TECHNICAL
RESEARCH
REPORT

SYSTEMS
RESEARCH
CENTER

*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
the University of Maryland,
Harvard University,
and Industry*

Auditory Representations of Acoustic Signals

by X. Yang, K. Wang and S.A. Shamma

TR 91-16r1

Auditory Representations of Acoustic Signals

Xiaowei Yang[a], Kuansan Wang[a], and Shihab A. Shamma[b]

(a) Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, MD 20742

(b) Electrical Engineering Department, Systems Research Center, and the University of Maryland Institute for Advanced Computer Studies. University of Maryland, College Park, MD 20742

The Mathematical Research Branch, NIDDDK, National Institutes of Health, Bethesda, MD 20892

Abstract

An analytically tractable framework is presented to describe mechanical and neural processing in the early stages of the auditory system. Algorithms are developed to assess the integrity of the acoustic spectrum at all processing stages. The algorithms employ wavelet representations, multiresolution processing, and the method of convex projections to reconstruct close replica of the input stimulus. Reconstructions using natural speech sounds demonstrate minimal loss of information along the auditory pathway. Furthermore, close inspections of the final auditory patterns reveals spectral enhancements and noise suppression that have close perceptual correlates. Finally, the functional significance of the various auditory processing stages are discussed in light of the model, together with their potential applications in automatic speech recognition and low bit-rate data compression.

Key words: Auditory System, Multiresolution Processing, Wavelet Transform, Convex Projections.

I. Introduction

The human auditory system possesses remarkable abilities to detect, separate, and recognize speech, music, and other environmental sounds. In recent decades, these capabilities have been the subject of theoretical and experimental research, particularly with a view towards applying auditory functional principles to the design and implementation of man-machine communication links. The basic premise of this research is that understanding cochlear function and central auditory processing will provide new insights into the nature and representation of complex sounds, and will motivate novel approaches to the problems of robust recognition of acoustic patterns.

Many strategies mimicking the structure of the peripheral auditory system have already been incorporated in systems for the analysis, synthesis, and transmission of acoustic signals. For instance, automatic speech recognition systems now often employ the bark or Mel-frequency scale [1–3], adaptive mechanisms [4,5], compressive nonlinearities, and synchrony at the output of the analysis filters [6,7]. The adoption of such auditory processes has usually led to significant improvements in performance over systems using more traditional parametric representations, such as LPC, Cepstral, or spectral coefficients, and their temporal derivatives [7–9].

Auditory approaches, however, often involve complex, multistage, nonlinear transformations which make analytical treatments intractable. This has made assessing performance improvements and their underlying causes very difficult and uncertain, and almost totally dependent on extensive experimentation. Therefore, in order to realize fully the potential benefits of applying auditory processing, it is essential that a viable analytical approach be developed to characterize the signal representations generated by such processing. Furthermore, it is likely that this will lead to a deeper understanding of the fundamental principles operative in the auditory system, and in other sensory systems such as vision and touch.

This paper presents a theoretical framework to describe the transformations that acoustic signals undergo in the early stages of the auditory nervous system. Within this framework, algorithms are developed to relate the acoustic signal to its auditory representations. Specifically, the algorithms utilize wavelet representations, multiscale processing, and the method of convex projections, to reconstruct the original signal from its intermediate and final auditory outputs. Consequently, a precise determination can be made of the spectral information preserved or enhanced through the auditory stages. While strongly motivated by auditory processing, the algorithms presented here are immediately applicable to other sensory processes such as visual and tactile signals. They are also closely related to a general class of algorithms to reconstruct a signal from reduced (or schematized) versions of its affine wavelet transform [10–14].

In the following section (section II), we present a brief description of the peripheral auditory stages both from a biophysical and an algorithmic perspective. In section III, the reconstruction algorithms are formulated and discussed in relation to other multiresolution processing algorithms. Section IV presents examples of reconstructions of natural speech sounds, together with a discussion of the properties of the reconstructed spectra from the point of view of noise robustness, data compression, and perceptual relevance.

II. Processing of Acoustic Signals in the Auditory System

Sound signals undergo a complex series of transformations in the early stages of auditory processing. Numerous descriptions of these processes exist ranging from detailed biophysical models to approximate computational algorithms [15–17]. All models, however, can be reduced to three stages: *analysis, transduction, and reduction.* In the following, we briefly elaborate on these three stages of processing, and then discuss an important limiting case which simplifies the formulation of the auditory model considerably.

II.A The Analysis Stage

When sound pressure waves impinge upon the eardrum of the outer ear, they cause vibrations that are transmitted via the middle ear to the fluids of the cochlea of the inner ear. These pressure waves in turn produce mechanical displacements in the membranes of the cochlea, specifically the so-called *basilar membrane.* The amplitude and time course of these vibrations reflect directly the amplitude and frequency content of the sound stimulus (Fig.1). There are two equivalent ways of viewing the patterns of basilar membrane displacements. The first is to focus on their spatial distribution along the length of the cochlea. Thus, vibrations evoked by a single tone appear as traveling waves that propagate up the cochlea (from base to apex), reaching a maximum amplitude at a particular point before slowing down and decaying rapidly. The point at which maximum displacement occurs depends on the frequency of the tone, with lower frequencies propagating further towards the apex of the cochlea. As such, the cochlea segregates incoming frequencies onto different spatial locations in a tonotopically ordered manner along its length.

A second more functional view of the cochlea is to think of it as a parallel bank of band-pass filters. Thus, at each point along the membrane, one can measure the displacement as a function of tone frequency, i.e., a transfer function. In the mammalian cochlea, the transfer functions are moderately well tuned, with center frequencies decreasing towards the apex of the cochlea. Above about 800 Hz in humans, the impulse responses of these "filters" are related to each other by a dilation. Consequently, along a logarithmic frequency, the transfer functions appear approximately invariant except for a translation, i.e., they maintain a constant Q-factor. It is therefore natural to interpret the outputs of the cochlear filters as an affine wavelet transform of the stimulus, and the continuous spatial axis of the cochlea as the *scale* parameter axis.

The above simplified view of the basilar membrane deviates from the real structure in many ways that may be consequential in some applications. For instance, we have ignored several nonlinear phenomena that play an important role in enhancing the sensitivity and tuning of the cochlear filters at lower sound levels. These phenomena, usually lumped under the term "active cochlear mechanisms", are less important when dealing with relatively broadband signals at moderate to high levels of intensity [16], as is the case for speech and other complex sounds. The other simplification concerns the view of cochlear filtering as strictly an affine wavelet transform. The actual frequency scale of the cochlea is not purely logarithmic below 800 Hz, but rather becomes progressively more linear, especially below 500 Hz [18]. The assumption of cochlear processing as a wavelet transform is adopted in this paper primarily because of its intuitive appeal in interpreting the spatial axis of the cochlea. It is, however, unnecessary for the validity of the theoretical treatments and algorithms formulated here, as we shall elaborate in the next section.

To summarize, an acoustic signal, $x(t)$, entering the ear produces a complex spatiotemporal pattern of displacements, $y_1(t;s)$, along the basilar membrane of the cochlea. To a first approximation, it is described by the following equation:

$$y_1(t;s) = h(t;s) *_t x(t),$$

where $h(t;s)$ represents the (finite energy) impulse response of the cochlear filter at location s along the cochlea ($s = 0$ is the base, and $s > 0$ towards the apex), $y_1(t;s)$ represents the output of the filter at s with input $x(t)$, and $*_t$ denotes the convolution operation with respect to time. If $y_1(t;s)$ is considered an affine wavelet transform of $x(t)$, then $h(t;s)$ become wavelets that are related to each other through a dilation, i.e., $h(t;s) = a^s h(a^s t; 0)$ for some $0 < a < 1$.

II.B The Transduction Stage

The mechanical vibrations along the basilar membrane are transduced into electrical activity along a dense, topographically ordered, array of auditory-nerve fibers. At each point, membrane displacements cause a local fluid flow which bends small filements (*cilia*) that are attached to transduction cells, called the *inner hair cells.* The bending of the cilia controls the flow of ionic currents through nonlinear channels into the hair cells. The ionic flow, in turn, generates electrical potentials across the hair cell membranes. Finally, these potentials are conveyed by the auditory-nerve fibers to the central auditory system. In the human auditory system, there are roughly 30,000 auditory-nerve fibers, innervating approximately 3000 inner hair cells along the length of the cochlea (3.5 cm). While an auditory-nerve fiber innervates only one inner hair cell, several fibers (up to 10) may converge onto one hair cell.

These three complex transduction stages – the fluid-cilia coupling, the ionic channels, and the membrane potentials – can be surprisingly well modeled by a three step process (Fig. 1): a velocity coupling stage (modeled by a time derivative), an instantaneous nonlinearity modeling the opening and closing of the ionic channels, and a lowpass filter with a relatively short time-constant ($< 0.3\ ms$) to describe the ionic leakage through the hair cell membranes. Detailed considerations of the biophysical bases of these models can be found in [15].

The intracellular hair cell potentials generated at the end of these stages are conveyed via the auditory-nerve fibers to the cochlear nucleus, the first station of the central auditory system. This is achieved through a series of transformations in which the intracellular potentials are first converted into stochastic trains of electrical impulses (firings) on the auditory-nerve. Detailed biophysical models of these transformations can be found in [16, 19]. More abstractly, the stochastic firings can be modeled as nonstationary point processes with instantaneous rates that approximately reflect the underlying intracellular hair cell potentials [20]. Recipient neurons in the cochlear nucleus then reconstruct estimates of the hair cell potentials by effectively computing the ensemble averages of activity in locally adjacent fibers [21].

From an information processing point of view, these complex transformations merely convey hair cell potentials to the cochlear nucleus. Consequently, in our functional model, they can be ignored all together. Such a simplifying assumption ignores the effects of the adaptive mechanisms operative at the hair cell/auditory-nerve junctions which might be important in describing the responses to the onset of sound [19]; They have also been found useful in some phonetic segmentation algorithms [6].

To summarize, the spatiotemporal patterns of basilar membrane vibrations, $y_1(t;s)$, are transduced into intracellular hair cell potentials (or equivalently, into instantaneous firing rates the auditory-nerve), $y_2(t;s)$, as follows:

$$y_2(t;s) = g(\partial_t y_1(t;s)) *_t w(t),$$

where $\partial_t y_1(t;s) = \partial y_1(t;s)/\partial t$ is the output of the fluid-cilia coupling; $g(\cdot)$ is an instantaneous sigmoidal nonlinearity of the form

$$g(u) = \frac{1}{1+e^{-\gamma u}} - \frac{1}{2} \tag{1}$$

where γ is the gain at the input of the nonlinearity; $w(t)$ is the impulse response of the lowpass filter (temporal smoothing window) due to the hair cell membrane. Note that, apart from the smoothing effect of the lowpass filter, the patterns at the output of this stage look similar to the basilar membrane vibrations with three basic changes:

1. Because of the velocity coupling, the *extrema* of $y_1(t;s)$ with respect to time become the *zero-crossings* of $y_2(t;s)$.

2. $y_2(t;s)$ patterns are compressed and approximately half-wave rectified by the nonlinearity. In the auditory-nerve, the dynamic range between threshold and saturation of

activity in a given fiber is limited to 30-40 dBs [22]. Thus, a sinusoidal vibration at a particular point on the basilar membrane may look more like a square wave firing rate on the nerve.

3. Temporal fluctuations of $y_2(t;s)$ (also known as the *phase-locked* activity) in any given fiber are limited to frequencies below 4-5 kHz because of the lowpass effect of the hair cell membranes. Above these frequencies (in mammals), the auditory nerve indicates the presence of a particular frequency in the sound stimulus by a steady increase in the firing rate at the appropriate filter output (much like the representation used in traditional spectrograms).

II.C The Reduction Stage: Spectral Estimation

The auditory-nerve transmits the sound evoked activity ($y_2(t;s)$) to the cochlear nucleus of the central auditory system. Information about various attributes of the stimulus, such as its timbre, pitch, temporal character, and location in space, are then extracted and processed along parallel pathways. In this report, we focus on the estimation by the auditory system of the short-time sound spectrum, a stimulus feature that plays a pivotal role in the recognition of different sounds and in other fundamental auditory tasks.

There are many ways by which a spectral estimate of the stimulus can be extracted from the patterns of auditory-nerve responses. (Please refer to [21] for a detailed review of these issues). We shall emphasize here a particularly simple and elegant scheme that is found in all sensory systems, and is implemented biologically by a neural network known as the *lateral inhibitory network (LIN).* In vision, the network exists in the retina, and functions to highlight regions in the image that are characterized by fast transitions in light intensity, such as edges and peaks [23]. In audition, it has been shown that exactly the same network can produce a "spectral" profile of the stimulus by rapidly detecting discontinuities along the spatial axis of the auditory-nerve patterns and integrating its outputs over a few milliseconds [21,24]. On the auditory-nerve, such discontinuities are not created by stationary patterns (e.g., the intensity profile of an image), but rather are due to instantaneous mismatches in the time-waveforms in different channels because of different frequencies, phases, or amplitudes [24]. Such a neural network is thought to exist in the anteroventral cochlear nucleus which receives direct input from the auditory nerve, and exhibits physiological and anatomical characteristics consistent with the structure and function of the LIN [21].

The simplest models of the *LIN* consist of a layer of nonlinear neurons that are mutually inhibited either in a feedback or a feedforward manner [21,25]. From a mathematical viewpoint, the *LIN* action can be effectively reduced to a series of three steps with three intermediate outputs $(y_3(t;s), y_4(t;s), y_5(t;s))$ (Fig.1):

1. *A derivative with respect to the spatial axis of the cochlea*: The spatial derivative models the lateral inhibitory influences among the LIN neurons, which essentially endow it with the sensitivity to, and enhancement of, spatial discontinuities in its input patterns. More realistically, this derivative is not pure, but rather is "leaky", i.e., is accompanied by local smoothing due to the finite spatial extent of the lateral interactions and/or the convergence of input fibers [21]. The output of this stage can be expressed as :

$$y_3(t;s) = \partial_s(g(\partial_t y_1(t;s))) *_t w(t) *_s v(s)$$

$$= (g'(\partial_t y_1(t;s)) \cdot \partial_s \partial_t y_1(t;s)) *_t w(t) *_s v(s),$$

where g' is the derivative of the sigmoidal nonlinearity with respect to its argument, $\partial_s \partial_t y_1(t;s)$ is the mixed partial derivative with respect to both time and space of the basilar membrane patterns (or, equivalently, of the wavelet transform of $x(t)$),

$*_s$ denotes the convolution operation with respect to space, and $v(s)$ is a spatial smoothing window reflecting the leakiness of the derivative.

2. *A half-wave rectifier*: This step models the threshold nonlinearity in the neuron models of the *LIN* network. The output of this stage is expressed as:

$$y_4(t;s) = \max(y_3(t;s), 0).$$

3. *A long time-constant (*$10 - 20$ ms*) integrator*: This step models primarily the fact that *central* auditory neurons (unlike auditory-nerve fibers) are unable to follow rapid temporal modulations (e.g., higher than a few hundered hertz) [26]. Rather, they signal a temporally integrated version of their output. The final output of the LIN is then:

$$y_5(t;s) = y_4(t;s) *_t \Pi_T(t),$$

where $\Pi_T(t)$ is a (possibly rectangular) window of duration $T(\approx 10 - 20\ ms)$.

Therefore, at the final output of the *LIN* we obtain a representation of the sound stimulus that, as we shall elaborate, approximately reflects its *short-time* amplitude spectrum. We shall call this pattern the final *auditory representation* of the signal. As mentioned earlier, the exact correspondence between this representation and the original short-time Fourier transform of the signal is difficult to see because of the complexity of the intervening transformations. However, experimental tests with automatic speech recognition systems have consistently demonstrated that auditory representations preserve all spectral information and may even highlight more perceptually useful features [8]. In the following sections, we shall illustrate this fact by reconstructing close replica of the stimulus spectrum from the outputs of the *LIN*.

