x_c, t) *_x w(x_c, \phi_c, s_c)]dt. \end{aligned}$$

By the central value theorem, we may approximate the leaky derivative and rewrite the above equation as

$$\gamma(x_c, \phi_c, s_c) = Tp(x_c, t_o) *_x w(x_c, \phi_c, s_c)$$

for some $t_o \in [t_1, t_1 + T]$. Assume that t_o occurs approximately when the tone sweep $p_{\rm fm}(x, t)$ traverses to the center of the RF (arbitrarily labeled as $x_c = 0$). Then, given that

$$P(k) = \frac{-1}{jvk - 1/\tau} \tag{9}$$

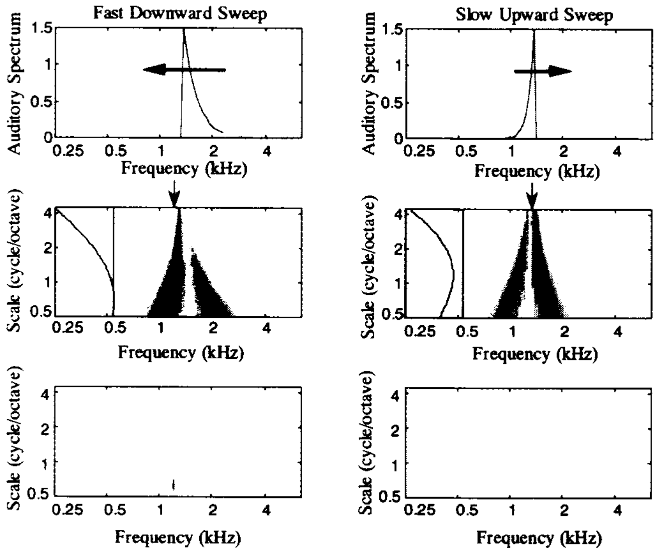


Fig. 5. Auditory spectra of two FM swept tones (top panels, with the moving directions indicated by arrows) and their corresponding cortical (middle) and sharpened (LIN-reduced) responses (bottom panels; see Section VI). The FM sweeps are of opposite directions and different rates ($v\tau = -0.26$ and 0.1 octaves in left and right columns, respectively). For $\tau = 16$ ms, these correspond to FM rates of 16.25 and 6.25 octaves/s, respectively. The detailed computer implementations are described in Appendices A and B. The reduced responses are sharpened versions of the cortical representation. As predicted by (10), the largest responses for each FM signal is close to a single point in the reduced representation. For the fast sweep (left), the largest response occurs at a lower scale, whereas the opposite is true for a slow sweep (right). In comparison, the cross sections along the s axis taken from the tonotopic locations marked by arrows in the middle panel are shown in the sidebars to indicate the maxima of the corresponding cortical responses. Since the direction of the sweep implies opposite spectral profile asymmetry, the downward (upward) sweep appears to have a positive (negative) phase, which is represented by blue (red) color in this figure.

the maximal response of $\gamma(x_c, \phi_c, s_c)$ can be obtained by matching the phase ϕ_c at the characteristic ripple frequency $k_c(= 1/|v_c|\tau)$, i.e.

$$v_c = -\tan\phi_c/k_c\tau. \qquad (10)$$

This equation suggests that neurons with broad tuning bandwidth (i.e., small k_c) will be selective to high FM rates, and the rate preference is inversely proportional to k_c or proportional to the tuning bandwidth. This is illustrated more clearly in the cross sections of the cortical responses to the FM sweeps (see the side figures in the middle panels in Fig. 5). Furthermore, FM directional sensitivity can be explained by noting that neurons whose RF have positive asymmetry ($\phi_c > 0$) prefer downward sweeps $v_c < 0$, and vice versa. In addition, since this equation is parameterized by $\tan\phi_c$, it implies that rate selectivity and directional sensitivity are best observed for ϕ_c far from 0, i.e., for neurons with asymmetric RF. These conclusions are consistent with current physiological data [37] and the propositions summarized in [10]. However, it cannot be overemphasized that our understanding of temporal processing mechanisms in biological systems is still in its infancy. The method employed in this section will most likely be proven to be a drastically simplified approximation of biological reality. Nevertheless, since this model is able to provide a consistent explanation to the observed phenomena, it seems to be a reasonable starting point for the design of future experimental paradigms.

VI. Interpreting the Cortical Representations

The multiscale cortical representation $r(\cdot)$ of the auditory spectrum is a highly redundant representation, whose presumed benefits include rendering perceptually significant features more explicit and robust. Such a claim can be partly based on the well-known advantages of of similar "wavelet-based" representations [29] and partly on our own findings from the mathematically analogous representation of the early auditory system [44]. However, in order to address these claims directly, it is necessary to gain more familiarity and intuition with the representation proposed by this cortical model. For instance, what features reflect the pitch and timbre of a vowel, the rate of an FM tone, or the descending formant of a consonant?

