Journal of Phonetics (1988) **16,** 55–76

A joint synchrony/mean-rate model of auditory speech processing

Stephanie Seneff

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

This paper describes a speech processing system that is based on properties of the human auditory system. A bank of critical-band filters defines the initial spectral analysis. Filter outputs are processed by a model of the nonlinear transduction stage in the cochlea, which accounts for such features as saturation, adaptation and forward masking. The parameters of the model were adjusted to match existing experimental results of the physiology of the auditory periphery. The output of this model is delivered to two parallel channels, each of which produces spectral representations appropriate for distinct subtasks of a speech recognition system. One path yields an overall energy measure for each channel that can be identified with the average rate of neural discharge. The outputs of this path appear to be useful for locating acoustic events and assigning segments to broad phonetic categories. In the other path, the extent of dominance of periodicities at each channel's center frequency is captured by a synchrony measure, which yields a spectral representation with enhanced spectral contrast, relative to the mean-rate spectrogram. The outputs of this stage show distinct formant peaks during sonorant regions, with smooth transitions over time, as well as preserving spectral prominences in the high-frequency region for fricatives and stops.

1. Introduction

The human auditory system, together with its central connections, is a speech recognizer with excellent performance. If a computational model could be designed that adequately reflects the transformations occurring in the auditory pathway, the resulting spectral representations should be superior to representations based on non-biological criteria commonly used in computer speech recognition algorithms. Due to a wealth of physiological data, particularly at the level of the auditory nerve, it is now possible to characterize many of the transformations that occur in the auditory periphery. Although many features of the auditory system have been characterized quite explicitly, it is still a difficult task to design a computer system that achieves a comparable level of performance, particularly when computational issues are taken into account.

Having chosen a design for the peripheral auditory model, the speech researcher is confronted with the task of modeling the processing which occurs in the more central regions of the auditory pathway. Beyond the periphery, the physiological properties of

0095–4470/88/010055 + 22 $03.00/0 © 1988 Academic Press Limited

the system are not nearly so well-defined, and therefore the criteria for the design are open to considerably more speculation. Instead of trying to match the responses of a given system, one can only try to create a reasonable processing strategy that yields "promising" representations given a general knowledge about neural processing and about the important features to be preserved in speech signals.

The following section briefly reviews those properties of the auditory periphery which are relevant to the processing of speech. In Section 3 an auditory model for speech processing is described in detail and referred to the relevant physiological data discussed in Section 2. Section 4 compares the results produced by the model with physiological responses for a number of different experimental paradigms. The final section describes a model for synchrony detection and illustrates various outputs of the computer model for speech signals.

2. Brief review of relevant features of auditory system

Auditory physiologists have gathered considerable data describing the response of mammalian auditory-nerve fibers to spectrally simple (Kiang, Watanabe, Thomas & Clark, 1965; Johnson, 1974, 1980; Smith & Zwislocki, 1975) as well as more complex signals, such as synthetic speech (Young & Sachs, 1979; Sachs & Young, 1980). From these data it is clear that some form of frequency analysis is performed and that this operation is heavily influenced by such nonlinearities as response saturation and both long- and short-term adaptation.

The dynamics of the response to non-steady-state signals are important aspects to be captured by any model of auditory processing. The "instantaneous" discharge rate of auditory-nerve fibers is often significantly highest during the initial 15 ms of acoustic stimulation and decreases thereafter, until it reaches a steady-state level approximately 50 ms after signal onset. This decrease in response rate is referred to as "adaptation" (Smith & Zwislocki, 1975). Typically, there is a very rapid initial decay in rate immediately after onset, followed by a slower decay to a steady-state level[1]. The "rapid adaptation" is attributed in part (if not in full) to the refractory property of auditory-nerve fibers (Johnson & Swami, 1983). The slower, "short-term" adaptation is attributed to a depletion of neurotransmitter in the synaptic region between the inner hair cell and associated nerve fibers (Eggermont, 1973). Another important response property, possibly related to adaptation, is "forward masking". This occurs when the response to a particular sound is diminished as a consequence of a preceding, usually considerably more intense signal (Harris & Dallos, 1979).

In addition to the dynamics of the gross temporal *envelope* of the response discussed above, another important aspect of neural firing behavior involves the detailed time course of the probabilistic response to each cycle of the input signal. Auditory-nerve fibers tend to fire in a phase-locked fashion to low-frequency periodic stimuli. In other words, the intervals between nerve firings tend to be integral multiples of the stimulus period.

The detailed temporal patterns in the neural response are a potential source of more specific information about the frequencies present in the input signal. In response to

[1]The low- and (to a lesser extent) medium-spontaneous rate fibers do not exhibit the rapid adaptation typical of the more numerous high-spontaneous units (Rhode & Smith, 1986). However, the present model attempts to simulate only the more general characteristics of auditory-nerve fiber response and thus will ignore such differences among nerve fiber populations.

sinusoidal stimuli, the spectrum of the response pattern contains energy at the input frequency and its harmonics. The harmonics are introduced primarily as a consequence of half-wave rectification. Fibers responsive to the high-frequency components of a signal tend to synchronize only to the modulation envelope of the signal, which is correlated with the signal's fundamental frequency. Thus, there will typically be some degree of synchronization to the fundamental frequency in the response of high-frequency fibers despite the fact that they are incapable of phase-locking to the frequency components lying within their response areas. Such envelope synchrony may be useful for pitch processing (Delgutte, 1980; Delgutte & Kiang, 1984*a*).

2.1. *Responses to speech-like stimuli*

Only recently have researchers begun to examine the nerve fiber response characteristics to complex stimuli that more closely resemble natural speech. Noteworthy are the studies by Young & Sachs (1979) and Sachs & Young (1980) on the responses of cat auditory-nerve fibers to steady-state synthetic vowels, and the work by Delgutte (1980), Miller & Sachs (1983), Sinex & Geisler (1983) and Delgutte & Kiang (1984*a*, *b*, *c*), on the responses to other speech-like stimuli such as formant transitions, fricatives and stop-consonants. These researchers observed response patterns that were consistent, in many ways, with those obtained with less complex stimuli.

Young and Sachs were particularly interested in addressing the issue of whether discharge rate alone is sufficient for vowel identification, or whether some form of synchrony measure is required at a higher stage in the auditory system to determine the formant frequencies. They studied a large population of fibers, and computed the mean-rate response, as well as period histograms to synthetic vowel stimuli presented over a range of sound pressure levels. They found that the formant information was almost completely obliterated from the rate response of most fibers at the higher amplitudes, due to the saturation of their discharge rate[2].

Young and Sachs also tested the adequacy of a synchronized response measure for vowel representation. The measure, "average localized synchronized rate" (ALSR) was evaluated for the frequencies corresponding to harmonics of the fundamental. It is computed by averaging the spectral amplitude of the period histograms at a given frequency, nf_0, over a group of fibers whose characteristic frequencies (CFs) are close to that harmonic. This representation yields a more robust representation of the formants over a wide range of amplitudes. However, the ALSR measure is also sensitive to spurious peaks in the spectral representation which are the consequence of cochlear nonlinearities such as rectification. These nonlinearities introduce substantial energy at the second harmonic of a strong peak. Srulovicz & Goldstein (1983) have explored, within a theoretical framework, a similar model for a "central spectrum" using an approach which complements the experimental results of Young and Sachs.

3. Peripheral auditory model

The analysis system consists of a set of 40 independent channels which collectively cover the frequency range from 130 to 6400 Hz. The bandwidth of the channels is approximately

[2]However, the rate-place profiles for the low- and, to a lesser extent, the medium-spontaneous fibers were shown to contain some information relevant to the formant frequencies (see Geisler, 1988, and Sachs, Blackburn & Young, 1988, in this volume).