II.D The Auditory Representation in the High-Gain Limit

When the auditory-nerve is driven at saturation levels, as is the case for speech stimuli at moderate sound levels, the interpretation of the output patterns $(y_3(t;s), y_4(t;s), y_5(t;s))$ can be somewhat simplified. Specifically, note that in the high-gain limit, i.e., as $\gamma \to \infty$, the nonlinearity $g(\cdot)$ becomes:

$$\lim_{\gamma \to \infty} g(u) = \theta(u) - \frac{1}{2}, \tag{2}$$

where $\theta(u)$ is the Heaviside step function. Consequently, the derivative of the sigmoidal nonlinearity $g'(\partial_t y_1(t;s))$ approaches distributionally a Dirac delta function, $\delta(\partial_t y_1(t;s))$, centered at the extrema of $y_1(t;s)$ [27,28]. The output $y_3(t;s)$ can therefore be re-interpreted as follows:

$$y_3(t;s) = (\delta(\partial_t y_1(t;s)) \cdot \partial_s \partial_t y_1(t;s)) *_t w(t) *_s v(s),$$

i.e., $y_3(t;s)$ is the sum of nonuniformly distributed samples centered at the extrema (with respect to time) of the original wavelet transform, and scaled by the values of the *mixed partial derivative* (or the curvature) of the wavelet transform around these points. $w(t)$ and $v(s)$ simply provide a localized average of the impulses.

Similarly, $y_4(t;s)$ becomes the sum of the *positively*-valued samples only, while $y_5(t;s)$ is the short-time average value of the positively-valued samples.

II.E Summary of Data Compression in the Auditory Processing Stages

In a simplified view of auditory processing, the acoustic signal is represented by a succession of increasingly smaller body of data. Thus at the basilar membrane stage, it is approximately represented by its affine wavelet transform $(y_1(t;s))$. Further on, the wavelet transform is replaced by relatively few samples located at its temporal extrema, and which evaluate its mixed partial derivative at these points ($y_3(t;s)$). In the next stage ($y_4(t;s)$), only a portion of these samples, the positively-valued samples, are retained. And finally, at the last stage ($y_5(t;s)$), all sample times are discarded in favor of keeping only a short-time average value of the positively-valued samples from each auditory output. Given these significant reductions in the data, it is important to ask how and what information does the auditory system preserve about the original spectrum of the stimulus. In the following sections, we shall elaborate on these issues through algorithms capable of reconstructing the stimulus from these different representations.

III. Reconstructing the Acoustic Signal from its Auditory Representations

In this section, algorithms are developed to reconstruct the acoustic signal from its representations at various stages of auditory processing. They are based on the method of convex projections as discussed in [29], and further developed and applied by [10,30]. We start by an overall outline of such algorithms, followed by more precise formulations, and end by a brief discussion of their functional significance.

III.A An Outline of the Reconstruction Algorithms

An input signal $x(t)$ is assumed to belong to a linear Hilbert subspace of a parent Hilbert space. Output signals $y_j(t;s), j = 1, \cdots, 5$ are generated at successive stages of auditory processing (Fig.1). For the sake of simplicity in presentation, we ignore the effects of the local spatiotemporal smoothing windows $w(t)$ and $v(s)$ in the following discussion. Therefore, in the *high-gain limit*:

$$y_1(t;s) = h(t;s) *_t x(t); \tag{3}$$

$$y_2(t;s) = \theta(\partial_t y_1(t;s)) - \frac{1}{2}; \tag{4}$$

$$y_3(t;s) = \delta(\partial_t y_1(t;s)) \cdot \partial_s \partial_t y_1(t;s); \tag{5}$$

$$y_4(t;s) = \max(y_3(t;s), 0); \tag{6}$$

$$y_5(t;s) = \frac{1}{T} \int_{t-T}^{t} y_4(\tau;s) d\tau. \tag{7}$$

These outputs can be further abstracted in terms of corresponding data vectors $V_j, j = 1, \cdots, 5$, defined at each s as (Fig.2):

- V_1 is the collection of amplitudes and locations of the extrema points of $y_1(t;s)$ with respect to time in all dilation channels at the output of the first stage.

- V_2 records only the locations of the extrema of $y_1(t;s)$ with respect to time, which are also the zero-crossings of the second stage output, $y_2(t;s)$. We also define an enlarged vector $V_{2'}$ which augments V_2 with the energy in each auditory channel.

- V_3 is similar to V_1, except that the amplitudes of the extrema are replaced by the mixed partial derivatives at the instants of the extrema, i.e., the samples of $y_3(t;s)$.

- V_4 retains only the samples of V_3 with positively-valued mixed derivatives. They are obtained from the samples of $y_4(t;s)$ at the output of the half-wave rectifier.

- V_5 records the short time-average of the samples of V_4 in each channel, i.e., samples of the output $y_5(t;s)$ at the final stage of auditory processing.

$V_j, j = 1, \cdots, 4$ can be regarded as sample points of the functionals $y_j(t;s) = (F_j x)(t)$, sampled at a nonuniform rate in time t; V_5 are samples of the functional $y_5(t;s) = (F_5 x)(t)$ which transforms a time function $x(t)$ into a spatial function (the final auditory spectral representation). In all of the above data vectors, including the effects of the two local spatiotemporal smoothing windows $w(t)$ and $v(s)$ entails simply scaling all sample values uniformly in proportion to the area under the appropriate window.

Our goal is to reconstruct the original signal $x(t)$ from each of these sample vectors based on the concept of convex projections. To do so, we first observe that each sample vector V_j can generally be satisfied by a set (A_j) of functions. For instance, the vector V_1 of extrema locations and amplitudes can be satisfied by many functions besides the specific $y_1(t;s)$ generated by the cochlear filters. However, if each of the sets $A_1, A_2, \cdots, A_n$ is a well-defined closed convex set, and if A_0 is some specific invertible transform of $x(t)$ (e.g., a wavelet transform Hx), then the transform Hx (and hence the signal x) can be determined from the intersection of such sets, i.e., $Hx \in \bigcap_{j=0}^{n} A_j$. Ideally, n is large enough that $H(x)$, or x, is uniquely represented in $A = \bigcap_{j=0}^{n} A_j$, i.e., A contains only one nonzero point.

In order to identify such intersection point(s), we perform a series of projections upon the sets $A_j, j = 0, \cdots, n$. Let P_j be the projection operator onto the individual convex set A_j. If every P_j is nonexpansive, then so is the composition operator $P = P_n P_{n-1} \cdots P_1 P_0$. Hence, any point in the intersection set $A = \bigcap_{j=0}^{n} A_j$ is a fixed point under the projection P, and a fixed point y in A can be found by repeated application of the composition projector as

$$y = \lim_{m \to \infty} P^m y^0 \tag{8}$$

where P^m denotes m successive compositions of P and y^0 is an arbitrary initial point in the Hilbert space. The desired $y = Hx$ exists if each $A_j, j = 0, \cdots, n$ is convex and closed.

III.B Feature Sets and Projections

The following are formal definitions of the subspaces and convex sets mentioned earlier. Let $\mathcal{B}$ denote the subspace of $L^2(R)$ composed of all $x(t)$ bandlimited to ω_c rad/s:

$$\mathcal{B} = \{x(t) : x(t) \in L^2(R), X(\omega) = 0 \text{ a.e. in } |\omega| > \omega_c\} \tag{9}$$

where $X(\omega)$ is the Fourier transform of $x(t)$. A Hilbert subspace $L^2(\mathcal{B})$ is defined as

$$L^2(\mathcal{B}) = \{y(t;s) : y(t;s) \in \mathcal{B} \text{ for } \forall s, \int_S \int_R |y(t;s)|^2 dt\, ds < \infty\} \tag{10}$$

where the inner product of $y(t;s)$ and $z(t;s)$ in $L^2(\mathcal{B})$ is defined as

$$< y, z >= \int_S \int_R y(t;s) z(t;s) dt\, ds, \tag{11}$$

where S is an index set of s.

The cochlear filters perform a transformation on the input signal $x(t)$, defined as:

$$Hx = \{h(t;s) *_t x(t) : \forall s \in S\} \tag{12}$$

where $h(t;s)$ represent the the impulse responses of the cochlear filters. If the transform is assumed to be an affine wavelet transform, then $h(t;s) = a^s h(a^s t; 0)$ $(0 < a < 1)$. In the frequency domain, this relationship becomes $H(\omega;s) = H(a^{-s}\omega; 0)$, and the inverse wavelet transform is written as

$$H^{-1}y = \int_S y(t;s) *_t h(-t;s) ds. \tag{13}$$

We shall assume the overall frequency response of the cochlear filters to be flat within the effective band so that it can be normalized as

$$\int_S |H(\omega;s)|^2 ds = 1, \forall \omega \in [-\omega_c, \omega_c]. \tag{14}$$

In the remainder of this discussion, we shall take the cochlear transform to be an affine wavelet transform. We emphasize, however, that *any unity-gain invertible transform* is applicable to the theory or algorithms developed here. Consequently, filters that more closely approximate the bark scale of the cochlea can be readily applied.

For a fixed wavelet transform H, a linear subspace A_0 of the Hilbert space consists of the affine wavelet transform of any finite-energy bandlimited signal:

$$A_0 = \{Hx : x(t) \in \mathcal{B}\}. \tag{15}$$

Obviously, $A_0 \subset L^2(\mathcal{B})$. It can be shown that, by Parseval's formula, the wavelet transform H is a norm-preserving operator, i.e.,

$$||Hx|| = ||x||.$$

This implies an isomorphism between A_0 and $\mathcal{B}$, and hence A_0 is convex and closed.

Let $V_1(\cdot;\cdot)$ be the vector of ordered extrema in time (location and amplitude) of the functional $y_1(t;s)$ for the input $x(t)$ to be reconstructed. For instance, $V_1(i;s)$ denotes the i-th extrema in time (location and amplitude) on the s-th channel of $y_1(t;s)$. For simplicity, we use V_1 to mean $V_1(\cdot;\cdot)$, and this rule applies also to $V_j, j = 2,3,4,5$. A set A_1 is related to the vector V_1 as

$$A_1 = \{y(t;s) : y(t;s) \in L^2(\mathcal{B}), y(t;s) \text{ has the same extrema as in } V_1\}. \tag{16}$$

In the *high-gain limit*, the compressive sigmoidal nonlinearity $g(\cdot)$ causes $y_2(t;s)$ to exhibit purely rectangular waveforms along the time axis (ignoring the smoothing effect of the lowpass filter $w(t)$). Because of the time-derivative (Fig.1), the zero-crossings of $y_2(t;s)$ (collected in vector V_2) represent the extrema locations of $y_1(t;s)$. Then, a set A_2 can be defined as

$$A_2 = \{y(t;s) : y(t;s) \in L^2(\mathcal{B}), y(t;s) \text{ has the same extrema locations as in } V_2\}. \tag{17}$$

Similarly, if V_3 is a vector of the ordered mixed derivatives of $y_1(t;s)$, evaluated at instants of the extrema of $y_1(t;s)$, i.e., $y_3(t;s) = \delta(\partial_t y_1(t;s)) \cdot \partial_s \partial_t y_1(t;s)$, then a set A_3 can be defined as

$$A_3 = \{y(t;s) : y(t;s) \in L^2(\mathcal{B}), y(t;s) \text{ has the same mixed differential}$$

$$\text{at the extrema of } y_1(t;s) \text{ as in } V_3\}. \tag{18}$$

V_4 is the same vector as V_3 except for eliminating the non-positive values of the mixed derivatives. Hence, a set A_4 can be defined as

$$A_4 = \{y(t;s) : y(t;s) \in L^2(\mathcal{B}), y(t;s) \text{ has the same positively-valued}$$

$$\text{mixed derivatives at the extrema of } y_1(t;s) \text{ as in } V_4\}. \tag{19}$$

V_5 is a highly compressed vector containing only the short-time average values of the samples of V_4 in each channel. Specifically, $y_5(t;s)$ is computed as follows:

$$y_5(t;s) = \frac{1}{T} \int_{t-T}^{t} \max(y_3(\tau;s), 0) d\tau \tag{20}$$

Vector V_5 simply takes the individual values of $y_5(t;s)$. Note that over a given frame T of time, one can sample $V_5(t;s)$ in t (with sampling rate T) to reduce data. Hence each frame is featured by $V_5(nT, s), n = 0, 1, \cdots$. Accordingly, A_5 is defined in terms of V_5 as

$$A_5 = \{y(t;s) : y(t;s) \in L^2(\mathcal{B}), y(t;s) \text{ has the average sample values as in } V_5\}. \tag{21}$$

In *Appendix I*, we prove the following three propositions.

Proposition 1. The feature sets $A_j, j = 0, \cdots, 5$ are closed.

Proposition 2. The closed feature sets $A_j, j = 0, \cdots, 3$ are convex.

Unfortunately, A_4 and A_5 are not convex, and hence convergence of (8) to a *unique* solution is not possible. Our examples, however, indicate that excellent reconstructions of the signal are still obtainable from V_4 and V_5.

Next we define the operators $P_j, j = 0, \cdots, 5$ as projections upon the corresponding sets $A_j, j = 0, \cdots, 5$, i.e., the image of $y \in L^2(\mathcal{B})$ under P_j is the element in A_j which is closest

(L^2 norm) to y. If A_j is a closed convex set, then the corresponding operator P_j is an orthogonal nonexpansive projection [29]. The operator P_0 from $L^2(\mathcal{B})$ onto A_0 is composed of the wavelet transform operator and the inverse transform operator:

$$P_0 = HH^{-1}. \tag{22}$$

Proposition 3. P_0 so defined is a nonexpansive orthogonal projection.

P_1 is the projection from $L^2(\mathcal{B})$ onto A_1. Mallat and Zhong [11] used the composition projection $P = P_1 P_0$ to reconstruct images. Note that, unlike in [11], our input signal $x(t)$ is restricted to be bandlimited. This restriction is necessary for set $A_j, j = 1, \cdots, 5$ to be closed under the natural norm derived from the inner product defined in (11). In Mallat and Zhong [11], there is no bandlimited condition on the input, and the closure of A_1 is guaranteed under the Sobolev norm.

Our P_1 is realized as follows. Let $y \in A_0$ and its image under P_1 be z. P_1 minimizes the difference of z and y in the mean-square sense (L^2 norm). Let $(t_{i;s}, a_{i;s})$ denote the location and amplitude of an extremum in $y_1(t; s)$ (recorded in V_1), and $e(t; s) = z(t; s) - y(t; s)$. Then $z = P_1 y$ minimizes

$$||\epsilon||^2 = \int_S \int_R |e(t; s)|^2 dt\, ds = \int_S \sum_i \int_{t_{i;s}}^{t_{i+1;s}} |e(t; s)|^2 dt\, ds \tag{23}$$

while it satisfies the conditions that (1) $dz(t; s)/dt$ not changing sign for $t \in [t_{i;s}, t_{i+1;s})$, and (2)$z(t_{i;s}; s) = a_{i;s}$.

The other P_j's are realized in a similar fashion. Thus, performing P_2 is equivalent to minimizing (23) subject to $dz(t; s)/dt$ not changing sign for $t \in [t_{i;s}, t_{i+1;s})$. Likewise, performing P_3 is equivalent to minimizing (23) subject to (1) $dz(t; s)/dt$ not changing sign for $t \in [t_{i;s}, t_{i+1;s})$ and (2) $\partial_s \partial_t z(t_{i;s}; s) = \beta_{i;s}$, where $(t_{i;s}, \beta_{i;s})$ denotes the time location and the value of the mixed derivative of an extremum of $y_1(t;s)$. Performing P_4 is equivalent to minimizing (23) subject to (1) $dz(t;s)/dt$ not changing sign for $t \in [t_{i;s}, t_{i+1;s})$ and (2) $\partial_s\partial_t z(t_{i;s};s) = \beta_{i;s}$, if $\partial_s\partial_t z(t_{i;s};s) > 0$. In this case, only positively-valued samples of the mixed derivative are checked. Finally, P_5 is equivalent to minimizing (23) subject to (1) $dz(t;s)/dt$ not changing sign for $t \in [t_{i;s}, t_{i+1;s})$ and (2) $\frac{1}{T}\sum_i max(\beta_{i;s}, 0) = V_5(t;s), \forall s$.