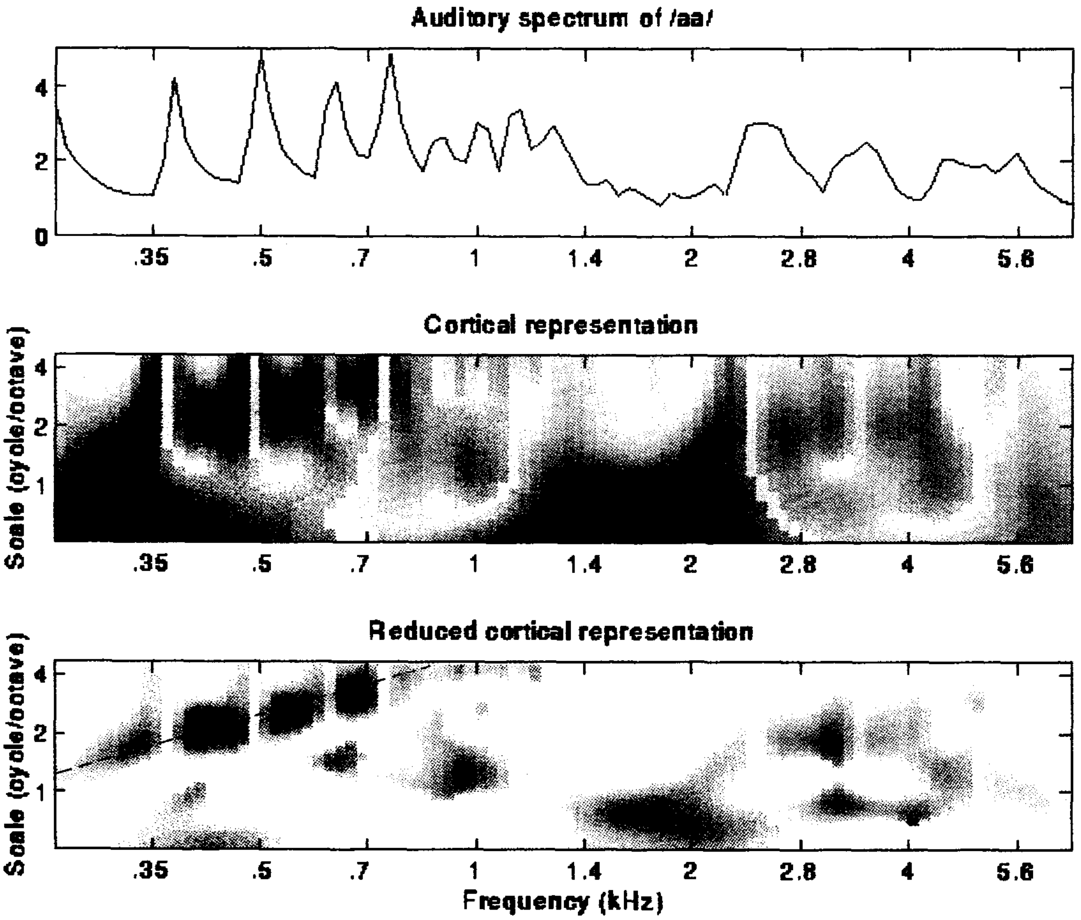
A. The Enhanced Cortical Representation

Because of the redundancy of the cortical representation $r(x, \phi, s)$, it is possible to generate a highly reduced version of it not only without significant loss of information but with some added advantages. For instance, as demonstrated earlier in the peripheral auditory system [44], extracting the output's peaks and edges coupled with its nonlinear compression can significantly enhance and stabilize its features. Such transformations can be readily accomplished here by *lateral inhibitory* interactions across the ϕ–s plane of the output representation. In fact, such cortical lateral inhibitory networks (LIN's) have been demonstrated in the visual cortex, where they act across channels tuned to different spatial frequencies [4].

The cortical outputs $r(x, \phi, s)$ and their enhanced versions are illustrated in Figs. 5 and 6 for FM tones and stationary vowels, respectively. The reduction is achieved by a recurrent LIN (which is further described in the Appendix), which is also commonly employed for feature extraction in image processing [20], [9]. As is evident, the reduced representation mostly preserves the regions with the largest curvature along the scale axis (which is called an edge in image processing). For the FM stimulus (Fig. 5), this occurs near $k_c = 1/|v_c|\tau$, where the ripple spectrum of the stimulus has a pole (9). Similarly, for the vowel stimuli (Fig. 6), the reduced representation suppresses all the uniformly activated outputs along the s axis (e.g., vertical stripes for vowel/iy/ at $x \approx 2$ and 2.8 kHz), preserving only the outputs near their maxima (i.e., at $s \approx 1.5$ cycles/octave and above). In a more abstract viewpoint, the salient features represented by the LIN-reduced cortical representations can be interpreted as a *feature set*

$$\mathcal{F} = \{(x_c, \phi_c, s_c)\text{: the cortical response } r(x_c, \phi_c, s_c) \text{ is an edge}\}.$$

