Parallel Auditory Filtering by Sustained and Transient Channels Separates Coarticulated Vowels and Consonants

Michael A. Cohen, *Associate Member, IEEE,* and Stephen Grossberg, *Member, IEEE*

***Abstract*—A neural model of peripheral auditory processing is described and used to separate features of coarticulated vowels and consonants. After preprocessing of speech via a filterbank, the model splits into two parallel channels, a sustained channel and a transient channel. The sustained channel is sensitive to relatively stable parts of the speech waveform, notably synchronous properties of the vocalic portion of the stimulus. It extends the dynamic range of eighth nerve filters using coincidence detectors that combine operations of raising to a power, rectification, delay, multiplication, time averaging, and preemphasis. The transient channel is sensitive to critical features at the onsets and offsets of speech segments. It is built up from fast excitatory neurons that are modulated by slow inhibitory interneurons. These units are combined over high-frequency and low-frequency ranges using operations of rectification, normalization, multiplicative gating, and opponent processing. Detectors sensitive to frication and to onset or offset of stop consonants and vowels are described. Model properties are characterized by mathematical analysis and computer simulations. Neural analogs of model cells in the cochlear nucleus and inferior colliculus are noted, as are psychophysical data about perception of CV syllables that may be explained by the sustained-transient channel hypothesis. The proposed sustained and transient processing seems to be an auditory analog of the sustained and transient processing that is known to occur in vision.**

I. Introduction: Early Auditory Filtering and Coarticulated Speech

RAPID spoken language is effortlessly produced and understood by normal humans despite the extraordinary demands that it makes upon motor, sensory, and cognitive mechanisms. Although some of the musculature of the human vocal apparatus moves quite rapidly compared with usual skeleto-muscular rates, the muscles for many tasks cannot keep up with the transmission rate of spoken speech. Hence, the phenomena of coarticulation during which significant overlap of articulator motions occur for adjacent speech segments. In particular, motion of the tongue and larynx for vowels frequently overlaps the motion for consonants. Often, segments are nasalized both before and after an utterance [16]. In order to model speech recognition in a biologically plausible fashion, it is necessary to account for the recognition of coarticulated speech productions. We here investigate a peripheral speech processing mechanism that partially disambiguates coarticulated vowels and consonants.

One conceivable method to do this is to operate on the short time output spectrogram of the speech waveform. Standard speech spectrographs are, however, not reliable transducers of phonetic information in natural speech environments; see [43] for a review. For example, the spectral pattern in standard speech spectrograms degrades markedly for female speakers, young children, and all speakers in noise. This lack of robustness is in contrast to human behavior: The speech of adult females and children is about as intelligible as adult male speech. Speech is often completely intelligible in the presence of noise whose power is the same as the speech signal.

Other standard preprocessors of speech input, such as homomorphic filtering (cepstral analysis) and linear predictive coding techniques, suffer similar degradation under natural conditions [25]. Since most speech recognition systems use such preprocessed data without feedback, they are inherently unstable and therefore unreliable under normal uncontrolled speaking conditions.

An alternative approach to early auditory processing takes its inspiration from data about speech perception that articulate major differences between vowel and consonant sounds [28]. The Fourier spectrum of a typical vowel consists of a series of sinusoidal components whose frequencies are integral multiples of a fundamental frequency and whose amplitudes depend upon the resonant formant patterns of the vocal tract configurations. In contrast, for nonvocalic sounds, such as stop consonants and fricatives, the waveforms lack a clear periodic quality, and have spectra that change more quickly and over briefer durations than vocalic segments. These differences raise the question of what types of mechanisms are used by the brain to efficiently process such different types of signals.

Data about how the eighth nerve works provide a starting point for our analysis [62]. It has long been known that eighth nerve cells, as recorded in a sedated animal, have a dynamic range of only 30 dB or three orders of magnitude. However, from psychophysical studies, it is also known that auditory

Manuscript received December 21, 1993; revised October 16, 1996. The work of M. A. Cohen was supported in part by the Air Force Office of Scientific Research under Grant AFOSR F49620-92-J-0225. The work of S. Grossberg was supported in part by the Advanced Research Projects Agency under Grant AFOSR 90-0083 and ONR N00014-92-J-4015, the Air Force Office of Scientific Research under Grant AFOSR F49620-92-J-0225, and the Office of Naval Research under Grant ONR N00014-95-1-0409. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

The authors are with the Center for Adaptive Systems and Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215 USA (e-mail: steve@cns.bu.edu; mike@cns.bu.edu).

Publisher Item Identifier S 1063-6676(97)04852-9.

perception has a dynamic range of 90 dB of Weber Law sensitivity. How can we explain this discrepancy? The classical view for speech as well as for noise is that this sensitivity is effected by a (usually unspecified) mechanism of recruitment or by gain control [51]. A series of experiments initiated by Young and Sachs [67] and Sachs and Young [63], and followed up in a detailed fashion by [20]–[24], suggests a significant modification of the classical view.

First Sachs and Young [63] and Young and Sachs [67] replicated classical work of Kiang *et al.* [42] using steady state synthetic vowels. Sachs and Young [63] plotted the average response rate of the eighth nerve fibers as a function of characteristic frequency (frequency of best response to a sinusoidal stimulus). They also studied the spectrum of the response of these cells to these sinusoidal stimuli. They found a roughly 30 dB of dynamic range before the response of the ensemble of eighth nerve fibers saturate to the steady state vowel stimulus. However, Young and Sachs [67], Sachs and Young [63], and Delgutte and Kiang [20]–[24] constructed a series of response measures, called *average localized synchrony measures* (ALSM) that extend the selectivity of eighth nerve fibers to a dynamic range of 60–90 dB in response to vowel-like sounds and enable recovery of critical features of the vowel spectra through a frequency range of 4 kHz. This enables recovery of the formant structure of a vowel at least through the second formant.

The response measures of Sachs and Young are constructed for each frequency ω as follows. The Fourier transform of the period histogram of the response of an eighth nerve cell is computed. This Fourier spectrum is multiplied by a narrow window whose frequency response is centered at the characteristic frequency ω of the fiber, and the product is integrated. This output is normalized by an appropriate statistic of the discharge rate of the fiber. Generally, the average rate or the root mean square rate is used. Finally, this response is averaged over different fibers whose characteristic frequencies lie within a critical band of ω.

Even though the overall response rate of individual fibers saturates in a 30 dB intensity range, ALSM measures maintain selectivity up to about 80 dB. Furthermore ALSM's are relatively insensitive to background noise, whereas rate codes are highly sensitive to noise. On the other hand, these ALSM measures are less effective in producing a reliable signature for either fricatives or stops. In the case of stop consonants in an environment of vowels, such as in the syllable /ida/, the ALSM measures are hardly changed by the presence of the consonant [20]–[24]. Thus the strengths of ALSM are balanced by important weaknesses.

From the perspective of engineering system design, the ALSM is computationally complex and requires large amounts of numerical precision, and is therefore hard to compute in real time. Operations based upon the short-time Fourier transform of the input waveform are also problematical from a biological standpoint. There is no evidence that any mammalian auditory system computes an analog of a Fourier transform. The problem of time frequency resolution of signals has been approached in the signal processing literature by the use of wavelet or Wigner–Wille transforms [39], [61]. Continuous short-time Fourier transforms use a time-frequency representation; continuous wavelet transforms use a time–scale space representation. Time-variant filters can be constructed with specific localization using the short-time Fourier transform or the wavelet transform. The construction proceeds by applying either of these two transforms followed by a multiplicative operation in the time–frequency or time–scale space representation, followed by an inverse transform. However, there is no clear guidance as to which member of the families of transforms to use for the problem of speech recognition. Furthermore, there is no evidence that the linear processing inherent in these schemes has favorable noise suppression properties, whereas the synchronous averaging carried out by the auditory system does [24], [25]. This paper constructs a filter that achieves a related time–frequency analysis, and that is motivated by auditory psychoacoustics and neurophysiological criteria. These detectors appear to have more favorable noise suppression qualities than many based on the short-time Fourier transforms or wavelet transforms referred to above, and also help to model the synchronous processing carried out by the auditory system.

The new model forms part of the front end of a self-organizing neural network architecture for real-time auditory and speech processing (see Fig. 1) from the periphery to the word recognition level that our group has been developing [3]–[6], [11]–[14], [27], [31], [32], [34], [36]. It is suggested that processing of the speech waveform splits into two channels, a sustained channel and a transient channel. The sustained channel processes slowly varying envelopes that reflect synchronous properties in the vocalic portion of the stimulus. The transient channel responds to critical features at onsets and offsets of speech segments. The model hereby helps to disambiguate coarticulated speech segments.

The present article illustrates how such a front end works by showing how it can separate transient and sustained signals for several key speech sounds, such as stop and vowel onsets and offsets, and frications. Further development of this front end will require that it be integrated into a larger architecture for auditory and speech processing that is now being assembled. This architecture includes a new model of pitch perception [14], of auditory scene analysis and source localization [27], and of variable-rate speech categorization [4], [36]. The sustained-transient filter described herein does not, in itself, accomplish speech recognition. Its role in the architecture can be clarified by the following examples.

Boardman, Grossberg, and Cohen [4] have proposed how to explain why the perception of CV syllables exhibits context effects whereby voice onset time (VOT) of a consonant and duration of a subsequent vowel interact. Percepts of /ba/ and /wa/ can, for example, depend on the durations of the consonant and vowel segments, with an increase in the duration of the subsequent vowel switching the percept of the preceding consonant from /w/ to /b/ [50], [55]. In their model, C and V inputs are hypothesized to be filtered by parallel auditory streams that, as in the present work, respond preferentially to transient and sustained properties of the acoustic signal. These streams are represented by working memories that adjust their processing rates to cope

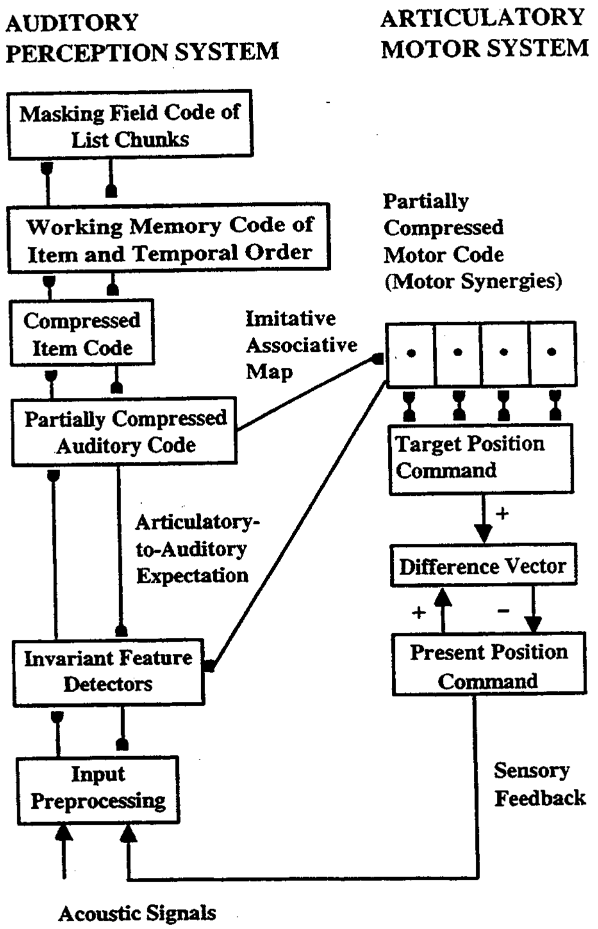

AUDITORY PERCEPTION SYSTEM
ARTICULATORY MOTOR SYSTEM
Masking Field Code of List Chunks
Working Memory Code of Item and Temporal Order
Partially Compressed Motor Code (Motor Synergies)
Compressed Item Code
Imitative Associative Map
Partially Compressed Auditory Code
Target Position Command
+
Articulatory-to-Auditory Expectation
Difference Vector
+
-
Invariant Feature Detectors
Present Position Command
Input Preprocessing
Sensory Feedback
Acoustic Signals

Fig. 1. Neural network architecture for auditory and speech processing from the periphery to the word recognition level includes the present filter as an early processing stage.

with variable acoustic input rates. More rapid transient inputs can cause greater activation of the transient stream which, in turn, can automatically gain control the processing rate in the sustained stream. An invariant percept obtains when the relative activations of C and V representations in the two streams remain approximately unchanged. The context effect may be simulated as a result of how different experimental manipulations affect this ratio. The role of the sustained-transient filter in this example is thus to enable parallel working memories to respond more selectively to sustained and transient properties, respectively, of acoustic waveforms, and to thereby enable these different types of signals to modulate each other's processing.