Thus, in general, the reconstruction at stage j is implemented using the composition operator $P = P_j P_0$, given the of set of constraints implied by the vector V_j. Of course, if vectors V_j are given at several stages, one can perform the composition operator $P = P_5P_4P_3P_2P_1P_0$, or any combination composition of the individual P_j's. However, some operations would be redundant since some of the sets are nested within others. For instance, $A_1 \subset A_2$, and $A_3 \subset A_4 \subset A_5$.

III.C Interpreting the Data Vectors V_j

In next section, we shall demonstrate how a signal $x(t)$ can be reconstructed by the method of iterated projections from any of the data vectors $V_j, j = 1, \cdots, 5$. From an encoding point of view, this suggests that one can replace the signal by its vector V_j representation, which may be useful for a variety of reasons. Among them are the efficiency of the representation (data compression), perceptual relevance, and robustness to noise. All reasons will be discussed later in section IV. Note, however, that it is intuitively apparent that for a given frame (T seconds) of signal, the corresponding amount of samples stored in V_j decreases as j increases (except V_2); V_5 has the highest compression rate.

Reconstruction from V_1 is in many respects similar to sub-band waveform coding. Thus, instead of the uniform sampling employed in sub-band coding, a nonuniform sampling at the extrema points on each dilation channel is adopted. Nevertheless, the Nyquist criterion for this nonuniform sampling is still satisfied because the number of extrema is roughly twice the number of periods of the highest frequency component, which guarantees the high fidelity of the reconstruction [31].

V_2 (or $V_{2'}$) is a considerablely more efficient code of the signal than V_1 since all extrema amplitudes are discarded. Nevertheless, we shall demonstrate that good quality reconstructions are still possible under certain conditions.

V_3 generally contains more information and data than V_2 since it consists of pairs of extrema locations and the mixed derivative at the extrema locations. It is, however, slightly more efficient than V_1. This is due to the overlap in the cochlear filters, which often causes the extrema of the waveforms from different dilation channels to line up in a small neighborhood, and to have comparable amplitudes. Consequently, the mixed partial derivatives $(\partial_s \partial_t y_1(t; s))$ in such regions vanish.

V_4 is approximately half as large as V_3 as it contains only the samples with positive mixed partial derivatives.

Finally, V_5 is the most efficient representation since, for a given time frame T, only one average value for each channel is retained. Signals constructed from this vector, however, do not contain the phase information because extrema locations in time are discarded by the averaging. Nevertheless, in speech applications, the intelligibility of the signal is not affected since the short-time amplitude spectrum of the original input $x(t)$ can still be faithfully represented. This statement is verified analytically in *Appendix II*, and illustrated experimentally in the next section.

IV. Examples of Reconstructions

The algorithms discussed above were used to reconstruct natural speech segments. In this section we discuss the details of the implementations, the fidelity of the reconstructions, and the implications of these algorithms for robust and efficient representations of signals.

IV.A Details of Implementations

As described previously, the first stage of auditory processing is assumed to be a linear bank of filters that, under specific conditions, can be considered an affine wavelet transform. Such a filter bank of 64 channels, uniformly spaced in s axis, was constructed by dilating a mother wavelet (dilation parameter $a = 0.9445$), which is the impulse response of a biophysical model of the basilar membrane [15] at the region tuned to approximately 1 kHz. The transfer characteristics of a portion of the filter bank are shown in Fig.3 where the filters with the highest and lowest frequency bands (corresponding to the base and apex of the basilar membrane) are indexed as $s = 0$ and $s = 63$, respectively. Note that the dilation here is chosen to preserve the maximum amplitude rather than the energy of the mother wavelet so that when the transfer functions of the 64 channels are added up, the total response is satisfactorily flat between 200 Hz and 5.9 kHz. This makes the wavelet transform together with its inverse counterpart act like a unity-gain bandpass filter. For all examples shown here, we select a finite interval of the signal ($T = 51.2$ ms), and all filter responses are then computed using standard FFT algorithms with a 20 kHz sampling rate. It is important to note here that any unity gain invertible transform (instead of the wavelet transform) would be adequate for the reconstruction algorithms described here. However, the fidelity of the reconstructions depends very much on such parameters of the filter bank as their number, bandwidth, shape, and the degree of overlap between adjacent filters. We shall touch briefly upon these issues in section V.

From the spatiotemporal outputs of the filter bank, we extract the data for each of the vectors $V_j, j = 1, \cdots, 5$, as needed. For the vectors V_1 and V_2 (or $V_{2'}$), we simply use, as appropriate, the values and locations of the extrema at each channel. For the other vectors $(V_j, j = 3, 4, 5)$, we need to evaluate the mixed derivative of $y_1(t; s)$. Since all waveforms are sampled uniformly in space and in time, we approximate the mixed derivative $\partial_t \partial_s y_1(t; s)$ by:

$$\partial_t y_1(t; k+1) - \partial_t y_1(t; k),$$

which by further discretization and appropriate normalization becomes:

$$[y_1(i+1; k+1) - y_1(i; k+1)] - [y_1(i+1; k) - y_1(i; k)].$$

For an extremum (with respect to time) of $y_1(t; s)$ at $t = i$ in the k-th channel, the above expression simplifies to

$$\partial_t \partial_s y_1(t; s) \approx y_1(i+1; k+1) - y_1(i; k+1)$$

In other words, evaluating the mixed derivative at an extremum in the k-th channel is equivalent to sampling the time derivative of the output of the $k+1$-st filter (Fig.2). This effective sampling of each channel by the extrema of the channel below it is in general nonuniform, as discussed in the last section. Once the mixed derivatives are evaluated, the data vectors $V_j, j = 3, 4, 5$, can be readily constructed.

The procedure for reconstructing the signal $x(t)$ from each of the data vectors $V_j, j = 1, \cdots, 5$ follows a very similar pattern of operations. For instance, given V_1 (the locations and amplitudes of the extrema of a particular signal to be reconstructed), the reconstruction algorithm starts with a random pattern in A_1 whose extrema have the same locations and amplitudes as in V_1. The projection P_0, which is a mapping from $L^2(\mathcal{B})$ to A_0, is implemented as an inverse wavelet transform followed by its counterpart. After P_0, the signal is passed through P_1 which, in terms of $L^2(\mathcal{B})$ norm, induces minimum adjustment while changing the locations and amplitudes of the extrema for each channel so as to match those in V_1. The resultant pattern is recursively put through P_0 and P_1 until a predefined condition is achieved, say, after a given number of iterations is reached or the adjustment between two successive iterations is small enough. This algorithm is very similar to one proposed by [11].

The same procedure is applied to reconstructions from other V_j vectors, except that each is based on the corresponding A_j domain. For example, in the reconstruction from V_3 data vector, the projection P_3 from A_0 to A_3 adjusts $y_1(t;s)$ to have its mixed derivatives satisfy V_3, based on the $L^2(\mathcal{B})$ norm. Because of the implementational difficulty, P_5 is not a true minimum mean-square distance operator. However, it still converges (at a slower rate).

IV.B Reconstructions of Speech Signals

We illustrate here the result of reconstructing two normally spoken vowels, /iy/ and /ae/. For each, the signal is compared to reconstructions, both in the time- and frequency-domains, from the data vectors $V_j, j = 1,3,4,5$ (Figs.4 and 5). Reconstructions from V_2 and $V_{2'}$ are considered separately (Fig.7). The fidelity of the spectral reconstructions are measured by the signal-to-noise (S/N) ratio defined as :

$$S/N = \frac{\int_\Omega |X(\omega)|^2}{\int_\Omega (|X(\omega)| - |\hat{X}(\omega|V_j)|)^2} \quad \text{where } \Omega/2\pi = [200, 5900]\text{Hz}$$

where $X(\omega)$ and $\hat{X}(\omega|V_j)$ are the spectra of the original and reconsctructed (from V_j) signals, respectively. The number of iterations applied in each case is also indicated. The algorithm is typically stopped after a 100 iteration or when the S/N exceeds 20 dB. For reasons explained later, the algorithm to reconstruct $\hat{X}(\omega|V_5)$ is stopped when the *maximum* S/N or a preset number of iterations is achieved.

It is evident from Figs. 4 and 5 that signal reconstructions from the the data vectors $V_j, j = 1, 3, 4$ are excellent, both in time and frequency domains. This clearly proves that up to the fourth stage, very little information about the detailed structure of the signal is lost through the auditory stages. This is despite significant reductions in the amount of data preserved in the V_j vectors. For instance, for a given frame of signal, e.g., vowel /iy/ in Fig.4**A** (51.2 ms or 1024 samples at 20 kHz sampling rate), the length of the data vectors decreases from about 10000 samples in V_1, to about 5000 samples in V_4. Little corresponding drop in the S/N occurs.

In the final stage, all phase information is lost as the locations in time of the $y_1(t; s)$ extrema are ignored. Nevertheless, the spectral reconstruction remains good despite a 16 fold reduction in the data rate (64 samples in V_5). The drop in the S/N ratio, however, does not imply necessarily a deterioration in the quality of the representation. Rather, as we elaborate later, the distortions reflect enhancements of perceptually important features in the spectrum which are inherent to the way information is processed in the auditory system.

Unlike other data vectors, the final auditory representations in V_5 can be compared directly to the acoustic spectrum (Fig.6). As shown in *Appendix II*, they are approximate estimates of the acoustic spectrum that are warped in frequency according to the following formula:

$$V_5(t; s) \leq 2(\frac{a^s\omega_0}{2\pi})^2|X(a^s\omega_0)|, \tag{24}$$

i.e., the normalized pattern, $V_5(t; s)/a^{2s}$, reflects the short-time amplitude spectrum of the input signal $x(t)$ in a dilated fashion $|X(a^s\omega_0)|$. However, more subtle changes in the amplitude spectrum occur, which are discussed in more detail in sections IV.D and V.

Finally, note that for all spectral reconstructions above, most details of the Fourier spectrum are reproduced (e.g., the harmonic structure), and not only the envelope of the spectrum as is the case, for instance, in *Linear Predictive Coding.* Consequently, such details as the pitch, voice quality, and timbre of the signal are all preserved well in this representation.

IV.C Reconstructions from Data Vectors V_2 and $V_{2'}$

Examining the reconstructions from data vectors V_2 and $V_{2'}$ raises important questions concerning the detailed spectral features preserved by the auditory representations and their robustness to noise. Fig.7 illustrates such reconstructions for the vowel /iy/. In both cases, the *fine* structure of the original spectrum is well reproduced. However, only $V_{2'}$ is able to convey accurately the correct amplitudes (and hence the spectral envelope) of the stimulus harmonics. The reason why V_2 fails is that zero-crossings alone cannot reflect the *absolute* levels of the underlying harmonics. Thus, in the low frequency regions where the filters are narrow, the responses (or zero-crossings) due to different harmonics are spatially well separated on different channels. Consequently, it is impossible to determine their absolute or relative levels without additional information. This is indeed the case in $V_{2'}$, where the channel energies are explicitly available.

The situation is different in the higher frequency region (> 3 kHz) where the bandwidths of the cochlear filters become broader. Here the responses due to neighboring stimulus harmonics overlap on the same channel, and the zero-crossings of the resulting compound waveform do reflect the *relative* (not absolute) levels of the interfering harmonics [32]. Consequently, reconstructions of the spectral envelope from V_2 improve.

Finally we emphasize, that succeeding data vectors $V_j, j = 3, 4, 5$ do not contain explicit channel energies as in $V_{2'}$. Instead, the zero-crossings are augmented by other *across channel* information, specifically the mixed-derivative values, that make it readily possible to determine the spectral envelope.

IV.D Spectral Enhancement and Noise Suppression in V_5

V_5 is the most important data vector in that it represents the final output upon which all higher auditory percepts are based. As shown in *Appendix II*, this vector can be treated directly as reflecting the amplitude spectrum of the original stimulus with two important qualifications:

1. Its spatial axis s is dilated (or, more generally, warped) relative to the original frequency axis (see *Appendix II*).
2. Its valleys are more depressed, giving the impression of enhanced harmonic peaks.

The causes for the enhancement of the peaks are discussed briefly in *Appendix II.* However, an intuitive argument can be based on the expression for V_5 derived earlier. From *Appendix II*, we have

$$V_5(t;s) = \frac{1}{T} \sum_i \max(\partial_s \partial_t y_1(t_{i;s};s), 0), \tag{25}$$

where $t_{i;s}$ are the extrema of $y_1(t;s)$ during $[t-T, t)$. The mixed derivative values in channel s depend strongly on the degree of *coherence* of the waveforms in nearby channels. For instance, consider two adjacent channels with *partially overlapping* transfer functions. If, on the one hand, the channels are dominated by a spectral component in the region of the overlap, their outputs are similar and the derivatives vanish. If, on the other hand, the filters are dominated by a component in the frequency region where they differ, the derivatives become sizable. Put more succinctly, the output V_5 at s is *enhanced* by spectral energy that drives it *differentially* from its neighbors, and is *suppressed* by spectral energy that drives it *coherently* with its neighbors.

The consequences of this interplay of between the two influences can be seen in a more exaggerated form in Fig.8. Here, a two-tone stimulus (1, 2.5 kHz) in broadband noise is applied to a bank of overlapping cochlear filters that are more broadly tuned, but also are highly asymmetrical (Fig. 8a). The reconstructed spectrum from V_5 (Fig. 8b) displays significant suppression of the noise on either side of the tone peaks. The width of the reconstructed peaks reflect the bandwidth of the differential filters $(\partial_s H(\omega; s))$ centered at $s = 31$ (1 kHz) and $s = 15$ (2.5 kHz). The surrounding suppression is due to the dominance by each of the tones of the patterns of zero-crossings in neighboring overlapping filters. Because of the asymmetric spread of the cochlear filters towards lower frequencies (Fig.8a), the suppression is more extensive on the high frequency side of each peak. Finally note that the suppression near the 2.5 kHz tone peak is more extensive than that at 1 kHz because of the broader bandwidths of the high frequency filters. A possible perceptual correlate of this side-band suppression is known as *masking* in the psychoacoustical literature [33].

Similar enhancements of the spectral harmonics due to suppression of surrounding valleys occur in the vowel reconstructions $\hat{X}(\omega|V_5)$ shown in Figs.4, 5, and 9. As mentioned earlier, such "deviations" of the reconstructed spectra from the original lead to lower S/N values. A consistent observation in our experiments is that enhancement of the reconstructed spectra (and hence "deterioration" of the S/N) increases for larger numbers of iterations. And that usually a higher S/N can be achieved after only a few iterations, i.e. before the algorithm converges (Fig.9). Consequently, one has the choice of stopping the iterations either at the closest approach to the original spectrum (maximum S/N), or at its most enhanced version.

It has long been recognized that spectral peaks and their neighborhoods are particularly important in the perception of acoustic stimuli [34,35]. As such, auditory representations more faithfully reflect our perception of speech and music than is indicated by the perceptually-blind and simple distortion measures used in Figs. 4,5, and 9. In fact, perceptually-sensitive metrics suggested in recent years have all been based on transformations of the Fourier spectrum that resemble those effectively needed to produce the final auditory spectrum V_5 (Fig.6), namely, a distortion of the frequency axis and an enhancement of the amplitude spectrum [3,35].