For each triplet, the first component specifies at what frequency of the auditory spectrum an edge occurs, while the second and third correspond to the local symmetry and bandwidth. Since each (x_c, ϕ_c, s_c) also indicates a spatial location in the cortical model, this viewpoint suggests a *topographic* information representation principle that is common to neural systems



(a)

Fig. 6. Auditory spectra of naturally spoken vowels /aa, iy/ (a), and the corresponding cortical representations (middle) and their LIN-reduced versions (b). The scale axis is labeled by the characteristic ripple (or spatial frequency) k_c of the cortical filters. The symmetry index is represented by colors in the manner detailed in Appendix A, and the strength of the response is denoted by the saturation of the color.

and many multidimensional signal processing frameworks. In such a scheme, the magnitude of the neural response is less important than where an edge occurs since, as demonstrated in the peripheral auditory system [44], the latter may be designed to be less vulnerable to distortion.

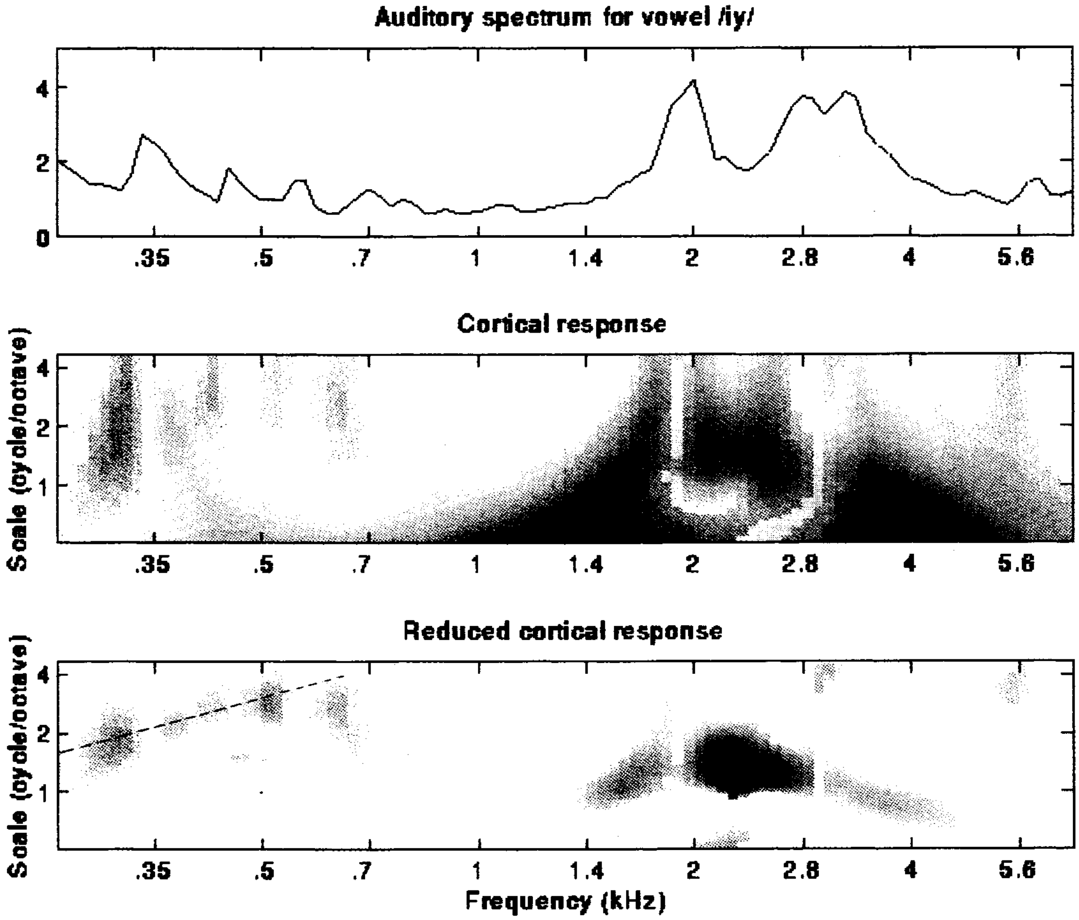
B. Cortical Representation of Speech Stimuli

Two examples of the cortical representation of speech signals are interpreted here in detail. In the first, which are the stationary vowels /aa, iy/ (as in 'hot' and 'heat,' respectively), the focus is on the features corresponding to the spectral shape and pitch of the signals. In the second, which is a continuous speech sequence, we emphasize the interpretation of the dynamic features in the response.

