Spectrum Restoration From Multiscale Auditory Phase Singularities by Generalized Projections

Taishih Chi, *Member, IEEE,* and Shihab A. Shamma, *Senior Member, IEEE*

***Abstract*—We examine the encoding of acoustic spectra by parameters derived from singularities found in their multiscale auditory representations. The multiscale representation is a wavelet transform of an auditory version of the spectrum, formulated based on findings of perceptual experiments and physiological research in the auditory cortex. The multiscale representation of a spectral pattern usually contains well-defined singularities in its phase function that reflect prominent features of the underlying spectrum such as its relative peak locations and amplitudes. Properties (locations and strength) of these singularities are examined and employed to reconstruct the original spectrum by using an iterative projection algorithm. Although the singularities form a nonconvex set, simulations demonstrate that a well-chosen initial pattern usually converges on a good approximation of the input spectrum. Perceptually intelligible speech can be resynthesized from the reconstructed auditory spectrograms, and hence these singularities can potentially serve as efficient features in speech compression. Besides, the singularities are very noise-robust which makes them useful features in various applications such as vowel recognition and speaker identification.**

***Index Terms*—Auditory model, convex projection, phase singularity, spectrum restoration.**

I. Introduction

SIGNAL discontinuities such as edges and peaks, have played a key role in the representation and encoding of signals, especially of audio and images [1]–[4]. The importance of these features stems primarily from their enhanced detectability by the human sensory system, and hence their perceptual role in interpreting scenes and sound [5], [6], and their efficiency as encoders of perceptually faithful versions of the underlying signal [7].

To define, detect, and process these features, several multiscale representations have been proposed and proven effective in image texture analysis, measurement of binocular disparity and image orientation in the field of early biological and computational visual processing [7]–[12]. These approaches typically involve the use of Gaussian-like filter banks followed by detection of the zero crossings of the second derivative to localize the edges. Much less investigated is the local phase of the filters' responses, which has been found moderately useful in binocular depth and disparity estimation problems [13]. A common difficulty with utilizing the phase is the existence of singularities in scale space at which the phase is discontinuous and ill-defined. While this singularity is usually avoided in applications [14], [15], it is possible that they may play a role in a robust representation of the signal, one akin to that played by the amplitude discontinuities (edges and peaks).

We examine here this possibility in the context of the auditory processing of complex sounds. Specifically, physiological and psychophysical evidence suggests that the auditory system analyzes and extracts a multiresolution representation of their input sound [16]–[19]. A model of this process has been developed and exploited in a variety of applications including the assessment of speech intelligibility and the perception of complex sounds [20]–[22]. In its simplified version [23], the model performs an affine wavelet transformation on the auditory spectrum of its input sound. As in vision, this representation contains singularities in scale space that reflect the shape of the input spectrum. We describe in this paper how these singularities can be exploited to reconstruct the auditory spectrum that generates them using iterative projection methods [24]–[26]. Such algorithms can also be used to reconstruct perceptually comparable sounds from the reconstructed auditory spectrum, but not necessarily the identical original waveforms [17].

This paper is organized as follows. In Section II, we describe the multiscale auditory representation of the input spectrum and explain how singularities in scale space are expressed and detected. In Section III, an iterative algorithm is formulated to reconstruct the original spectra from certain parameters of the singularities such as their locations, gradients (strength), and the energies at the scales where they occur. A critical factor in acceptable reconstructions is the choice of the initial (starting) spectrum to invert. Consequently, in Section IV we propose procedures to estimate initial approximations of the signal spectrum from singularities that contain critical features of the desired spectrum. In Section V, we demonstrate the robustness of the singularities by conducting a vowel recognition task. The recognition performance by singularity features is compared with the performance by the mel-frequency cepstral coefficient (MFCC) features. We end in Section VI with a summary and brief discussions of their potential applications.

II. Multiscale Cortical Processing and Singularities

The multiscale model of the auditory cortex integrates findings from a wide range of physiological and psychoacoustic sources. The details of these experimental findings and their interpretation in the context of the model are available elsewhere [16], [18], [19]. Most relevant for our purposes here is the topographic

Manuscript received July 9, 2004; revised November 2, 2004 and June 10, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

T. Chi is with the Department of Communication Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: tschi@cm.nctu.edu.tw).

S. A. Shamma is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: sas@isr.umd.edu).

Digital Object Identifier 10.1109/TSA.2005.860828

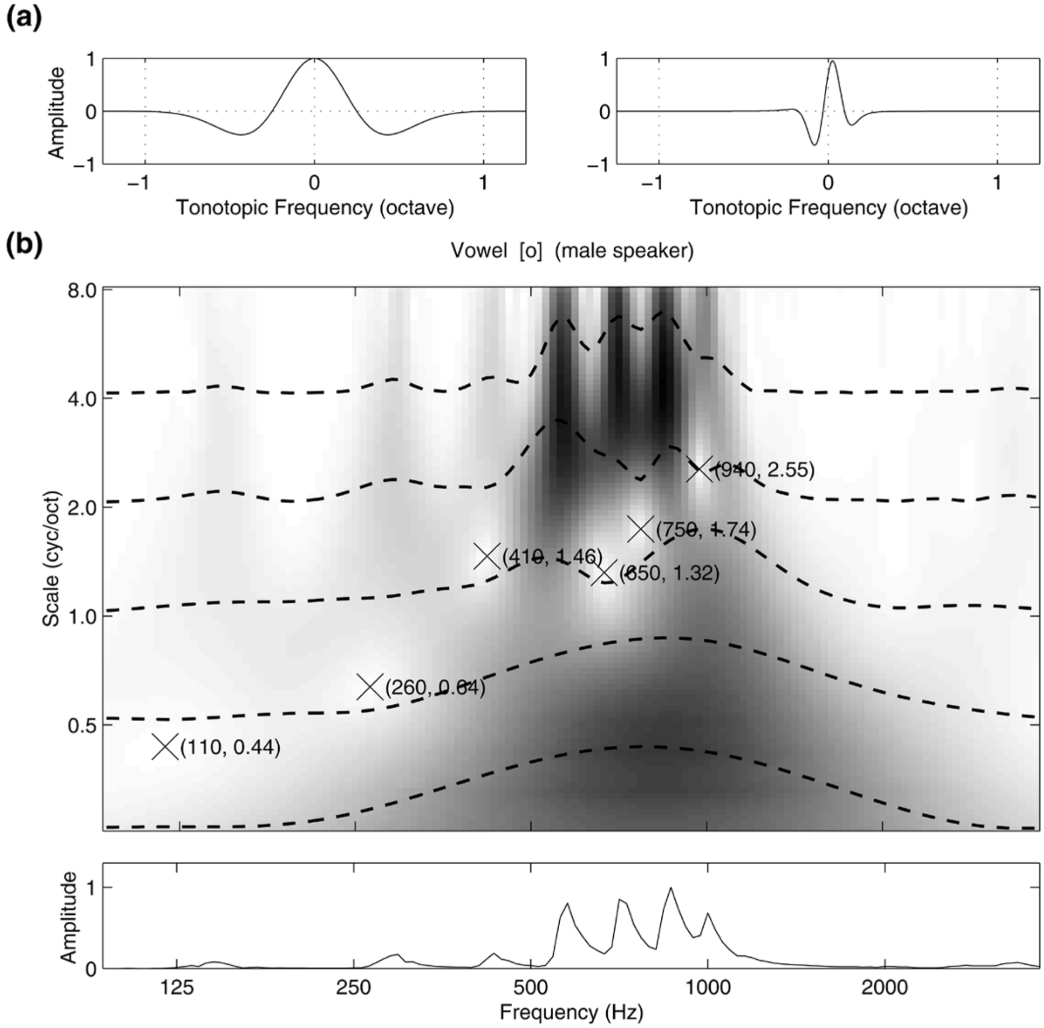


Fig. 1. Representative receptive fields and multiscale representation of the vowel. (a) Two representative RFs. The RF in left panel is tuned to $\Omega_c = 1\ \mathrm{cycle/octave}$ and $\phi_c = 0$. RF in right panel is tuned to $\Omega_c = 4\ \mathrm{cycle/octave}$ and $\phi_c = \pi/4$. (b) A multiscale representation of the vowel [*o*] (as in "home"). The abscissa is the tonotopic (log) frequency (x) axis. The ordinate is the scale (Ω) axis. Several cross-sectional profiles of the magnitude response are shown at different scales ($\Omega = .25$,.5, 1.0, 2.0 and 4.0 cyc/oct). Some of the singularities are marked by an "X" symbol. Their locations are given in the parenthesis as (frequency in hertz, scale in cyc/oct).

organization of neuron responses in primary auditory cortex (AI) to various stimulus features [27]. For instance, unit responses exhibit an organized distribution of their frequency tuning or "best frequencies" (BF), local symmetry above and below BF, and the local bandwidth around the BF (see review of these data in [28], [29]). To capture these organizational principles, the cortical model assumes that the receptive field (RF) of a neuron could be characterized by three parameters : best frequency (BF), bandwidth, and asymmetries. Arrays of neurons tuned to different BF's, bandwidths, and asymmetry, would then effectively compute a multiscale representation of the input spectrum. Therefore, from a mathematical point of view, functions of arrays of neurons can be modeled by a complex wavelet transform as performing a multiscale analysis on the input stimulus. A brief review of this multiscale analysis model of the AI is given below and much more detailed validation and discussions about this model can be found in [23]. Furthermore, such multiscale cortical representation has already been validated by successful applications in the manipulations of sound percepts [30], [31].