V. General Discussion

Given the auditory model's ability to preserve and enhance the acoustic spectrum, a question arises as to the functional significance of the specific sequence of transformations, filter shapes, nonlinear compression, and the LIN that the auditory system invokes to generate its final representation V_5. The answers here are particularly useful when one attempts to re-design the auditory processing stages for various engineering applications such as low bit-rate data compression and automatic speech recognition. In the following, we shall first briefly discuss the presumed rationale behind the general properties of the initial cochlear transformations. Then we relate the specific shape of the cochlear filters to the functional role of the LIN and the compressive nonlinearity. The section ends with a brief discussion of the potential engineering applications of auditory processing concepts.

V.A Cochlear Analysis as a Wavelet Transform

Why does the auditory system perform a spectral transformation on its input sound? And why is an approximate affine wavelet transform useful? Spectral decomposition of the acoustic stimulus in the cochlea offers two immediate benefits. The first is that it converts a purely time-varying signal to a spatially distributed pattern of activity along the cochlea. This is vital for the sensory nervous system in general since its ability to process spatially distributed patterns (through richly interconnected neural networks) is superior to its ability to manipulate rapidly varying temporal signals [36]. The second benefit is that the spectrum provides a more direct access to a fundamental characteristic feature of the signal, namely the resonances of its source. For instance, vocal tract resonances largely distinguish speech vowels, while those of musical instruments influence their perceived timbre.

Spectral representations, both in parametric (e.g., LPC) and nonparametric (e.g., spectrogram) forms, have been extensively used in sound analysis. Most commonly, the encoded spectrum can be interpreted as one generated by a bank of closely spaced bandpass filters of equal bandwidth, i.e., of a constant resolution along the frequency axis [37]. A fundamental implication of this choice of a filter bank is the fixed width of the short-time window within which the signal is analyzed. Since window width is inversely proportional to the bandwidth of the filters, the need for fine frequency resolution (narrower filters) must be balanced against that for fast dynamic response (fine time resolution) and hence for broader filters [37].

Multiscale decompositions like the affine wavelet transform offer a partial way out of this dilemma. This is because the filter bank implied by such a transform exhibits progressively broader bandwidths at higher frequencies. As such, not one, but a range of window durations are used to analyze the signal. Thus, rapidly varying signals (e.g., acoustic transients) are effectively analyzed with shorter windows than those of the slower components and events. An analogous rationale underlies the use of multiscale decomposition in image processing to preserve both the global features (course resolution) and the finer details (high resolution) of an image [38].

Cochlear filters, as mentioned earlier, conform to this multiresolution scheme at all sound frequencies. For frequencies $>$ 800 Hz, the transform is strictly an affine wavelet transform since the frequency axis is logarithmic. For lower frequencies, the increase in filter bandwidths is less rapid (non-logarithmic), probably because a higher premium is placed on preserving spectral resolution in order to extract the low order harmonics (see section IV). From an algorithmic point of view, this deviation simply entails an appropriate adjustment in the normalization and frequency warping that V_5 implies in relation to the sound spectrum (see Eq.24 and *Appendix II).*

V.B Multiscale Processing and Spectral Estimation

Despite the multiresolution decomposition, the narrowest cochlear filters remain relatively broad and highly overlapping. Thus, if the auditory system were to discard the detailed form of the wavelet transform in its spectral estimate by, for instance, measuring the average output power of each filter (as done in spectrogram displays), it would face two important limitations [24,39]: (1) Poor spectral resolution, and (2) an almost total loss of the spectral envelope at high sound levels because of the limited dynamic range of the auditory-nerve responses. Instead, the auditory system *preserves* the wavelet transform through the phased-locked activity on the auditory-nerve (section II.B), using it centrally via the LIN to extract a well resolved and stable spectral estimate.

Intuitively, the LIN extracts a spectral estimate by correlating the outputs across different scales of the cochlear decomposition. Specifically, it detects regions along the scale (frequency) axis of the decomposition at which neighboring outputs are highly mismatched. This means that large LIN outputs occur only when a signal component passes through the *difference (or differential) filters ($\partial_s H(\omega; s)$)* between adjacent channels. It can be seen from (*Appendix II*; Eq.32) that if these differential filters are both narrow and non-overlapping, then the LIN effectively removes the redundancy in the wavelet representation due to the extensive overlap of the cochlear filters.

It is apparent from the above arguments that the spectral resolution of the final auditory representation V_5 is largely determined by the bandwidth of the differential filters $\partial_s H(\omega; s)$, *and not of the cochlear filters directly.* Nevertheless, the detailed shapes of the cochlear filters, and hence the properties of the wavelets, play an important role in the analysis.

For instance, in order to combine both a fine spectral resolution (i.e., use sharp differential filters) and a good dynamic response (i.e., use relatively broad filters), it is desirable to invoke highly asymmetrical filters similar to those depicted in Fig.8. And this is exactly what the auditory system employs at its analysis stages.

Finally, we comment on the role of the compressive nonlinearity in relation to the LIN spectral estimate. Compression does not influence the resolution of the encoded spectrum. Instead, it affects the detailed shape of the wavelet transform through its compression of the channel responses. Conceptually, compression makes explicit the fact that the acoustic spectrum extracted by the LIN is encoded *not uniformly*, but rather at discrete points (samples) near the extrema of the wavelet transform. This, in turn, is directly responsible for the suppression and enhancement effects which endow the auditory representations V_5 with superior robustness in noise *(Appendix II)*. Such waveform compression as a way of suppressing interference has long been used in amplifier design [40]. Note, however, that the auditory system is able to employ compression only because it preserves the wavelet transform from its analysis filters through the phase-locked activity on the auditory-nerve.

V.C Potential Applications of the Auditory Representations

The final auditory representations in V_5 are essentially spatial patterns that can be interpreted and applied in different ways. Since, as discussed above, the exact form of the resulting patterns is strongly influenced by the parameters of the filter bank, it is possible to tailor the representation to the specific needs of the task at hand.

For instance, V_5 may simply be used to reconstruct an accurate replica of the original spectrum. This would be useful in speech applications where exact voice and timbre quality are to be reproduced. In this case, it is best to use narrower filters, lessen the overlap (to reduce suppression), and increase the number of channels to obtain a full and fine frequency coverage.

In automatic recognition systems or in the study of acoustic features of speech phonemes, V_5 may be interpreted as an *auditory spectrum*, a pattern that reflects our perceptual weighting of the acoustic spectrum. As such, it is best to mimic closely the broad and asymmetric form of the cochlear filters. The *auditory spectrum* can, in turn, be viewed as a one-dimensional pattern to be subjected to further enhancements and multiresolution analyses, as is likely done in the higher auditory and other nervous centers of the brain [10, 41,42].

Finally, V_5 may be used in data compression applications, where it might be sufficient to encode only the envelope of the spectrum using fewer and broadly tuned channels. In fact, even with accurate reconstructions of the detailed spectrum (as in Figs.4, 5, and 9), the bit rates needed are quite low. For instance, at the V_5 sample rate of 1250 samples/sec, and an average 3 bits/sample, we need only 3.7 Kbits/sec to encode speech at a signal-to-noise ratio of 8.8 dB.

VI. Summary

We have presented an analytically tractable framework to describe acoustic signal processing in the early stages of the auditory system. Algorithms were developed to assess the integrity of the acoustic spectrum at all processing stages. The algorithms employed wavelet representations, multiresolution processing, and the method of convex projections to reconstruct close replica of the input stimulus. Reconstructions using synthetic stimuli and natural speech sounds demonstrated minimal loss of information along the auditory pathway. Furthermore, close inspections of the final auditory patterns revealed spectral enhancements and noise suppression that have close perceptual correlates. Finally, the functional significance of the various auditory processing stages was discussed in light of the model, together with their potential applications.

Acknowledgement: This work is supported by grants from the Office of Naval Research, The Naval Research Laboratory, and the Air Force Office of Scientific Research. The authors are members of the System Research Center which is partially funded by an NSF grant (NSFD CD 8803012). We are also grateful to Mr. Y. Pati for helpful discussions, and to three anonymous reviewers for their detailed and valuable criticisms.

Appendix I

We shall prove three propositions in this appendix:

Proposition 1. $A_j, j = 0,1,2,3,4,5$ are closed sets.

Proposition 2. $A_j, j = 0,1,2,3$ are convex sets.

Proposition 3. The operator from $L^2(\mathcal{B})$ onto A_0 defined as $P_0 = HH^{-1}$ is a nonexpansive orthogonal projection.

Proposition 1. $A_j, j = 0,1,2,3,4,5$ are closed sets.

Proof: Since $\int_S |H(\omega;s)|^2 ds = 1$, $\forall\omega$, then by Parseval's formula we have

$$||Hx||^2 = \int_S \int_R |h(t;s) *_t x(t)|^2 dt\, ds = \frac{1}{2\pi} \int_S \int_\Omega |H(\omega;s)|^2 |X(\omega)|^2 d\omega\, ds$$

$$= \frac{1}{2\pi} \int_\Omega |X(\omega)|^2 d\omega = \int_R |x(t)|^2 dt = ||x||^2.$$

H therefore is an invertible norm-preserving operator, which implies an isomorphism between A_0 and $\mathcal{B}$ (recall $\mathcal{B}$ is a linear subspace of a Hilbert space). Therefore, A_0 is a closed (and convex) set. To show the closure of A_1, we first prove the following Lemma .

Lemma 1. If sequence $\{z_n\}$ converges to z in L^2, then for any $b > 0$, any $\delta > 0$, and any t_0, there exist a t and an $N(b,\delta)$, s.t. $|t - t_0| \leq \delta$ and $n > N(b,\delta)$ imply $|z(t) - z_n(t)| < b$.

Proof of lemma 1: Since $\{z_n\} \longrightarrow z$ in L^2, then for a given $\epsilon = 2\delta b^2 > 0$, there exists an N s.t. when $n \geq N$, we have

$$\int_R |z(t) - z_n(t)|^2 dt < \epsilon.$$

Suppose that for a given $b > 0$, $\delta > 0$, and t_0, there is no t in $|t - t_0| \leq \delta$, s.t. $|z(t) - z_n(t)| < b$. Then in $|t - t_0| \leq \delta$, $\forall t$, we have $|z(t) - z_n(t)| \geq b$. This implies

$$\int_R |z(t) - z_n(t)|^2 dt \geq \int_{\{t : |z(t) - z_n(t)| \geq b\}} |z(t) - z_n(t)|^2 dt \geq 2\delta b^2 = \epsilon$$

which creates a contradiction. Hence the lemma is proved.

Let a sequence $\{z_n\} \in A_1$ converge to z. It is clear that z is also bandlimited, i.e., $z \in \mathcal{B}$. We need to prove $z(t; s)$ has the same extrema as $z_n(t; s), \forall s$. It suffices to show that, for any s, $\{z_n(t)\}$ converges pointwise to $z(t)$. Because $z_n(t)$'s are energy-limited and bandlimited with the same bandwidth, the sequence $\{z_n(t)\}$ is equicontinuous. That is, there exists a δ_1 independent of n, such that $|t - t_0| < \delta_1$ implies $|z_n(t) - z_n(t_0)| < b$. Since both z and $\{z_n\}$ are bandlimited signals, then they are are continuous. Suppose first that $\{z_n(t)\}$ convergence to $z(t)$ is not pointwise. Then there exists at least one t_0, such that $|z(t_0) - z_n(t_0)| > 3b$ holds for an infinite number of n. Let $n > N(b, \delta)$, and use Lemma 1. Then for any $\delta < \delta_1$, there is a $t \neq t_0$ in $|t - t_0| < \delta$ such that

$$|z(t) - z(t_0)| \geq |z(t_0) - z_n(t_0)| - |z(t) - z_n(t)| - |z_n(t) - z_n(t_0)| > 3b - b - b = b,$$

This implies that $z(t)$ is not continuous at t_0 — a contradiction. Therefore, A_1 is closed.

A sequence $\{z_n\} \in A_2$ records only the locations of the extrema. However, all arguments in the previous proof remain valid. Hence $\{z_n\} \longrightarrow z$ in L^2 implies that $\{z_n(t)\}$ converges to $z(t)$ pointwise since they are all bandlimited. So $dz(t)/dt$ has the same zero-crossings and $z(t)$ has the same signs of the second derivatives (with respect to t) at the zero-crossings as $z_n(t)$, which indicates that A_2 is closed.

This argument can be generalized to the following. Consider a bandlimited sequence $\{z_n\} \longrightarrow z$ in L^2, and all $z_n(t)$'s have the same properties. Since $\{z_n(t)\}$ converges to $z(t)$ pointwise, $z(t)$ also holds these properties. Hence, A_3, A_4, A_5 are closed sets.

Proposition 2. $A_j, j = 0, 1, 2, 3$ are convex sets.

Proof: A_0 is a linear space, hence convex. For $A_j, j = 1, 2, 3$, all maximum and minimum locations are fixed; If z_1, $z_2 \in A_j$, and $0 \leq \lambda \leq 1$, the locations of maxima and/or minima does not change in $\lambda z_1 + (1 - \lambda)z_2$. Similarly, specific measures at these locations (e.g., the amplitude values for A_1 or the mixed partial derivative values for A_3) remain unchanged too. Furthermore, a linear combination of bandlimited signals is still bandlimited. Therefore, $\lambda z_1 + (1 - \lambda)z_2 \in A_j$, and all above sets are convex. These conditions do not hold for A_4 and A_5.

Proposition 3. The operator from $L^2(\mathcal{B})$ onto A_0 defined as $P_0 = HH^{-1}$ is a nonexpansive orthogonal projection.

Proof: We need only to show it is an orthogonal projection. The nonexpansive property automatically holds since A_0 is closed and convex.

It is easy to see that P_0 is linear and bounded. Let $y = P_0 z$, $\forall z \in L^2(\mathcal{B})$. We need to prove $< y, z - y >= 0$, $\forall z \in L^2(\mathcal{B})$. Since

$$y(t; r) = [\int_S z(t; s) *_t h(-t; s) ds] *_t h(t; r) \tag{26}$$

where $*_t$ denotes the convolution with respect to t, therefore,

$$< P_0 z, P_0 z >=< y, y >= \int_S \int_R |y(t; r)|^2 dt\ dr$$

$$= \int_S \int_R y(t; r) [\int_S z(t; s) *_t h(-t; s) ds] *_t h(t; r) dt\ dr \tag{27}$$

which in the frequency domain is equivalent to

$$< y, y >= \frac{1}{2\pi} \int_S \int_\Omega |Y(\omega; r)|^2 d\omega \; dr$$

$$= \frac{1}{2\pi} \int_S \int_\Omega \bar{Y}(\omega; r)[\int_S Z(\omega; s)\bar{H}(\omega; s)ds]H(\omega; r)d\omega \; dr$$

$$= \frac{1}{2\pi} \int_S \int_S \int_\Omega \bar{Y}(\omega; r)Z(\omega; s)\bar{H}(\omega; s)H(\omega; r)d\omega \; dr \; ds \tag{28}$$

where $\bar{X}$ denotes the complex conjugate of X. On the other hand, every point in A is a fixed point under P_0, i.e., $\forall y \in A_0$,

$$y(t; r) = [\int_S y(t; s) *_t h(-t; s)ds] *_t h(t; r), \; \forall t \in R, \; \forall r \in S. \tag{29}$$

Hence, we have

$$< P_0 z, z >=< y, z >= \int_S \int_R z(t; r)y(t; r)dt \; dr$$

$$= \int_S \int_R z(t; r)[\int_S y(t; s) *_t h(-t; s)ds] *_t h(t; r)dt \; dr$$

$$= \frac{1}{2\pi} \int_S \int_\Omega Z(\omega; r)[\int_S \bar{Y}(\omega; s)H(\omega; s)ds]\bar{H}(\omega; r)d\omega \; dr.$$

Exchanging r with s and rearranging the above equation, we obtain

$$< y, z >= \frac{1}{2\pi} \int_S \int_S \int_\Omega \bar{Y}(\omega; r)Z(\omega; s)\bar{H}(\omega; s)H(\omega; r)d\omega \; dr \; ds \tag{30}$$

which is (28).