1) Stationary Vowels: Spectral shape is the most important factor in perceiving the timbre of a vowel. It is evident from Fig. 6 that this shape is dominated by the spectral peaks, i.e., the overall formant structure and the underlying harmonicity in the low-frequency region (usually <1 kHz). These features are explicitly analyzed in the cortical representation in terms of their local symmetry and bandwidth.

Consider, for example, the reduced representation of the vowel /aa/ (Fig. 6). The formants are relatively broad in bandwidth and, thus, are represented in the low scale regions: generally, ≤ 2 cycles/octave. For instance, the region of the third formant (at approximately 2.5 kHz) is approximately represented by edges at scales 0.5 and 2 cycles/octave. The higher scale corresponds to the third formant peak (which is approximately 0.25 octave in width). The lower scale captures the broad and skewed distribution of energy due to the combined third and fourth formant peaks. In contrast, the fine structure of the spectral harmonics is only visible at high scales (usually >1.5–2 cycles/octave).

The ripple (or spatial) phase information in this representation is encoded by the color of the response (see the Appendix). It provides a description of the local energy distribution in the auditory spectrum. For example, the tonotopic locations at which zero phase (yellow regions) responses occur closely reflect the positions of the peaks in the auditory spectrum, and the negative (red) and positive (blue) phase regions indicate whether the nearest spectral peak is at a higher or lower frequency. For instance, the spectral peak of vowel/aa/ at 3.25 kHz is not resolved at the broad scale, i.e., there is no yellow color at $s \approx 1$ at this frequency. Instead, it is regarded as a trough because it is flanked by two stronger peaks. However, the peak and its surrounding narrow valleys are resolved at



(b)