A. Multiscale Cortical Model

The input spectral profile to the cortex is extracted in the early auditory pathway (from cochlea to midbrain) and is referred to as the "auditory spectrum" in this study [20]. Functions of arrays of cortical neurons with receptive fields (RF's) centered at different frequencies along the tonotopic (logarithmic) frequency axis x, and with a range of bandwidths and asymmetries can be modeled as performing zero-lag cross-correlations between RF's and the input auditory spectrum [23]. Fig. 1(a) illustrates two examples of such RF's. The asymmetries of RF's can be modeled by sinusoidally interpolating a symmetric seed function $h(\cdot)$ and its Hilbert transform. Therefore, the RF of neuron c can be formulated as [23]

$$\begin{aligned}\mathcal{RF}(x; x_c, \Omega_c, \phi_c) = h(x - x_c; \Omega_c) \cos\phi_c \\ - \hat{h}(x - x_c; \Omega_c) \sin\phi_c \quad (1)\end{aligned}$$

where the $h(x; \Omega_c)$ is a real, even function (i.e., $h(x - x_c; \Omega_c) = h(x_c - x; \Omega_c)$) and with Ω_c (in cycle/octave) as the bandwidth parameter, and ϕ_c is the characteristic phase (in radians) which determines the asymmetry of the RF. $\hat{h}$ denotes the Hilbert transform of the function h (i.e., $\hat{h}$ is an odd function and $\hat{h}(x - x_c; \Omega_c) = -\hat{h}(x_c - x; \Omega_c)$). The exact shape of this even function is not important as long as it can manifest the lateral-inhibitory structure, i.e., a central excitatory (positive) band symmetrically flanked by inhibitory (negative) side bands. The RF's shown in Fig. 1(a) are based on a Gabor function formulation [23]. The left panel shows an RF tuned to $\Omega_c = 1$ cycle/octave and $\phi_c = 0$, while the RF in the right panel is tuned to $\Omega_c = 4$ cycle/octave and $\phi_c = \pi/4$. The response of the neuron tuned to (x_c, Ω_c, ϕ_c) for an auditory spectrum $y(x)$ is computed as the inner product of the $\mathcal{RF}$ and $y(x)$ [23]

$$\begin{aligned} r(x_c, \Omega_c, \phi_c) &= \langle y(x), \mathcal{RF}(x; x_c, \Omega_c, \phi_c)\rangle && (2)\\ &= a(x_c, \Omega_c)\cos(\psi(x_c, \Omega_c) - \phi_c) && (3)\end{aligned}$$

where

$$\begin{aligned} a(x_c, \Omega_c) &= \left\{\langle y(x), h(x_c - x; \Omega_c)\rangle^2 \right.\\ &\quad \left. + \langle y(x), \hat{h}(x_c - x; \Omega_c)\rangle^2\right\}^{1/2} && (4)\\ \psi(x_c, \Omega_c) &= \arctan\frac{\langle y(x), \hat{h}(x_c - x; \Omega_c)\rangle}{\langle y(x), h(x_c - c; \Omega_c)\rangle} && (5)\end{aligned}$$

are called the characteristic amplitude and the characteristic phase of the response, respectively; and $\langle \cdot, \cdot \rangle$ denotes the inner product.

The above characteristic amplitude and phase responses of neuron c can be computed by a complex wavelet transform as follows. Assume an analytical function h_w is defined as

$$h_w(x; \Omega_c) = h(x; \Omega_c) + j\hat{h}(x; \Omega_c).$$

Then the linear convolution of input $y(x)$ and function h_w can be derived as

$$\begin{aligned} z(x_c, \Omega_c) &= y(x) * h_w(x; \Omega_c)|_{x=x_c} && (6)\\ &= a(x_c, \Omega_c)e^{j\psi(x_c, \Omega_c)} && (7)\end{aligned}$$

with the same characteristic amplitude $a(x_c, \Omega_c)$ and the characteristic phase $\psi(x_c, \Omega_c)$ as in (4) and (5). In other words, the output amplitude and phase responses of neuron c can be computed by the complex wavelet transform (6) with a mother wavelet $h_w(x)$ and for different Ω_c

$$h_w(x; \Omega_c) = \Omega_c h_w(\Omega_c x).$$

An example of this multiscale representation is shown in Fig. 1(b) for the auditory spectrum of the vowel $[o]$ (as in "home"). The auditory spectrum is depicted at the bottom and the corresponding multiscale magnitude response $(a(x, \Omega))$ is displayed above it. The superimposed dashed lines are the magnitude $(a(x, \Omega_c))$ at each of the different scales Ω_c. RF's with the widest bandwidths (i.e., at the lowest scale of $\Omega_c = 0.25$) smooth over the details of the vowel spectrum and hence capture only its major outlines. Such outlines are referred to as the "global shape" of the spectrum in this study. Meanwhile, RF's with progressively narrower bandwidths display finer response features (peaks and valleys). Consequently, the magnitude responses at different scales simultaneously represent the *local energy* of the vowel spectral pattern at various degrees of resolution.

B. Occurrence of Singularities

The scale space defined by the complex wavelet transform (6) is analytic with a number of isolated zeros $(z(x, \Omega) = 0)$. Zeros of the magnitude $a(x, \Omega)$ are marked by an "X" symbol in Fig. 1(b). The phase response $\psi(x, \Omega)$ is also differentiable except at the zeros of $a(x, \Omega)$ where the complex response passes the origin in the complex plane. At these points, phase discontinuities occur and jump by π. These points are called singularities.

Fig. 2 demonstrates the signal behavior around the singularity at (260 Hz, 0.64 cyc/oct) in Fig. 1(b). The three panels in Fig. 2(a) illustrate the behavior of $z(x, \Omega)$ as a function of x in the complex plane at Ω scales above (left panel), at (middle panel), and below (right panel) the singularity, respectively. Fig. 2(b) panels illustrate the derivative of the phase function ($(\partial/\partial x)\psi(x, \Omega)$, also known as "*local frequency*") of $z(x, \Omega)$ at the same three scales as those of Fig. 2(a). It is evident that away from the singularity, this derivative remains rather smooth. However, its absolute value increases sharply as the singularity is approached from the right and left (along the x-axis). Furthermore, it undergoes a rapid change of sign along the scale axis. In a continuous representation of this scale-space plot, $(\partial/\partial x)\psi(x, \Omega)$ tends to ∞ or $-\infty$ as the singularity is approached. This implies a numerical instability in the neighborhood (along both x and Ω axes) of the singularities and is the rationale of avoiding such regions in measuring binocular disparity or image velocity [14], [15].

C. Strength of Singularity

Here we define the strength of a singularity and discuss how it can be used as a measure of its significance. In the following analysis, we assume a singularity appears at (x_0, Ω_0) in the scale space. Therefore

$$\begin{aligned} \Re\{z(x_0, \Omega_0)\} &= 0 && (8)\\ \Im\{z(x_0, \Omega_0)\} &= 0 && (9)\end{aligned}$$

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

The real part of the bandpass analytical signal $z(x, \Omega_0)$ can be expressed as [32]

$$g(x) \triangleq \Re\{z(x, \Omega_0)\} = \frac{1}{\pi}\int_{\Omega_a}^{\Omega_b} |G(\Omega)|\cos(\Omega x + \angle G(\Omega))d\Omega \quad (10)$$

where $G(\Omega)$ is the Fourier integral of $g(x)$ and Ω_a, Ω_b are the cutoff frequencies of the corresponding bandpass filter centered at Ω_0. The discrete-time implementation of (10) gives

$$g(n) = \sum_{i; \Omega_a \leq \Omega_i \leq \Omega_b} a_i \cos(\Omega_i n + \phi_i) \quad (11)$$

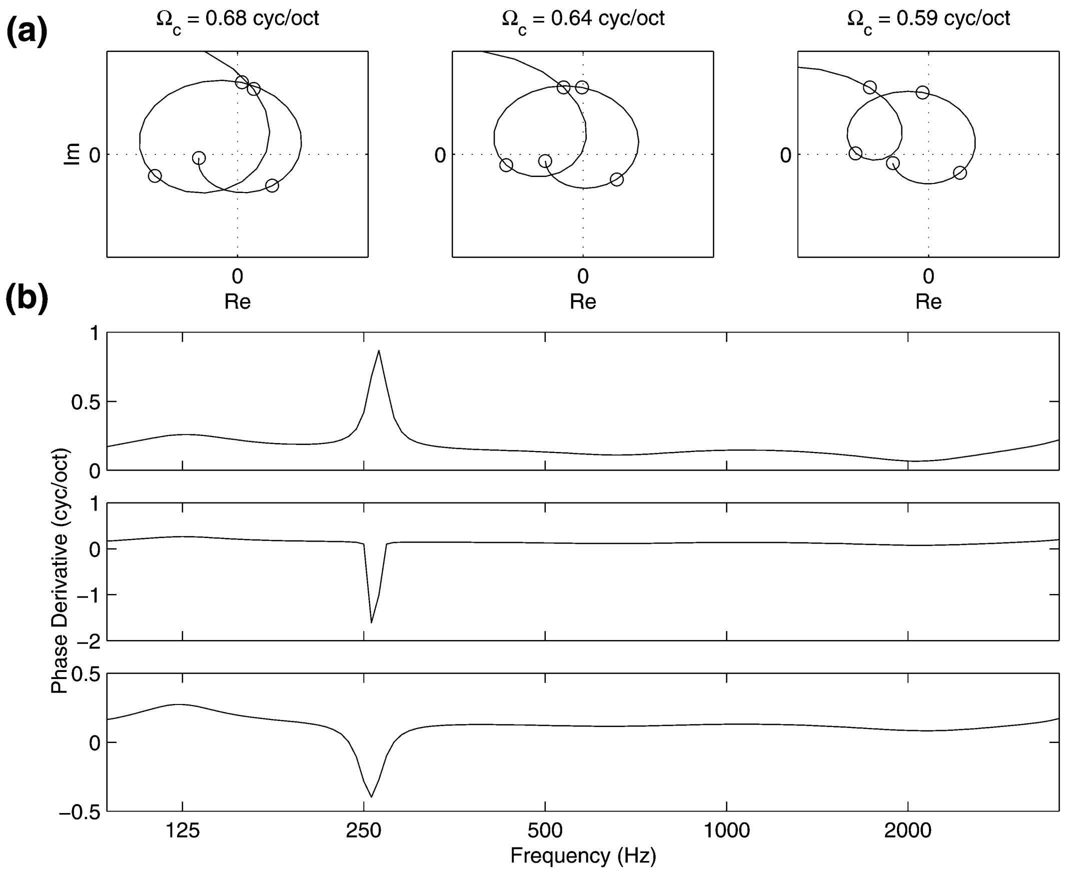


Fig. 2. Cortical response and the phase derivative near a singularity. (a) The evolution of response through a singularity from high scale ($\Omega_c = 0.68$ cyc/oct; left panel) to low scale ($\Omega_c = 0.59$ cyc/oct; right panel). The center panel is at approximately the scale ($\Omega_c = 0.64$ cyc/oct) where the magnitude response passes through origin indicating the occurrence of a singularity. The symbol "o" marks points at 90, 120, 160, 214 and 285 Hz along the x axis. $\mathrm{Re}(\mathrm{Im})$ indicates the real (imaginary) part of a complex number. (b) The derivative of the phase functions at three scales as selected in (a) ($\Omega_c = 0.68, 0.64$ and 0.59 cyc/oct). The top (bottom) panel is at the scale right above (below) the scale of singularity.