Mann and Soli [48] have provided additional experimental support for the hypothesis that consonants gain-control vowels. As in the Miller and Liberman [50] study, they showed that the succeeding vowel in a consonant/vowel (CV) pair can influence categorization of the initial consonant. In addition, they showed that, if consonant and vowel order is reversed, then the vowel has little effect on consonant classification in vowel/consonant (VC) pairs. Mann and Soli ruled out articulatory cues by constructing artificial CV and VC pairs in which the C and V sounds were reversed. They concluded that "current auditory processing models, such as backward recognition masking, preperceptual auditory storage, or models based on linguistic factors, do not account for the observed asymmetries" (p. 399). The present hypothesis that parallel sustained and transient channels exist and appropriately interact can thus help to explain a variety of basic speech data that previous models cannot handle.

The gain-controlled working memories do not, however, generate temporally distinct or recognized events on their own. Several additional problems need to be analyzed to understand how these context effects lead to speech recognition. How are consonant and vowel features temporarily stored in a working memory in such a way that a subsequent event, such as a change in vowel duration, can alter the percept of a preceding consonant before it reaches conscious awareness? Why does the conscious percept take so long to emerge that the duration of a subsequent vowel *can* influence the percept of a preceding consonant? Why is the consonant not already consciously perceived before the vowel is fully presented? Finally, how are these several processes designed to ensure that language can be understood even if it is spoken at different rates?

Grossberg, Boardman, and Cohen [36] suggested partial answers to these questions by modeling how VC–CV syllables are categorized when they are presented with variable silence intervals between the two consonants. Repp [57] showed that the categorical curve representing the probability of perceiving */ib/–/ba/* instead of */iba/*, as a function of silence interval, was shifted by a silence interval of 150 ms from the curve representing the probability of perceiving */ib/–/ga/* instead of */iga/* as a function of silence interval. Why is this shift so large? We propose that it is large for the same reason that the duration of a subsequent vowel can influence the percept of a prior consonant; namely, conscious speech perception is not the result of a bottom-up filtering process alone. Rather, it emerges as a result of a nonlinear resonance that develops more slowly between bottom-up and top-down signals. Thus, a bottom-up filter like the sustained-transient filter is not designed to compute phonemic boundaries on its own. The difference between fusion and temporal separation, as in the */ib/–/ba/* to */iba/* and */ib/–/ga/* to */iga/* distinctions, depends also upon the intervention of top-down processes.

In particular, after preprocessors such as the sustained-transient filter operate upon individual acoustic segments, acoustic events in the model are represented as spatial patterns of activation across one or more working memories. These working memories can temporarily store a series of preprocessed sounds. Their temporally evolving patterns are categorized by a competitive learning or self-organizing feature map network [30], [31], [44], [47]. In this network, the working memory activation pattern at any time generates output signals that are processed by an adaptive filter. The filter generates inputs to a second level of nodes, or cell populations, that categorize the patterns that are active in working memory. Category nodes are chosen by lateral inhibitory, or competitive, interactions. Only the node, or nodes, that receive the largest input—or close to largest inputs—from the adaptive filter win the competition. Adaptive weights, or long-term memory traces, in the filter pathways undergo learning only if they input to a winning node. Learning is designed to encode the ratio of activations across the working memory nodes. This is how category nodes in the model become sensitive to the sustained/transient ratio in the */ba/* or */wa/* percept.

Why does this classification process take so long that the duration of a vowel can influence the percept of the preceding consonant? Why is not the consonant already classified before the vowel is fully presented? Grossberg [31], [34] proposed that the perceptual event is not bottom-up activation of a category node *per se*. Rather, when a category node is activated, it releases learned top-down signals to the working memory. These top-down signals represent the prototype of the chosen category. The prototype is matched against the working memory pattern, and can hereby reorganize it by generating a focus of attention that selects the feature pattern that is expected by the prototype from the total activation pattern.

As this matching process takes hold, it reactivates consistent category nodes via the bottom-up filter. The amplified category nodes, in turn, reinforce their top-down prototype signals. This bottom-up and top-down exchange of amplified matching signals generates a resonant state within the system. The resonant state evolves on a slower time scale than bottom-up activation. The resonant state, rather than bottom-up activation *per se*, is assumed to subserve the conscious speech percept. The resonant state is also assumed to trigger any new learning of categories in the bottom-up filter and of prototypes in the top-down expectation. Hence, this resonant event has been called an *adaptive resonance*, and the larger theory of which it is a part has been called *adaptive resonance theory* (ART) [9], [33], [35].

Within ART, the brain's sensitivity to the sustained/transient ratio is ascribed to the fact that the resonance takes hold slowly enough that the duration of the vowel has a chance to influence the final CV percept. Carpenter and Grossberg [10], Cohen, Grossberg, and Stork [13], Grossberg [34], Grossberg, Boardman, and Cohen [36], and Grossberg and Stone [37] have used ART mechanisms to explain a variety of other data about speech and language perception and production.

In summary, the sustained-transient filter described herein is proposed to help set up some key working memory distinctions and to gain-control working memory representations so that resonances with these working memories can extract invariant acoustic and speech properties.

II. Filters, Synchrony, and Transients

We model the response of the basilar membrane by a filterbank of linear filters with the filter shape of a bank of cochlear neurons. The output of this filterbank is passed through a simple rectifying nonlinearity. Remarkably, the response properties of eighth nerve cells in broadband stimuli can to first order be modeled by such a filterbank with considerable success [8], [17]. However, the data cited above of Sachs and Young [63] and of Delgutte and Kiang [20]–[24] suggest that further processing is necessary at higher levels to account for the stability of vowel recognition at signal levels from 60–90 dB above threshold and in noise. Their work indicates that some sort of synchronous nonlinear short-time averaging is used to provide stable recognition of vocalic stimuli.

Each sustained channel is modeled by a coincidence detector that computes the following operations: i) the output of each cochlear filterbank is passed through a power function and rectified; ii) the rectified output is passed through two parallel channels, one with a delay, and the output of both channels are multiplied; iii) the product is exponentially time averaged; iv) the average is scaled by the frequency of the input. This latter operation, which is known as *preemphasis*, compensates for the known increasing sensitivity to high frequencies in the mechanical spectrum of the outer and inner ears [54], and produces more phonetically reliable spectrograms. This output is plotted as a cochlear spectrogram of the sustained channels. Outputs across the sustained channels can also be pooled to obtain a measure that is sensitive to the gross sustained characteristics of the input.

In order to detect phonemic boundaries, and to distinguish between different types of consonant segments, rapid onsets and offsets need to be detected in the speech waveform. The transient channel accentuates onset and offset information in different frequency bands in the speech waveform. In order to detect transient information in a specified frequency range, a transient detector is applied to the output of a set of cochlear filters in this range. For example, pooling low-frequency offset information enables detection of rapid offsets of vowels, indicating the start of a consonantal segment. Pooling outputs of offset detectors in a higher frequency region enables detection of the offset of a consonantal burst, as shown below in Section V. Pooling low-frequency onset information enables detection both of the onset of a vowel and the offset of the an immediately preceding burst. The high-frequency transient detectors in the model are sensitive to fricative stimuli. The sustained detectors and all the other transient detectors have no such sensitivity.

In summary, the transient channel is in many ways computationally complementary to the sustained channel. The transient channel is sensitive to rapid changes in auditory signals at a cost in frequency selectivity. The sustained mechanism is much more sensitive to frequency information at a cost in temporal resolution. The transient channel thus excels in detecting aperiodic auditory signals, whereas the sustained channel focuses upon periodic or synchronous signals. These complementary sensitivities are processed within parallel, but distinct, representations that help to spatially and temporally disambiguate coarticulated consonants and vowels.

The complementary properties of these parallel channels clarify a sense in which spectrograms, in themselves, do not provide a natural or complete representation of the information contained within naturally occurring auditory or speech signals. It is therefore quite difficult for even trained human subjects to actively retrieve phonetic representations from spectrograms. The model presented in this article suggests that combinations of separately filtered sustained and transient information are used by listeners to achieve phonetic discrimination and recognition.

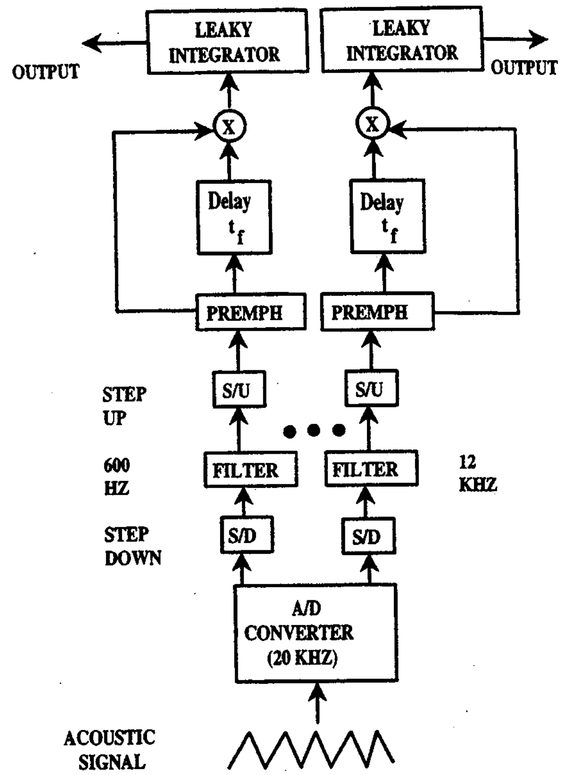


Fig. 2. Sustained detector. The acoustic speech waveform is passed through an A/D converter sampled at 20 kHz to maintain frequency resolution. The signal is then stepped down to a variety of rates, and passed through a filterbank of "cochlear filters" of length 512 for each selected frequency. The output rate of the filters was stepped up to 20 kHz to ensure equivalent processing across channels. The output of each filterbank channel is preemphasized and halfwave rectified. It is then delayed by a period equal to the reciprocal of the center frequency of the filter. An autocorrelation with the given delay is performed.

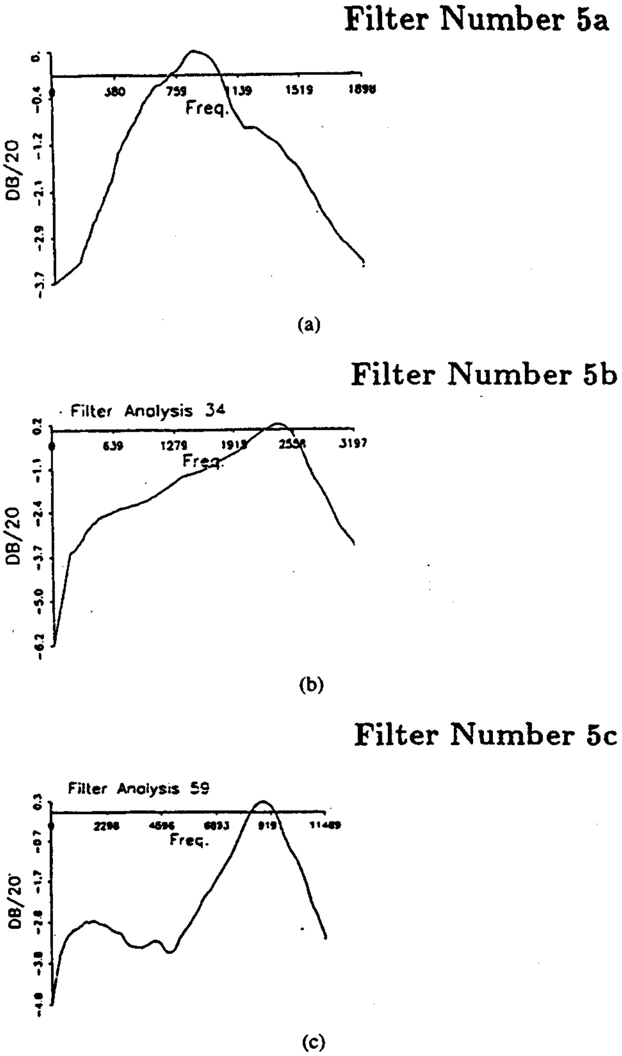


Fig. 3. Plot of dB of intensity attenuation of the cochlear filters. (a) This "cochlear filter" has a center frequency of 834 hz with reconstruction data in the range from 160 to 1700 Hz. Notice the relatively symmetric falloff. This filter has broader high frequency falloff than low frequency cutoff. This is only characteristic of the symmetric responses. (b) Midrange cochlear filter with a frequency response that is characteristic of a typical cochlear filter. Its center frequency is 2350 Hz with reconstruction data in the range from 170 Hz to 3 kHz. Note the sharp high frequency falloff, which is considerably sharper than the low frequency data. This is typical of most of the amplitude response on the basilar membrane. (c) Filter with center frequency of 8444 Hz with reconstruction data ranging from 300 Hz to 1.1 kHz. Notice again the broad low-frequency tails and the sharp high-frequency falloff.