Appendix II

In this appendix, we verify analytically the statement that V_5 reflects the amplitude spectrum of the input signal. Consider the expression for V_5 from (20) earlier:

$$V_5(t;s) = \frac{1}{T}\int_{t-T}^{t} \max(\partial_s\partial_\tau y_1(\tau;s)\cdot\delta(\partial_\tau y_1(\tau;s)),0)d\tau = \frac{1}{T}\sum_i \max(\partial_s\partial_t y_1(t_{i;s};s),0), \quad (31)$$

where $t_{i;s}$ are the extrema of $y_1(t;s)$ during $[t-T,t)$. Replace

$$\partial_s\partial_t y_1(t_{i;s};s) = \frac{1}{2\pi}\int X(\omega)\partial_s H(\omega;s)\omega e^{j\omega t_{i;s}}d\omega \quad (32)$$

in the above equation. Let $H(\omega;s) = |H(\omega;s)|e^{j\theta_H}$. Then it can be shown that the differential filter can be expressed as

$$\partial_s H(\omega;s) = |\partial_s H(\omega;s)|e^{j(\theta_H(\omega;s)+\theta_\delta(\omega;s))} \quad (33)$$

where

$$\theta_\delta(\omega;s) = \arctan(\frac{|H(\omega;s)|\partial_s\theta_H(\omega;s)}{|\partial_s H(\omega;s)|}). \quad (34)$$

Let the spatial derivative of the filter at s, $|\partial_s H(\omega;s)|$, be narrow and centered at ω_s so that

$$|\partial_s H(\omega;s)| \simeq \delta(\omega-\omega_s)+\delta(\omega+\omega_s), \quad (35)$$

then

$$\partial_s\partial_t y_1(t_{i;s};s) = \frac{1}{2\pi}|X(\omega_s)\omega_s|(e^{j(\theta(\omega_s)+\theta_\delta(\omega_s)+\omega_s t_{i;s})} + e^{j(\theta(-\omega_s)+\theta_\delta(-\omega_s)-\omega_s t_{i;s})})$$

$$= \frac{1}{\pi}|X(\omega_s)\omega_s|\cos(\theta(\omega_s)+\theta_\delta(\omega_s)+\omega_s t_{i;s}), \quad (36)$$

where $\theta(\omega_s)$ is the sum of phases of $X(\omega)$ and $H(\omega;s)$ evaluated at ω_s. $V_5(t;s)$ becomes

$$V_5(t;s) = \frac{1}{\pi T}|X(\omega_s)\omega_s|\sum_i \max(\cos(\theta(\omega_s)+\theta_\delta(\omega_s)+\omega_s t_{i;s}),0). \quad (37)$$

When a single tone at frequency $f = \omega/2\pi$ *dominates* the s channel, its extrema are uniformly distributed on the t axis, and hence the interval between a pair of maximum and minimum is $\lambda = 1/f$ (see also comment at the end of this appendix). Consequently, $t_{i;s} = t_{0;s} + i\lambda$ where $t_{0;s}$ is the first extremum of $y_1(t; s)$ starting at $t - T$, and the number of $t_{i;s}$ in $[t - T, t)$ is equal to $T|\omega|/\pi$. Therefore, $\cos(\theta(\omega_s) + \theta_\delta(\omega_s) + \omega_s t_{i;s}) = \cos(\theta(\omega_s) + \omega_s t_{0;s} + \theta_\delta(\omega_s) + i\pi\omega_s/\omega)$. Note now that the $\omega_s t_{0;s}$ is an effective lead that is purely due to the phase of the signal and filter, i.e., $\omega_s t_{0;s} = -\theta(\omega_s)$. $V_5(t; s)$ therefore is simplified to

$$V_5(t; s) = \frac{|\omega_s|}{\pi T}|X(\omega_s)| \sum_i \max(\cos(\theta_\delta(\omega_s) + i\pi\omega_s/\omega), 0). \tag{38}$$

Let the differential phase $\theta_\delta(\omega)$ of the filters vanish at $\omega = \omega_s$. We shall later support this assertion. Consider now the case where a channel at s is driven by components at $\omega = \omega_s$. Then the extremum points $t_{i;s}$ are mainly due to frequency ω_s, and the equation above simplifies to

$$V_5(t; s) = \frac{|\omega_s|}{\pi T}|X(\omega_s)| \sum_i \max(\cos(i\pi), 0) = 2(\frac{\omega_s}{2\pi})^2|X(\omega_s)|, \tag{39}$$

i.e., $V_5(t; s)$ is a representation of the amplitude spectrum at s, and in this case $V_5(t; s)$ reaches its maximum. Suppression appears in nearby channels because they are driven by frequencies other than their own center frequency, i.e., it is effectively like driving the s channel by $\omega \neq \omega_s$. The $V_5(t; s)$ becomes instead:

$$V_5(t; s) = \frac{|\omega_s|}{\pi T}|X(\omega_s)| \sum_i \max(\cos(i\pi\omega_s/\omega), 0), \tag{40}$$

which for $\omega_s/\omega \neq 1$ is less than the earlier estimate of $V_5(t; s)$. That is, a channel at s will transmit an *attenuated* estimate of the spectral energy through its differential filter.

Note that $V_5(t;s)$ is a dilated estimate of the spectrum $X(\omega_s)$. This is because from (35), we have

$$|\partial_s H(\omega; s+\sigma)| = \delta(a^{-\sigma}\omega - \omega_s) = a^{\sigma}\delta(\omega - a^{\sigma}\omega_s), \omega \geq 0,$$

which results in

$$V_5(t; s+\sigma) = 2(\frac{a^{\sigma}\omega_s}{2\pi})^2 |X(a^{\sigma}\omega_s)|.$$

This implies

$$V_5(t;s) = 2(\frac{a^{s}\omega_0}{2\pi})^2 |X(a^{s}\omega_0)|. \tag{41}$$

Hence, the normalized $V_5(t;s)/a^{2s}$ represents the amplitude spectrum of the input signal $x(t)$ in a dilation fashion $|X(a^s\omega_0)|$.

The differential phase of the filters, $\theta_\delta(\omega)$

Here we argue that the differential phase of the filters is very small, or in effect vanishes at $\omega = \omega_s$, i.e., $\theta_\delta(\omega_s) = 0$. We assume that the filter impulse response $h(t;s)$ is appropriately discretized in time t, denoted by $h(n;s)$, and its z-transform can be expressed as

$$\hat{H}(z;s) = \sum_{n=0}^{\infty} h(n;s)z^{-n} = \exp\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|\frac{e^{-j\omega}+z^{-1}}{e^{-j\omega}-z^{-1}}d\omega\}.$$

Let $G(z;s) = \ln \hat{H}(z;s) = \sum_{n=0}^{\infty} g(n;s)z^{-n}$. Since $h(n;s)$ is a minimum phase function, $G(z)$ is analytical outside unit circle. Thus, we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|\frac{e^{-j\omega}+z^{-1}}{e^{-j\omega}-z^{-1}}d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|d\omega + \sum_{n=1}^{\infty}\frac{1}{\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|e^{jn\omega}d\omega z^{-n}$$

where the fact that $\frac{e^{-j\omega}+z^{-1}}{e^{-j\omega}-z^{-1}} = 1 + 2(e^{j\omega}z^{-1} + e^{j2\omega}z^{-2} + \cdots)$ is used. Therefore,

$$g(0;s) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|d\omega;$$

$$g(n;s) = \frac{1}{\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|e^{jn\omega}d\omega = \frac{1}{\pi}\int_{-\pi}^{\pi} \ln|H(\omega;s)|\cos(n\omega)d\omega, \quad n \geq 1.$$

The last equality is due to the fact that $h(n;s)$ is real. Thus, from

$$H(\omega;s) = \hat{H}(z;s)|_{z=e^{j\omega}} = e^{G(z;s)}|_{z=e^{j\omega}}$$

we can express the phase response of the filter as their Hilbert transform relation:

$$\theta_H(\omega;s) = Im[G(e^{j\omega};s)] = -\sum_{n=0}^{\infty} g(n;s)\sin(n\omega)$$

$$= \frac{-1}{\pi}\int_{-\pi}^{\pi} \ln|H(\lambda;s)|(\sum_{n=0}^{\infty}\cos(n\lambda)\sin(n\omega))d\lambda. \tag{42}$$

Hence, its partial derivative with respect to s is

$$\partial_s\theta_H(\omega;s) = \frac{-1}{\pi}\int_{-\pi}^{\pi} \frac{\partial_s|H(\omega;s)|}{|H(\lambda;s)|}(\sum_{n=0}^{\infty}\cos(n\lambda)\sin(n\omega))d\lambda < \infty, \forall\omega.$$

However, from (35) we know that the dominator in (34) is very large at $\omega = \omega_s$. Therefore, the ratio in Eq.34 is very small at ω_s.

Comment regarding the use of the term "dominance"

The term *dominance* is used here in a narrow technical sense exemplified by the following phenomena. When two tones of significantly unequal amplitudes are added together, the intervals of successive zero-crossings tend to cluster disproportionately around the interval of the larger tone. For instance, one can easily demonstrate that adding a very small amplitude interfering tone to a sinewave causes only a jitter in the zero-crossing intervals, and that a minimum threshold amplitude is needed before any new intervals reflecting the frequency of the interfering tone can appear. This phenomena has been described previously in the auditory experimental literature as *synchrony capture* or *synchrony suppression* [16, 27,43].

References:

[1] Kai-Fu Lee, in *Automatic Speech Recognition: The Development of the SPHINX*, Kluwer Academic, Boston, MA, 1989.

[2] H. Hermansky, K. Tsuga, S. Makino & H. Wakita, "Perceptually based processing in automatic speech recognition," Proc. IEEE-ICASSP (37.5), Tokyo, 1986.

[3] A. Bladon, "Using auditory models for speaker normalization in speech recognition," *Proceedings of the Symposium on Speech Recognition*, Montreal (1986).

[4] R. Lyon, "A computational model of binaural localization and separation," *IEEE Proc. ICASSP* , Boston, Mass. (1983).

[5] J. Cohen, "Applications of an auditory model to speech recognition," *J. Acoust. Soc. Am.* 85 (1989), 2623–2629.

[6] S. Seneff, "A joint synchrony/mean-rate model of auditory processing," *J. Phonetics* 16(1) (1988), 55–76.

[7] O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," *J. Phonetics* 16(1) (1988), 109–124.

[8] W. Byrne, J. Robinson & S. Shamma, "The Auditory Processing and Recognition of Speech," *Proc. DARPA Workshop on Speech Recognition* (Nov. 1989).

[9] B. Juang, L. Rabiner & J. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust. Speech Sig. Proc.* 35 (1987), 947–954.

[10] S. Mallat, "Multifrequency Channel Decompositions of Images and Wavelet Models," *IEEE Trans. Acoust. Speech Sig. Proc.* 37(12) (1989), 2091–2110.

[11] S. Mallat & S. Zhong, "Complete Signal Representations with Multiscale Edges," *Robotics Research Technical Report* No. 483 (1989).

[12] B. Escudie & B. Torresani, "Wavelet Analysis of Asymptotic Signals," *Tech. Report of the Centre de Physique Theorique at Marseille* (1990).

[13] J. Lienard & C. d'Alessandro, "Wavelets and Ganular Analysis of Speech," in *Wavelets , Time-Frequency Methods and Phase Space*, J. Combes, A. Grossman & Ph. Tchamitchian, eds., Springer Verlag, 1989, 158–163.

[14] R. Hukin & R. Damper, "Testing an Auditory Model by Resynthesis," *Proc. ESCA-Eurospeech'89*, Paris (1989).

[15] S. A. Shamma, R. Chadwick, J. Wilbur, J. Rinzel & K. Moorish, "A biophysical model of cochlear processing: intensity dependence of pure tone responses," *J. Acoust. Soc. Am.* 80 (1986), 133–145.

[16] L. Deng, C. D. Geisler & S. Greenberg, "A composite model of the auditory periphery for the processing of speech," *J. of Phonetics* 16(1) (1988), 93.

[17] S. Greenberg, "Acoustic transduction in the auditory periphery," *J. Phonetics* 16(1) (1988), 3–18.

[18] D. J. Hermes & J. C. van Gestel, "The frequency scale of speech intonation," *J. Acoust. Soc. Am.* 90(1) (1991), 97–102.

[19] L. A. Westerman & R. L. Smith, "Rapid and short term adaptation in auditory nerve responses," *Hear. Res.* 15 (1984), 249–260.

[20] W. M. Siebert, "Frequency discrimination in the auditory system: place or periodicity mechanisms?," *Proc. IEEE* 58 (1970), 723–730.

[21] S. Shamma, "Spatial and Temporal Processing in Central Auditory Networks," in *Methods in Neuronal Modelling*, C. Koch & I. Segev, eds., MIT Press, Cambridge, 1989.

[22] M. B. Sachs & E. D. Young, "Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate," *J. Acoust. Soc. Am.* 66 (1979), 470–479.

[23] H. K. Hartline, *Studies on Excitation and Inhibition in the Retina*, Rockefeller University Press, New York, 1974.

[24] S. A. Shamma, "Speech processing in the auditory system. II: Lateral inhibition and the processing of speech evoked activity in the auditory-nerve ," *J. Acoust. Soc. Am.* 78 (1985), 1622–1632.

[25] I. Morishita & A. Yajima, *Kybern.* 11 (1972 Analysis and simulation of networks of mutually inhibiting neurons), 154–165.

[26] C. Schreiner & J. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between fields. ," *Hearing Res.* 32 (1988), 49–64 .

[27] S. A. Shamma & K. Morrish, "Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea," *J. Acoust. Soc. Am.* 81(5) (1987), 1486–1498.

[28] M. J. Lighthill, in *Fourier Analysis and Generalized Functions*, Cambridge University Press, 1973.

[29] D. C. Youla & H. Webb, "Image Restoration by the Method of Convex Projections: Part 1 - Theory," *IEEE Trans. Med. Imaging* MI-1 (2) (1982), 81–94.

[30] M. I. Sezan & H. Stark, "Image Restoration by the Method of Convex Projections: Part 2 - Applications and Numerical Results," *IEEE Trans. Med. Imaging* MI-1 (2) (1982), 95–101.

[31] J. L. Yen, "On Nonuniform Sampling of Bandwidth-Limited Signals," *IRE Trans. On Circuit Theory* (Dec. 1956).

[32] B. Logan, "Information in the zero-crossings of band pass signals," *Bell Systems Tech. Journ.* 56 (1977), 510.

[33] B. J. Moore, in *Psychology of Hearing*, Academic Press, 1982.

[34] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," *Proc. ICASSP* 82 (1982), 1278.

[35] P. Assmann & Q. Summerfield, "Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency ," *J. Acoust. Soc. Am.* 85 (1988), 327 .

[36] S. A. Shamma, "Hearing as Seeing: The Role of Space and Time in Auditory Processing," in *Proceedings of Symposium on Analysis and Modelling of Neural Systems*, Berkeley, CA, 1990.

[37] J. B. Allen & L. R. Rabiner, "A unified theory of short-time spectrum analysis and synthesis," *Proc. IEEE* 65 (1977), 1558–1564.

[38] J. Koenderink, "The structure of images," *Biological Cybernetics* (1984).

[39] S. A. Shamma, "Speech processing in the auditory system. I: Representation of speech sounds in the responses of the auditory-nerve," *J. Acoust. Soc. Am.* 78 (1985), 1612–1621.

[40] W. B. Davenport & W. L. Root, in *An Introduction to the Theory of Random Signals and Noise*, IEEE Press, 1987, 307.

[41] S. Shamma, J. Fleshman & P. Wiser, "Receptive Field Organization in Primary Auditory Cortex: Spectral Orientation Columns ," *Systems Research Center Tech. Report (TR 90-46)* (1990).