and

$$\hat{g}(n) = \Im\{z(n,\Omega_0)\} = \sum_{i;\Omega_a \le \Omega_i \le \Omega_b} a_i \sin(\Omega_i n + \phi_i) \quad (12)$$

where a_i and ϕ_i are the normalized magnitude and phase of the ith frequency component at Ω_i. Accordingly, the singularity appears at (n_0, Ω_0) in the discrete scale space. To simplify notation, the constraint $\Omega_a \le \Omega_i \le \Omega_b$ will be omitted in the following analysis.

Equations (11) and (12) around n_0 can be approximated as

$$g(n) \approx -\sum_i a_i \Omega_i \sin(\Omega_i n_0 + \phi_i)(n - n_0) \quad (13)$$

$$\hat{g}(n) \approx \sum_i a_i \Omega_i \cos(\Omega_i n_0 + \phi_i)(n - n_0) \quad (14)$$

by the first order Taylor series expansions. Therefore, the local energy $E(n)$, which is defined as $E(n) \triangleq |z(n,\Omega_0)| = \sqrt{g^2(n) + \hat{g}^2(n)}$, around n_0 can be derived as

$$E(n)^2 \approx (n-n_0)^2 \cdot \left[\sum_i a_i^2 \Omega_i^2 + 2\sum_{i,j}\sum_{;i\neq j} a_i a_j \Omega_i \Omega_j \cos((\Omega_i - \Omega_j)n_0 + (\phi_i - \phi_j))\right]. \quad (15)$$

If the bandpass filter is narrow-band $(\Omega_i \simeq \Omega_j \simeq \Omega_0)$

$$E(n) \approx |n - n_0| \cdot \Omega_0 \cdot \sqrt{\sum_i a_i^2 + 2\sum_{i,j}\sum_{;i\neq j} a_i a_j \cos(\phi_i - \phi_j)}. \quad (16)$$

Therefore

$$\begin{aligned}\frac{\Delta^2 E(n_0)}{\Omega_0} &\approx 2 \cdot \sqrt{\sum_i a_i^2 + 2\sum_{i,j}\sum_{;i\neq j} a_i a_j \cos(\phi_i - \phi_j)} \\ &\le 2 \cdot \sqrt{\sum_i a_i^2 + 2\sum_{i,j}\sum_{;i\neq j} a_i a_j} \\ &= 2 \cdot \sum_i a_i \end{aligned} \quad (17)$$

where $\Delta^2 E(n) \triangleq E(n+1) - 2E(n) + E(n-1)$ is the second order difference of the discrete signal $E(n)$.

The significance (or "strength") of a singularity is defined by its scale-normalized second order difference. As shown in (17), this quantity $\Delta^2 E(n_0)/\Omega_0$ roughly corresponds to the absolute energy of the spectrum $y(x)$ at scale Ω_0. Intuitively, the second order difference along the tonotopic (log) frequency axis at the singularity reflects the depth and steepness of the surrounding function. However, since singularities at high scales tend to be steeper than those at low scales due to the broader bandwidths of the RFs at these scales, it is necessary to normalize this measure by the scaleΩ_0.

III. Reconstruction From Singularities

To what extent can singularities be used to reconstruct the input spectral profile? Specifically is it possible to have a stable, perceptually faithful reconstruction only from such information as the locations and strength of the most significant singularities?

A. Previous Studies

Numerous insights into the challenges of reconstructing from singularities can be gleaned from previous extensive studies of nonlinear inverse problems especially in image processing applications. For instance, numerous algorithms have been implemented for reconstruction from multiscale edges, especially within the zero-crossing framework [12], [33], [34]. Although multiscale zero-crossings have been proven to be complete under certain conditions (e.g., when the input pattern is a polynomial function [33], [35] or an irreducible band-limited function [36]), they cannot characterize a *general* function uniquely [37]. Instead, approximations of the input signals can be recovered with additional information such as the average values between any pair of consecutive zero crossings [12] or the gradient along the zero crossing boundaries [33], [38], [39]. These gradients in fact are related to the gradients around our singularities (and hence the strength of a singularity as we shall discuss next).

Another example of a related nonlinear inverse problem is the phase retrieval problem—restoring original signal from the magnitude of its Fourier transform [25]. Most applications have focused on the two-dimensional image restoration problems and the signal extrapolation problems [26], [40], and several error-reduction algorithms have been proposed [41], [42] in conjunction with these algorithms. Nevertheless, mathematical convergence of these algorithms is not generally achievable [43]–[45], but can be significantly improved by combining different algorithms [42], [46].

Although both types of problems above lack closed-form solutions, iterative procedures have nevertheless demonstrated stable reconstruction results. In a similar vein, our reconstruction algorithm described below is iterative, employing a generalized projection procedure which was originally used to solve image restoration problems [24].

B. Reconstruction Algorithm

The iterative algorithm below reconstructs an approximation of the input spectrum from the positions and gradients of the singularities. The basic idea is to project an initial pattern between two domains (spectrum and scale space) while satisfying constraints applied in each domain.

Let $f(x) \in \mathbf{L}^2(\mathbf{R})$ denote the input spectrum and $J(x)$ be the set of functions which result in the singularities with the same locations and (magnitude) gradients as the ones from $f(x)$. Our purpose is to find a member in the $J(x)$ set to approximate $f(x)$. Let $(x_n^j)_{n\in Z}$ denote the abscissae where singularities from $f(x)$ occur at scale Ω_j. The singularity constraints at the scale space for $J(x)$ can be decomposed into two conditions.

1) At each scale Ω_j, the singularities from $J(x)$ located at $\left(x_n^j\right)$ has the same gradients as the ones from $f(x)$.
2) At each scale Ω_j, the singularities from $J(x)$ are located at $\left(x_n^j\right)$.

Condition 1 is not convex due to the fact that the (magnitude) gradients are calculated from the magnitude which is not convex. In [7], a similar nonconvex constraint of local maxima as condition 2 stated above was approximated by a convex constraint. However, this approach cannot be adopted for the set of local magnitude minima such as the set of singularities. Nevertheless, the projection method can still be applied even to inverse problems with nonconvex constraint sets [24] such as the signal restoration problems with Fourier transform magnitude constraints [25], [26] or wavelet transform magnitude constraints [47]. In such case, the convergence of the projection strongly relies on the initial starting point of the algorithm.

In addition to the location and gradients of each singularity, the energy $(\sum_n E(n;\Omega_0))$ at the scale Ω_0 where the singularity occurs is also needed. The proposed reconstruction algorithm can be summarized as:

1) Estimate an impulse-type initial spectrum based on the locations and gradients of known singularities by the properties derived in Section IV.
2) Calculate the complex multiscale response associated with the input spectrum by (6).
3) Apply scale space magnitude constraints: condition 1 and 2 stated above and the energies at the scales where singularities occur.
4) Identify the undesired singularities generated at step 2.
5) Calculate the spectrum by the inverse wavelet transform.
6) Apply spectrum domain constraints (smoothing certain part of the spectrum to eliminate the undesired singularities occurred in the scale space followed by half-wave rectification).

Repetitive application of steps 2 to 6 defines the algorithm.

An example of a reconstruction of a natural vowel spectrum is given in Fig. 3. The top panel shows the multiscale representation of the vowel /o/ as in Fig. 1(b). The locations of the six most significant singularities of the original spectrum are indicated by the crosses. Center panel shows the reconstructed spectrum (solid line) superimposed against the original spectrum (dashed line). The corresponding multiscale representation of the reconstructed spectrum is illustrated in the bottom panel. The derived initial impulse-type pattern (darker solid lines) to start the reconstruction is also shown in the center panel. Note, the reconstruction errors at high $(> 1000$ Hz) and low $(< 125$ Hz) frequency ranges are obviously seen due to the absence or weakness of singularities at those regions. Ideally, the gradients (both sides) of each singularity implicitly encodes the absolute (see Section II-C) and relative energy levels of the surrounding harmonic peaks if the bandwidths of the bandpass cortical filters are sufficiently narrow with respect to the scale axis. However, since our cortical filters are relatively broad (with 1 octave 3 dB bandwidth), this information is diluted, and hence including in addition the energies at the scales where singularities occur would be necessary to yield a stable reconstruction [48].