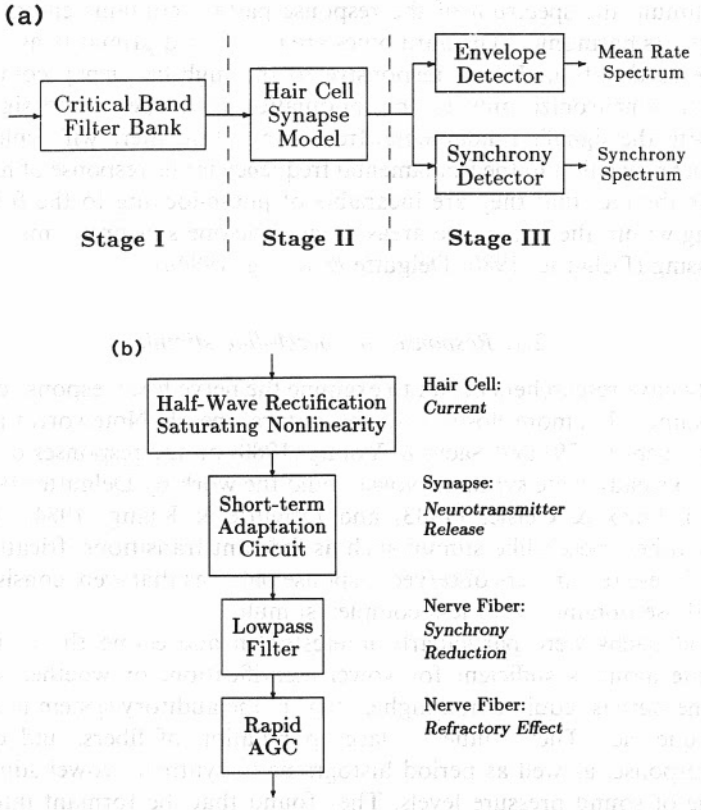

Figure 1. (a) The computer model. (b) The subcomponents of Stage II with suggested auditory system affiliations indicated at right.

0.5 Bark[3]. Although a larger number of channels would provide superior spatial resolution of the cochlear output, the amount of computation time required would be increased significantly. Thus, practical considerations of the model's design have kept the number of channels to the minimum required to provide the resolution required to produce a clear representation of the speech spectrum. In the future it may be possible to increase the number of channels and keep the computation time down by implementing the model in hardware.

The model is illustrated in Fig. 1(a). Each channel consists of a linear critical-band filter, followed by a nonlinear stage (Stage II), intended to capture the prominent features of the transformation from basilar membrane vibration to the probabilistic response properties of auditory-nerve fibers. The Stage II outputs include the detailed waveshape of the probabilistic response to individual cycles of the input stimulus. The nerve responses are never reduced to spike trains, as would be the case for single neurons. Rather, the outputs represent the probability of firing as a function of time for an ensemble of similar fibers acting as a group. The outputs are delivered to two parallel,

[3]A Bark corresponds to the width of one critical band, which is a unit of frequency resolution and energy integration derived from psychophysical experiments. A critical band is equal to approximately $f_c/6$ for frequencies greater than 1 kHz and becomes somewhat broader (on a logarithmic scale) in the low-frequency range. A concise definition is provided by Zwicker (1961).

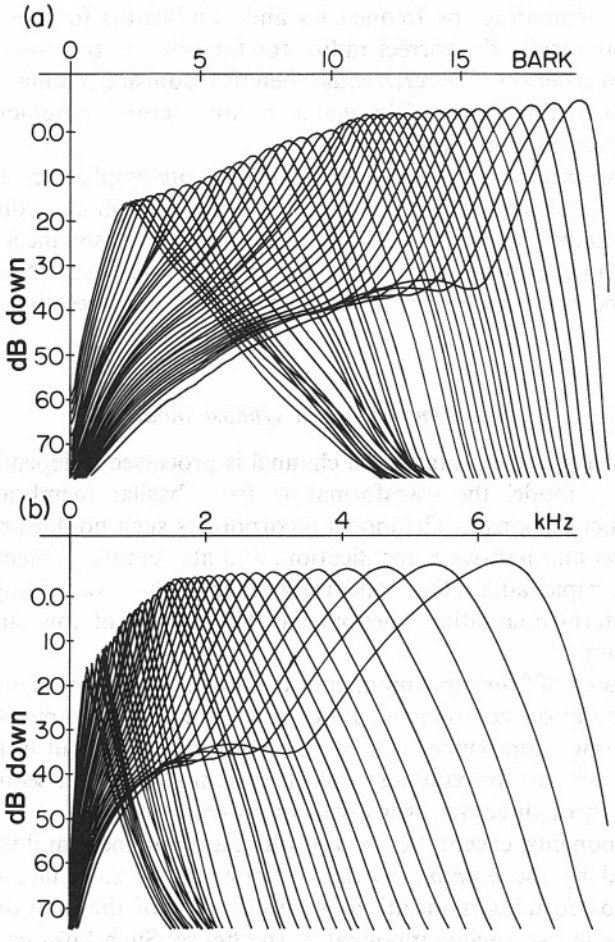

Figure 2. Frequency response characteristics of the filter bank plotted along (a) a Bark scale (Zwicker, 1961) and (b) a linear frequency scale.

non-interacting modules. One module determines the envelope amplitude, corresponding to the average discharge rate response. The other module measures the extent to which information near the center frequency (CF) of the linear filter dominates the output (i.e. determines the "synchronous response").

3.1. *Filter bank design*

The frequency response characteristics of the filters are shown in Fig. 2(a), plotted on a Bark-frequency scale (Zwicker, 1961), and in Fig. 2(b) on a linear scale. The analog speech signal is initially band-limited to 6.5 kHz and sampled at 16 kHz. In the interest of efficiency the filters were implemented as a cascade of complex high-frequency zero pairs (anti-resonances), with taps to individual tuned resonators after each zero pair. The high-frequency zeros serve to filter out energy above resonance, and help to produce a steep cutoff on the high-frequency side of the filter. The high-CF filters have broad low-frequency tails, such as are observed in neural data (Kiang *et al.*, 1965). The filters

were designed by estimating the frequencies and bandwidths for the zeros, and then determining automatically the correct radius for the poles in the z-plane to match the critical-bandwidth criterion. The zeros could then be readjusted manually to improve the match to the desired filter shape. The details of this iterative interactive filter design process are discussed in Seneff (1984, 1985).

In traditional spectral analysis, speech is typically pre-emphasized prior to Fourier analysis. Some form of pre-emphasis can also be justified from an auditory standpoint. It has been determined experimentally that broad outer-ear resonances should result in a 10–20 dB boost in energy between about 1.5 and 5.0 kHz (Yost & Nielsen, 1977). The gains of the filters in the model are set so as to reflect these resonances, as shown in Fig. 2.

3.2. *Inner-hair-cell/synapse model*

Following the linear-filtering stage, each channel is processed independently through a nonlinear stage to model the transformation from basilar membrane vibration to auditory-nerve fiber responses. The model incorporates such nonlinearities as dynamic range compression and half-wave rectification, and also captures effects such as short-term adaptation, rapid adaptation, and forward masking. No attempt was made to model any long-term adaptation phenomena. The output of this stage represents a probability of firing.

The model consists of four subcomponents, as shown in Fig. 1(b): a half-wave rectifier, a short-term adaptation component, a lowpass filter, and a rapid Automatic Gain Control (AGC). Each component will be described in more detail in this section. The numerical values used for the parameters of the system are given in Section 4, since these were determined through comparisons with physiological data.

All of the components, except the lowpass filter, are nonlinear and therefore the final output is affected by the *ordering* of the components. A particular ordering can be justified in part by forming associations with elements of the auditory apparatus, as suggested to the right of each component in the figure. Such links can also aid in the design of each individual component.