[42] S. A. Shamma & G. Chettiar, "A Functional Model of Primary Auditory Cortex: Spectral Orientation Columns," *System Research Center Tech. Rep. (TR 90-47)* (1990).

[43] E. Javel, J. McGee, E. J. Walsh & G. R. Farley, "Studies of synchrony suppression in normal and hearing-impaired cats," in *Mechanisms of Hearing*, L. Aitkin & W. Webster, eds., Monash University Press, Clayton, 1983.

Figure Legends

Figure 1

Early stages of processing in the auditory system.

A: Block diagram of the three basic stages in auditory processing.

B: Quasi-anatomical sketches of the auditory stages.

C: Examples of the response patterns at various stages of processing.

D: Mathematical models of the different auditory stages. See section II in the text for details of derivations.

Figure 2

Schematics of the abstraction of the auditory outputs into the data vectors $V_j, j = 1, \cdots, 5$ at different stages of processing. $y_1(t; s)$ responses in three channels are shown (at $k - 1$, k, and $k + 1$). Dashed arrows symbolize samples of the $y_1(t; s)$ waveforms. Sample values are reflected by the length of the arrow; Sample locations are marked by the **x**'s. In the first stage, V_1 samples the amplitude and locations of the extrema of y_1. In V_2, only the locations of the $y_1(t; s)$ extrema are retained. In V_3, sample values at the extrema reflect the mixed-derivative values of $y_1(t; s)$. In V_4, only the positively-valued samples in V_3 are retained. V_5 contains an average value of the samples at each channel.

Figure 3

Top: Amplitude profiles of the cochlear filters on a linear frequency axis. Only 32 filters are shown. The ordinate is in linear units.

Bottom: Amplitude profiles of the cochlear filters on a logarithmic frequency axis. Only

21 filters are shown. The ordinate is in linear units.

Figure 4

Reconstructions of a naturally spoken vowel /iy/. For each, the number of iterations and resulting spectral signal-to-noise ratio are indicated.

A: Waveforms of the original vowel (top) and of reconstructions from data vectors $V_j, j = 1,3,4,5$. The ordinate is in linear units.

B: Corresponding Fourier spectra of the original and reconstructed waveforms. The ordinate is in logarithmic units (dB's).

Figure 5

Reconstructions of a naturally spoken vowel /ae/. Details are as in Fig.4 legend.

Figure 6

Comparing the dilated Fourier spectra of the vowels /iy/ and /ae/ against their auditory representations in $\bar{V}_5(\cdot;s) = V_5(t;s)/a^{2s}$. Both the frequency axis and the *scale* axis are indicated on the abscissa. The ordinate is in logarithmic units.

Figure 7

Reconstructions of the acoustic spectrum of the vowel /iy/ (*top trace*) from data vectors V_2 (*middle trace*) and $V_{2'}$ (*bottom trace*). The ordinate is in logarithmic units (dB's).

Figure 8

Fig 8a: Schematic plot of cochlear filters with highly asymmetric shapes. Top plot illustrates the filters at locations $s = 31$ and $s = 15$ which have center frequencies at

approximately 1 and 2.5 kHz. Bottom trace shows the differential filters associated with each of the above channels. Plots have been locally smoothed by a three-point triangular window. The ordinate is in linear units.

Fig.8b: Noise suppression in the auditory representation V_5. *Top trace* represents the amplitude spectrum of a two-tone stimulus in broadband noise. *Bottom trace* is its reconstruction from V_5, showing side-band suppression. All ordinates are in logarithmic units (dB's).

Figure 9

Enhancement of spectral peaks in the auditory representation V_5. In each box, the original spectrum of the indicated vowel (*top trace*) is juxtaposed against reconstructions from V_5 at two different numbers of iterations (*bottom traces*). Note the enhanced representation of the harmonic peaks in the reconstructed spectra. The ordinates are in logarithmic units (dB's). The patterns are shifted upwards relative to each other for illustrative purposes.

Fig. 1

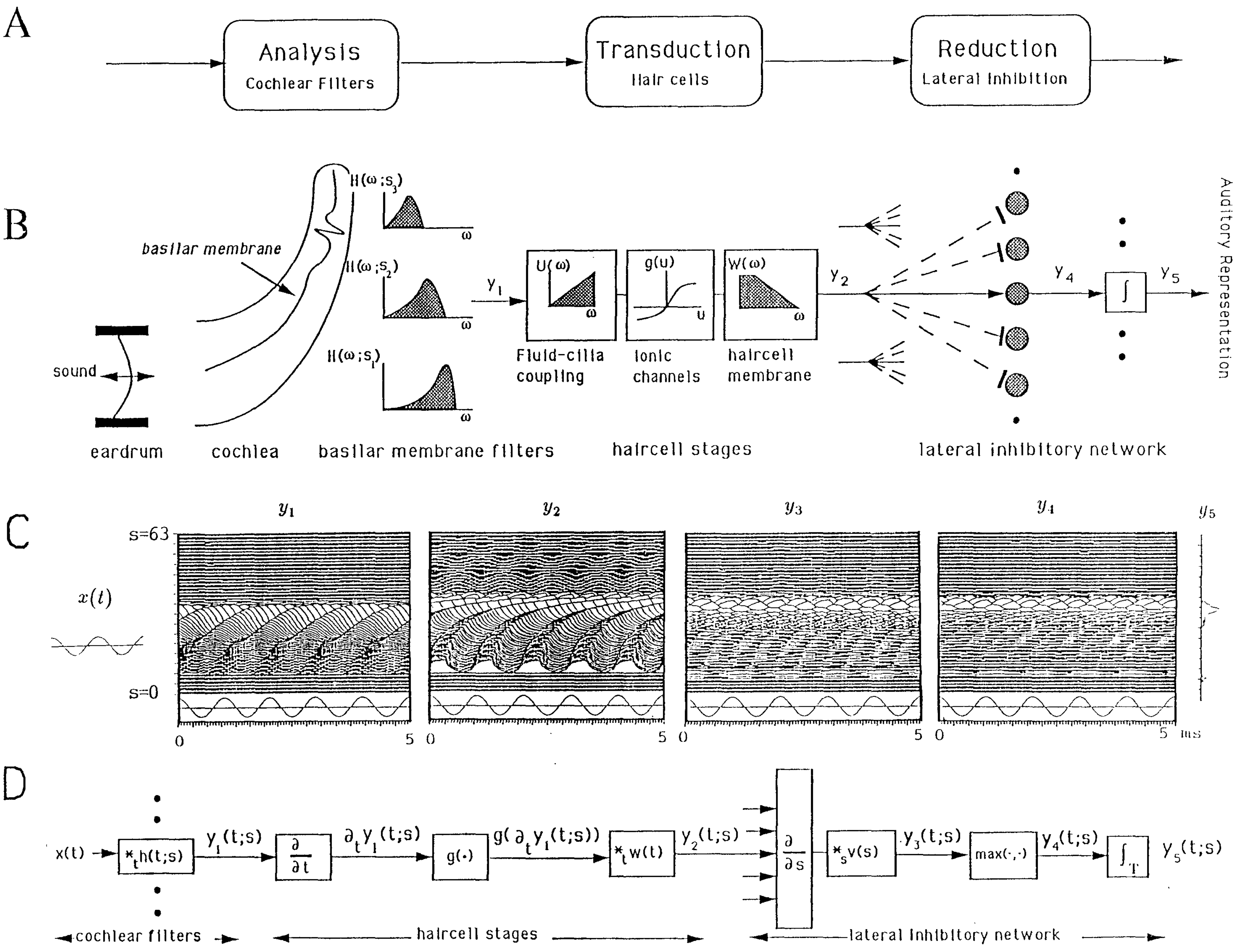

A
Analysis
Cochlear Filters
Transduction
Hair cells
Reduction
Lateral Inhibition
B
basilar membrane
H(ω;s3)
H(ω;s2)
H(ω;s1)
ω
sound
y1
U(ω)
g(u)
W(ω)
u
Fluid-cilia coupling
Ionic channels
haircell membrane
y2
y4
y5
∫
Auditory Representation
eardrum
cochlea
basilar membrane filters
haircell stages
lateral inhibitory network
C
y1
y2
y3
y4
y5
s=63
s=0
x(t)
0
5
ms
D
x(t)
*t h(t;s)
y1(t;s)
∂/∂t
∂t y1(t;s)
g(·)
g(∂t y1(t;s))
*t w(t)
y2(t;s)
∂/∂s
*s v(s)
y3(t;s)
max(·,·)
y4(t;s)
∫T
y5(t;s)
cochlear filters
haircell stages
lateral inhibitory network

Fig. 1

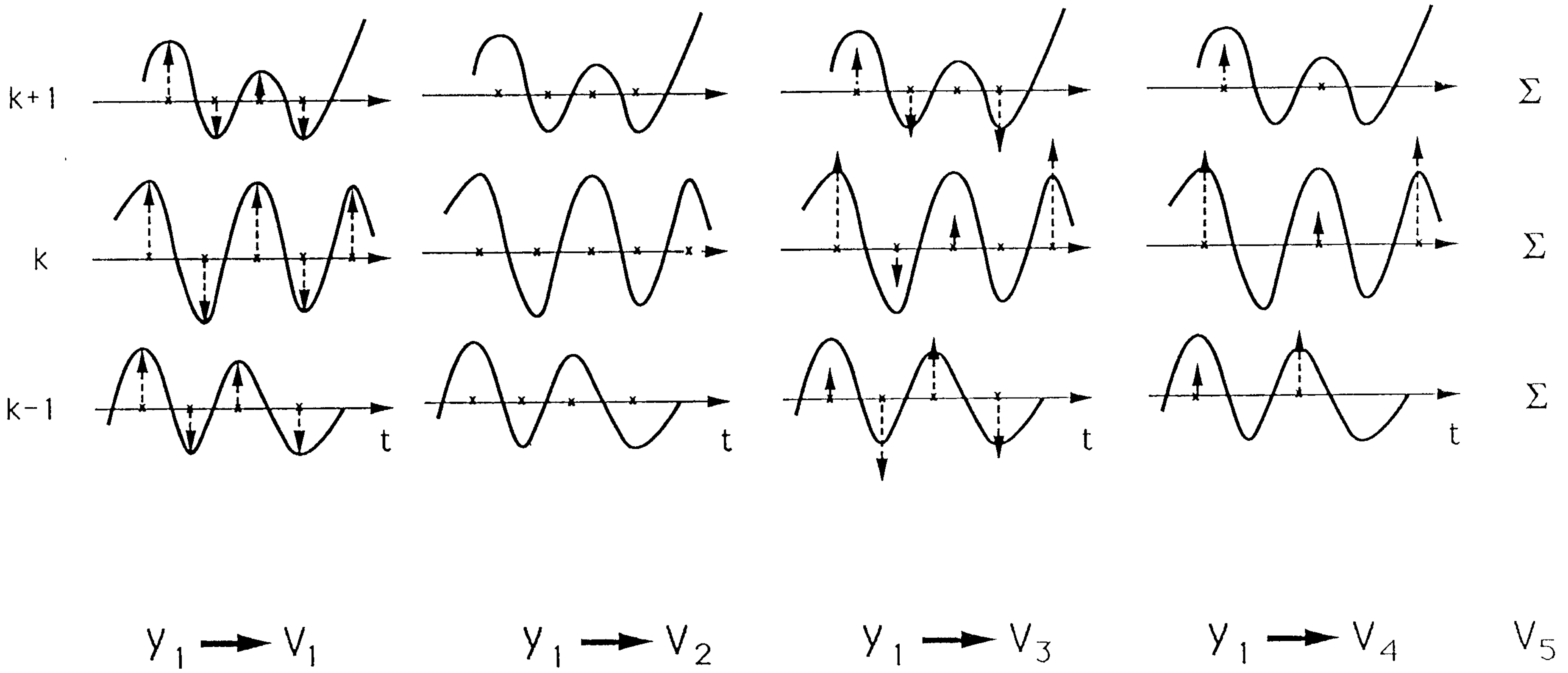
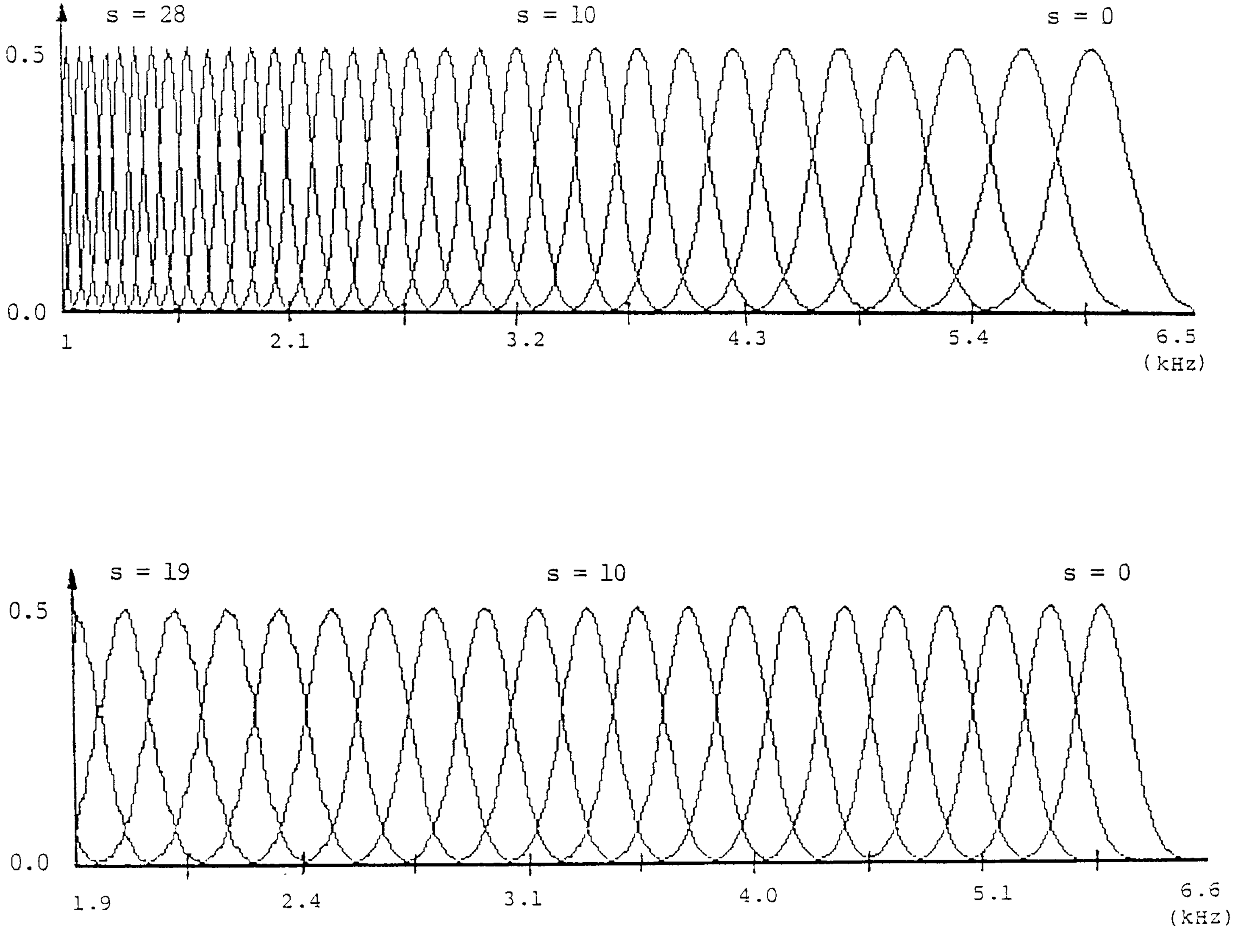