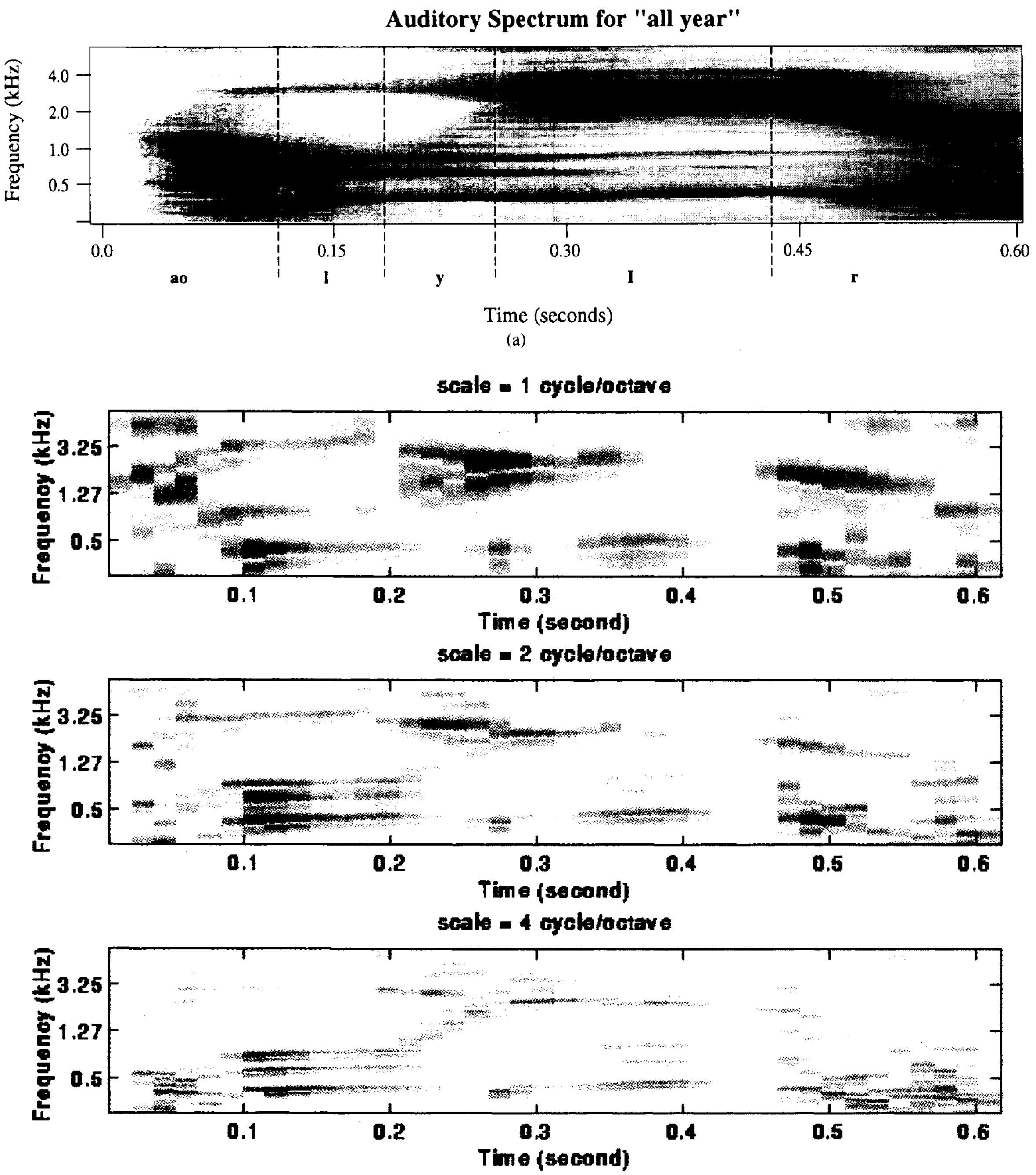
Fig. 6. *(continued).*

a higher scale corresponding to twice the spatial frequency ($s \approx 2$ cycles/octave). A similar "double-scale" representation occurs near 600–700 Hz, where the fine harmonic structure is represented at the higher scales, whereas the format structure (which is evident in the envelope of the harmonic peaks) is captured by the lower scales.

The pitch of the vowel complex can be discerned in the distinctive regular appearance of the harmonic portion of the spectrum in the cortical representation. Specifically, on a logarithmic frequency axis, a harmonic series appears as a train of peaks with progressively closer spacing or, in effect, a logarithmically increasing spatial frequency. Since the scale is a logarithmic mapping of spatial frequency, this trend is resolved as a straight line in a tonotopic-scale plane, which is outlined by the dashed lines in Fig. 6. A change in the fundamental frequency, corresponding to a change in the pitch of the vowel, simply results in a translation of the harmonic peak pattern along the tonotopic axis. Therefore, a harmonic series with a higher fundamental frequency appears again as a straight line with the *same* slope as before but horizontally shifted to the right (e.g., as in the case of the vowel /aa/ compared to /iy/). If a harmonicity detection mechanism is based on such a straight line with this particular slope, it should be able to interpolate or extrapolate the missing components to determine the fundamental frequency, much like the auditory perception of pitch [27].

A Segment of Continuous Speech Spectral transitions are important cues for sound classification and perception [30]. Such temporal variations of spectral energy distributions can be readily described in terms of the FM response characteristics of the cortex model discussed earlier (Section V). In Fig. 7, we illustrate an integrated *scalogram* of the adapted cortical representation of a speech segment ("all year") uttered by a female speaker. The representation is shown only at three scales: broad, intermediate, and fine at $s = 1, 2$, and 4 cycles/octave, respectively. The adaptation enhances the onsets and nonstationary portions of the signal by relatively suppressing the stationary spectral segments.

As with the vowel spectra (Fig. 6), the symmetry and local bandwidths of the spectral energy distribution at each time frame are represented by the different colors and scales. For instance, near the onset, the harmonic structure of the spectrum is seen clearly in the highest scale, whereas the formant structure is more concentrated in the lower two scales. An additional feature that is evident in this figure is the differential coloring of the spectral transitions. For example, the upward slow glide in the /y/ region ($s = 4$ panel) and the downward transition from /i/ to /r/ ($s = 2$ panel) are



(a)

(b)

Fig. 7. Auditory spectrogram (a) and the adapted cortical scalogram (b) of a segment of continuous speech "all year." The scale, symmetry, and tonotopic metrics are represented in the same manner as in Fig. 5 and described in Appendix A.

represented by the negative (red) and positive (blue) symmetry indices, respectively. Furthermore, because of their relatively slow rates, these transitions are more visible at the higher scales ($s = 2$ and 4). Such a representation is therefore able to provide an integrated representation of spectral scale, symmetry, and the direction and rate of transitions.

VII. Discussion and Conclusion

This paper presents a mathematical model of the functional organization of the primary auditory cortex. It is concluded that the central auditory system analyzes the acoustic spectrum along three feature axes: the spectral symmetry on the ϕ axis, the local bandwidth on the scale s axis, and the frequency components on the tonotopic x axis. This analysis can be viewed as a local Fourier transformation of the spectrum, which itself is a local Fourier transform of the acoustic signal.

Such double transform analysis bears the same merits as those of the conventional discrete cosine transform (DCT) or, more generally, the cepstrum-based frameworks that are often employed for spectral shape analysis. Usually, the cepstrum coefficients are obtained by taking the (inverse) Fourier transform on the logarithmic LPC or Fourier power spectrum of the signal. The high-order coefficients represent the rapidly varying properties in the spectrum, corresponding to features with high spatial frequency such as the harmonic peaks. The lower order coefficients describe the general trend of the spectral shape and, hence, can be associated with components having low spatial frequency such as the formants. From this perspective, the cepstral analysis is similar to the cortical approach. However, these two approaches differ fundamentally in that the cortical representation is *local* along the tonotopic axis, whereas the cepstral approaches are global in nature. For example, the first- and second-order cepstrum coefficient describe the overall spectral tilt and spectral compactness of the signal over all frequencies. Consequently, spectral shape information is scattered over all cepstrum coefficients and, hence, must be considered collectively and not individually.