The hair-cell current response, as measured for amphibians, shows a distinct directional sensitivity (Hudspeth & Corey, 1977). It is not clear if the electrical current is a direct link in the response mechanism; nonetheless, it is tempting to assume that half-wave rectification first occurs in the hair cell and, hence, this is the first component in the model. There seems to be no evidence for short-term adaptation in hair-cell current or voltage responses; therefore it is generally assumed that this effect is introduced in the synaptic region between the hair cell and the nerve fiber (Eggermont, 1973). The logical ordering is therefore to place this component second.

The AGC is assumed to be affiliated with the refractory phenomenon of nerve fibers; therefore, this component should be placed late in the series. Such an affiliation implies that the rapid-adaptation component of responses to onsets is due to the refractory phenomenon, a hypothesis proposed by Johnson & Swami (1983).

The lowpass filter is associated with the gradual loss of synchrony in nerve-fiber responses as stimulus frequency is increased. There are probably several loci where further synchrony loss is introduced; for example, ion diffusion can be viewed as a lowpass process. The filter must follow the half-wave rectifier, because it only makes sense after signal energy has been preserved through a d.c. component. If the filter

precedes the adaptation circuit, the time constants of adaptation become significantly dependent on signal frequency. Therefore, it was decided to place the filter just before the AGC.

The model for the instantaneous half-wave rectifier is defined mathematically as follows:

$$\begin{aligned} y &= 1 + A \tan^{-1} Bx \quad x > 0 \\ &= e^{ABx} \quad\quad\quad\quad x \leqslant 0 \end{aligned} \tag{1}$$

The parameter B can be viewed as an input gain, or, alternatively, as a mechanism for setting the operating range of the channel. This function is exponential for negative signals, linear but shifted (by a "spontaneous" rate of unity magnitude) for small positive signals, and compressive for larger signals, saturating at $1 + A\pi/2$. It is based on the measured hair-cell current responses as a function of a fixed displacement of the cilia as determined in the frog saculus by Hudspeth & Corey (1977).

The model for short-term adaptation is very similar to one proposed by Goldhor (1985). It consists of two separate mechanisms that influence the concentration of a substance, which could be thought of as a neurotransmitter or an ion. A model "membrane" allows flow of a supply from a source region at a rate that is proportional to the concentration gradient across the membrane, with a proportionality constant, μ_a. However, channels in this membrane are closed whenever the concentration in the supply region is too small (i.e. when the concentration gradient is *negative*). The substance is also lost through natural decay at a rate that is proportional to its concentration within the region, with a proportionality constant μ_b. Mathematically, the process can be expressed as follows:

$$\begin{aligned} dC(t)/dt &= \mu_a[S(t) - C(t)] - \mu_b C(t) \quad C(t) < S(t) \\ &= -\mu_b C(t) \quad\quad\quad\quad\quad\quad\quad\quad C(t) \geqslant S(t) \end{aligned} \tag{2}$$

where $C(t)$ is the concentration of the substance within the region, and $S(t)$ is the concentration in the source region. The output of this system is the flow rate across the membrane, $\mu_a[S(t) - C(t)]$, which controls the probability of firing of the nerve fiber. A discrete realization is achieved by approximating d/dt by a first difference in time.

Goldhor showed that such a model, when applied using the *envelope* of the stimulus as the source concentration, $S(t)$, obeys certain linear response properties of short-term adaptation that have been observed for auditory data (Smith & Zwislocki, 1975) (see Section 4 for details). When a high-amplitude signal turns on abruptly, the flow rate is initially very high and then decreases exponentially, with a time constant, $\tau_1 = 1/(\mu_a + \mu_b)$, to a steady-state value. After the signal is turned off, the concentration gradient becomes negative, and the flow rate remains zero until $C(t)$ decays (exponentially with a time constant $\tau_2 = 1/\mu_b$) to the spontaneous concentration level. Thus, the time constant for recovery after offsets is longer than that for adaptation after onsets, a feature which also resembles the auditory-nerve response (Harris & Dallos, 1979).

Our system uses the same model, except that the detailed cycle-by-cycle behavior of the input signal is preserved in $S(t)$. In this case, the channel opens and closes for each period of the stimulus, and an adapted response is obtained only after the amount of substance gained while $S(t)$ is greater than $C(t)$, is exactly equal to the amount lost during the remaining portion of the cycle. One consequence is that the effective time constant

for adaptation lies somewhere between the "open" time constant, τ_1, and the "closed" time constant, τ_2. The time constant for recovery, on the other hand, remains equal to τ_2.

The output of the adaptation stage is next processed through a lowpass filter that achieves two important effects: it reduces synchrony to high-frequency stimuli and it smooths the square-wave shape encountered in the half-wave response for saturating stimuli. The lowpass filter was realized as a cascade of n_{LP} leaky integrators, each with an identical time constant τ_{LP}. The two parameters, n_{LP} and τ_{LP}, were adjusted to match

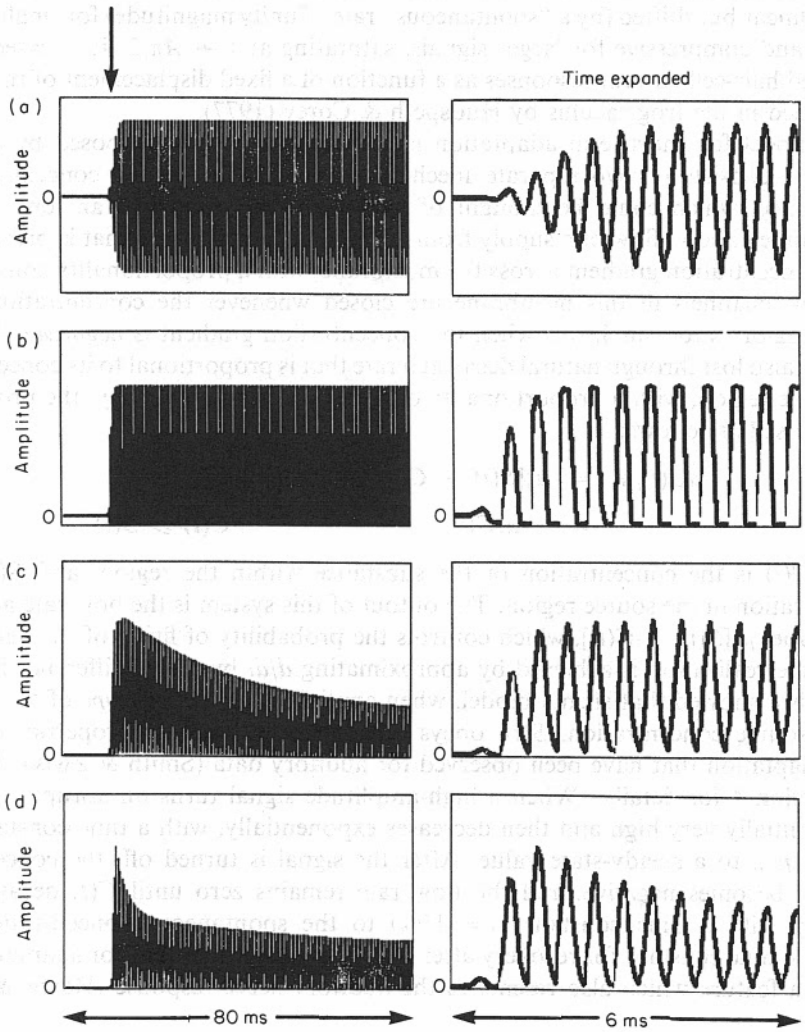

Figure 3. Responses at CF for the intermediate stages of the inner-hair-cell/synapse model in response to a 2-kHz signal presented at a high sound-pressure level: (a) after passing through a critical band filter, (b) after half-wave rectification, (c) after short-term adaptation and lowpass filtering, and (d) after the AGC. The arrow marks the center of the time-expanded region on the right.

available data on synchrony as a function of frequency (Johnson, 1974). The equation in the discrete domain for the resulting transfer function is:

$$H(z) = \left(\frac{1 - \alpha}{1 - \alpha z^{-1}}\right)^{n_{LP}}, \tag{3}$$

where α is the pole location on the real axis of the z-plane such that $\alpha^n = \exp(-1)$ at a sample count, n, corresponding to τ_{LP} ms.