Fig. 4 illustrates the original, reconstructed (frame-by-frame basis) and lowpass filtered (cut-off frequency at about 32 Hz) reconstructed spectrogram of the word "away" spoken by a male

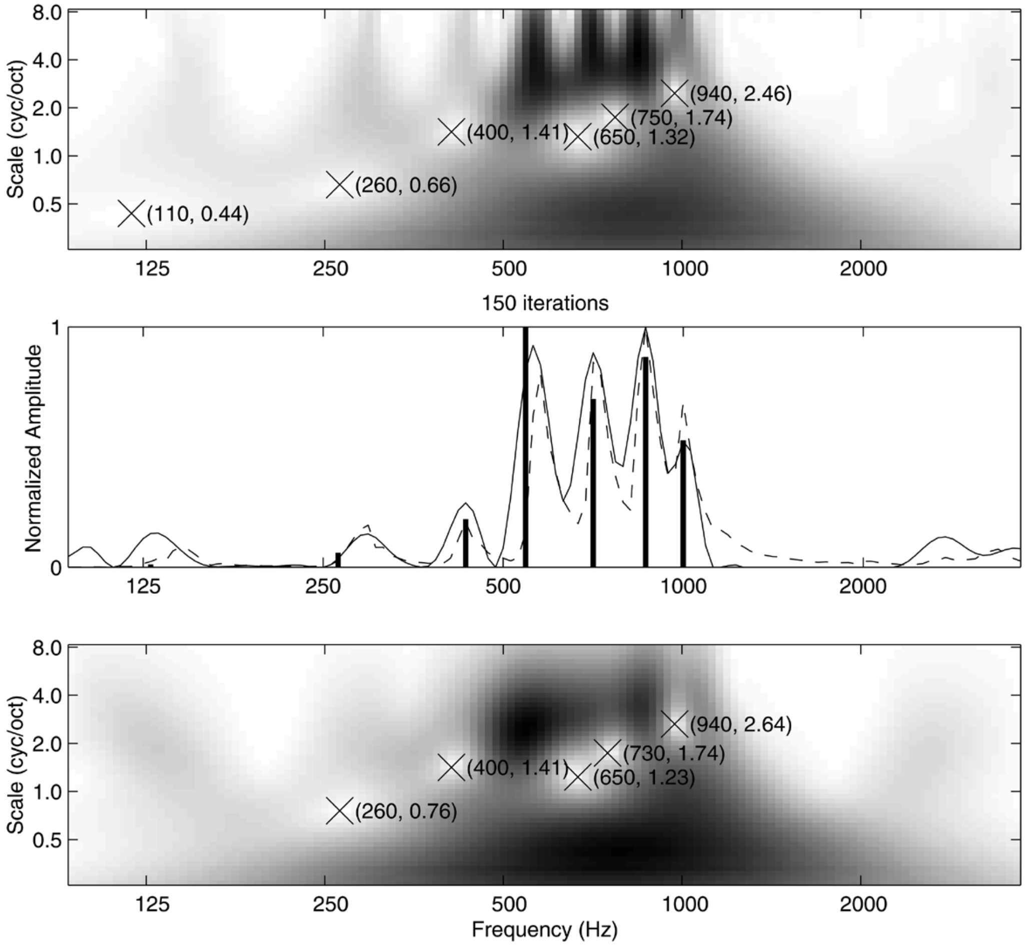


Fig. 3. Example of a reconstructed spectrum from phase singularities. The locations and gradients of the six strongest singularities (shown in top panel) combined with their scale energies are used to reconstruct the original spectrum (dashed line in center panel). The reconstructed spectrum after 150 iterations is indicated by the solid line in the center panel, and the corresponding multiscale magnitude response of the reconstructed spectrum is shown in the bottom panel. The initial impulse-type pattern (see Section IV-D) to initiate the reconstruction algorithm is superimposed upon the original and reconstructed spectra in the center panel.

speaker, respectively. In this example, seven strongest singularities per frame are used to reconstruct the spectrogram. Similar to the reconstruction shown in Fig. 3, errors are apparent at the high and low frequency regions and around the peaks with low peak-to-valley ratio. To resolve such weak peaks, one has to consider more singularities per frame or include the higher scale singularities during the reconstruction. The reconstructed acoustic signals from the auditory spectrograms in Fig. 4 are available at http://www.isr.umd.edu/CAAR/pubs.html; and the iterative algorithm used to invert the auditory spectrogram back to the acoustic signal is described in [17], [48].

The quality of the reconstructed phrase was estimated using the “Perceptual Evaluation of Speech Quality” package (PESQ) [49] as an indicator of the Mean Opinion Score (MOS) of the signal in Fig. 4. The PESQ scores of the three reconstructed signals (from top to bottom) are 4.26 (toll quality), 2.81 (synthetic quality) and 3.00 (professional quality) [50].

IV. Determining the Initial Approximate Spectrum

For the reconstruction procedure to converge to stable and accurate patterns, it is essential that its initial spectral pattern be broadly consistent with the location and strengths of singularities. In this section, we discuss in more detail what properties of the initial pattern can be gleaned from the singularities, and how to generate such a pattern for the reconstruction. These insights and properties are readily evident from a cursory inspection of the singularities in Fig. 1(b). For instance, (1) singularities appear between adjacent peaks of the spectrum, and (2) their locations depend on the spacing between peaks and their relative amplitudes. To elaborate on this relationship, we analyze in detail the singularities associated with a simple abstract pattern consisting of two impulses located along the x axis and separated by a distance d as shown in Fig. 5(a). In the following deterministic analysis, the function $h(x)$ is implemented by the second derivative of a Gaussian function $(-e^{-x^2/2})$. The even function $h(x) = (1 - x^2)e^{-x^2/2}$ and odd function $\hat{h}(x)$ are shown in Fig. 5(b).

A. Spacing Between Peaks

Let the input spectrum $I(x)$ in Fig. 5(a) be expressed as

$$I(x) = \delta(x-a) + \delta(x-b)$$

where $b - a = d$. Hence, the output z can be written as

$$\begin{aligned} z(x,\Omega) &= I(x) * h_w(x;\Omega) \\ &= h_w(x-a;\Omega) + h_w(x-b;\Omega). \end{aligned} \tag{18}$$

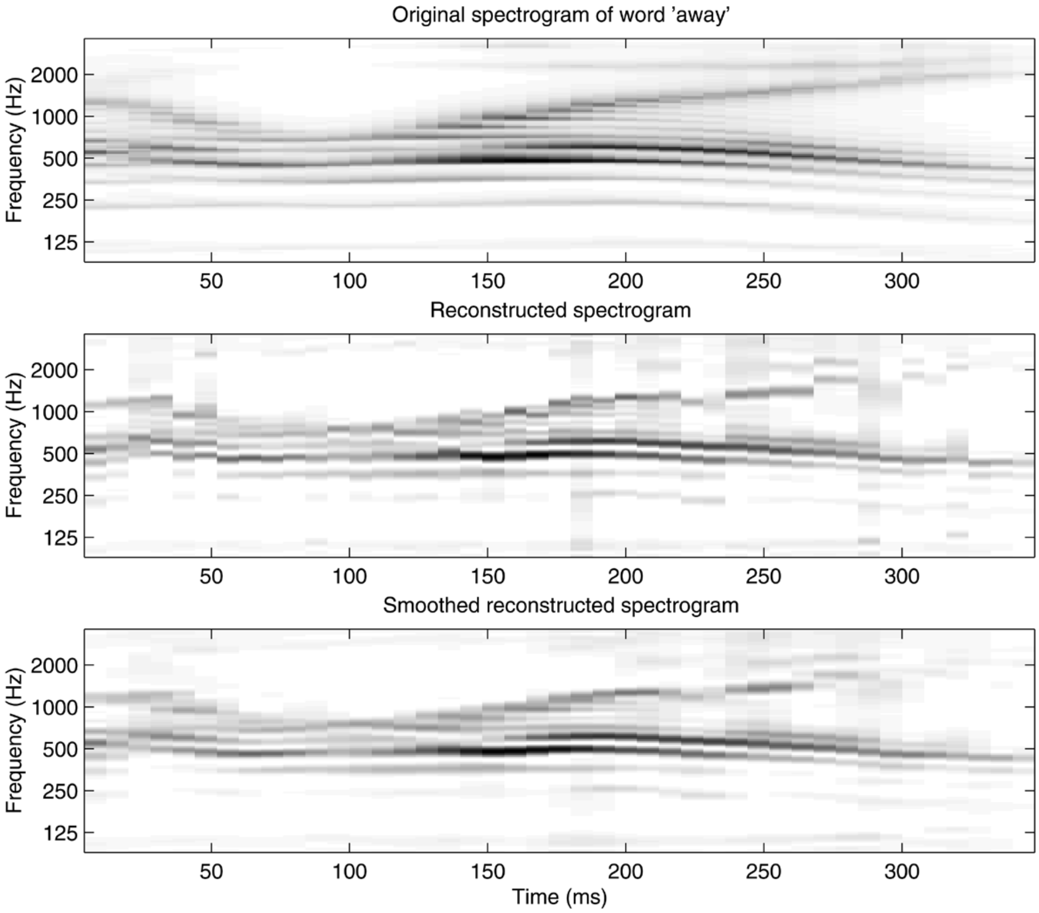


Fig. 4. Reconstruction of the spectrogram of the word “away” extracted from TIMIT corpus. (*Top panel*)—Original spectrogram. (*Center panel*)—Reconstructed spectrogram. (*Bottom panel*)—Smoothed (lowpass filtered at 32 Hz) reconstructed spectrogram. The reconstructed acoustic signals of these spectrograms are available at http://www.isr.umd.edu/CAAR/pubs.html.