Fig. 2

Fig. 3



s = 28
s = 10
s = 0
0.5
0.0
1
2.1
3.2
4.3
5.4
6.5
(kHz)
s = 19
s = 10
s = 0
0.5
0.0
1.9
2.4
3.1
4.0
5.1
6.6
(kHz)

Fig. 4A

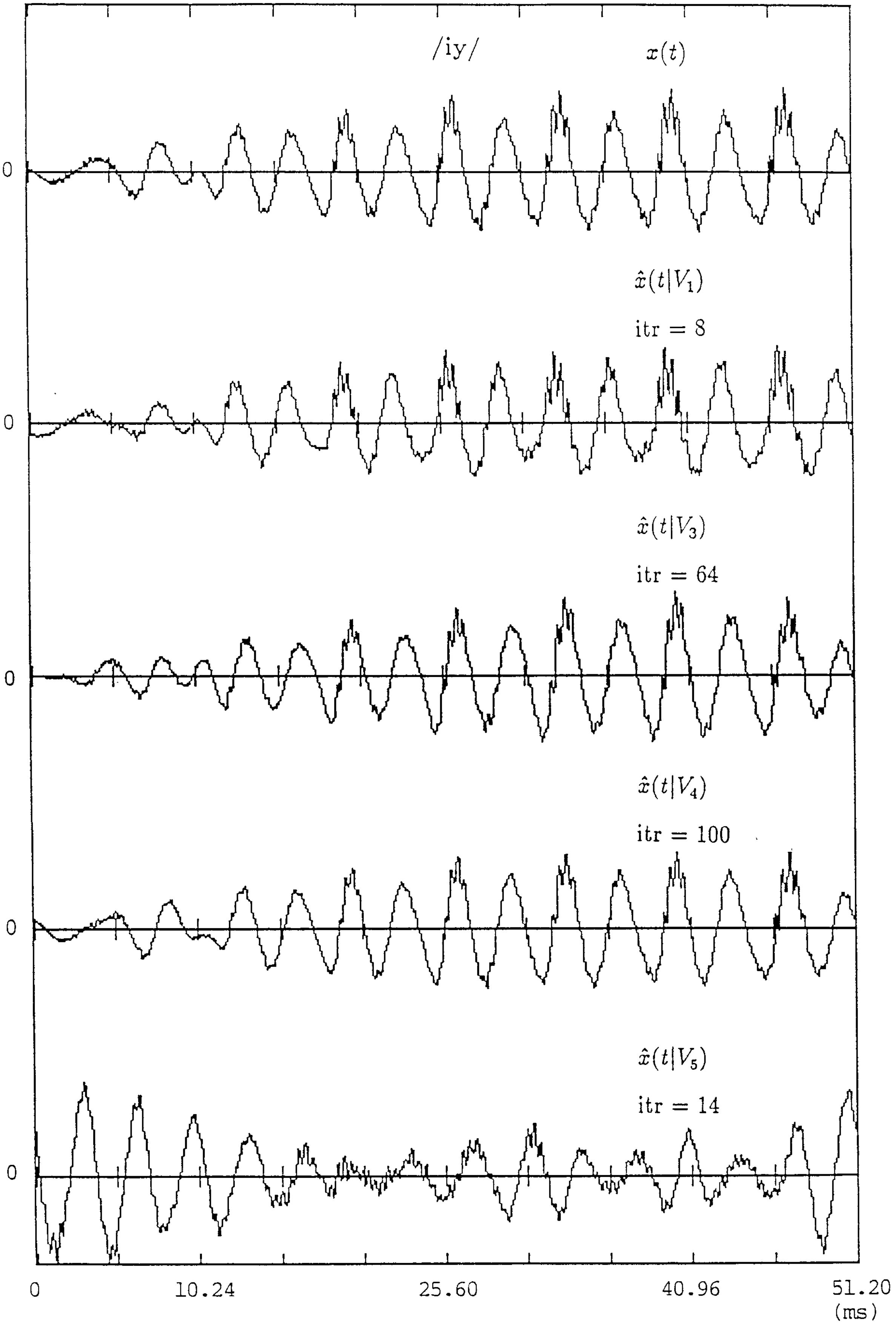
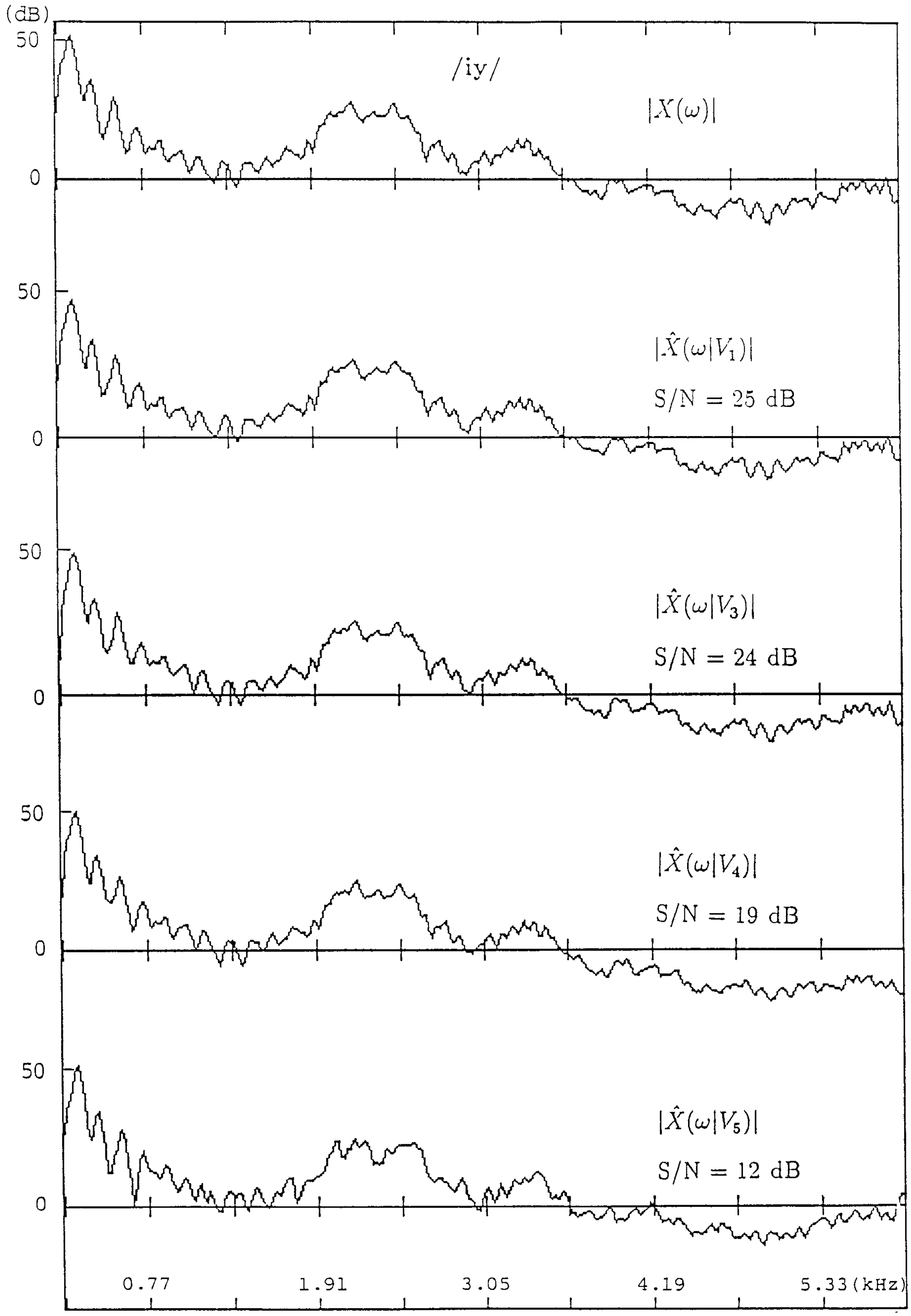

/iy/
$x(t)$
0
$\hat{x}(t|V_1)$
itr = 8
0
$\hat{x}(t|V_3)$
itr = 64
0
$\hat{x}(t|V_4)$
itr = 100
0
$\hat{x}(t|V_5)$
itr = 14
0
0
10.24
25.60
40.96
51.20
(ms)

Fig.4B



(dB)
50
0
/iy/
$|X(\omega)|$
50
0
$|\hat{X}(\omega|V_1)|$
S/N = 25 dB
50
0
$|\hat{X}(\omega|V_3)|$
S/N = 24 dB
50
0
$|\hat{X}(\omega|V_4)|$
S/N = 19 dB
50
0
$|\hat{X}(\omega|V_5)|$
S/N = 12 dB
0.77
1.91
3.05
4.19
5.33(kHz)

Fig. 5A

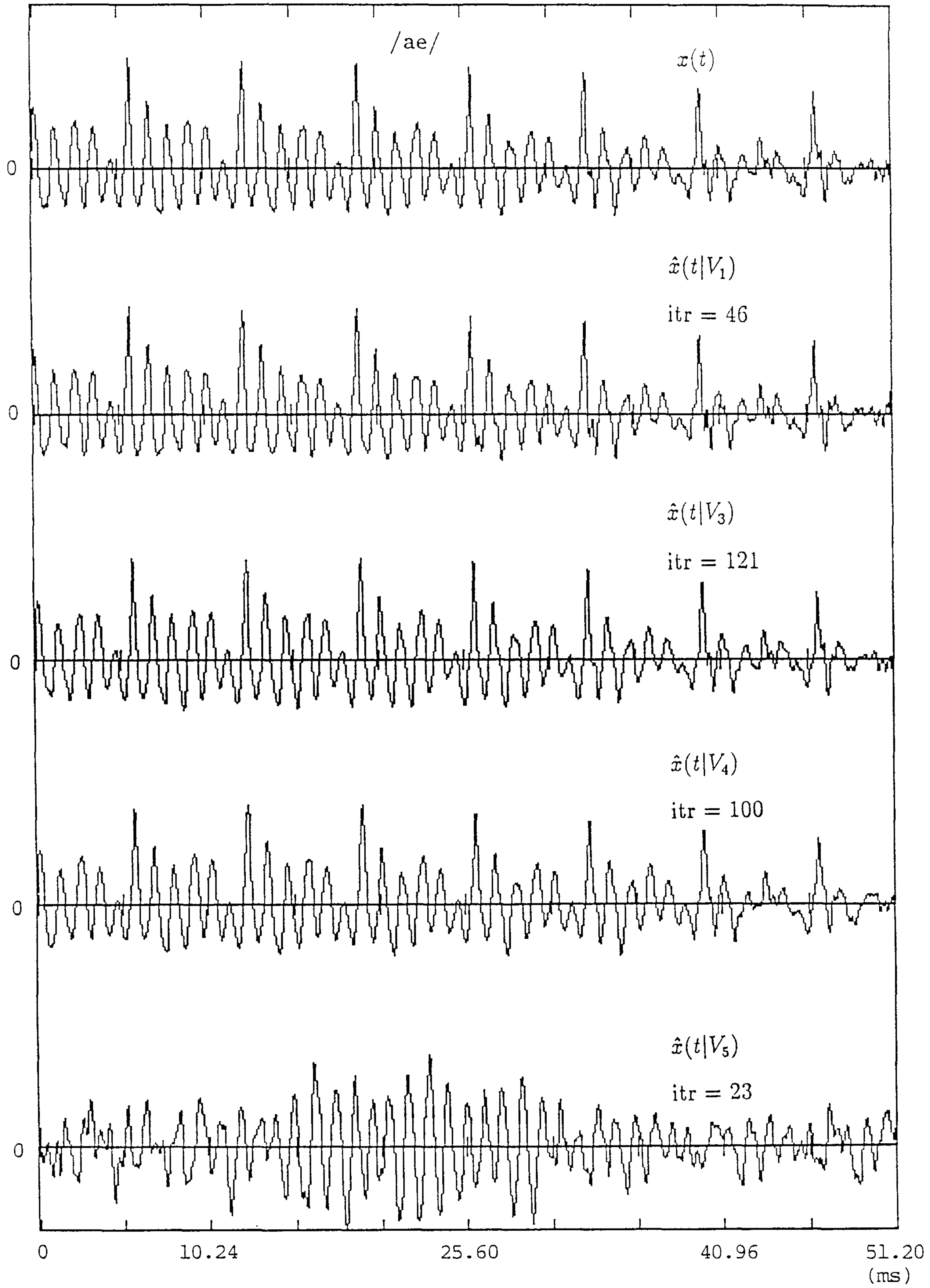

/ae/
$x(t)$
0
$\hat{x}(t|V_1)$
itr = 46
0
$\hat{x}(t|V_3)$
itr = 121
0
$\hat{x}(t|V_4)$
itr = 100
0
$\hat{x}(t|V_5)$
itr = 23
0
0
10.24
25.60
40.96
51.20
(ms)