The final component is the rapid AGC, which is defined as follows:

$$y[n] = \frac{x[n]}{1 + K_{AGC}\langle x[n]\rangle}, \tag{4}$$

where K_{AGC} is a constant and $\langle\ \rangle$ symbolizes "expected value of", obtained by processing $x[n]$ through the first-order lowpass filter, with time constant τ_{AGC}. This equation resembles in form the formula obtained theoretically by Johnson & Swami (1983) as a steady-state solution for a simple model of the refractory effect, where it is assumed that a response is locked out for a time interval Δ after a spike occurs:

$$y(t) = \frac{x(t)}{1 + \int_{t-\Delta}^{t} x(\alpha)d\alpha}. \tag{5}$$

Figure 3 shows the outputs of intermediate stages of the 2-kHz channel in response to a high-amplitude tone at CF. The envelope of the response over a long time interval is shown on the left, and the detailed wave shapes near tone onset are shown on the right. Figure 3(a) shows the response after only the linear filter of Stage I. Figure 3(b) shows the response after the instantaneous half-wave rectifier. The square-wave shapes introduced here are lost after the lowpass filter. The effects of the short-term adaptation component are apparent in the envelope response on the left in Fig. 3(c). The final AGC further alters the dynamics of the onset, to produce a trend quite typical of auditory-nerve fibers, as shown in Fig. 3(d).

4. Comparison of the model with physiological data

The system described above contains a number of parameters that can be adjusted according to a specific set of criteria based on relevant physiological data. The degree to which the model agrees with the relevant physiology provides a measure of the system's ability to adequately describe the essential properties underlying speech coding in the auditory periphery. The following physiological properties are considered to be significant with respect to speech representation in the auditory nerve:

(1) Temporal envelope of nerve-fiber discharge rate as a function of signal amplitude level, particularly during the initial 40 ms following stimulus onset (Delgutte, 1980).

(2) Forward masking effects as a function of masker sound-pressure level (Delgutte, 1980).

(3) Period histogram responses in steady-state conditions for single-formant stimuli, as a function of stimulus amplitude (Delgutte, 1980).

(4) Dynamic properties of discharge response to amplitude increments (Smith & Zwislocki, 1975).

(5) Synchrony falloff characteristics as a function of signal frequency (Johnson, 1974).

TABLE I. Fixed parameter values used for experiments

Half-wave			Adaptation		Lowpass		AGC	
A	B	G_{HW}	τ_1	τ_2	τ_{LP}	n_{LP}	τ_{AGC}	K_{AGC}
10	65	2.35	15 ms	120 ms	0.04 ms	4	3 ms	0.002

The parameters of the system were adjusted to match all of the above criteria as well as possible. Several iterations through the matching process were necessary to achieve convergence. Some surprising results emerged from the exercise. Most remarkable was that the τ_2 parameter of the Goldhor adaptation model had to be set to a much larger value than was anticipated in order to match the forward-masking data. Another discovery was that although the short-term adaptation component and the AGC component interact in a complex way, it is possible to set their parameters so that the equal-increment criterion imposed by the Smith and Zwislocki paradigm is reasonably well matched. Each of the above criteria is discussed in turn. For each example the relevant physiological data are compared with the response of the model.

In all instances the parameters of the model, as empirically determined, were set at fixed values, which are shown in Table I. Parameter B of the half-wave rectifier, an input gain term, is based on the assumption that the input speech signal has been normalized to a maximum amplitude of 1.0. The output of the half-wave rectifier was multiplied by a gain term, G_{HW}, which was adjusted to yield a final output that could be equated with firing rate. The lowpass filter has a very gradual falloff as a function of frequency. The response is down by 3 dB at 2 kHz, by 9 dB at 4 kHz and by 13 dB at 6 kHz.

4.1. *Tone onsets*

Delgutte (1980) plotted the envelope of the discharge pattern as a function of time in response to a sinusoidal signal presented over a large range of sound-pressure levels (Fig. 4). The experimental paradigm was simulated for the computer model and the resulting responses are shown in the right-hand column. Onset response characteristics are largely dominated in the model by the parameters of the rapid AGC component.

4.2. *Forward masking*

Delgutte's (1980) plot for a forward masking experiment are shown in Figure 5 (left), along with the results of the computer model (right). The plots are given as a function of adapter sound-pressure level, with the test-tone level held fixed. The main controlling factor of forward masking in the model is τ_2 of the short-term adaptation model.

4.3. *Formant period histograms*

Delgutte (1980) obtained plots of the period histograms of steady-state responses to a single-formant, vowel-like stimulus (i.e. a pulse train of frequency F_0 was passed through a resonator whose center frequency was set to F_1). Figure 6 compares the period histograms from a fiber (CF = formant frequency) with those produced by the model for signals presented over a wide range of sound-pressure levels. For both the physiological data and the model output, the bandwidth of the response appears to be broader

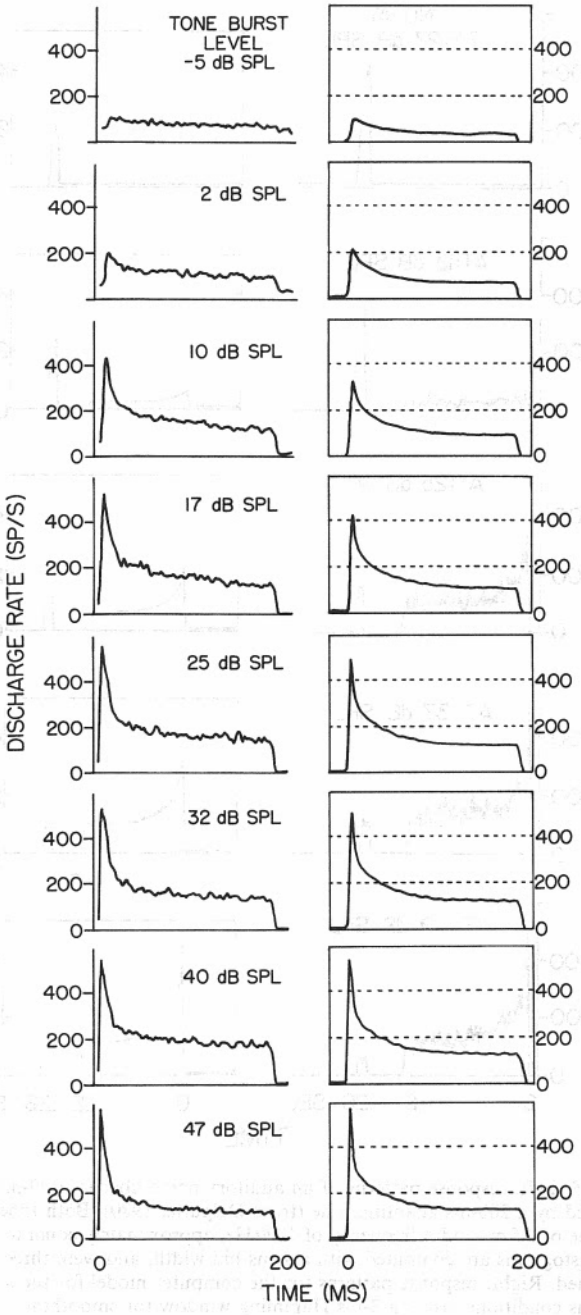

Figure 4. Response patterns of an auditory-nerve fiber to a tone burst as a function of signal amplitude (from Delgutte, 1980). The 180-ms burst has a rise/fall time of 2.5 ms, and a frequency of 770 Hz, approximately equal to the fiber CF. The post-stimulus-time (PST) histogram was computed with a bin width of 1.4 ms and then smoothed with a three-point smoother. Response patterns generated by the model for the same stimulus conditions. The response was smoothed with a 4.2-ms Hamming window.