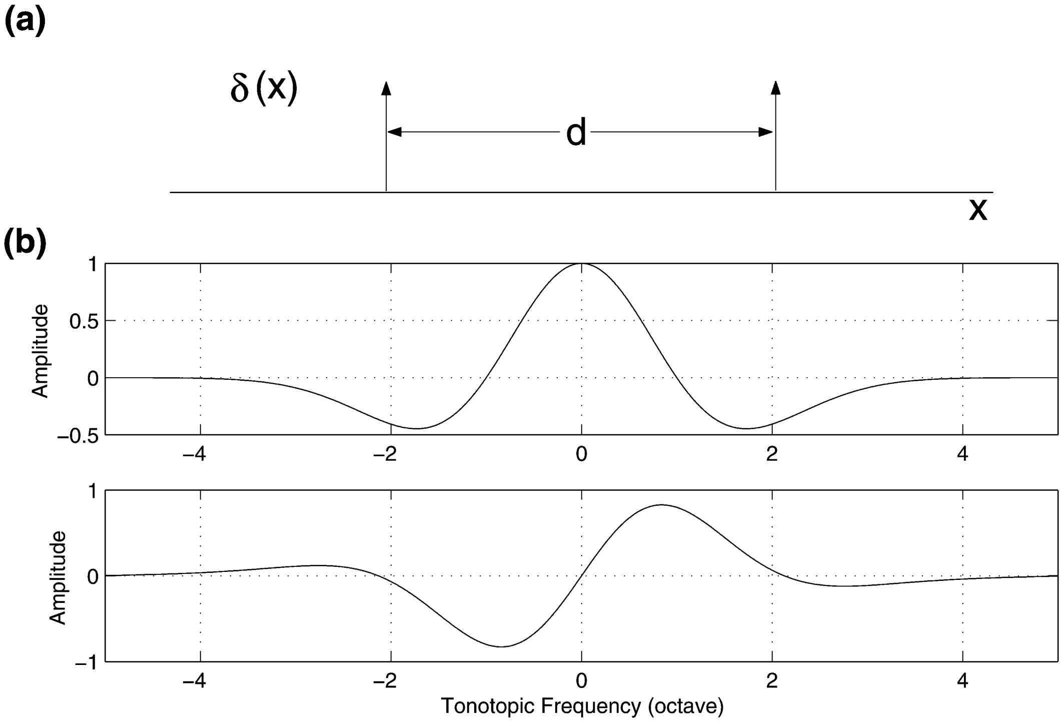


Fig. 5. Impulse-type stimulus and the shape of mother wavelet. (a) The two-impulse stimulus with spacing d and unity amplitude. (b) The real ($h(x)$, an even function) and imaginary part ($\hat{h}(x)$, an odd function) of the complex mother wavelet are shown in top and bottom panels, respectively.

Assume a singularity emerges at location (x_0, Ω_0) for input $I(x)$, then the real and imaginary parts of $z(x_0, \Omega_0)$ should equal to zero simultaneously

$$h(x_0 - a; \Omega_0) + h(x_0 - b; \Omega_0) = 0 \tag{19}$$
$$\hat{h}(x_0 - a; \Omega_0) + \hat{h}(x_0 - b; \Omega_0) = 0. \tag{20}$$

Since the $\hat{h}$ is an odd function, $x_0 = (a+b)/2 = a + d/2$ is an obvious solution for (20). Substituting this solution in (19), and combining the fact that $h(x)$ is an even function, the solution for Ω_0 should satisfy the following equation:

$$h\left(\frac{d}{2}; \Omega_0\right) = h\left(-\frac{d}{2}; \Omega_0\right) = 0. \tag{21}$$

A nontrivial solution for (21)—d equals to the width of the excitatory (positive) band of function h—can then be deduced by observing $h(x)$ in Fig. 5(b).

In summary, for a pattern with two impulses spaced d apart, the singularity occurs at the center of the two peaks $(x_0 = (a + b)/2)$ and at the scale whose excitatory band width equals to d. Note, if the input is only a single impulse (a special case with condition $d = 0$), the singularity is defined to occur at the location of the impulse and at infinite scale.

B. Relative Amplitude of Peaks

In this section, we consider the case of unequal amplitudes. Let the amplitudes be 1 and $(1 + \triangle c)$, where $\triangle c \geq 0$ parameterizes the relative amplitude of the two impulses. Therefore, (19) and (20) can be restated as

$$h(x_0 - a; \Omega_0) + (1 + \triangle c) \cdot h(x_0 - b; \Omega_0) = 0 \tag{22}$$
$$\hat{h}(x_0 - a; \Omega_0) + (1 + \triangle c) \cdot \hat{h}(x_0 - b; \Omega_0) = 0. \tag{23}$$

The solution to (23) is of the form $x_0 = (a + b)/2 + \triangle x$, where $\triangle x$ is due to the effect of $\triangle c$. Substituting it into (22), we get

$$h\left(\frac{d}{2} + \triangle x\right) + (1 + \triangle c) \cdot h\left(-\frac{d}{2} + \triangle x\right) = 0 \tag{24}$$

at scale Ω_0. Derive the Taylor series expansions of function $h(x)$ about two points $x = (d/2)$, $-(d/2)$ up to second order and substitute the expansions into (24), we get

$$\triangle x = \frac{h'\left(\frac{d}{2}\right) \triangle c}{h''\left(\frac{d}{2}\right)\left[1 + \frac{\triangle c}{2}\right]} \tag{25}$$

where $h'(d/2) = -h'(-d/2) < 0$ and $h''(d/2) = h''(-d/2) > 0$. This solution should be verified to satisfy (23) as well, but this verification can be omitted based on one of Logan's theorems [51] (see discussions in Section VI-A for details).

If $\triangle c \ll 1$,

$$\frac{\triangle x}{\triangle c} = \frac{h'\left(\frac{d}{2}\right)}{h''\left(\frac{d}{2}\right)} < 0 \tag{26}$$

which indicates $\triangle x$ is inversely proportional to $\triangle c$ when $\triangle c$ is small. In other words, the singularity moves toward the peak with the smaller amplitude when the amplitudes of the two peaks are slightly different. Fig. 6 shows the displacement of singularity $(\triangle x)$ as a function of the discrepancy between peaks' amplitudes $(\triangle c)$. The solid line demonstrates the actual result while the dashed line shows the approximation by (26). Although (26) is derived for $\triangle c \ll 1$, this approximation holds well even for $\triangle c \simeq 0.6$ as shown in Fig. 6.

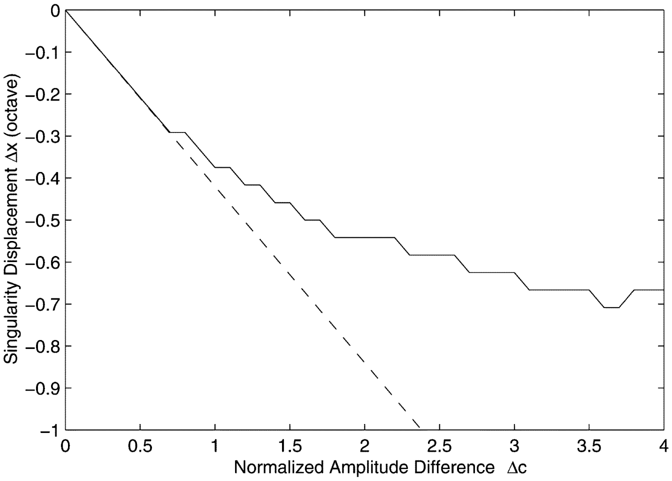


Fig. 6. Displacement of singularity along the tonotopic (log) frequency (x) axis as a function of the relative amplitude of two peaks. The solid line depicts the simulation result while the dashed line denotes the approximation for small $\triangle c$. The displacement $\triangle x$ is normalized to the case of $d = 1$ octave.

C. Width of Peaks

We discuss next the effect of having broader peaks on the location of singularity. A more realistic model of the spectral peaks would be an impulse function $(\delta(x))$ convolved with a Gaussian function (e^{-qx^2}). Therefore, the input spectrum of two peaks $I_w(x)$ becomes

$$I_w(x) = [\delta(x-a) + \delta(x-b)] * e^{-qx^2}$$

and the output is

$$z(x, \Omega) = I(x) * e^{-qx^2} * h_w(x; \Omega) \tag{27}$$

which is equivalent to the output associated with the impulse-type input $I(x)$ but filtered by a modified filterbank $h_m(x; \Omega)$ where

$$h_m(x; \Omega) = e^{-qx^2} * h_w(x; \Omega).$$

Therefore, the real part of the mother wavelet of this modified filterbank can be derived as

$$\begin{aligned} \Re\{h_m(x)\} &= e^{-qx^2} * \frac{d}{dx^2}[-e^{-px^2}] \\ &= \frac{d}{dx^2}[e^{-qx^2} * -e^{-px^2}] \\ &= \frac{d}{dx^2}\left[-\sqrt{\frac{\pi}{p+q}} \cdot e^{-((p \cdot q)/(p+q))x^2}\right] \end{aligned} \tag{28}$$

by the usage of the following definite integral

$$\int_{-\infty}^{\infty} e^{-m^2x^2 \pm nx} dx = \frac{\sqrt{\pi}}{m} \cdot e^{n^2/4m^2} \quad [m > 0]$$

where $p = 1/2$ (our original filterbank). Equation (28) also depicts the fact that convolution of two Gaussian functions is still a