III. A Model Cochlear Filterbank

We construct the simplest filterbank that models the peripheral auditory transduction that is needed to provide inputs to the sustained and transient channels. A schematic of this filterbank is shown in Fig. 2. The acoustic waveform was recorded by an analog to digital converter sampling at 20 kHz. The output of the converter was transformed to a slower sampling rate by low pass filtering and undersampling, thereby maintaining resolution in the filterbank, as in Crochiere and Rabiner [15]. The smoothed and filtered data were then stepped back up to 20 kHz, so that the output of each channel was at the same frequency and so is directly comparable. This interpolation was accomplished by lowpass filtering a sequence consisting of the scaled input data interspersed between substrings of zeros of a fixed length. A bank of filters of length 512 whose attenuation was the same as the measured frequency response of the cat basilar membrane was constructed [45].

This amplitude response does not, however, specify the response of the filter uniquely [52]. The phase shift at each frequency also needs to be represented. There is a unique filter that has the shortest phase delay at any given frequency for a fixed amplitude response. Such a filter is called a *minimal phase* filter. Our filters were constructed to be minimal phase using the Durbin algorithm [40]. It is known that the basilar membrane response in the linear range exhibits minimal phase behavior [65], [68]. The output of these filterbanks form the time-varying input to the sustained and transient channels. Representative filtershapes are shown in the following figures. These filtershapes are characteristic of cochlear response and can be seen in psychophysical studies as well [66]. Note the relatively symmetric lowpass and highpass responses in Fig. 3(a). The high frequency falloff on the skirts in Fig. 3(b) and (c) is much sharper in accord with physiological data.

Spectrograms were constructed from the filterbank using a digital spectrographic package constructed by the Speech Recognition Group, Carnegie Mellon University (CMU) [1].

These spectrograms are compared with the sustained and transient channel outputs below.

IV. The Sustained Channel

Fig. 2 is a schematic of the filtered output of one sustained detector. The input of this system is the output of the cochlear filters constructed above. Let $F_i(t)$ be the output of the ith filter at time t. This input is preemphasized and halfwave or fullwave rectified. The signal function used to carry out these operations is of the form

$$f(x) = (x^+)^\alpha \quad \text{or} \quad f(x) = (x^*)^\alpha \tag{1}$$

where $x^+ = \max(x - T, 0)$ and $x^* = |x|$. Whether halfwave rectification (x^+) or fullwave rectification (x^*) is chosen will be made clear by context. Parameter α is chosen so that $0 < \alpha < 1$; $\alpha = 0.5$ in the simulations.

Each signal function output $f[F_i(t)]$ is delayed and autocorrelated. The delay τ_i is chosen to be the reciprocal of the center frequency of the filter. The output of the filterbank S_i thus has the form

$$S_i = \omega^\gamma \int_0^t e^{-\beta(t-u)} f[F_i(u + \tau_i)] f[F_i(u)]\, du \tag{2}$$

which is approximated as

$$S_i = \omega^\gamma \sum_{i=1}^{p[t/p]} e^{-\beta(t-ip)} f\{F_i(ip + [\tau_i/p]p)\} f[F_i(ip)] \tag{3}$$

where $[z]$ denotes the largest integer less than z, p is the sampling period, and $f(x)$ is defined as in (1). Spectrograms were constructed that represent the output of the sustained channel filterbank, so as to be able to compare with the control CMU spectrogram.

The sustained channel spectrogram is constructed as follows. For each sustained channel filter, a band is created starting at the average of the center frequency of the prior and current filter and ending at the average of the center frequency of the current and following filter, using a halfwave rectified signal function. Since the center frequencies of the filters are monotone increasing, and have roughly the same shape, this band is approximately the response area of each individual filter. The output of this filterbank is displayed as follows. The response magnitude of the entire stimulus is scaled to the maximal response of all the channels. The output of each filterbank is compared to this maximum and the amplitude of response of the filterbank is quantized in 100 steps. Since the printer output at each point is black or white only, multiple levels are simulated by making the probability of blackening a point proportional to the quantized output level of the channel. Some representative outputs of the sustained channel are summarized in the following figures.

Fig. 4(a) exhibits the response of the CMU spectrogram to the vowel */ae/*. The waveform is plotted above and the spectrum is plotted below the waveform. The higher frequency formants at about 3 and 4 kHz are present in this spectrograph largely because of the use of a compressive nonlinearity $\{1 + 1/5 \log[I(\omega, t)/\bar{I}]\}^3$, where $I(\omega, t)$ is the short time spectral energy at the frequency ω computed from a Hamming-windowed fast Fourier transform (FFT), and $\bar{I}$ is the maximal intensity of the output of the frequency response within a small window centered around the coordinate on the CMU spectrogram.

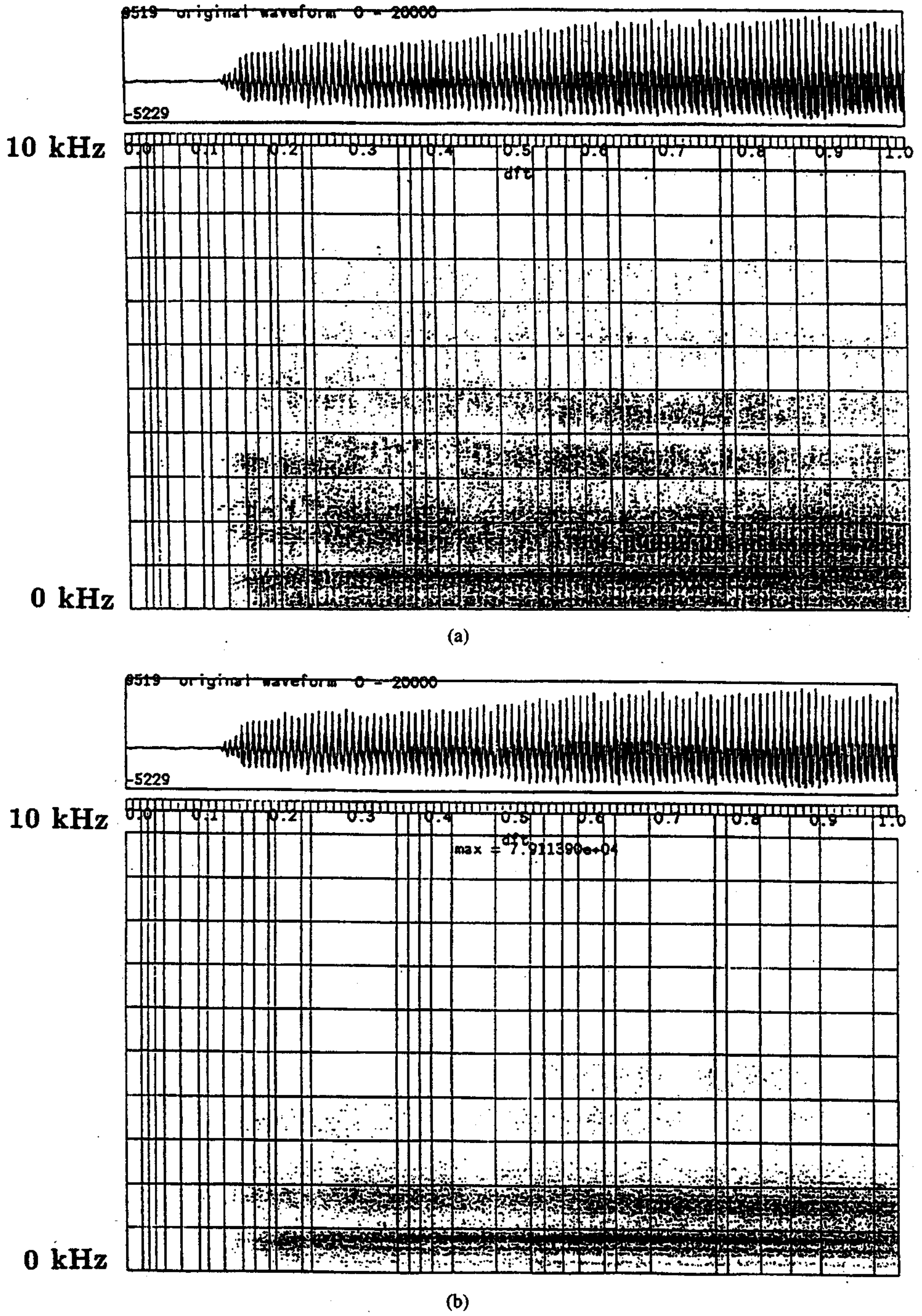
Fig. 4(b) plots the response of the sustained detector for the vowel */ae/*. The sustained spectrogram that is constructed preserves the formant structure of the vowel at least up to about 3 or 4 kHz. Energy at higher frequencies is present but the higher frequency formants are poorly localized in time.

Fig. 5(a) plots the response of the CMU spectrogram to */aepae/*. Notice the onset burst of the voiceless stop *p*. It contains considerable energy and has large amounts of broadband energy. Fig. 5(b) shows how the sustained channel attenuates markedly this broadband energy. Thus, the sustained channel is sensitive to the shape and formant structure of vowels while it attenuates transients such as stop consonants.

To better gauge the global properties of the sustained channel, we pool across channels so as to be able to observe the total detector output of the entire bank of detectors across the entire spectrum; that is, we let

$$S(t) = \sum_{i=1}^{N} S_i(t). \tag{4}$$

We compare the response to */stop/* of the pooled output (4) of the sustained channel, the output (3) of the individual sustained channels, and the output of the CMU control spectrogram. In the CMU control spectrogram for */stop/* plotted in Fig. 6(a), the large burst of frication energy for the first 0.1 s corresponds to */s/* starting the syllable. The silent gap of about 0.1 s indicates a stop fricative cluster */st/*, which is followed by the energy in vowel */Λ/*. The burst of broadband energy at about 0.6 s indicates the final stop */p/*. Fig. 6(b) shows the response of the individual sustained detectors to the word */stop/*. Notice that the formant structure is broadly preserved but that the shape of the third and fourth formants, appearing between 3–4 kHz, are blurred. Notice as with */aepae/* that the response to the bilabial plosive */p/* is almost completely obliterated. Thus, the sustained detector can obliterate the consonantal burst of */p/* independent of the vocalic environment. The */p/* burst is attenuated in both the environment between */ae/* and the environment following the vowel */o/* (phonetic */a/*). Fig. 7 plots the output of the pooled sustained detectors to */stop/*. Notice the large attenuation of the frication and burst. To maximize synchronous response, we let S_i in (3) take fullwave rectified input from the filters F_i. Since we are always pooling a positive signal, there is no direct cancellation of the noise. However, the output of a given sustained detector S_i is correlated with energy at multiples of the best frequency of the given channel S_i. This correlation effectively attenuates incoherent energy in the signal relative to the coherent response, which is always passed through the detector maximally. Since a coherent signal may equally well be obtained by temporal



(a)

(b)

Fig. 4. (a) Steady-state representation of the vowel /ae/ plotted using the CMU spectrogram. (b) The response of the sustained channel to the vowel /ae/. The parameters chosen are $\beta = 0.001$, $T = 2000$, $\alpha = 0.5$, and $\gamma = 1$.

coincidence during the negative going as well as the positive going part of the acoustic wave, full wave rectification takes advantage of this fact. This reflects the biological fact that the phase locking of the inner hair cells need not all take place during the same phase of a given spectral component of the signal.