In many areas of speech processing, spectral shape has been shown to provide a more complete representation of the speech signal than conventional approaches such as formant frequency analysis [1], [28], [46] and can be closely associated with speech production processes [6]. Spectral shape information, which is already present in the representation of earlier auditory stages, is simply made more explicit at the cortical level. The multiscale representation provides a complimentary perspective on the spectral shape, with coarse scales enhancing overall trends and fine scales resolving the local spectral details. While investigating the features across all the scales certainly gives a complete description of the spectral shape at a tonotopic location, it is nevertheless redundant. The feature extraction process of the LIN provides a means for enhancing the most important scales. Specifically, the basic function of the LIN is to detect the feature edges of the representation, which were postulated by [20] to contain all the information necessary to reconstruct the original signal. Although such an assertion is not verified, ample empirical studies have widely demonstrated its usefulness in practical applications [17], [18], [45].

Finally, the cortical representation is potentially capable of encoding dynamic features of the spectrum, such as the direction and rate of spectral peak transitions. While such transitions are known to be important cues for speech analysis, recognition, and perception [7], [16], [30], [39], their exact role is still uncertain because the dynamics of the spectral change are usually intimately connected with other acoustic features, such as the position of formants. Therefore, more elaborate models of these dynamic spectral effects must await further data from physiological and psychoacoustical experiments.

Appendix A
Implementation of the Cortex Model

For the computer simulations of the cortex model, we choose the seed function $h_m(x)$ to be the negative second derivative of a normal Gaussian function (Fig. 4). The rest of the RF's are then generated by dilating and sinusoidally interpolating $h_m(x)$ as described in Section III. The scale or spatial frequency axis covers the range from 0.5 cycles/octave to 4.6 cycles/octave at the resolution of a 20 channel-per-octave increase in spatial frequency, i.e., a dilation factor equivalent to 0.05/octave or 3.5% is used. This resolution is sufficient to estimate from the psychoacoustic experiments (Section IV). Choosing such a fine resolution is purely for the purpose of implementing the convolution in (12) (see Appendix B) with FFT; this task is much easier if the number of channels is some power of 2 and matches the number of phase channels. Similarly, the phase axis is discretized into 64 channels, covering -180 to $180°$.

The tonotopic axis is discretized in the same way as in the peripheral model [44] at 20 channel/octave, covering the frequency range from 250 to 6.7 kHz. In the 2-D display, the cortical response is shown in the form of $a(x, s)$ and $\psi(x, s)$, in which the phase is indicated by colors in the following manner:

$-3\pi/4 - -\pi/4$	red
$-\pi/4 - \pi/4$	yellow
$\pi/4 - 3\pi/4$	blue
other	purple.

The magnitude of $a(x, s)$ is linearly represented by the saturation of the color.

Appendix B
Implementation of the LIN

The LIN model for feature extraction implemented here is a recurrent network that can be described as

$$\tau\frac{dq}{dt} + q(x,\phi,s,t) = A(\phi,s) * y(x,\phi,s,t) + r(x,\phi,s,t) \quad (11)$$

$$y(x,\phi,s,t) = g(q(x,\phi,s,t)) \quad (12)$$

where $g(\cdot)$ is a nonlinear function accounting for the finite output dynamic range of the neurons, $A(\phi, s)$ models the inhibitory connections among neighboring neurons on the ϕ-s plane, and the steady state output of $y(x, \phi, s, t)$ is the LIN-reduced cortical response. In this paper, a Mexican hat [20], or the second derivative of Gaussian function with σ equal to five channels, is used to model the inhibitory connection matrix $A(\phi, s)$ of the LIN in (12). The nonlinear function $g(\cdot)$ is implemented with a half-wave rectifier. Similar to the implementation of [34], the differential equation in (12) is approximated in the discrete time difference equation. The steady-state output $y(\cdot)$ is obtained by iteratively applying $A(\phi, s)$ to the intermediate outputs. All initial conditions are set at rest. In most cases, it is found that 40 iterations are usually enough to reach the steady state.