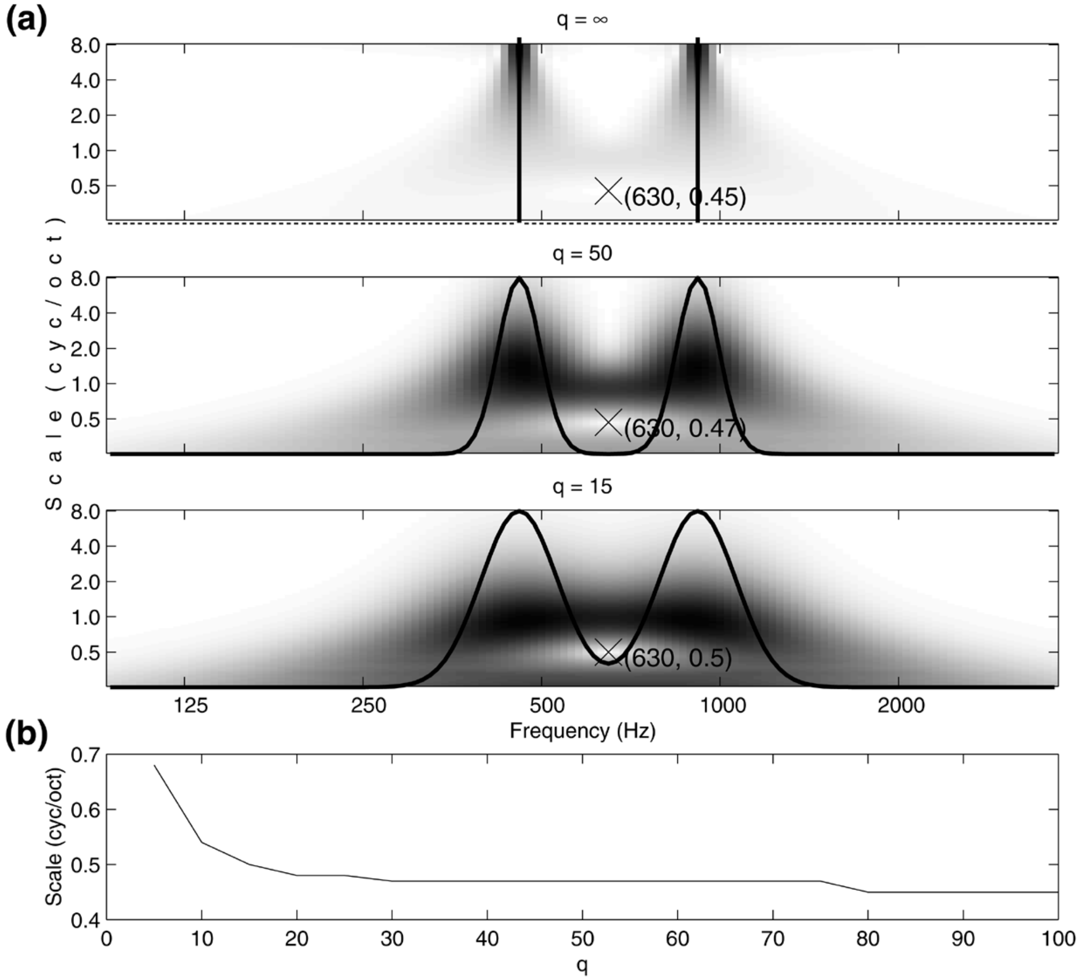


Fig. 7. Displacement of singularity along the scale (Ω) axis as a function of the width of the surrounding two peaks. (a) The singularity moves along the scale axis due to the change of the surrounding peaks' width. The three panels from top to bottom depict the slight upward move of the singularity (from 0.45, 0.47 to 0.5 cyc/oct) with the increasing width of the peaks (parameter q from ∞, 50 to 15). The artificial spectra (solid lines) are superimposed on the multiscale magnitude responses in each panel. (b) The scale of the singularity varies as a function of the width of the peaks, which is characterized by parameter q.

Gaussian function. Since $p > ((p \cdot q)/(p+q))$, the modified filterbank has a wider excitatory band than our original filterbank. In other words, the width of the excitatory band of the modified filterbank at scale Ω_0 is now wider than the spacing d between the two peaks. Hence, the singularity will occur at higher scale than Ω_0 in the scale space defined by the modified filterbank.

In general, the width of the peaks in which we are interested is smaller than the spacing between the two peaks, i.e., $q \gg p$. In such a case, the mother wavelet of the modified filterbank can be approximated by

$$\Re\{h_m(x)\} = \frac{d}{dx^2}\left[-\sqrt{\frac{\pi}{q}} \cdot e^{-px^2}\right] \tag{29}$$

which has the same decay factor p as the original filterbank. This approximation implies that the scale of the singularity is not affected much by widening the harmonic peaks as shown in Fig. 1(b). Fig. 7(a) demonstrates the slight shifts as the width of the peaks increase. Fig. 7(b) shows the scale of singularity as a function of the width parameter q. As expected, the singularity occurs almost at the same scale even for significantly wide peaks.

D. Constructing the Initial Pattern

The above analysis focuses on the location of the singularity and the factors which move the singularity along both x and Ω axes. In addition, while the location of the singularity due to two narrow peaks signifies the spacing d between the two peaks regardless of their absolute amplitudes, the quantity $\Delta^2 E(n_0)/\Omega_0$ in (17) (defined as the "strength" of singularity) serves to indicate the energy in the spectrum at these peaks, and hence the significance of these peaks. This quantity is a crucial indicator which helps to weed out numerous spurious peaks due to various noise sources.

As indicated in Section III-B, the reconstruction result strongly relies on the initial pattern for the projection between nonconvex sets. Procedures to estimate an initial impulse-type pattern based on the locations and gradients of the singularities can be summarized as follows.

1) Determine the spacing d between two impulses which surround the singularity at highest scale by the property derived in Section IV-A.
2) Adjust the relative amplitudes of these two impulses according to the gradients of that singularity.

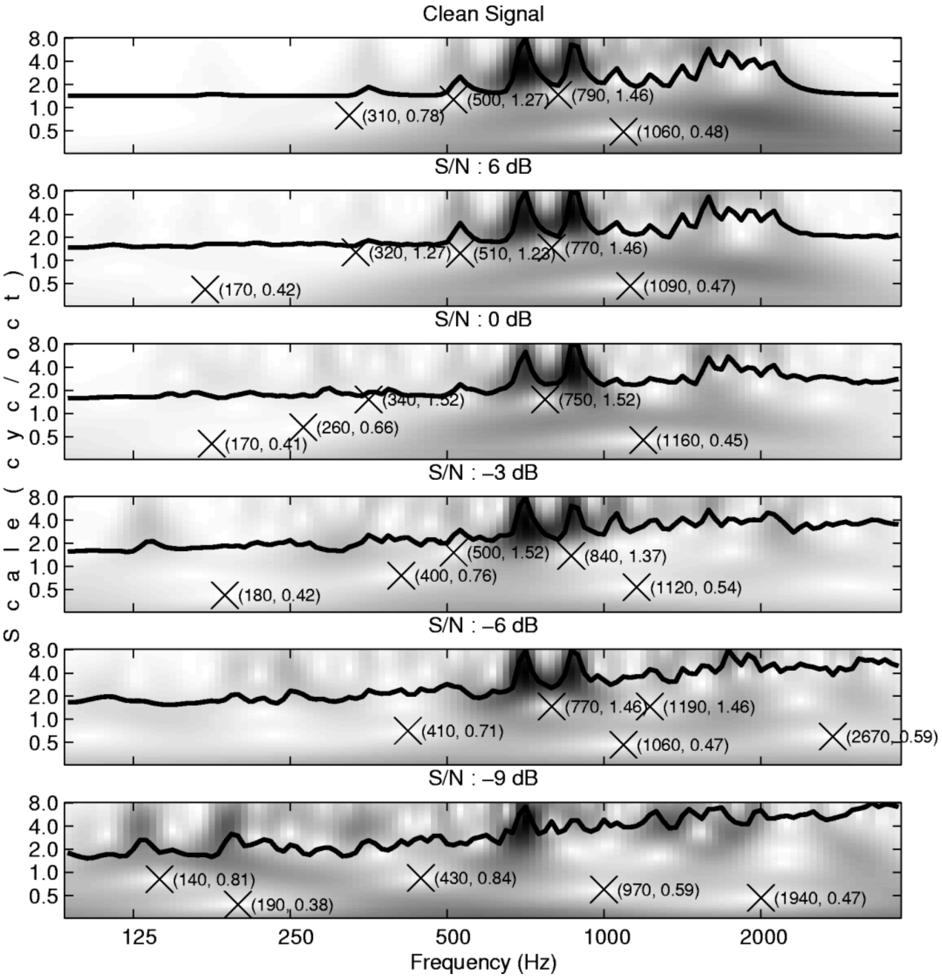


Fig. 8. Low scale singularities under noisy conditions. The five strongest singularities below 1.8 cyc/oct are plotted for spectra at different signal-to-noise ratios. The auditory spectra (solid lines) are superimposed on the multiscale magnitude responses in each panel.

3) Determine the exact locations of these two impulses by the property derived in Section IV-B.
4) Determine the location (Section IV-A) and amplitude (Section IV-B) of the next impulse from the location of the next closest singularity to the resolved singularities.

Step 4 is repeatedly applied until all singularities are considered. The constructed initial impulse pattern for the reconstruction example in Fig. 3 is generated using this procedure. As is evident, the locations and amplitudes of the peaks of the original spectrum between 500 and 1000 Hz are well estimated from the most significant singularities.

V. Robustness of Singularities

The representation of spectra by their singularities has several applications that we plan to pursue in the future. One example is the de-noising and robust representation of spectral patterns in noisy environments. This is illustrated in Fig. 8, which depicts the singularities of the vowel $[a]$ under clean and various SNRs. As shown in Fig. 8, the strongest singularities of the clean vowel, e.g., those two occurring at low and medium scales near $x = 1060$ and 790 Hz (top panel), do not move much over a wide range of SNR's. Specifically, the strongest singularity $(x = 1060\ \mathrm{Hz})$ captures information about the *global shape* of the auditory spectrum at 0.48 cyc/oct, which reflects the spacing between the first and second formants and their relative amplitudes, and remains near the same location down to SNR's of -6 dB. Another strong singularity at 790 Hz, which occurs at a *median scale* $(\sim 1.5\ \mathrm{cyc/oct})$ and encodes the prominent harmonic peaks with highest peak-to-valley ratios, moves only slightly with increasing noise level. By contrast, the small and closely-spaced peaks due to the additive white noise generate many *weak* singularities (white spots) at high scales. In de-noising applications, these singularities can be separated from those of the clean vowel spectrum in scale space by thresholding both the scales and strength of singularities.