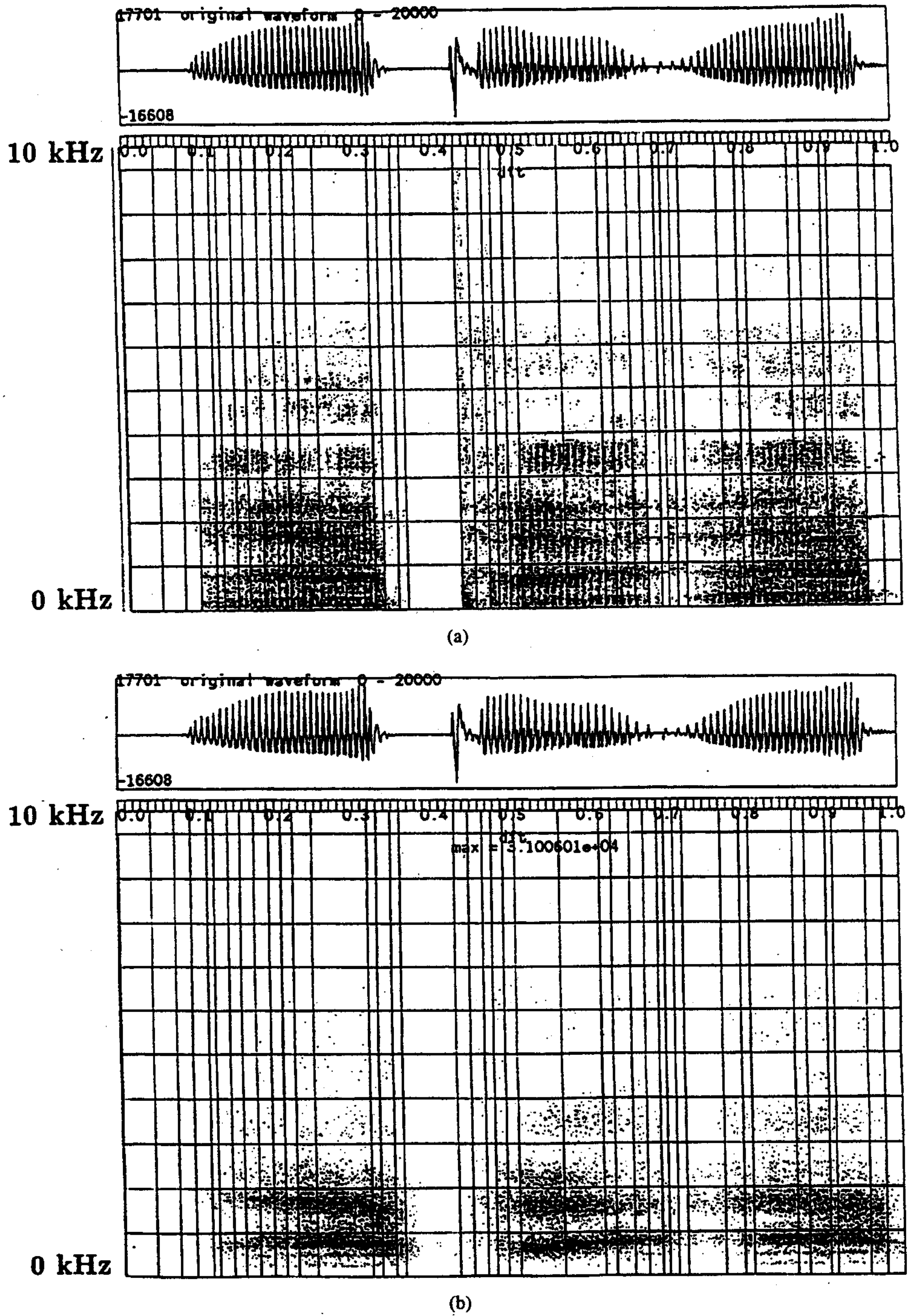
V. The Transient Channel

A simple neural circuit that produces a transient output signal utilizing a feedforward inhibitory interneuron was introduced in Grossberg [29]. The simplest equations that realize key properties of such a circuit are

$$\frac{dx}{dt} = -\delta x + I - \epsilon y \tag{5}$$

$$\frac{dy}{dt} = \zeta(-y + I) \tag{6}$$

where I is an input, x is the activity of an output cell, and y is the activity of a slow feedforward interneuron that inhibits the output cell. In more general detectors of this type, the



(a)

(b)

Fig. 5. (a) CMU spectrograph response to the phoneme sequence /aepae/. (b) Sustained channel response to /aepae/.

inhibitory threshold is higher than the excitatory threshold, but the inhibitory gain is larger than the excitatory gain [29].

This equation is analyzed by Laplace transforms in the Appendix for both the discrete and continuous variants that were simulated. In this analysis, the impulse response of this linear time-invariant system is constructed and parameterized in terms of the step response g of the system, the energy G^2, and two decay rates δ and ϵ. The Fourier transforms of the discrete system used in simulations and of the idealized continuous system are also constructed. This analysis shows that these systems act as "bandpass differentiators." By this we mean that they take the derivative (or the first difference) of the input over a range of temporal frequencies that always includes zero. The range of frequencies in which this occurs is controlled by the parameters α and γ. However, this range is lowpass, as shown by (A12) and (A24) in the Appendix. Such a system acts as a reliable change detector in a broad frequency range. However, it is important in processing speech stimuli to be able to respond to changes in response as rapidly as possible in a relatively narrow band portion of the frequency range, while remaining insensitive to changes in energy at other frequency regions of the short time spectrum.

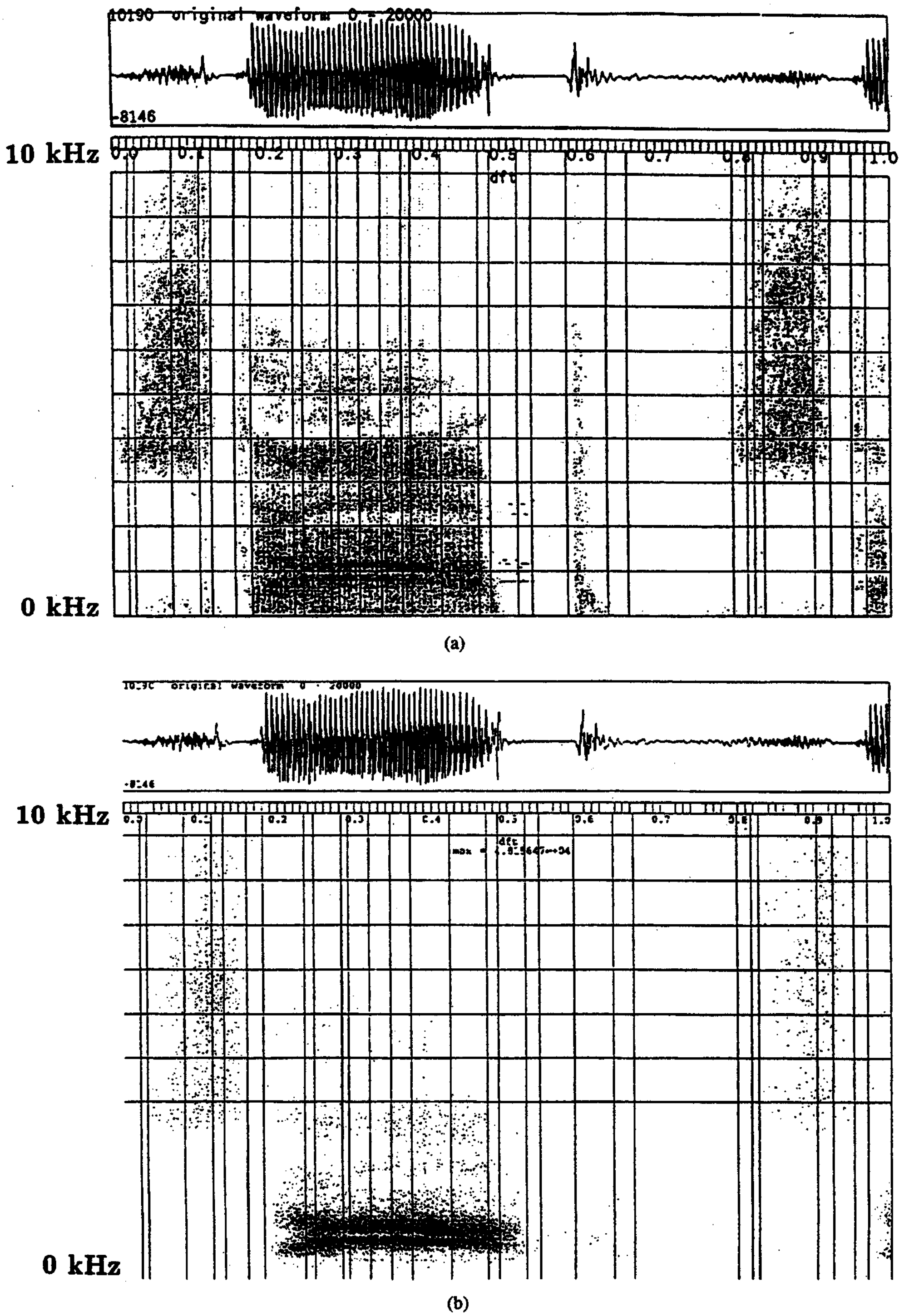


Fig. 6. (a) Control spectrum of one second of the utterance */stop/*. (b) Sustained detector response to the segment */stop/*. Parameters are $T = 0, \beta = 0.001, \gamma = 1$, and $\alpha = 0.5$.

For example, it is well known that an important cue to place of articulation of stop consonants are formant transitions within relatively narrow frequency regions [18], [46]. Rapid onsets or offsets of energy in high-frequency regions are important features in the speech waveform in distinguishing the affricate */tch/* from the fricative */sh/* [19], [22], [38]. We wish to have a measure that changes relatively slowly compared with the rate of change of individual spectral components, but which responds to rapid energy changes in a particular frequency region. The simplest transient detector, because it is linear and because its DC gain $g = 0$, responds equally well to positive and negative going waves in the speech waveform. If we simply average the changes over these frequency bands, then the transients may appear to cancel out even though there are significant energy changes in the frequency regions of interest. Furthermore, it is well known that onsets and offsets

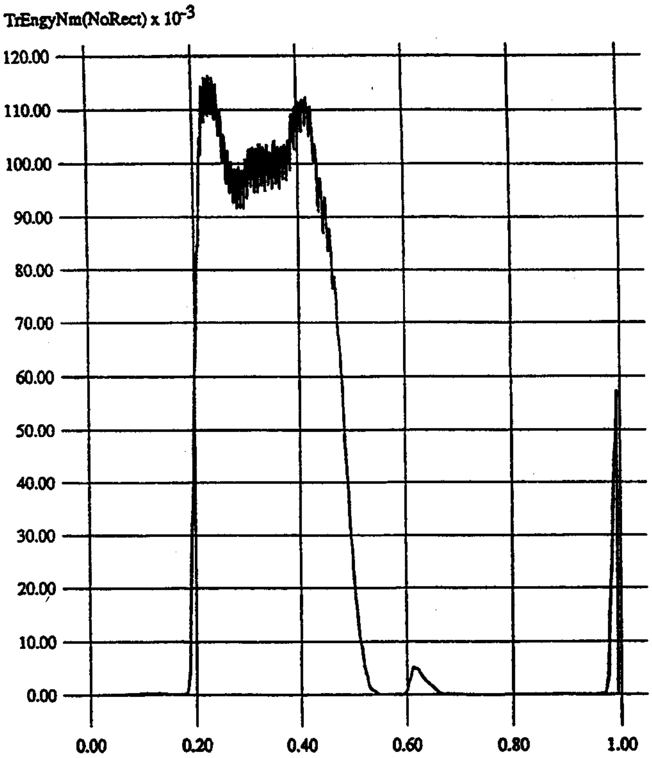


Fig. 7. Attenuation of the initial fricative and the final consonantal burst in /stop/ compared with the control spectrogram in Fig. 6(a). However, the energy profile for the vowel in stop is well maintained.

of the speech waveform signal different phonetic events [46]. We thus wish to have distinct detectors that are sensitive to onsets and to offsets. In addition, the speech waveform varies over a large dynamic range, so it is useful to compress the output of the individual filterbanks so as to maintain sensitivity to the entire stimulus range.