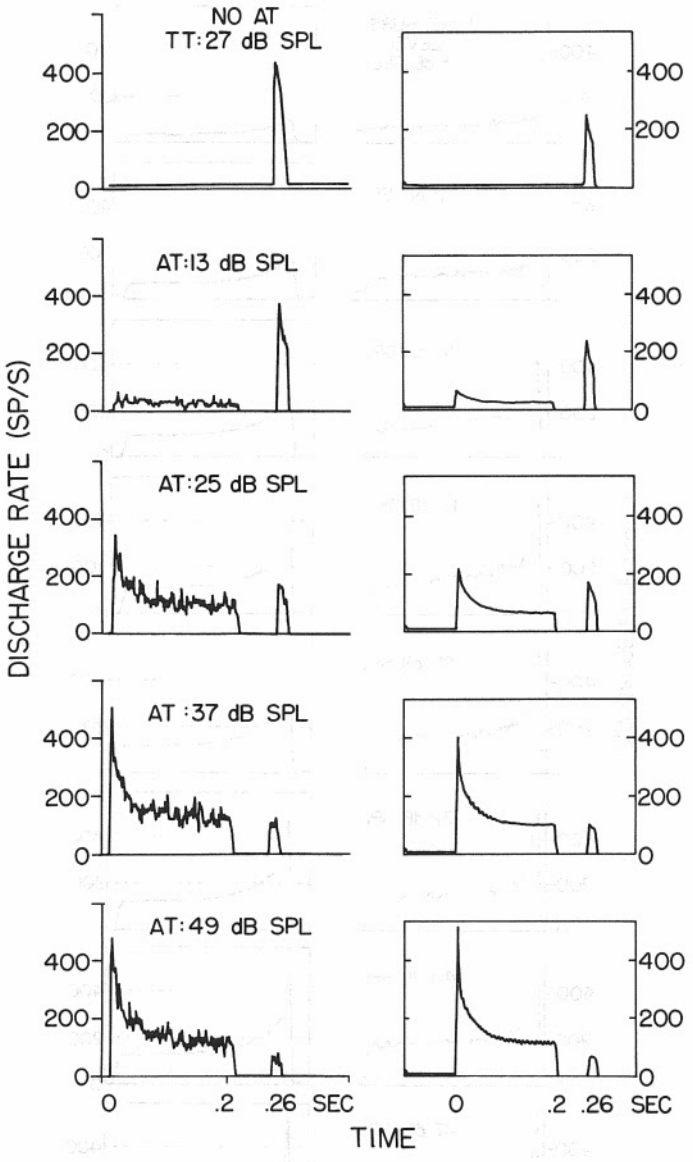

Figure 5. Left: response patterns of an auditory-nerve fiber to a 20-ms test tone preceded by a 200-ms adapting tone (from Delgutte, 1980). Both tones have a rise time of 2.5 ms and a frequency of 1220 Hz, approximately equal to the fiber CF. Histograms are computed with a 1-ms bin width, and were three-point smoothed. Right: response patterns for the computer model for the same stimulus conditions, using a 3-ms Hamming window for smoothing.

(i.e. the response decays more rapidly with each period) at intermediate amplitudes than at higher amplitudes, where saturation effects are dominating the response. Such domination at high signal amplitudes may well be related to the phenomenon of two-tone suppression (Sachs & Abbas, 1976; Javel, Geisler & Ravindran, 1978). The half-wave rectifier is the controlling factor in this steady-state phase-locked response characteristic, although the short-term adaptation component also plays a role.

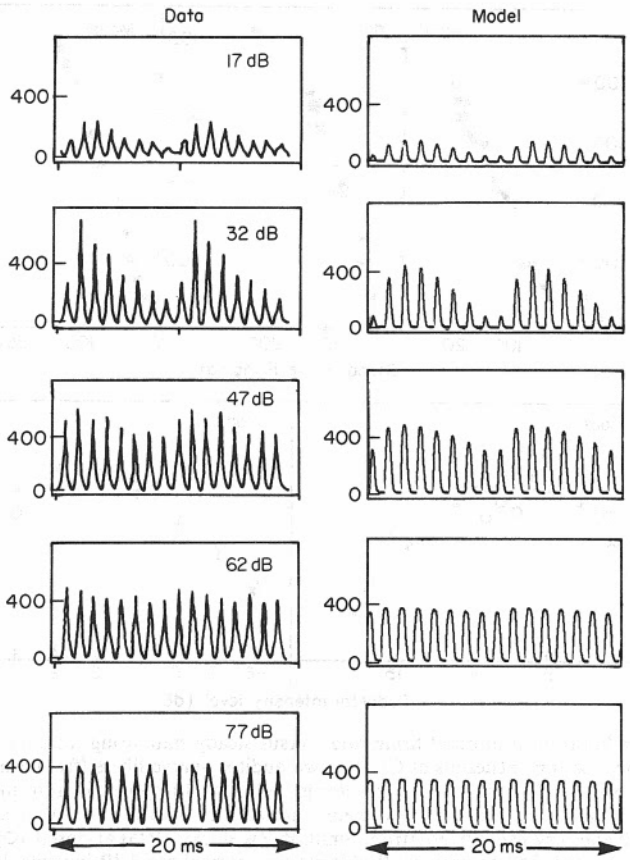

Figure 6. Left: response patterns of an auditory-nerve fiber to a single-formant synthetic stimulus as a function of signal amplitude. The stimulus has a formant frequency of 800 Hz, approximately equal to the fiber CF. Formant bandwidth is 70 Hz and the fundamental frequency is 100 Hz. The 10-ms period histogram, computed with a 50 μs bin width, is repeated twice in each case to show two pitch periods of the response. Right: response patterns generated by the model for the same stimulus conditions. The responses, in this case, are unsmoothed.

4.4. *Incremental responses*

Smith & Zwislocki (1975) measured the discharge rate of auditory-nerve fibers in response to abrupt increments of a sinusoid's amplitude. The amplitude, I, was incremented by an amount δI at a time $\tau = 150$ ms after initial onset. A post-stimulus-time histogram of the response was computed, and a difference between the response immediately preceding (R_τ^-) and following (R_τ^+) the amplitude increment was designated the "steady-state incremental response". This incremental response, defined as $IR = R_\tau^+ - R_\tau^-$, was then compared with an "onset incremental response", which is defined as the difference between the response to an onset signal at amplitude $I + \delta I$ and the response to a signal at amplitude I. Two important observations were: (1) the steady-state and onset IRs were approximately the same for signals of low-to-moderate sound-pressure level, and (2) the ratio of the response at signal onset, R_0, to the response during the steady-state portion, R_τ^-, was approximately equal to 2.5, regardless of the