In this paper, the output of the LIN can be approximately predicted as follows. First note that the cortical selectivity $r(\cdot)$ along each ϕ axis is always a sinusoid. Hence, the feature edge enhanced by LIN on the ϕ axis is the maximum located at $\psi(x, s)$, where it assumes the value $a(x, s)$. Therefore, effectively, it is the $a(x, s)$ of each s that is involved further in the local competition in the LIN, i.e., the 2-D inhibition (on the ϕ-s plane) can be realized with only a 1-D lateral inhibition on $a(x, s)$ along the s axis. Since s can be associated with logarithmic spatial frequency, the local magnitude response at each spatial location is a local "magnitude spectrum" of the input. Therefore, the lateral inhibition may be interpreted as detecting the edges in the magnitude of the auditory spectrum. To illustrate this argument, consider the FM stimulus in Fig. 5. From (9), this pattern has a pole at $k_c = 1/|v|\tau$. Its magnitude spectrum can therefore be approximated in the Bodé plot as being flat for $k < k_c$ and having a 3-dB/octave decay for $k > k_c$. Conceptually, the corner frequency k_c is where the magnitude spectral profile has the largest curvature, and hence, the LIN detects an edge only at the scale corresponding to k_c. According to (7)

$$\begin{aligned} s_c &= (\log k_c - \log k_m)/\log \alpha \\ &= (-\log v - \log k_m \tau)/\log \alpha \end{aligned}$$

which is equal to 0.5 and 2 cycles/octave in this example. This value is close to the output computed from the full LIN simulation (Fig. 5).

References

[1] S. E. Blumstein and K. N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Amer.*, vol. 67, no. 2, pp. 648–662, Feb. 1980.

[2] W. Byrne, J. Robinson, and S. A. Shamma, "The auditory processing and recognition of speech," in *Proc. Speech Natural Language Workshop*, Oct. 1989, pp. 325–331.

[3] B. M. Calhoun and C. E. Schreiner, "Spatial frequency filters in cat auditory cortex," in *Proc. 23rd Ann. Mtg. Soc. Neurosci.*, Washington, DC, Nov. 1993.

[4] R. L. De Valois and K. K. De Valois, *Spatial Vision*. New York: Oxford, 1990.

[5] A. Dobbins, S. W. Zucker, and M. S. Cynader, "End-stopped neurons in the visual cortex as a substrate for calculating curvature," *Nature*, vol. 329, pp. 438–441, Oct. 1987.

[6] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. New York: Springer-Verlag, 1972.

[7] S. Furui, "On the tole of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, no. 4, pp. 1016–1025, Oct. 1986.

[8] D. M. Green, "Frequency and the detection of spectral shape change," in *Auditory Frequency Selectivity*, B. J. C. Moore and R. D. Patterson, Eds. New York: NATO ASI Series, 1986, pp. 351–360.

[9] D. M. Green, *Profile Analysis: Auditory Intensity Discrimination*. New York: Oxford, 1988.

[10] P. Heil, R. Rajan, and D. R. Irvine, "Sensitivity of neurons in cat primary auditory cortex to tones and frequency-modulated stimuli II: Organization of response properties along the 'isofrequency' dimension," *Hearing Res.*, vol. 63, nos. 1/2, pp. 135–156, Nov. 1992.

[11] P. Heil, R. Rajan, and D. R. Irvine, "Sensitivity of neurons in cat primary auditory cortex to tones and frequency-modulated stimuli I: Effects of variations of stimulus parameters," *Hearing Res.*, vol. 63, nos. 1/2, pp. 108–134, Nov. 1992.

[12] P. Heil and H. Scheich, "Spatial representation of frequency-modulated signals in the tonotopically organized auditory cortex analogue of chick," *J. Comparative Neurology*, vol. 322, no. 4, pp. 548–565, Aug. 1992.

[13] D. A. Hillier, "Auditory processing of sinusoidal spectral envelopes," Ph.D. Thesis, Sever Inst. of Technol., Washington Univ., St. Louis, MO, 1991.

[14] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *J. Physiol.*, vol. 195, pp. 215–243, 1968.

[15] E. R. Kandel and J. H. Schwartz, *Principles of Neural Science*. New York: Elsevier, 1985.

[16] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden markov model," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[17] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 14, no. 7, pp. 710–732, July 1992.

[18] S. G. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 2091–2110, Dec. 1989.

[19] B. S. Manjunath and R. Chellappa, "A unified approach to boundary perception: Edges, textures, and illusory contours," *IEEE Trans. Neural Networks*, vol. 4, no. 1, pp. 96–108, 1993.

[20] D. Marr, *Vision*. New York: W. H. Freeman, 1982.

[21] J. R. Mendelson and M. S. Cynader, "Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation," *Brain Res.*, vol. 327, pp. 331–335, 1985.

[22] M. M. Merzenich, P. L. Knight, and G. L. Roth, "Representation of cochea within the primary auditory cortex in cat," *J. Neurophysiol.*, vol. 28, pp. 231–249, 1975.

[23] A. R. Moller, "Analysis of the auditory system using pseudo-random noise," in *Advanced Methods of Physiological System Modeling, vol. 1*, V. Z. Marmarelis, Ed. Los Angeles: Biomed. Simulations Resource, Univ. of Southern Calif., 1987.

[24] B. C. J. Moore and B. R. Glasberg, "Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech," Scandinavian audiology supplementum 32, Stockholm, Sweden, 1989; distributed by Almqvist and Wiksell.