This suggests the following refinement, which produces a set of transient detectors sensitive to differing frequency ranges. To restrict sensitivity to a relatively small range of frequency bands, sum the output of the transient detectors linked to a small bank of cochlear filters. To obtain sensitivity to directional changes in the signal, halfwave rectify the positive or negative going signal. To obtain a detector that is simultaneously sensitive to both positive and negative going changes in the speech waveform, fullwave rectify the output. To obtain a change detector active in a small frequency range but insensitive to the direction of change, fullwave rectify the output of a small range of frequency channels. Wherever multiple detectors contribute to the filter, its output is normalized by the number of filters. Equations (A10), (A11), and (A18)–(A23) in the Appendix show that the slow rate constant ζ in (6) can be chosen sufficiently small so that averaging of transients takes place over a considerable interval of time. No additional averaging of the transient channels is needed to smooth the short time gains.

The operations used in the transient detector are schematized in Fig. 8 and described mathematically in the Appendix. The discrete variant of the detector, averaged over frequency channels i to j, is

$$T_{ij} = (i-j)^{-1} \sum_{k=i}^{j} T_k \tag{7}$$

where the transient detector for frequency channel i is given by

$$\begin{aligned} T_i[kp] = {} & \Gamma[\eta, \kappa] H \left[\sum_{j=0}^{k} (1-e^{-\kappa}) e^{-\kappa(k-j)} F_i^+(jp) \right. \\ & \left. -(1-e^{-\eta}) e^{-\eta(k-j)} F_i^+(jp) \right]. \end{aligned} \tag{8}$$

Expression

$$\Gamma[\eta, \kappa] = \frac{\sqrt{(1+e^{-\eta})(1+e^{-\kappa})[1-e^{-(\eta+\kappa)}]}}{\sqrt{2}\,(e^{-\eta}-e^{-\kappa})} \tag{9}$$

Equation (9) is determined from (A20) with $G^2 = 1$, $g = 0$, F_i^+ is the output of the ith cochlear filterbank, p is the sampling period, and H takes one of the following three forms. If $H(x) = \max(x, 0)$, then T_{ij} is called a *positive* transient detector. If $H(x) = \max(-x, 0)$, then T_{ij} is called a *negative* transient detector. Finally, if $H(x) = |x|$, then T_{ij} is called a *composite* detector. By (8), T_i is a rectified discrete convolution $s * F_i^+$ of the halfwave rectified output of the ith cochlear filter F_i^+ with a discrete representation, s, of the transient circuit defined by (5) and (6). The discrete representation, s, of the transient detector used in (8) is defined in the Appendix and displayed there in (A18).

Properties of the transient detectors as applied to representative data are summarized in Figs. 9–13. Fig. 9(a) illustrates a composite transient detector. Fig. 10 plots the output of the composite transient fullwave detector when the entire frequency range is pooled. This detector peaks at the onsets and offsets of consonant bursts for the utterance /stop/ and thus can be used as a generalized change detector, when followed by a simple threshold.

Although this detector responds to onsets and offsets of consonants, and therefore serves as an important general cue, it pools over too large a frequency range to distinguish between the onsets and offsets of differing segments. Because the detector is fullwave rectified, it also cannot distinguish between onsets and offsets of the stimulus. To detect changes between onsets and offsets and to selectively detect changes in differing frequency regions that are known to be useful in detecting different stop consonants [46], adjacent input channels are pooled and passed through the transient detectors, as in (7), and the output is summed and normalized. When pooling is done using high-frequency cochlear input, the detector is a high-frequency transient detector. When pooling is done using low-frequency cochlear channels, the detector is a low-frequency transient detector. The outputs of adjacent high- and low-frequency negative transient detectors are displayed in response to the syllable /stop/ in Fig. 11.

Fig. 9(b) shows how the negative transient detector might be used to distinguish between stop and vowel offset, when pooling over distinct frequency ranges. Fig. 11(a) plots the

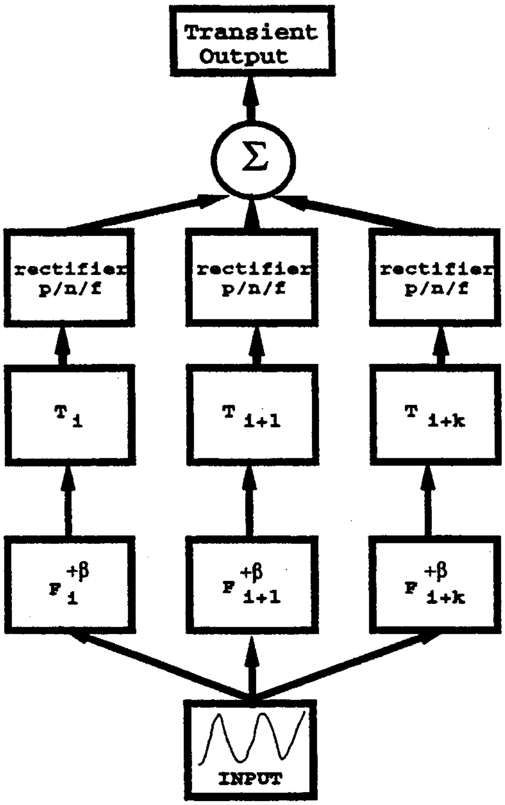


Fig. 8. Transient detector. See text for details.

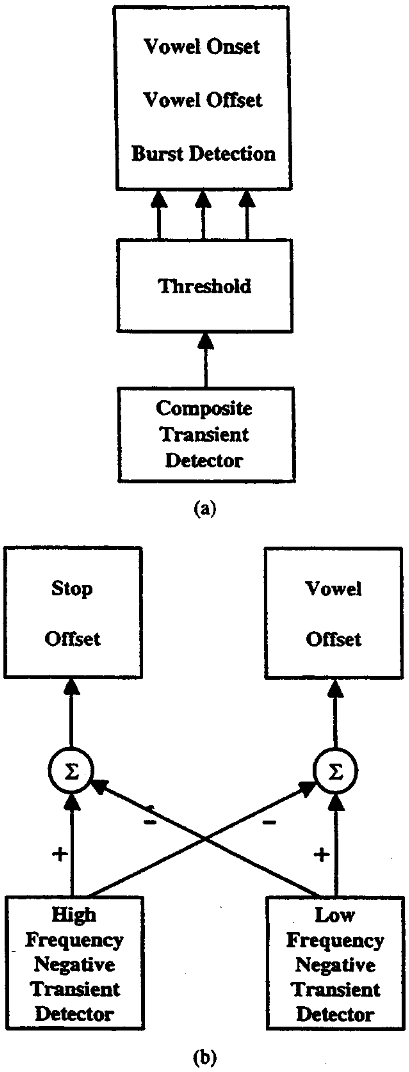

Vowel Onset
Vowel Offset
Burst Detection
Threshold
Composite Transient Detector
(a)
Stop Offset
Vowel Offset
Σ
Σ
+
-
-
+
High Frequency Negative Transient Detector
Low Frequency Negative Transient Detector
(b)

Fig. 9. (a) Composite transient detector. This detector is sensitive to vowel onsets and offsets and consonantal bursts. (b) Opponent interactions between high- and low-frequency detectors are sensitive to stop and vowel offsets.

output of the negative transient detector to input from the low-frequency cochlear filters. Note that a relatively large response at the offset of the vowel and to the release burst of /*t*/. Because the passive decay of the transient offset is slower than the onset these peaks are relatively broad. Note that the response to the vowel offset is much larger than the offset of the /*t*/ burst at the beginning of /*stop*/. Fig. 11(b) plots the pooled output of this detector to higher frequency cochlear filterbank inputs. Comparison of Fig. 11(a) and (b) show that a major difference between the response of the two detectors is the large offset burst of the stop /*t*/ in the segment /*stop*/ by the negative transient detector of high-frequency channels, and conversely, the relatively large response to the vowel offset by the low-frequency channels. Thus, as shown in Fig. 9(b), a detector that halfwave rectifies the difference of the output of the scaled low-frequency negative transient channels from the high-frequency channels will be sensitive to stop offset. Conversely, a detector that halfwave rectifies the difference of the output of the high-frequency transient channels from the low-frequency ones will be sensitive to vowel offset. Together, these detectors define an opponent processor.

The responses and some possible uses of the positive transient detectors are now considered. Fig. 12 illustrates a number of the uses of these detectors. Fig. 13(a) plots the output of the positively rectified transient detectors using the low-frequency cochlear input to the segment /*stop*/. Note the sharp response to the onset of the vowel. Thus, the offset of a burst and the onset of the immediately following vowel can be distinguished by the difference in response of the positive and negative frequency transient detectors. The positive low-frequency transient detector responds to the onset of the vowel, while the negative high-frequency transient detector responds to the offset burst. Fig. 13(b) plots the response of the positive transient detector to the output of high-frequency channels.

Fig. 13(a) and (b) show that the response of both positive high- and low-frequency transient detectors are relatively large *only* during the stop burst /*p*/. Observe the relatively large output at 0.6 s in both the low- and high-frequency transient detectors to the stop burst /*p*/. Thus, the multiplicative detector illustrated in Fig. 12(a) can detect the onset of /*p*/, which occurs at the time 0.6 s in Fig. 13(a) and (b).

During the fricative stimuli at times 0–0.15 s, the positive high-frequency transient detector shows a large sustained response, as illustrated in Fig. 13(b). However, Fig. 13(a) shows that the response of the positive low-frequency transient detector is attenuated in the same region. Thus, the halfwave rectified difference in the output of these two detectors can be used as a cue for fricative consonants. Such a mechanism is

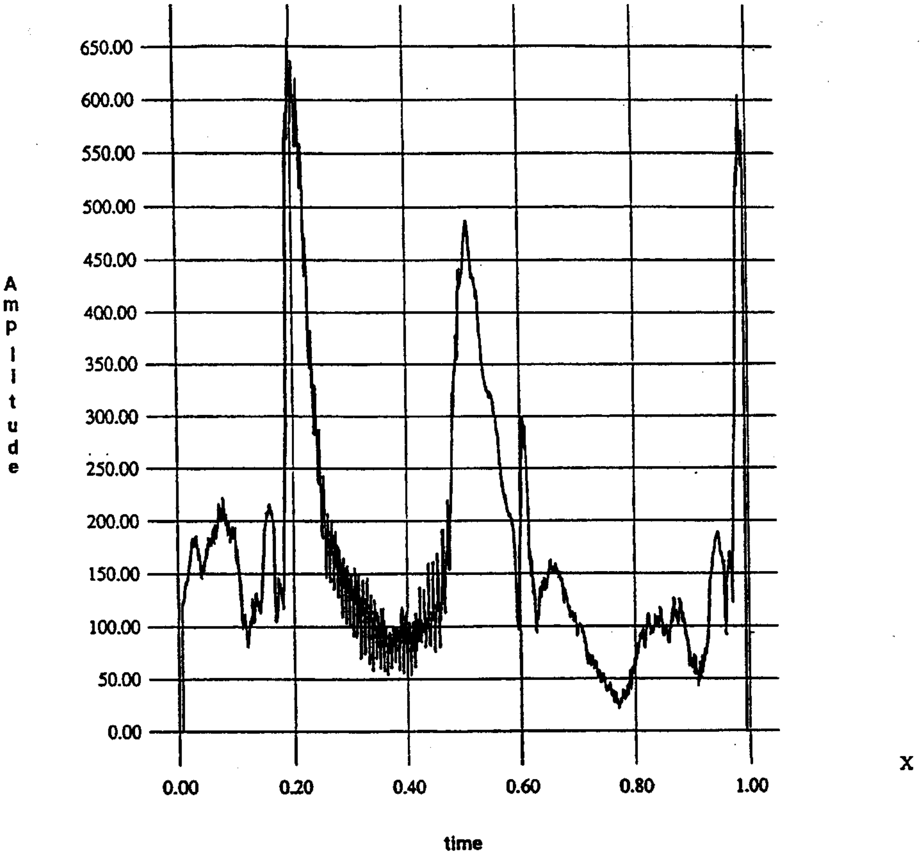


Fig. 10. A composite transient detector responds mainly to the onset and offset of the /t/ stimulus and to the onset of the /p/ stimulus. There is also a significant response to the vowel offset. Parameters are $\alpha = 0.01$, $\gamma = 0.001$, and $\beta = 0.4$. In this case, all filters in the filterbank are pooled.