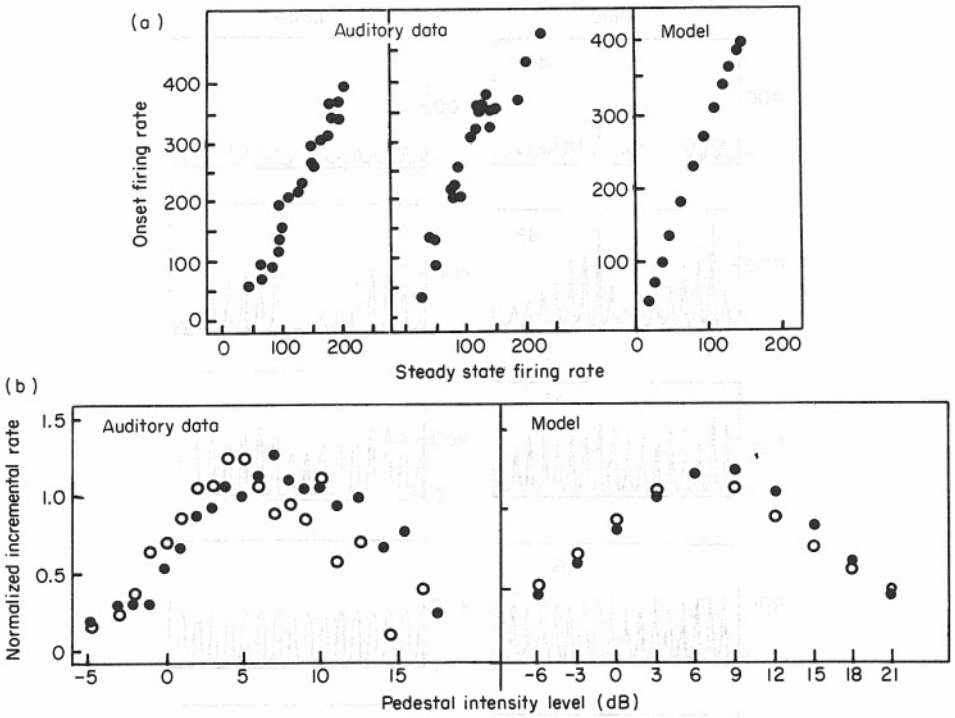

Figure 7. (a) Plots of onset firing rates versus steady-state firing rates, in response to tone pedestals at CF, for two auditory-nerve fibers (from Smith & Zwislocki, 1975) and for the computer model. Model response is from the 2 kHz channel. (b) Left: plots of median normalized 3 dB incremental responses for 10 auditory-nerve fibers (from Smith & Zwislocki, 1975) at onsets (○) and at steady-state (●) conditions. Right: plots of normalized 3 dB incremental responses for model at onsets (○) and at steady-state (●) conditions.

onset intensity level, *I*. This is the most difficult result to match with the model. The rapid AGC and the short-term adaptation tend to impose opposing constraints on the outputs. It is possible to obtain a fairly constant ratio of onset to steady-state response magnitude, but this ratio was consistently too large (3.0 instead of 2.5), as shown in Fig. 7(a). For the parameter settings shown in Table I, the 3-dB onset incremental response of the model was slightly larger than the 3-dB steady-state incremental response for low-amplitude signals. This response ratio became significantly smaller for more intense signals—a result which is in close agreement with the physiological data, as shown in Fig. 7(b).

4.5. *Synchrony falloff*

Johnson (1974) provided a specific definition for a "synchronization index" applied to the period histograms of auditory-nerve fiber responses to sinusoidal signals. This index is the same as the normalized Fourier coefficient which is defined as:

$$S_f = \frac{A(F_0)}{A(0)}, \tag{6}$$

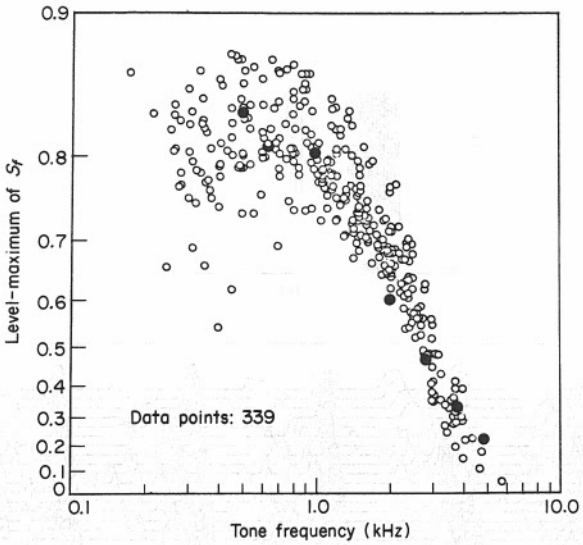

Figure 8. Scatter diagram of synchronization index (from Johnson, 1974) as a function of signal frequency (339 measurements from 233 units), compared with the model synchrony data (●).

where S_f is the synchronization index, $A(f)$ is the amplitude of the spectrum of the period histogram at frequency f, and F_0 is the signal frequency. Johnson measured S_f for a large number of fibers, using signals which did not always correspond to the fiber CF, and obtained the plot shown in Fig. 8. Points obtained by applying the same definition of synchrony to the model outputs are superimposed on Fig. 8 as closed circles. The primary component controlling the synchrony falloff in the model is the lowpass filter.

5. Output of the model for speech signals

Figure 9 shows an example of the Stage II outputs for a short segment of a male speaker's voiced speech, during the [e] of the word "make". Figure 9(a) is a wideband spectrogram of the signal, with a vertical bar indicating the point in time to which the channel outputs, shown in Fig. 9(b), refer. The 50-ms time-window includes approximately five pitch periods. The peaks are skewed slightly to the left for low frequencies, a feature that is present in the physiological data as well (Johnson, 1974). Figure 9(c) shows the output of the channel whose CF is nearest the vowel's second formant. A prominent component near the formant frequency is evident, in addition to the "envelope" periodicity at the fundamental frequency. Such formant periodicity is utilized by the synchrony algorithm in Stage III.

Figure 10 compares the outputs of Stages I and II outputs for the word "description" spoken by a female speaker. Each waveform is the output of one of the 40 channels, smoothed and downsampled to a 5-ms frame rate. The low-frequency channels are displayed at the bottom of the Figure. It is essential to represent Stage I outputs by a log-magnitude scale in place of a linear-magnitude representation; otherwise the formant peaks would overwhelm the remainder of the spectrum. A log-magnitude scale also corresponds to traditional analysis methods. Because of the saturating nonlinearity in the half-wave rectifier, as well as in the final AGC, a log representation is not appropriate

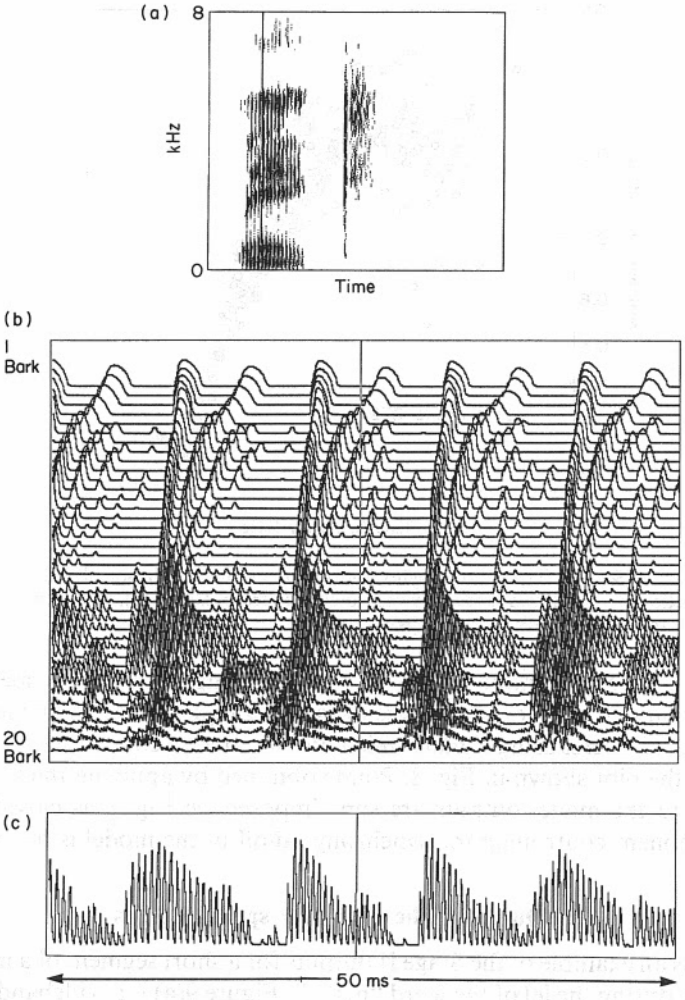

Figure 9. (a) Wideband spectrogram of the word "make", spoken by a male speaker. (b) Stage II outputs of 40 channels, with the *lowest* frequency channel at *top*, for five pitch periods during the vowel [e] at the time of the vertical bar in (a). (c) Output of a single channel near the frequency of the second formant at the same point in time as in (b).

for the outputs of Stage II. Magnitude at this level corresponds to "mean discharge rate", which is computed by dividing the number of spikes by the signal duration and scaling the result in units of seconds^{-1}. A phonetic transcription is provided below the channel outputs to facilitate segmentation.