[25] D. D. M. O'Leary, "Do cortical areas emerge from a protocortex?" *Trends Neuronsci.*, vol. 12, no. 10, pp. 400–406, Oct. 1989.

[26] R. Plomp, *Aspects of Tone Sensation: A Psychophysical Study*. New York: Academic, 1976, pp. 85–110, ch. 6.

[27] R. Plomp, *Aspects of Tone Sensation: A Psychophysical Study*. New York: Academic, 1976, pp. 111–142, ch. 7.

[28] B. H. Repp and H.-B. Lin, "Acoustic properties and perception of consonant release transients," *J. Acoust. Soc. Amer.*, vol. 85, no. 1, pp. 379–396, Jan. 1989.

[29] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, pp. 14–38, Oct. 1991.

[30] S. Rosen and A. Fourcin, "Frequency selectivity and the perception of speech," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. New York: Academic, 1986, pp. 373–487, ch. 7.

[31] C. E. Schreiner and M. L. Sutter, "Functional topography of cat primary auditory cortex: distribution of integrated excitation," *J. Neurophys.*, vol. 64, no. 5, pp. 1442–1459, Nov. 1990.

[32] ——, "Topography of excitatory bandwidth in cat primary auditory cortex: Single-neuron versus multiple-neuron recordings," *J. Neurophysiol.*, vol. 68, no. 5, pp. 1487–1515, Nov. 1992.

[33] S. A. Shamma, "Speech processing in the auditory system II: lateral inhibition and the central processing of speech evoked activity in the auditory nerve," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1622–1632, Nov. 1985.

[34] ——, "The acoustic features of speech sounds in a model of auditory processing: Vowels and voiceless fricatives," *J. Phonetics*, vol. 16, pp. 77–91, 1988.

[35] ——, "Spatial and temporal processing in central auditory networksm," in *Methods in Neural Modeling*, C. Koch and I. Segev, Ed. Cambridge, MA, MIT Press, 1989.

[36] S. A. Shamma and G. Chettiar, "A functional model of primary auditory cortex: Spectral orientation columns," Tech. Rep. TR90–47, Syst. Res. Center, Univ. of Maryland, 1990.

[37] S. A. Shamma, J. W. Fleshman, P. W. Wiser, and H. Versnel, "Organization of response areas in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 69, no. 2, pp. 367–383, Feb. 1993.

[38] S. A. Shamma, H. Versnel, and N. Kowalski, "Organization of primary auditory cortex evident in responses to rippled complex sound stimuli," in *Proc. 23rd Ann. Mtg., Soc. Neurosci.*, Washington, DC, Nov. 1993.

[39] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 64, no. 5, pp. 1358–1368, Nov. 1978.

[40] W. Strange, J. J. Jenkins, and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 695–705, 1983.

[41] N. Suga, "Feature extraction in the auditory system of bats," in *Basic Mechanisms of Hearing*, A. R. Møller, Ed. New York: Academic, 1973, pp. 675–744.

[42] S. Vranic-Sowers, "Modeling the perception of profile changes," Ph.D thesis, Dept. of Elect. Eng., Univ. of Maryland, College Park, 1993.

[43] K. Wang, "Neural networks that recognize phonemes by their acoustic features," Master's Thesis, Univ. of Maryland, Dec. 1989.

[44] K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 421–435, July 1994.

[45] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Information Theory* (Special Issue on Wavelet Transforms and Multiresolution Signal Analysis), vol. 38, no. 2, pp. 824–839, Mar. 1992.

[46] S. A. Zahorian and A. J. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Amer.*, vol. 94, no. 4, pp. 1966–1982, Oct. 1993.

Kuansan Wang received the B.S. degree in 1986 from National Taiwan University and the M.S. and Ph.D. degrees in 1989 and 1994, respectively, from the University of Maryland, College Park, all in electrical engineering.

From 1988 to 1992, he was an SRC Fellow at the Systems Research Center at the University of Maryland. From 1992 to 1994, he was a research assistant at the Institute for Systems Research, University of Maryland. Since May 1994, he has ben with the Speech Research Department of AT&T Bell Labortories, Murray Hill, NJ. His research interests are auditory modeling, neural networks, and speech analysis and recognition.

Shihab A. Shamma received the B.Sc. degree in electrical engineering from Imperial College, London University, in 1976 and the M.S., Ph.D., and M.A. degrees from Stanford University, Stanford, CA, in 1977, 1980, and 1980, respectively.

From 1981 to 1984, he performed postdoctoral research into the physiology of hearing and the mathematical modeling of the nervous systems at Stanford and the National Institutes of Health, Bethesda, MD. In 1984, he joined the Department of Electrical Engineering, University of Maryland, College Park, where he is currently a Professor. His research interests have focused on theoretical and experimental studies of sound processing in the auditory system.