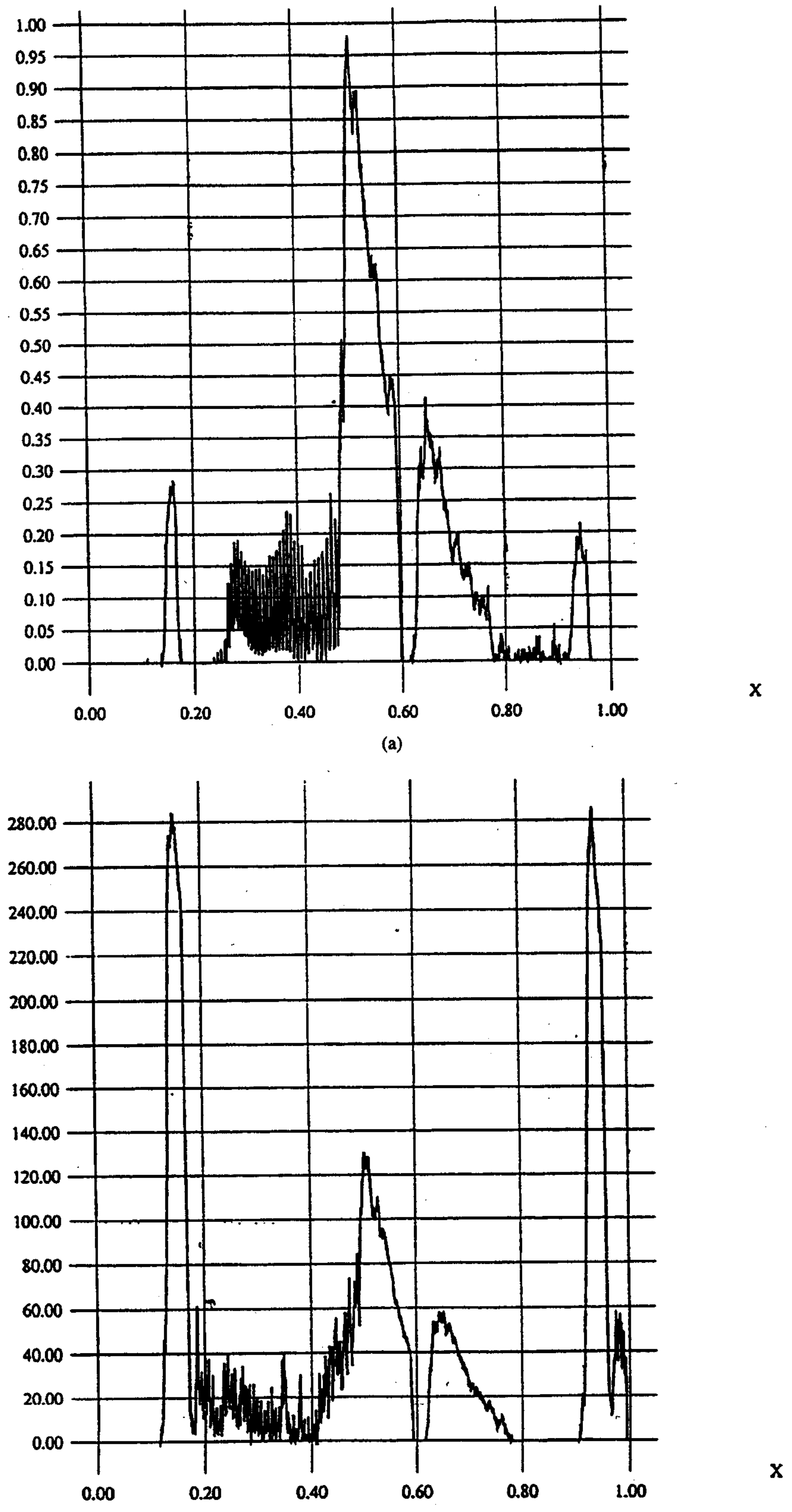
sketched in Fig. 12(b). In contrast, the positive low-frequency transient detector has a large response at the outset of the vowel /o/ at times 0.19–0.21 s in Fig. 13(a). At these times, the response of the positive high-frequency transient detector is much attenuated. Thus, the halfwave rectified difference between the low-frequency transient detector and the high-frequency transient detector can be used as a cue to vowel onset, as illustrated in Fig. 12(b). Together, these differenced high– and low-frequency channels comprise an opponent processing module for the computation of frications and vowel onsets.

VI. Concluding Remarks

This paper describes a neural model of how peripheral auditory detectors can facilitate automatic separation of coarticulated consonants and vowels during normal speech. Based on data about eighth nerve dynamics, a sustained detector channel is described that can discriminate synchronous vocalic quality, while suppressing transient information in the speech waveform. The sustained channel operates in parallel with a transient detector channel that can discriminate the onsets and offsets of fricatives and stop consonants, as well as detect vowel onsets and offsets.

The response properties of model transient detectors sensitive to high-frequency pooled cochlear input are quite similar to those of onset L cells, named by Pfeiffer [53]. These cells are found mainly in the posterior ventral cochlear nucleus, but they have also been found in the anterior ventral cochlear nucleus and the dorsal cochlear nucleus. The poststimulus time histogram of these cells shows a large response at onset followed by a much smaller but discernible response at later instants of time, when stimulated by tone bursts near the best frequency of the cell. A similar response at the onset of the vowel is anticipated for onset L cells selective for the formant frequency of a presented vowel. The poststimulus time histogram of the cells as reported by [59] to a short tone burst is similar in shape to the response of the on transient detector. Thorough investigation of the response properties of these cells, using synthetic or real speech, does not appear to be have been undertaken. Moreover, [7] have found that onset L cells often respond more vigorously to upward than downward linear FM ramps, or vice versa. Reference [60] reports replicating these results but found less directional selectivity than reported by [7]. Formant transitions are an important cue to consonant identity, and cells which detect unidirectional FM sweeps should be important for discriminating differing consonants. Further refinements of the parallel sustained and transient model detectors, such as introducing selectivity to FM ramps, may be used to achieve efficient segmental identification.

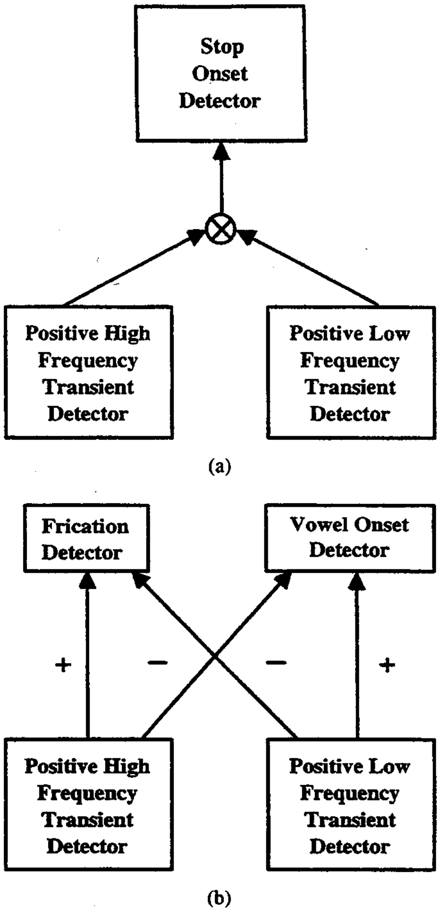
There does not appear to be much physiological evidence of cells that are synchronized to short time periodicities as posited by (2) and (3). However, Schriener and Langner [64] have found cells in which such periodicity detection



(a)

(b)

Fig. 11. (a) Response of a negative transient detector to low frequency output. Parameters are $\alpha = 0.01$, $\gamma = 0.001$, and $\beta = 0.4$. Frequency channels whose starting filter is number $i = 9$ and ending filter is number $j = 18$ are pooled in (7). This corresponds to a center frequency range from 400 to 850 Hz. (b) Response of a negative transient detector to high frequency output. Parameters are as in (a). Frequency channels whose starting filter is number 45 and ending filter is number 59 are pooled. This corresponds to a center frequency range from 4100–8800 Hz.


Stop Onset Detector
Positive High Frequency Transient Detector
Positive Low Frequency Transient Detector
Frication Detector
Vowel Onset Detector
+
−
−
+
Positive High Frequency Transient Detector
Positive Low Frequency Transient Detector

(a)

(b)

Fig. 12. (a) A product of the positive high and low frequency transient detectors can detect stop onsets. (b) Opponent processing between outputs of the positive high and low frequency transient detectors can detect frications and the vowel onsets.

occurs in the inferior colliculus of the cat. The measured best modulation frequencies of these cells are less than 1 kHz. It is conceivable that such cells could form the basis of the coincidence detection scheme modeled here by detecting periodicities at multiples of the period of an individual filter.

The interaction of these detectors with higher level cognitive attentional factors is yet to be addressed. Attentionally modulated feedback influences the neural response to an auditory stimulus as early as the receptor level [54], and influences processing of the speech waveform. Delgutte and Kiang [20]–[24] have shown that adaptation at the eighth nerve itself influences the short time response to speech stimuli in an anesthetized cat. Higher level effects also must be taken into account, but are not modeled by such peripheral mechanisms. For example, Assmann, Nearey, and Hogan [2] have shown effects of preceding and following consonants and speaking rate on the shape and perception of the intervening vowel. Miller [49] has shown significant effect of speaking rate on the perception of both stops and consonants, Repp [56], [58] has shown that detecting doubled stop consonants (*/raged/*, */ragged/*), and consonant clusters [(*/stop/*), (*/sop/*)] depends upon the prior adaptive state. Several of these higher types of processes have been analyzed as part of the larger auditory processing architecture in Fig. 2 of which the present filter is a part.