To demonstrate the efficiency and robustness of the strong singularities in encoding spectral shapes, we compared their performance in an automatic vowel classification task to that of the widely-used mel-frequency cepstral coefficients (MFCCs) under different SNR's. Both of these sets parameterize the overall spectral shape, a feature that has been shown to be better correlated with vowel identity than formant frequencies [52]. In the experiment described below, a database of clean vowels (subdivided into a *training* and *test* sets) was encoded by their corresponding MFCC's and low-scale singularities.

TABLE I
NUMBERS OF VOWELS FROM ALL DIALECT REGIONS OF THE TRAIN AND TEST SET OF THE TIMIT CORPUS (MALE SPEAKERS)

IPA Symbol	Typical Word	Training Set	Test Set
a	(hot)	1488	547
æ	(bat)	2500	842
$\wedge$	(but)	1089	419
$\supset$	(bought)	1051	366
ϵ	(bet)	1940	739
∂	(ago)	1203	460
I	(bit)	1812	624
i	(beet)	2247	977
U	(foot)	169	83
u	(boot)	222	82

A Bayes classifier was then trained using the training set, and the performance of the two feature sets was then tested and compared under progressively worse SNR's using the test set.

A. Database and Vowel Parameters

The speech material used in this study is a subset (male speakers) of the TIMIT corpus (additional information may be found in the printed documentation from National Institute of Standards and Technology NIST# PB91-100 354). Unlike the isolated-CVC(Consonant-Vowel-Consonant)-word databases used in vowel recognition tests [52], [53], TIMIT is a continuous speech corpus which is closer to a conversational speech corpus. All ten American English vowels (> 64 ms) by male speakers were extracted (viz., [a], [æ], [$\wedge$], [$\supset$], [ϵ], [∂], [I], [i], [U], and [u]) for the recognition task regardless of their quality, context, speakers' dialect regions, and probable mislabeling (vowels are extracted from 4380 sentences by 438 male speakers from 8 major dialect regions of the United States). However, only vowels having singularities between 250 and 1500 Hz and below 1 cyc/oct are actually used in training and testing. This constraint was based on the typical formant frequencies (F_1 and F_2) for the vowels shown in [54]. The numbers of vowels extracted from the training and test set of TIMIT corpus are listed in Table I. Since the population of vowels [U] and [u] was much smaller than the populations of other vowels, they were dropped from this study.

The extracted vowel signal was first pre-emphasized with transfer function $(1 - 0.9z^{-1})$ and windowed with a 16 ms Hanning window with 8 ms overlap between frames. The magnitude spectrum was then processed by a mel-scale filter-bank, and the resulting log-energy profile was cosine transformed to produce the mel-frequency cepstral coefficients [55]. The MFCC feature vector consisted of the coefficients averaged over the entire signal duration of the vowel (except for the first set of coefficients).

Similarly, the singularity features were computed from the auditory spectrograms of the same preemphasized signals [20]. Then the singularities of the multiscale representation of the averaged spectrum were detected. Here, the location and the gradients of the single strongest singularity between 250 and 1500 Hz and below 1 cyc/oct were extracted as the singularity feature vector. Note that this feature vector only consisted of 4 parameters (frequency, scale of the singularity and gradients at both sides).

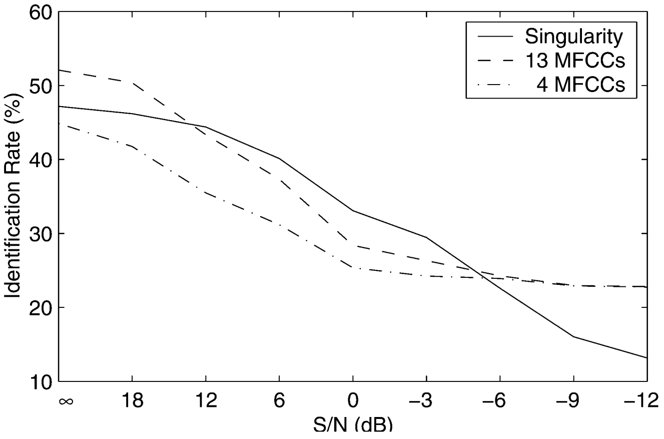


Fig. 9. Identification rate of vowels for singularity and MFCC feature sets at different SNRs.

B. Classifier

Since the main purpose of this experiment is to compare the robustness of the singularities and MFCC's in encoding noisy vowel spectra, we adopted a particularly simple uni-modal multivariate Gaussian (instead of a multi-modal) classifier to provide as direct and straightforward comparison between the two feature sets as possible. This choice is partly the reason why only one singularity (strongest) is utilized since the distributions of two or more singularities are poorly approximated by a uni-modal distribution.

Therefore, all feature sets for vowel i (i = 1–8) were assumed to be multivariate Gaussian distributed with mean $\mathbf{p}_i$ and covariance matrix Σ_i. The *a priori* probability for vowel i ($Prob(i)$) was estimated as

$$\mathrm{Prob}(i) = \frac{N_i}{\sum_{i=1}^{8} N_i}$$

where N_i is the number of vowel i in the training set (see Table I). The classifier used in this experiment was the Bayes classifier which minimizes the probability of error. The decision rule for the test feature vector $\mathbf{p}$ was given by

$$\check{i} = \arg\min_i D_i(\mathbf{p}) \tag{30}$$

where $\check{i}$ is the assigned label for the test vector $\mathbf{p}$ and distance function $D_i(\mathbf{p})$ is [52]

$$D_i(\mathbf{p}) = (\mathbf{p}-\mathbf{p}_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{p}-\mathbf{p}_i) + \ln \|\mathbf{\Sigma}_i\| - 2 \ln \mathrm{Prob}(i). \tag{31}$$

C. Results and Discussion

White Gaussian noise was added to simulate different SNR conditions for comparison. The correct identification rate was defined as

$$\%\text{identification rate} = \frac{\#\text{ of correctly identified vowels}}{\text{total }\#\text{ of vowels}} \times 100.$$

The results are plotted in Fig. 9 as a function of SNR levels both for the MFCC (13 and 4 coefficients) and singularity (4 parameters) feature sets. Basically, the performance of the strongest singularity and the MFCC(13) features are about equal down to low SNR's $(-6\ \mathrm{dB})$, and both are consistently and significantly better than the truncated MFCC(4) set. We note, however, that the overall *absolute* performance level at high SNR's $(\approx 50\%)$ is low compared to published identification rates [52], [53]. This is partly due to the use of a simple uni-modal classifier,[1] and partly due to the relative complexity and variability of the TIMIT database. Finally, it is likely that significantly better robustness can be achieved with more singularities, although it is essential then to "include" in the assessment some knowledge of the relative locations of these singularities in the structure of the classifier (e.g., a multimodal formulation).

VI. Summary and Discussion

We have described singularities in the scale space generated by a multiscale model of the auditory cortex. The singularities are parameterized by their location and significance (in terms of their local gradients), and are shown to be sufficient to reconstruct the original input spectral pattern that gave rise to them when combined with the energies at the scales where they occur. Also presented is a method to estimate an initial spectral pattern which yields stable results upon convergence of the reconstruction algorithm. In this section, we discuss further some properties of the singularities and their potential applications.

A. Completeness of the Set of Multiscale Singularities

As stated in Section III-A, the zero crossings in scale space only form a *complete* representation for *certain restricted classes of signal*. For instance, Logan defined the *free zeros* as the zeros shared by the function itself and its Hilbert transform and showed a bandpass signal whose bandwidth is less than 1 octave and has no free zeros can be determined by the multiscale zero crossings within a constant multiplier [51]. According to its definition, a free zero of a bandpass signal becomes a singularity in this study when the analytical form of the bandpass signal is considered. In addition, our cortical filters have bandwidth broader than 1 octave. Therefore, no conclusions can be drawn from Logan's theorems regarding the completeness of the set of multiscale singularities. Another one of Logan's theorems is relevant to the analysis in Section IV-B about the movement of the singularities along the x axis. It states that moving a free zero (real part of a singularity) of a bandlimited signal moves the corresponding zero of its Hilbert transform (imaginary part of the singularity) in the same way [51]. That is, the movement $\triangle x$ in (25) should satisfy both conditions in (22) and (23) simultaneously.

In summary, the set of *multiscale* zero crossings does not in general provide a complete representation of the signal. Therefore, the set of singularities (which is a small subset of the set of all zero crossings) would not be complete either. In other words, a mathematically identical spectral reconstruction from the singularity set is not possible. By adding gradients of singularities and the energies at the scales where singularities occur, a perceptually adequate reconstruction is potentially achievable as demonstrated in Section III-B.

B. Future Work

All the properties discussed in Section IV are based on the relationship between *a pair* of spectral peaks and the resulting singularity. Specifically, a singularity carry the essential information about those surrounding peaks including amplitude (characterized by the strength and location of the singularity along the log frequency x axis) and spacing (characterized by the scale where singularity occurs). A more accurate estimate of the initial pattern, which takes into account the effects of all peaks on a singularity (and not just the two surrounding ones), shall yield a better reconstruction result. For instance, using an impulse-type spectrum but now composed of impulses at the actual locations with actual amplitudes of the desired spectral peaks, exhibited faster convergence and lower mean squared error. Therefore, to have better reconstruction, the effects of other spectral peaks on the singularity in addition to the surrounding peaks must be investigated and incorporated in the future in generating better initial spectral estimates.

To obtain a satisfactory reconstruction, we have used additional information about the singularity, specifically, its local gradients (or strength). It is, however, possible that other parameters may also suffice to give acceptable reconstruction. For example, the low minima of the envelope (i.e., the singularities) of narrow-band signals are approximately hyperbolic in shape [57]. Hence, the neighborhood of each singularity can be parameterized by two local quantities: the location of the focus and its eccentricity. In such a case, the location of the focus and the slopes of the asymptotes of the hyperbola can serve as efficient features in spectral analysis.