Transitions from one phonetic segment to the next are more clearly delineated by onsets and offsets in the Stage II representation. All segment boundaries, except those associated with [rI], are well delineated in the Stage II representation. The closure intervals for both the [k] and [p] are flat valleys in the Stage II representation. There is clear evidence for forward masking here, particularly in the low-frequency region of the [pʃ] segment. The vowel [I] masks the low-frequency noise not only during the [p]-closure interval but also during the subsequent [ʃ]. Such masking phenomena should enhance

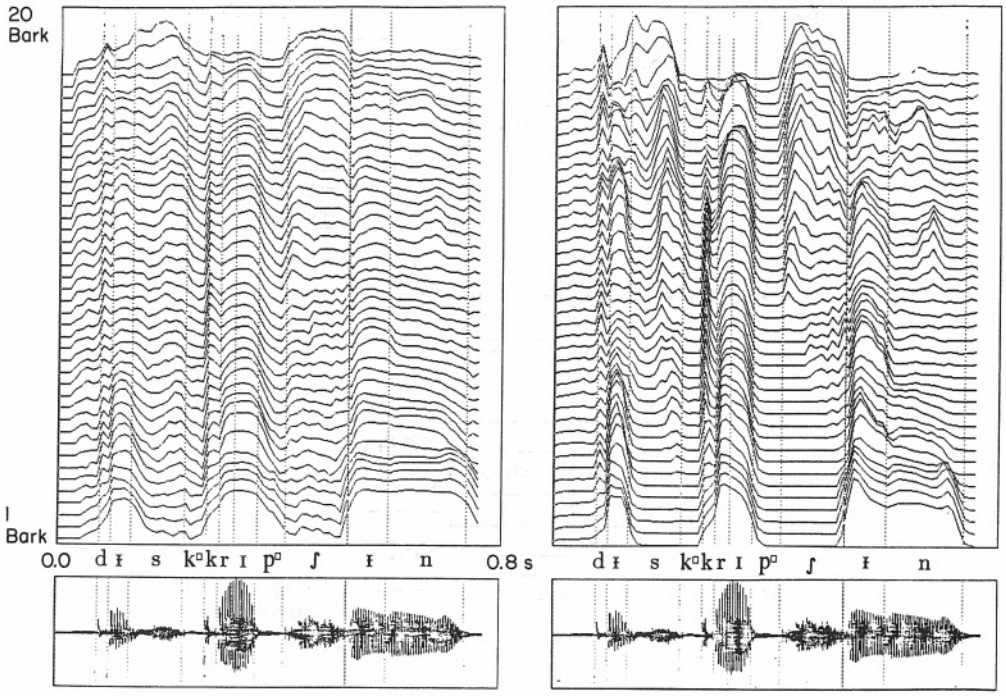

Figure 10. Left: log-magnitude response of Stage I outputs for the word "description" spoken by a female speaker, with the *lowest* frequency channel at the *bottom*. Right: magnitude response of Stage II outputs for the same word. The original waveform is shown below in each case. The dotted vertical lines denote the phonetic boundaries.

the contrast between vowels and fricatives. The boundary between the [ɪ] and the final [n] is very difficult to see in the Stage I representation. However, the Stage II nonlinearities serve to delineate this boundary. The stop-burst onsets for the [d] and the [k] are also much more sharply delineated after Stage II.

5.1. *The synchrony spectrogram*

The Stage II outputs, smoothed and downsampled, appear to be an excellent representation for locating transitions between phonemes, and thus could provide an adequate basis for phonetic segmentation. They may also be useful for broadly categorizing the resulting segments as *fricatives*, *closures*, *weak sonorants*, *vowels*, and so forth. However, these outputs, when displayed as a spectrogram, do not provide a precise estimate of the formant frequencies. This is to be expected because of the saturating nature and resulting limited dynamic range of auditory-nerve fibers. During the vocalic segments, many channels in the vicinity of the formant frequencies are responding at the saturation level and, as a consequence, the formant peaks become very broadly distributed.

The Stage II outputs do, however, contain significant information about the formant frequencies, which is discarded by the smoothing process. Such information is available as a dominant periodicity in the temporal response pattern. The ALSR calculation of Young & Sachs (1979) capitalized on such periodicity. We have chosen a somewhat different measure aimed at a similar goal. This "Generalized Synchrony Detector"

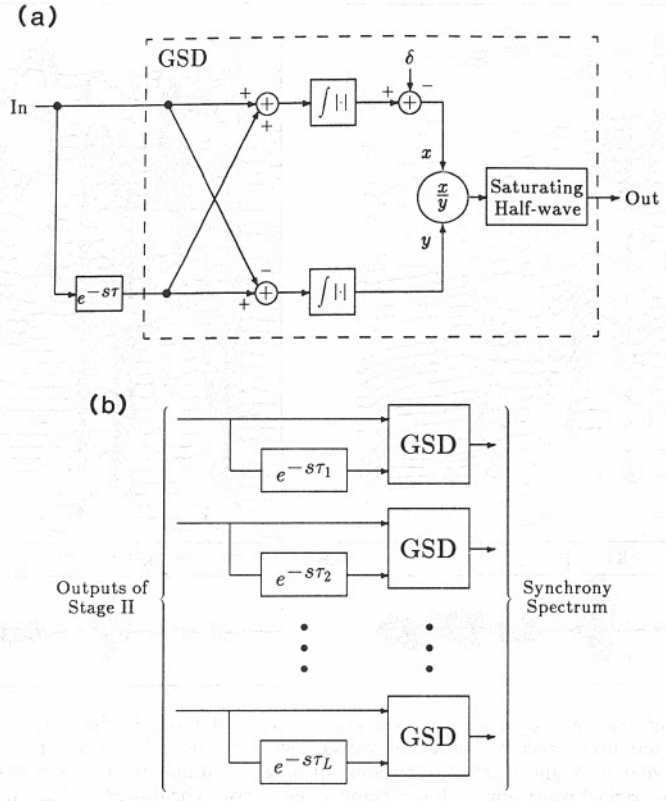

Figure 11. (a) Schematic diagram of the Generalized Synchrony Detector (GSD). (b) The synchrony branch of Stage III. Each channel output in Stage II is processed through a GSD tuned to the center frequency of the corresponding peripheral filter in Stage I. See text for details of GSD processing.

(GSD) was selected from a number of different possibilities because it offers certain advantages in representing the speech spectrum used to identify the phonetic content of the utterance. Our goal was to produce as clean a spectral representation as possible, one which would preserve prominent peaks at the formant resonances, while significantly reducing features of the spectrogram associated with the glottal excitation. We also sought to normalize for amplitude. Although pitch and loudness are certainly important perceptual attributes of the speech signal, we believe that a collection of distinct spectral representations, each of which is adapted to a specific assigned task, is to be preferred over a single complex representation which preserves all of the signal attributes.

The GSD is based on the ratio of the estimated magnitude of a sum waveform to the estimated magnitude of a difference waveform, as shown in Fig. 11(a). The inputs to the sum and difference computation are the GSD input signal and a delayed version of the input signal, with the delay period corresponding to the frequency to which the GSD is tuned. When the input to the GSD is perfectly periodic with the delay period, the magnitude of the difference waveform is zero. Hence, the ratio can become infinitely large during perfect synchrony. To constrain the response to be within reasonable limits, a final saturating nonlinearity is applied. In addition, a threshold is subtracted from the

numerator in order to preclude a response from very weak signals. This threshold is set to a level slightly greater than the spontaneous discharge rate.