Appendix
A Laplace Transform Analysis of Transient Detector Properties

In this appendix, we analyze the simplest version of the Grossberg [29] transient detector in (5) and (6). This system is analyzable via Laplace transforms. For any real w, we let $\hat{w} = \int_0^\infty e^{-st} w(t)\,dt$. For systems with zero initial output in both z and y. We can write the Laplace transforms of (5) and (6) as

$$(s+\delta)\hat{x} = \hat{I} - \epsilon\hat{y} \tag{A1}$$
$$(s+\zeta)\hat{y} = \zeta\hat{I}. \tag{A2}$$

It follows that

$$\hat{x} = \hat{I}\left(\frac{1}{s+\delta}\right)\left(1 - \frac{\epsilon\zeta}{s+\zeta}\right). \tag{A3}$$

The impulse response of this system can be written

$$h(t) = ae^{-\delta t} + be^{-\zeta t} \tag{A4}$$

where

$$\begin{aligned} a &= 1 - \frac{\epsilon\zeta}{\zeta-\delta} \\ b &= \frac{\epsilon\zeta}{\zeta-\delta} \\ a+b &= 1. \end{aligned} \tag{A5}$$

The step response g is $\int_0^\infty h(t)\,dt$ to a positive step input. The energy G^2 is $\int_0^\infty h^2(t)\,dt$. By (A4)

$$g = \frac{a}{\delta} + \frac{b}{\zeta} \tag{A6}$$
$$G^2 = \frac{a^2}{2\delta} + \frac{b^2}{2\zeta} + \frac{2ab}{\delta+\zeta}. \tag{A7}$$

Solving for a, and b in terms of g and G, we obtain

$$a = \frac{\delta}{\delta-\zeta}\left[\sqrt{2G^2(\delta+\zeta) - g^2\zeta\delta} - g\zeta\right] \tag{A8}$$
$$b = \frac{\zeta}{\zeta-\delta}\left[\sqrt{2G^2(\delta+\zeta) - g^2\zeta\delta} - g\delta\right]. \tag{A9}$$

This choice of parameters allows us to fix the step response and the energy gain to have desired values for an appropriate choice of rate constants. In simulations, we chose $g = 0$ and $G = 1$ to guarantee a transient response and to normalize the energy. Without loss of generality we can also assume $\zeta > \delta$. Then

$$h(t) = \frac{\sqrt{2(\delta+\zeta)}}{\zeta-\delta}\left(\zeta e^{-\zeta t} - \delta e^{-\delta t}\right). \tag{A10}$$

The frequency response of the system is

$$\hat{h}(\omega) = \frac{\sqrt{2(\delta+\zeta)}i\omega}{(\delta+i\omega)(\zeta+i\omega)} \tag{A11}$$

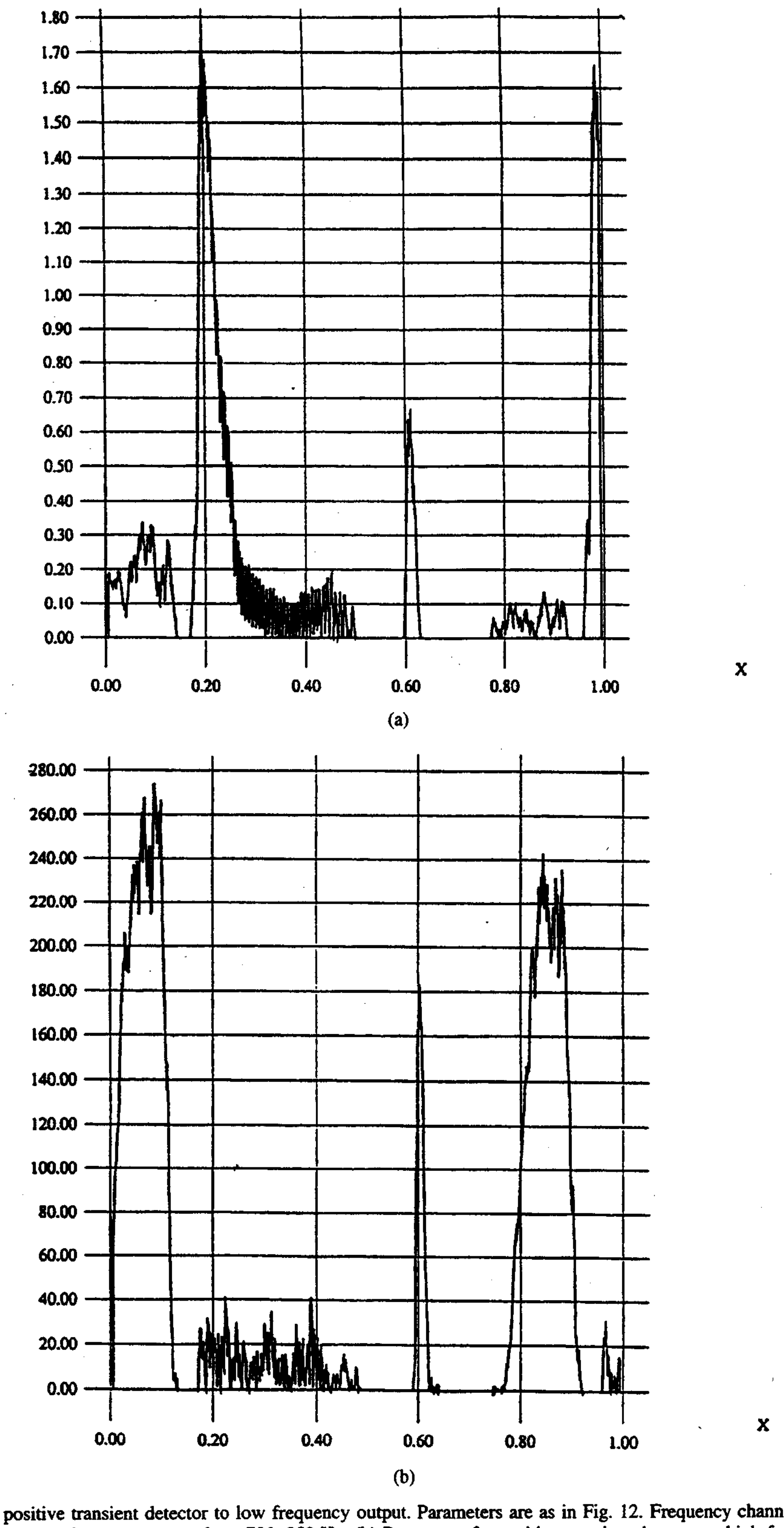


Fig. 13. (a) Response of a positive transient detector to low frequency output. Parameters are as in Fig. 12. Frequency channels between filter 9 and 18 are pooled. This corresponds to a center frequency range from 700–850 Hz. (b) Response of a positive transient detector to high-frequency output. Parameters are as in Fig. 12. Frequency channels between filter 45 and 59 are pooled. This corresponds to a center frequency range from 4100–8800 Hz.

and the modulus of the response is

$$|\hat{h}(\omega)| = |\omega| \sqrt{\frac{2(\delta + \zeta)}{(\delta^2 + |\omega|^2)(\zeta^2 + |\omega|^2)}}. \tag{A12}$$

Thus, at low frequencies, the system acts like a differentiator. At high frequencies, the system acts like a simple lowpass filter with an attenuation of 3 dB per octave, since the falloff is asymptotically linear with frequency. This enables the detector to be sensitive to transients in a prescribed lowpass frequency range. The frequency of the peak response is $\omega_{\max} = \sqrt{\delta\zeta}$.

Our simulations use a discrete impulse invariant representation. Specifically

$$s[n] = e\delta^n + f\zeta^n \tag{A13}$$

for suitable parameters e, f, $\delta < 1$, ζ. This is a discrete version of $h(t)$ as defined in (A4). As in the continuous case, we let g be the step response and G^2 be the energy of the system. Expressing G^2, and g in terms of e, f, δ, ζ we obtain

$$g = e(1-\delta)^{-1} + f(1-\zeta)^{-1} \tag{A14}$$

and

$$\begin{aligned} G^2 = {} & e^2(1-\delta^2)^{-1} + f^2(1-\zeta^2)^{-1} \\ & + 2ef(1-\delta\zeta)^{-1}. \end{aligned} \tag{A15}$$

Let $u = (1-\delta)/(1+\delta)$, $v = (1-\zeta)/(1+\zeta)$. Solving for e and f in terms of g, G^2, δ, ζ yields

$$e = (1-\delta)\left[\frac{\sqrt{G^2(u+v) - g^2uv} - gv}{u-v}\right] \tag{A16}$$

and

$$f = (1-\zeta)\left[\frac{\sqrt{G^2(u+v) - g^2uv} - gv}{v-u}\right]. \tag{A17}$$

Letting $G = 1$ and $g = 0$, one obtains

$$s[n] = \frac{\sqrt{(1+\delta)(1+\zeta)(1-\delta\zeta)}}{\sqrt{2}(\delta-\zeta)}\,[(1-\zeta)\zeta^n - (1-\delta)\delta^n] \tag{A18}$$

if $\zeta < \delta$. This expression can also be written in the form

$$s[n] = \Gamma(\eta, \kappa)[(1-e^{-\kappa})e^{-n\kappa} - (1-e^{-\eta})e^{-n\eta}] \tag{A19}$$

where

$$\Gamma(\eta, \kappa) = \frac{\sqrt{(1+e^{-\eta})(1+e^{-\kappa})[1-e^{-(\eta+\kappa)}]}}{\sqrt{2}(e^{-\eta} - e^{-\kappa})} \tag{A20}$$

$$\eta = -\log\zeta \tag{A21}$$

and

$$\kappa = -\log\delta. \tag{A22}$$

The expression $\Gamma(\eta, \kappa)$ in (A20) is used in (8) above.

The discrete Fourier transform of this impulse response can be written

$$\hat{s}(\omega) = \frac{\sqrt{(1+\delta)(1+\zeta)(1-\delta\zeta)}(1-e^{-i\omega})}{\sqrt{2}(1-\delta e^{-i\omega})(1-\zeta e^{-i\omega})}. \tag{A23}$$

Note the similarity in form of the discrete and continuous impulse responses, as shown by (A23) and (A11). The amplitude of this impulse response can be written

$$\begin{aligned} |\hat{s}(\omega)| = {} & \left|\sin\left(\frac{\omega}{2}\right)\right| \\ & \cdot \left\{\sqrt{\frac{2(1+\delta)(1+\zeta)(1-\delta\zeta)}{[1+\delta^2 - 2\delta\cos(\omega)][1+\zeta^2 - 2\zeta\cos(\omega)]}}\right\}. \end{aligned} \tag{A24}$$

As before, for frequencies near zero the system acts as a differentiator. Also, the system can be set up to attenuate the frequency response at higher frequencies. Equation (A23) shows that the system can be written as a convolution of a first difference followed by two first-order lowpass filters. Therefore, the system has more flexibility than the first difference which forms part of this system. In the discrete case

$$\omega_{\max} = \begin{cases} \cos^{-1}\{1 - 1/2[\delta^{-1/2} - \delta^{1/2}], & \text{if } [\delta^{-1/2} - \delta^{1/2}] \\ \quad [\zeta^{-1/2} - \zeta^{1/2}]\}, & [\zeta^{-1/2} - \zeta^{1/2}] < 4 \\ \pi & \text{otherwise} \end{cases} \tag{A25}$$

where $\omega_{\max}$ is the frequency of maximal response.

This frequency analysis shows that the transient detector is globally "bandpass" in character: The gain of the filter ramps up at low frequencies and falls as the frequency increases. In a scalable lowpass frequency range, the system acts as a differentiator by responding to the differential of intensity. It is sensitive to a relatively narrow band portion of the frequency range, while remaining insensitive to changes in energy at other regions of the short time spectrum.

ACKNOWLEDGMENT

The authors thank C. E. Bradford and D. J. Meyers for their valuable assistance in the preparation of the manuscript.

REFERENCES

[1] D. A. Adams and R. Bisiani, "The Carnegie-Mellon University distributed speech recognition system," *Speech Technol.*, pp. 14–23, Mar./Apr, 1986.

[2] P. F. Assmann, T. M. Nearey, and J. T. Hogan, "Vowel identification: Orthographic, perceptual and acoustic aspects," *J. Acoust. Soc. America*, vol. 71, pp. 975–989, 1982.

[3] I. Boardman and D. Bullock, "A neural network model of serial order recall from short-term memory," in *Proc. Int. Joint Conf. Neural Networks*, Seattle, WA, vol. II, pp. 879–884, 1991.

[4] I. Boardman, S. Grossberg, and M. A. Cohen, "Neural dynamics of perceptual order and context effects for variable-rate CV syllables," Tech. Rep. CAS/CNS-TR-94-037, Boston Univ., Boston, MA, 1994.

[5] G. Bradski, G. A. Carpenter, and S. Grossberg, "Working memory networks for learning temporal order with application to three-dimensional visual object recognition," *Neural Comput.*, vol. 4, pp. 270–286, 1992.

[6] ——, "STORE working memory networks for storage and recall of arbitrary temporal sequences," *Biolog. Cybern.*, vol. 71, pp. 469–480, 1994.

[7] R. Britt and A. Starr, "Synaptic events and discharge patterns of cochlear nucleus cells: Frequency modulated tones," *J. Neurophys.*, vol. 39, pp. 179–194, 1976.

[8] L. H. Carney and T. C. T. Yin, "Temporal coding of resonances by low-frequency auditory nerve fibers: Single-fiber responses and a population model," *J. Neurophys.*, vol. 60, pp. 1653–1677, 1988.