We have shown in Section V that the low-scale singularity preserves the overall shape of the vowel spectrum (including formant locations and relative amplitudes). Since such parameters are highly correlated with vocal tract shape and length, the lower scale singularities might be good at parameterizing a simple vocal tract model (e.g., as in [58]). By contrast, median-scale singularities $(\sim 1.5\ \mathrm{cyc/oct})$ capture well information about the harmonics, i.e., the pitch in speech or music and aspects of the voice quality, which have proven valuable in speaker identification problems. For instance, analogous parameters of pitch (e.g., pitch value, averaged pitch, pitch contours and jitter) have been successfully used in speaker identification tasks [59]–[62]. Furthermore, recent efforts at combining pitch and MFCC's have yielded promising improvements in performance of speaker identification systems [63]. Therefore, contours of median-scale singularities combined with low-scale singularities could similarly be used to identify speakers.

[1] It has been shown that a classifier based on a multi-modal Gaussian mixture model (GMM) achieves 25% more correct identifications over the uni-modal Gaussian classifier for a speaker identification task [56].

References

[1] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman, 1982.

[2] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Jun. 1986.

[3] G. J. Brown, "Computational auditory scene analysis: A representational approach," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 1992.

[4] L. S. Smith, "Onset-based sound segmentation," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 729–735.

[5] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *J. Neurophysiol.*, vol. 42, pp. 1083–1107, 1979.

[6] B. Girod, "The information theoretical significance of spatial and temporal masking in video signals," *Proc. SPIE*, vol. 1077, pp. 178–187, 1989.

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

[8] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proc. R. Soc. Lond. B*, vol. 204, pp. 301–328, 1979.

[9] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B*, vol. 207, pp. 187–217, 1980.

[10] M. R. M. Jenkin and A. D. Jepson, "The measurement of binocular disparity," in *Computational Processes in Human Vision : An Interdisciplinary Perspective*, Z. W. Pylyshyn, Ed. Westport, CT: Ablex, 1988.

[11] A. Witkin, "Scale space filtering," in *Proc. Int. Joint Conf. Artificial Intell.*, 1983.

[12] S. Mallat, "Zero-crossings of a wavelet transform," *IEEE Trans. Inform. Theory*, vol. 37, no. 4, pp. 1019–1033, Jul. 1991.

[13] T. Sanger, "Stereo disparity computation using gabor filters," *Biol. Cybern.*, vol. 59, pp. 405–418, 1988.

[14] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *GVGIP: Image Understand.*, vol. 53, no. 2, pp. 198–210, 1991.

[15] C.-J. Westelius, H. Knutsson, J. Wiklund, and C.-F. Westin, "Phase-based disparity estimation," in *Vision as Process: Basic Research on Computer Vision Systems*, J. L. Crowley and H. I. Christensen, Eds. New York: Springer-Verlag, 1995, pp. 157–178.

[16] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, no. 3, pp. 1220–1234, 2001.

[17] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, 2005.

[18] J. Simon, D. A. Depireux, and S. A. Shamma, "Representation of complex spectra in auditory cortex," in *Proc. 11th Int. Symp. Hearing*, A. R. Palmer, A. Ress, A. Q. Summerfield, and R. Meddis, Eds., London, U.K., 1998, pp. 513–520.

[19] H. Versnel and S. A. Shamma, "Ripple analysis in the ferret auditory cortex: III. Topographic and columnar distribution of ripple response parameters," *J. Aud. Neurosci.*, vol. 1, no. 2, pp. 271–285, 1995.

[20] T. Chi, Y. Gao, C. G. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, no. 5, pp. 2719–2732, 1999.

[21] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, no. 2-3, pp. 331–348, 2003.

[22] R. Carlyon and S. Shamma, "An account of monaural phase sensitivity," *J. Acoust. Soc. Amer.*, vol. 114, no. 1, pp. 333–348, 2003.

[23] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 382–395, Sep. 1995.

[24] A. Levi and H. Stark, "Image restoration by the method of generalized projections with application to restoration from magnitude," *J. Opt. Soc. Amer. A*, vol. 1, no. 9, pp. 932–943, 1984.

[25] M. H. Hayes, "The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 2, pp. 140–154, Apr. 1982.

[26] J. R. Fienup and C. C. Wackerman, "Phase-retrieval stagnation problems and solutions," *J. Opt. Soc. Amer. A*, vol. 3, no. 11, pp. 1897–1907, 1987.

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

[28] S. A. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in the ferret auditory cortex: I. Response characteristics of single units to sinusoidally rippled spectra," *J. Aud. Neurosci.*, vol. 1, no. 2, pp. 233–254, 1995.

[29] S. A. Shamma and H. Versnel, "Ripple analysis in the ferret auditory cortex: II. Prediction of unit response to arbitrary profiles," *J. Aud. Neurosci.*, vol. 1, no. 2, pp. 255–270, 1995.

[30] P. Ru and S. A. Shamma, "Presentation of musical timbre in the auditory cortex," *J. New Music Res.*, vol. 26, no. 2, pp. 154–169, 1997.

[31] 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. ICASSP*, 2003, pp. 517–520.

[32] A. Papoulis, *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962.

[33] R. Hummel and R. Moniot, "Reconstructions from zero crossings in scale space," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 2111–2130, Dec. 1989.

[34] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inform. Theory*, vol. 38, no. 3, pp. 617–643, Mar. 1992.

[35] A. L. Yuille and T. A. Poggio, "Scaling theorems for zero crossings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 1, pp. 15–25, Jan. 1986.

[36] S. Curtis, S. Shitz, and A. Oppenheim, "Reconstructions of nonperiodic two-dimensional signals from zero crossings," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 6, pp. 890–893, Jun. 1987.

[37] Z. Berman, "The Uniqueness Question of Discrete Wavelet Maxima Representation," Inst. Syst. Res, Univ. Maryland, TR 91-48, 1991.

[38] A. L. Yuille and T. A. Poggio, "Fingerprints theorems," *Proc. Amer. Assoc. Artif. Intell.*, pp. 362–365, 1984.

[39] J. J. Koenderink, "The structure of images," *Biol. Cybern.*, vol. 50, pp. 363–370, 1984.

[40] A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," *IEEE Trans. Circuits Syst.*, vol. CAS-22, no. 9, pp. 735–742, Sep. 1975.

[41] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[42] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Appl. Opt.*, vol. 21, pp. 2758–2769, 1982.

[43] R. H. T. Bates, "Uniqueness of solutions to two-dimensional Fourier phase problems for localized and positive images," *Comput. Vis., Graph., Image Process.*, vol. 25, pp. 205–217, 1984.

[44] M. H. Hayes, "The unique reconstruction of multidimensional sequences from Fourier transform magnitude or phase," in *Image Recovery: Theory and Application*, H. Stark, Ed. New York: Academic, 1987, pp. 195–230.

[45] J. H. Seldin and J. R. Fienup, "Numerical investigation of the uniqueness of phase retrieval," *J. Opt. Soc. Amer. A*, vol. 7, no. 3, pp. 412–427, 1990.

[46] Z. Mou-yan and R. Unbehauen, "Methods for reconstruction of 2-D sequences from Fourier transform magnitude," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 222–233, Feb. 1997.

[47] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3549–3554, Dec. 1993.

[48] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 3, pp. 824–839, Mar. 1992.

[49] "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for end-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," ITU-T, ITU-T Recommend. P. 862, 2001.

[50] M. Delprat, A. Urie, and C. Evci, "Speech coding requirements from the perspective of the future mobile systems," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, 1993, pp. 89–90.

[51] B. Logan, "Information in the zero crossings of bandpass signals," *Bell Syst. Tech. J.*, vol. 56, no. 4, pp. 487–510, 1977.

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

[53] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, pp. 3099–3111, 1995.

[54] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, no. 2, pp. 175–1849, 1952.

[55] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[56] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[57] N. M. Blachman, "The shape of low minima of the envelope of narrowband Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1007–1009, May 1996.

[58] P. Ru, T. Chi, and S. Shamma, "The synergy between speech production and perception," *J. Acoust. Soc. Amer.*, vol. 113, no. 1, pp. 498–515, 2003.

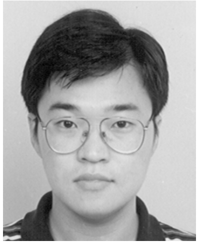
[59] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6, pp. 1687–1697, 1972.

[60] ——, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[61] C. R. J. Jr, T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *Proc. ICASSP*, 1995, pp. 325–328.

[62] F. Farahani, P. G. Georgiou, and S. S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," in *Proc. ICASSP*, 2004, pp. 89–92.

[63] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy, "Toward combining pitch and mfcc for speaker identification systems," in *Proc. Eurospeech*, 2001, pp. 2825–2828.



Taishih Chi (M'03) received the B.S. degree from National Taiwan University in 1992 and the Ph.D. degree from the University of Maryland, College Park, in 2003, both in electrical engineering.

From 1994 to 1996, he was a Graduate School Fellow at the University of Maryland, College Park. From 1996 to 2003, he was a Research Assistant at the Institute for Systems Research, University of Maryland. From August 2003 to June 2005, he was a Research Associate at the University of Maryland. He joined the Department of Communication Engineering, National Chiao-Tung University, Hsnichu, Taiwan, in July 2005. His research interests are in neuromorphic auditory modeling, soft computing, and speech analysis.

Shihab A. Shamma (SM'94) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1980.

He joined the Department of Electrical Engineering, University of Maryland, College Park, in 1984, where his research has dealt with issues in computational neuroscience and the development of microsensor systems for experimental research and neural prostheses. Primary focus has been on uncovering the computational principles underlying the processing and recognition of complex signals (speech and music) in the auditory system, and the relationship between auditory and visual processing. Other researches include the developmerit of photolithographic microelectrode array for recording and stimulation of neural signals, VLSI implementation of auditory processing algorithms, and development of algorithm for the detection, classification and analysis of neural activity from multiple simultaneous sources.