Fig. 5B

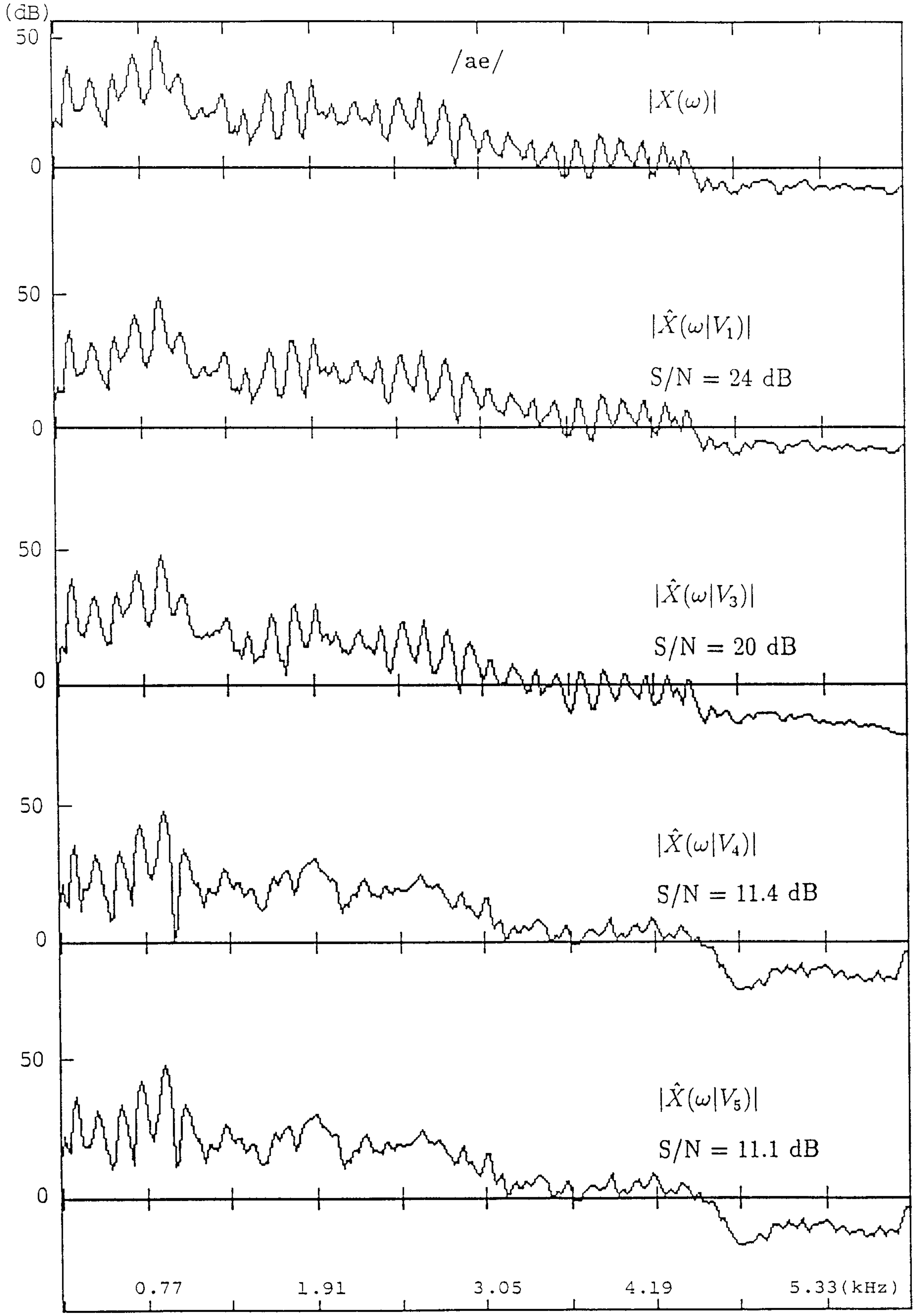
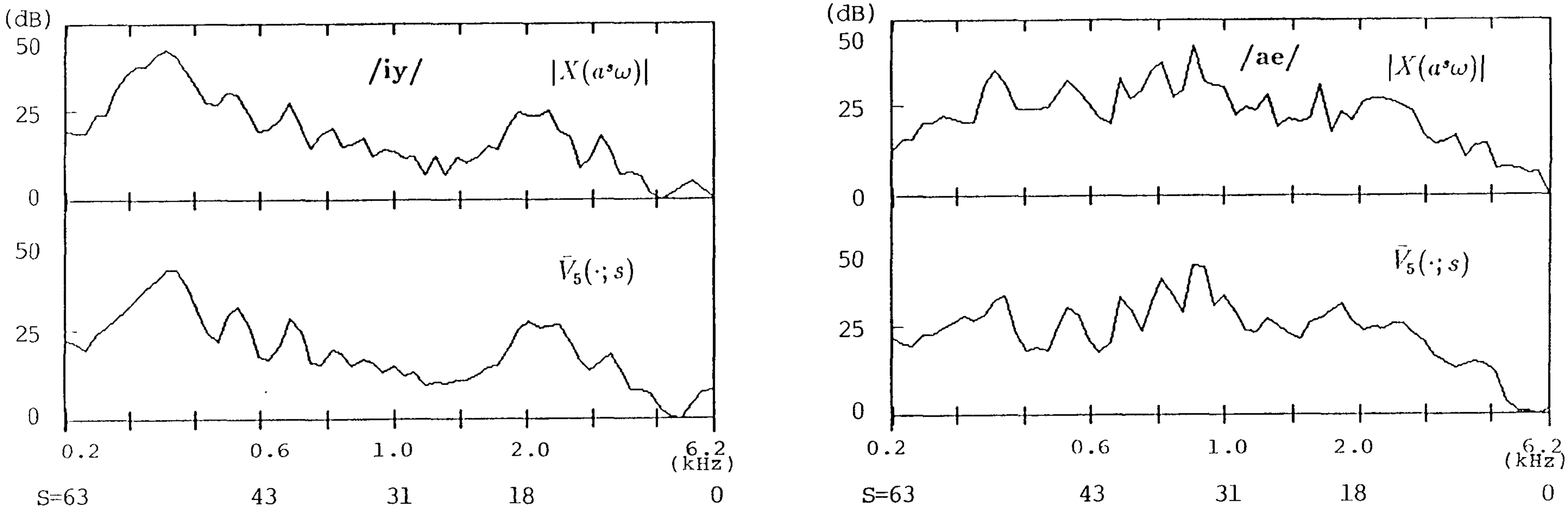

(dB)
50
0
/ae/
$|X(\omega)|$
50
0
$|\hat{X}(\omega|V_1)|$
S/N = 24 dB
50
0
$|\hat{X}(\omega|V_3)|$
S/N = 20 dB
50
0
$|\hat{X}(\omega|V_4)|$
S/N = 11.4 dB
50
0
$|\hat{X}(\omega|V_5)|$
S/N = 11.1 dB
0.77
1.91
3.05
4.19
5.33(kHz)

Fig.6



(dB)
50
25
0
50
25
0
/iy/
$|X(a^s\omega)|$
$\bar{V}_5(\cdot;s)$
0.2
0.6
1.0
2.0
6.2
(kHz)
S=63
43
31
18
0
(dB)
50
25
0
50
25
0
/ae/
$|X(a^s\omega)|$
$\bar{V}_5(\cdot;s)$
0.2
0.6
1.0
2.0
6.2
(kHz)
S=63
43
31
18
0

Fig. 6

Fig. 7

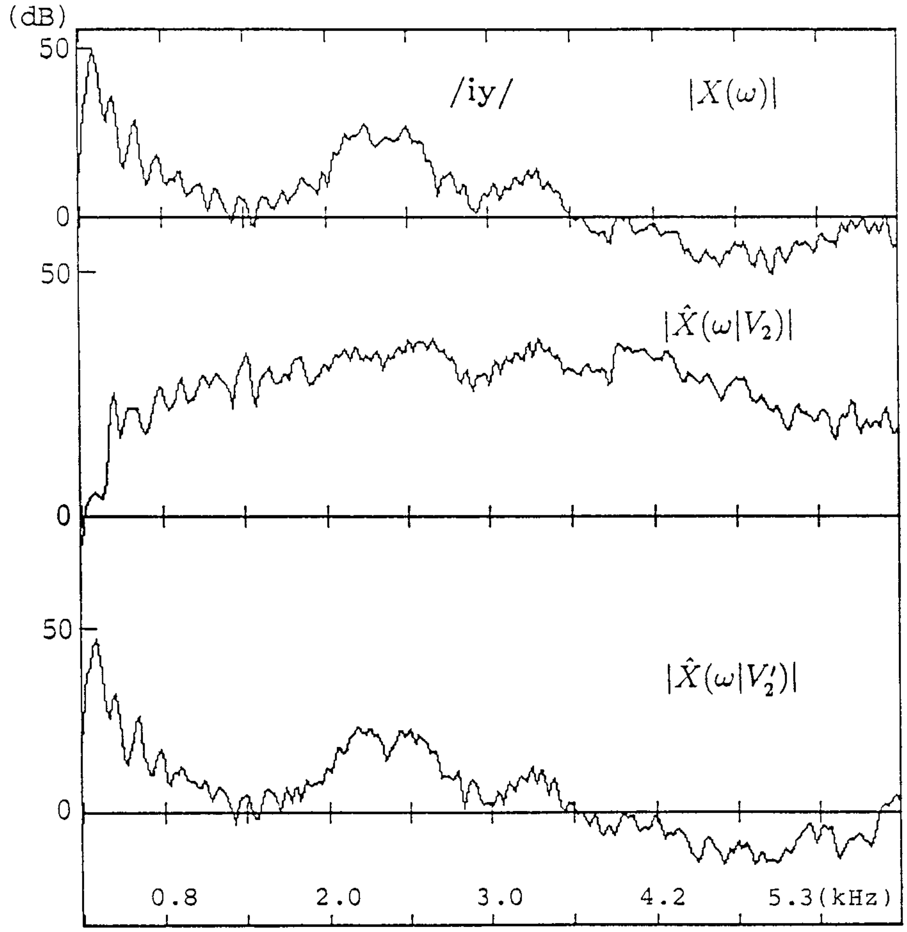

(dB)
50
0
/iy/
$|X(\omega)|$
50
0
$|\hat{X}(\omega|V_2)|$
50
0
$|\hat{X}(\omega|V_2')|$
0.8
2.0
3.0
4.2
5.3(kHz)

Fig. 8a

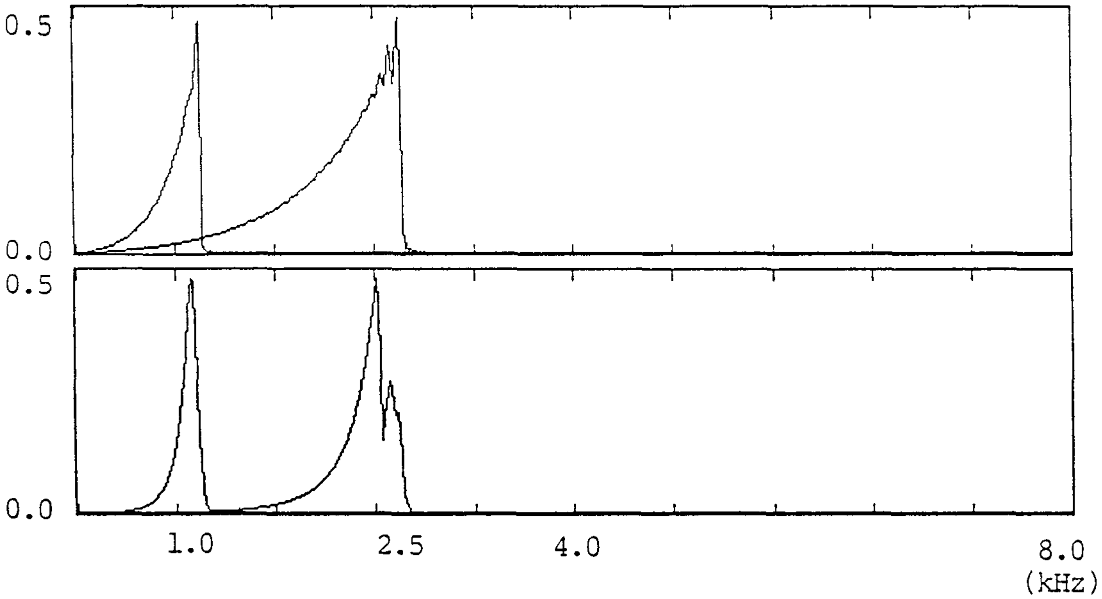


Fig. 8a

Fig. 8b

Fig. 8b

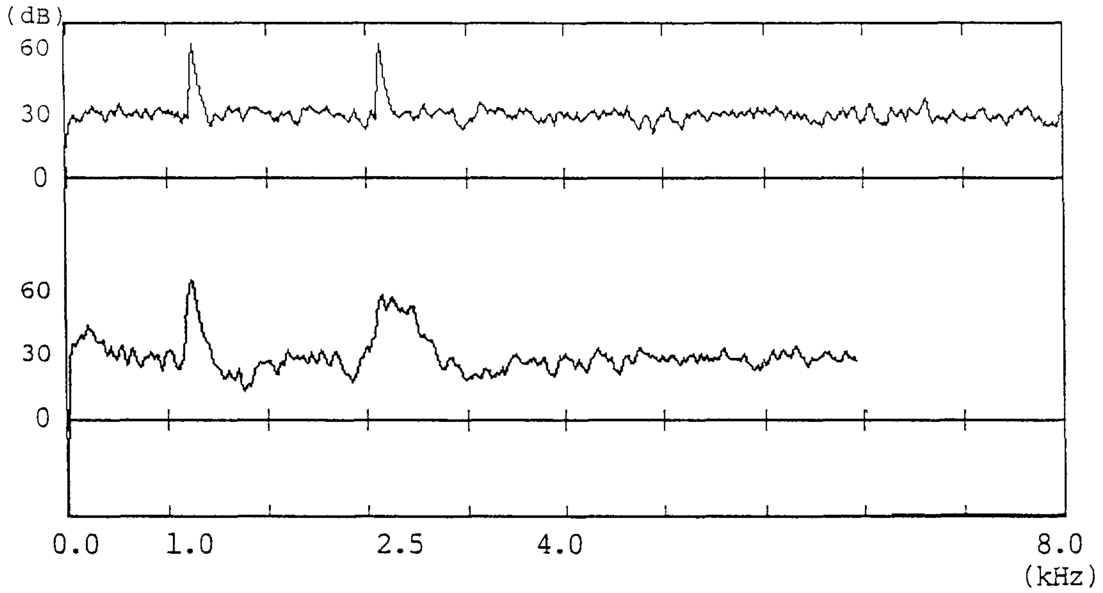
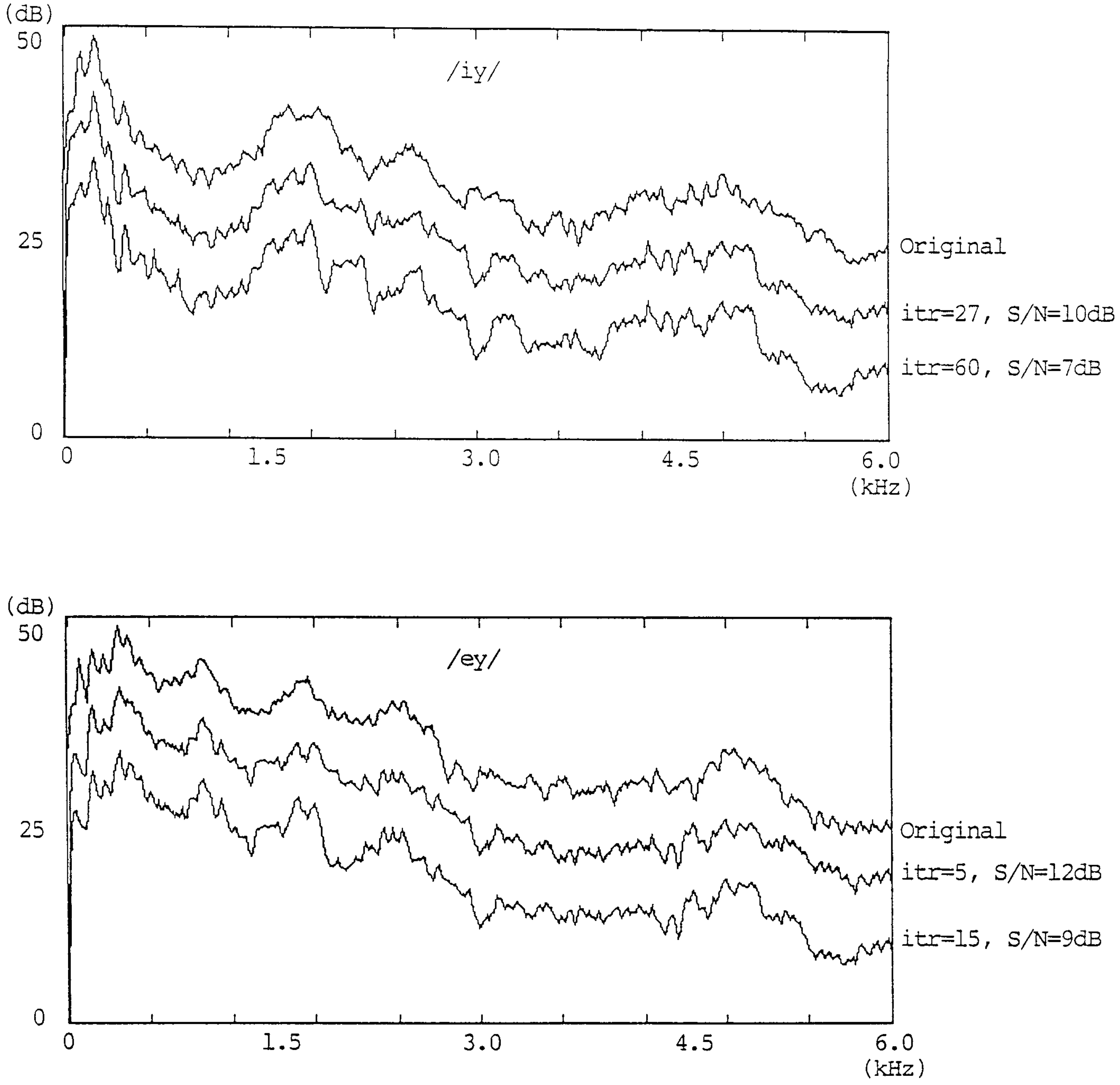


Fig. 9


(dB)
50
25
0
/iy/
Original
itr=27, S/N=10dB
itr=60, S/N=7dB
0
1.5
3.0
4.5
6.0
(kHz)
(dB)
50
25
0
/ey/
Original
itr=5, S/N=12dB
itr=15, S/N=9dB
0
1.5
3.0
4.5
6.0
(kHz)