[9] G. A. Carpenter and S. Grossberg, Eds., *Pattern Recognition by Self-Organizing Neural Networks*. Cambridge, MA: MIT Press, 1991.

[10] ——, "Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions," *Trends Neurosci.*, vol. 16, pp. 131–137, 1993.

[11] M. A. Cohen and S. Grossberg, "Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory," *Human Neurobiol.*, vol. 5, pp. 1–22, 1986.

[12] ——, "Masking fields: A massively parallel architecture for learning, recognizing, and predicting multiple groupings of patterned data," *Appl. Opt.*, vol. 26, pp. 1866–1891, 1987.

[13] M. A. Cohen, S. Grossberg, and D. Stork, "Speech perception and production by a self-organizing neural network," in *Evolution, Learning, Cognition, and Advanced Architectures,* Y. C. Lee, Ed. Hong Kong: World Scientific, 1988.

[14] M. A. Cohen, S. Grossberg, and L. L. Wyse, "A spectral network model of pitch perception," *J. Acoust. Soc. Amer.,* vol. 98, pp. 862–879, 1995.

[15] R. E. Crochiere and L. R. Rabiner, *Multivariate Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[16] R. Daniloff, *The Physiology of Speech and Hearing: An Introduction.* Englewood Cliffs, NJ: Prentice-Hall, 1980.

[17] E. de Boer and H. R. de Jongh, "On cochlear encoding: Potentialities and limitations of the reverse correlation technique," *J. Acoust. Soc. Amer.,* vol. 63, pp. 115–135, 1978.

[18] P. C. Delattre, A. Lieberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Amer.,* vol. 27, pp. 769–773, 1955.

[19] P. C. Delattre, A. M. Lieberman, F. S. Cooper, and L. J. Gerstman, "Formant transitions and loci as acoustic correlates of place of articulation in American fricatives," *Studia Linguistica,* vol. 16, pp. 104–121, 1962.

[20] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve I: Vowel-like sounds," *J. Acoust. Soc. Amer.,* vol. 75, pp. 866–878, 1984a.

[21] ——, "Speech coding in the auditory nerve II: Processing schemes for vowel-like sounds," *J. Acoust. Soc. Amer.,* vol. 75, pp. 879–886, 1984b.

[22] ——, "Speech coding in the auditory nerve III: Voiceless fricative consonants," *J. Acoust. Soc. Amer.,* vol. 75, pp. 887–896, 1984c.

[23] ——, "Speech coding in the auditory nerve IV: Sounds with consonant like dynamic characteristics," *J. Acoust. Soc. Amer.,* vol. 75, pp. 897–906, 1984d.

[24] ——, "Speech coding in the auditory nerve V: Vowels in background noise," *J. Acoust. Soc. Amer.,* vol. 75, pp. 907–918, 1984e.

[25] O. Ghitza, "Temporal nonplace information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment;" *J. Phonet.,* vol. 16, pp. 109–123, 1988.

[26] D. A. Godfrey, N. Y. S. Kiang, and B. E. Norris, "Single unit activity in the posteroventral cochlear nucleus of the cat," *J. Compar. Neurol.,* vol. 162, pp. 247–268, 1975.

[27] K. K. Govindarajan, S. Grossberg, L. L. Wyse, and M. A. Cohen, "A neural network model of auditory scene analysis and source segregation," Tech. Rep. CAS/CNS-TR-94-039, Boston Univ., Boston, MA, 1994.

[28] S. Greenberg, "Speech processing: Auditory models." in *Encyclopedia of Language and Linguistics,* R. E. Asher and J. M. Y. Simpson, Eds. New York: Pergamon, 1994, pp. 4206–4227.

[29] S. Grossberg, "Neural pattern discrimination," *J. Theoret. Biol.,* vol. 27, pp. 291–337, 1970, repr. G. A. Carpenter and S. Grossberg, Eds., *Pattern Recognition by Self-Organizing Neural Networks.* Cambridge, MA: MIT Press, 1991, pp. 111–156.

[30] ——, "Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors," *Biolog. Cybern.,* vol. 23, pp. 121–134, 1976.

[31] ——, "A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans," in *Progress in Theoretical Biology,* vol. 5, R. Rosen and F. Snell, Eds. New York: Academic, 1978a. Repr. S. Grossberg, *Studies of Mind and Brain.* Boston, MA: Kluwer, 1982, pp. 498–639.

[32] ——, "Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models?" *J. Math. Psychol.,* vol. 3, pp. 199–219, 1978b.

[33] ——, "How does a brain build a cognitive code?" *Psycholog. Rev.,* vol. 87, pp. 1–51, 1980.

[34] ——, "The adaptive self-organization of serial order in behavior: Speech, language, and motor control," in *Pattern Recognition by Humans and Machines, Vol. 1: Speech Perception.* E. C. Schwab and H. C. Nusbaum, Eds. New York: Academic, pp. 187–294, 1986. Repr. S. Grossberg, Ed., *The Adaptive Brain, II: Vision, Speech, Language, and Motor Control.* Amsterdam, The Netherlands: Elsevier/North-Holland, 1987, pp. 311–400.

[35] ——, "The attentive brain." *Amer. Scientist,* vol. 83, pp. 438–449, 1995.

[36] S. Grossberg, I. Boardman, and M. A. Cohen, "Neural dynamics of variable-rate speech categorization," Tech. Rep. CAS/CNS-TR-94-038, Boston Univ., Boston, MA, 1996.

[37] S. Grossberg and G. O. Stone, "Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance," *Psycholog. Rev.,* vol. 93, pp. 46–74, 1986.

[38] J. M. Heinz and K. N. Stephens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Amer.,* vol. 33, pp. 589–596, 1961.

[39] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Mag.,* vol. 9, pp. 21–69, Apr. 1992.

[40] L. B. Jackson, *Digital Filters and Signal Processing.* Boston, MA: Kluwer, 1986.

[41] B. M. Johnstone, R. Patuzzi, and G. K. Yates, "Basilar membrane measurements and the travelling wave," *Hearing Res.,* vol. 22, pp. 147–153, 1981.

[42] N. Y. S. Kiang, T. Watanabe, E. E. Thomas, and L. F. Clark, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve.* Cambridge, MA: MIT Press, 1965.

[43] D. H. Klatt, "Speech processing strategies based on auditory models," in *The Representation of Speech in the Peripheral Auditory System,* R. Carlson and B. Granström, Eds. New York: Elsevier, 1982, pp. 181–196.

[44] T. Kohonen, *Self-Organization and Associative Memory.* New York: Springer-Verlag, 1984.

[45] C. Lieberman, "Unpublished cat cell response data," communicated by B. Delgutte, 1988.

[46] P. Lieberman and S. E. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics.* Cambridge, U.K.: Cambridge Univ. Press, 1988.

[47] C. von der Malsburg, "Self-organization of orientation sensitive cells in the striate cortex," *Kybernetik,* vol. 14, pp. 85–100, 1973.

[48] V. Mann and S. D. Soli, "Perceptual order and the effect of vocalic context on fricative perception," *Perception Psychophys.,* vol. 49, pp. 399–411, 1991.

[49] J. L. Miller, "Contextual effects in the discrimination of stop consonants and semivowels," *Perception Psychophys.,* vol. 28, pp. 93–95, 1980.

[50] J. L. Miller and A. M. Lieberman, "Some effects of later-occurring information on the perception of stop consonant and semivowel," *Perception Psychophys.,* vol. 25, pp. 457–465, 1979.

[51] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* New York: Academic, 1989.

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

[53] R. R. Pfeiffer, "Classification of response patterns of spike discharges for units in the cochlear nucleus: Tone-burst stimulation," *Exp. Brain Res.,* vol. 1, pp. 220–235, 1966.

[54] J. O. Pickles, *An Introduction to the Physiology of Hearing, Second Edition.* New York: Academic, 1988.

[55] D. A. Pisoni, T. D. Carrell, and S. J. Gans, "Perception of the duration of rapid spectrum changes in speech and nonspeech signals," *Perception Psychophys.,* vol. 34, pp. 314–322, 1983.

[56] B. H. Repp, "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants," *Perception Psychophys.,* vol. 24, pp. 471–485, 1978.

[57] ——, "A range-frequency effect on perception of silence in speech," Status Rep. Speech Research SR-61, Haskins Labs, New Haven, CT, pp. 151–165, 1980.

[58] ——, "Bidirectional contrast effects in the perception of VC-CV sequences," Tech. Rep. SR-63/64, Haskins Labs, New Haven, CT, 1989.

[59] W. S. Rhode and P. H. Smith, "Encoding timing and intensity in the ventral cochlear nucleus of the cat," *J. Neurophys.,* vol. 56, pp. 261–287, 1986a.

[60] ——, "Physiological studies on neurons in the dorsal cochlear nucleus of the cat," *J. Neurophys.,* vol. 56, pp. 287–307, 1986b.

[61] D. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.,* vol. 8, pp. 14–38, 1991.

[62] M. B. Sachs, R. L. Winslow, and C. Blackburne, "Representation of speech in the auditory periphery," in *The Auditory Periphery,* G. M. Edelman, W. E. Gall, and W. M. Cowan, Eds. New York: Wiley, 1988, pp. 747–775.

[63] M. B. Sachs and E. D. Young, "Encoding of steady state vowels in the auditory nerve: Representations in terms of discharge rate," *J. Acoust. Soc. Amer.,* vol. 66, pp. 470–479, 1979.

[64] C. E. Schreiner and G. Langner, "Periodicity coding in the inferior colliculus of the cat, II: Topographical organization," *J. Neurophys.,* vol. 60, pp. 1823–1840.

[65] S. A. Shamma, "Neural networks in speech processing and recognition," in *Proc. IEEE Int. Conf. Neural Networks.* M. Caudill and C. Butler, Eds., San Diego, CA, 1987, pp. 397–405.

[66] L. L. M. Votgen, "Pure tone masking: A new result from a new method." in *Facts and Models in Hearing.* E. Zwicker and E. Terhardt, Eds. New York: Springer-Verlag, 1974.

[67] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers," *J. Acoust. Soc. Amer.,* vol. 66, pp. 1381–1403, 1979.

[68] G. Zweig, R. Lipes, and J. R. Pierce, "The cochlear compromise," *J. Acoust. Soc. Amer.,* vol. 59, pp. 975–982, 1976.

Michael A. Cohen (A'90) was born on November 17, 1947 and grew up in Queens, NY. He received an undergraduate degree in pure mathematics, with a minor in history, from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1969.

He left MIT to study psycholinguistics at Harvard University, Cambridge. He then switched to experimental psychology at Harvard, focusing on auditory and visual psychophysics. After a brief post-doctoral fellowship in New York, he became a Research Assistant Professor at the Department of Cognitive and Neural Systems, Boston University (BU), Boston, MA, where he later obtained tenure. During his first decade at BU, he was also Chief Systems Administrator for the department. His published work includes areas as diverse as auditory psychoacoustics, visual psychoacoustics, neural networks, dynamical systems, child language development, subjective probability, and measurement theory.

Stephen Grossberg (M'96) is Wang Professor of Cognitive and Neural Systems at Boston University (BU), Boston, MA, and Professor of mathematics, psychology, and biomedical engineering at BU. He is the founder and Director of the Center for Adaptive Systems, as well as the founder and Chairman of the Department of Cognitive and Neural Systems. He founded and was first President of the International Neural Network Society. His work focuses on the design principles and mechanisms that enable the behavior of individuals to adapt successfully in real time to unexpected environmental changes. He pioneered competitive learning and self-organizing feature maps, adaptive resonance theory, masking fields, gated dipole opponent processes, associative outstars and instars, associative avalanches, nonlinear cooperative-competitive feedback networks, boundary contour and feature contour systems, and vector associative maps. Such models have been used both to analyze and predict interdisciplinary data about the mind and brain, as well as to suggest novel architectures for technological applications.

Dr. Grossberg founded and is co-editor-in-chief of the Neural Network Society's journal, *Neural Networks*. He is also an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS, *Neural Computation*, and *Brain Research*, among other journals. He received the 1991 IEEE Neural Network Pioneer award, the 1992 INNS Leadership Award, and the 1992 Thinking Technology Award of the Boston Computer Society.