Figure 11(b) shows how the GSD is used to compute a synchrony spectrogram directly from the Stage II outputs. Each Stage II channel output is processed through a GSD tuned to the center frequency of the corresponding auditory filter in Stage I. Thus, if there is a prominent peak in the signal at a particular frequency, f, it will show up as a periodicity in the Stage II waveforms. Only the channel whose CF is closest to f will specifically detect the "correct" periodicity; its response will be correspondingly large. The output of adjacent channels will be significantly smaller because their tuned periodicity is inappropriate to that of the dominant signal.

This particular definition of synchrony was chosen for a number of reasons. First, because it measures a *periodicity* rather than a *frequency*, it avoids the problem of detecting synchrony to the second harmonic of a strong peak, such as was the case for the ALSR strategy[4]. Second, because the difference waveform in the denominator is balanced by a sum waveform in the numerator, this is effectively an *energy-normalized* scheme. Such normalization has the added advantage of significantly reducing temporal fluctuations in the response due to the envelope of the glottal excitation, which can be viewed as unwanted noise for this part of the recognition task. Finally, the algorithm is computationally simple, involving components that could reasonably be computed by neuron-like elements.

Harmonic structure due to the glottal excitation is usually completely obliterated in the synchrony spectrogram for male voices, but typically retained in the first-formant region for female voices. Harmonics between F_1 and F_2 are usually suppressed, because prominent energy at the first-formant frequency in the channel output destroys synchrony to the intermediate harmonics. Pitch striations over time are usually absent, due to the amplitude normalization process. Peaks at the formant frequencies are much narrower than in the envelope representation, thus making the synchrony spectrum more suitable for making fine distinctions.

The features of the synchrony branch of the system are illustrated in Fig. 12. A wideband spectrogram, an envelope spectrogram and a synchrony spectrogram are presented for the word "hesitate" spoken by a female speaker. The latter two are shown on a Bark-frequency scale (Zwicker, 1961). It is clear that the formant peaks are not well preserved in the envelope spectrogram, due mainly to the limited dynamic range of the nerve fibers. The formant resonances in the vowels are captured well by the synchrony measure. Furthermore, spectral peaks in regions of little energy, such as the initial [h] and the schwa, are enhanced relative to the wideband spectrogram. Perhaps surprisingly, the spectral prominences for obstruents in the high-frequency regions (i.e. in the [t] and the [z]) are accentuated by the synchrony algorithm, in spite of the fact that a good deal of synchrony to the stimulus frequency has been lost in Stage II. A possible explanation is that the synchrony measure incorporates energy at d.c., as well as energy at the CF. Any strong energy concentration in the signal at high frequencies is mostly converted to d.c. energy, which is passed by the synchrony measure. Prominent peaks in the input waveform well below the CF of high-frequency filters appropriately reduce the

[4]The GSD *does* detect synchrony at *half* the frequency of an input stimulus. This is a problem only for filters in the first formant region, since the high-frequency auditory filters typically have very steep slopes on the high-frequency side, such that input signals at twice the CF rarely trigger a response above the spontaneous rate.

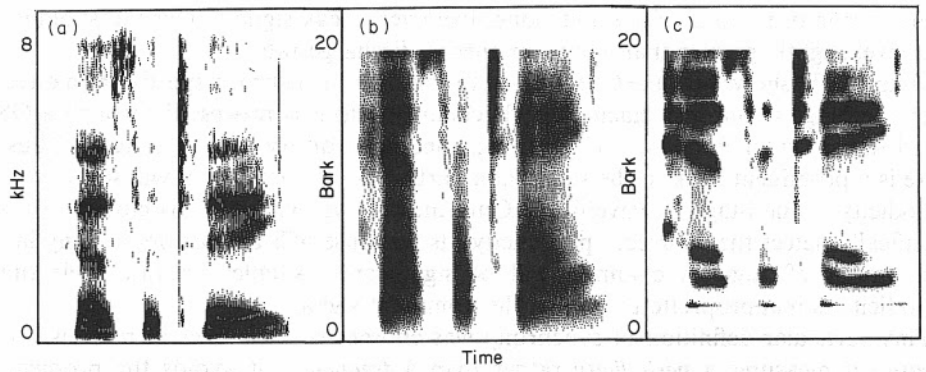

Figure 12. (a) Wideband spectrogram for the word "hesitate", spoken by a female speaker. (b) "Mean-rate" spectrogram for the same word, obtained by processing Stage II outputs through an envelope-detection scheme, corresponding to the top path of Stage III. (c) "Synchrony" spectrogram for the same word, obtained by processing Stage II outputs through a synchrony-detection scheme, corresponding to the bottom path of Stage III. The wideband spectrogram is shown on a linear-frequency scale, whereas the other two are displayed on a Bark scale.

synchronous response of such filters because the synchrony present in the signal is not of the appropriate frequency for these channels.

6. Summary and conclusions

This paper describes a relatively simple model for auditory processing of speech signals, which attains a reasonably good match to measured auditory responses for a number of different experimental paradigms. The model offers the hope of elucidating further the nature of the auditory response to speech. In addition, we anticipate that representations obtained from such a model will be well-suited to applications in computer speech recognition.

It is surprising that this model is capable of yielding such a close match to the Smith and Zwislocki data which show a constant ratio of onset to steady-state response and a close-to-equal incremental response characteristic for onset and steady-state conditions. Both the Goldhor adaptation model (when applied to a periodic signal rather than to the gross temporal envelope) and the AGC are nonlinear elements, yet a cascade of the two components results in an apparently linear overall response.

The model used for the AGC is a poor approximation of the refractory effect as it is currently understood. First, Equation (5) is only valid for steady-state conditions, and only exact for signals that are periodic with respect to Δ. Second, a leaky integrator yields an averaging window for $\langle x[n] \rangle$ that is exponential in shape, whereas a rectangular window is a much better approximation to the recovery function. Nonetheless, the value for K_{AGC} that was determined experimentally to best match auditory data is 0.002. This value corresponds to a 2-ms lockout period, which is of the correct order of magnitude. Perhaps a more realistic model for the refractory effect that would be appropriate during onsets, as well as steady states, would result in a better match to the dynamics of the onset envelope response.

It is not clear at what level of the auditory pathway a neural processing mechanism analogous to the Generalized Synchrony Detector should be sought. Nonetheless, such

a mechanism could be realized using simple units that are at least feasible neurologically. If the input to the GSD were a sequence of pulses instead of a waveform, then the difference waveform in the denominator would reduce to an XOR gate, with a suitably narrow time window over which the delayed input and the undelayed input "coincide". The division and half-wave rectification are functionally similar to an excitatory/inhibitory unit. This unit or "cell" would have a minimal response threshold, related to the silence threshold in Fig. 11(a), and a saturation level.

It is still premature to suggest that an auditory-based speech analysis system will pay off in speech recognition. There are emerging, however, strong indications that auditory-based representations are interesting and worthy of further study. We are now becoming more confident in the validity of the computer models, such that they may reveal interesting effects in auditory speech processing, which may lead the way to appropriate later-stage speech recognition strategies. We have described here a computer model that produces two distinct spectral-like representations for the speech signal, one based on the average discharge rate and the other based on the synchronous response. Several researchers in the speech group at M.I.T. are pursuing recognition strategies based on these representations. The mean-rate response outputs have been used successfully for locating acoustic boundaries and for making broad category decisions (Glass & Zue, 1986). Preliminary results using these outputs for syllable detection in continous speech are encouraging. The synchrony spectrogram has been applied to speaker-independent vowel recognition in continuous speech (Seneff, 1987). Preliminary results indicate superior performance with minimal computational load for the recognition stage.

The design of this system was influenced by interaction with several people. Among these are Bertrand Delgutte, Rich Goldhor, Don Johnson, Camp Searle, Ken Stevens, Tim Wilson and Victor Zue. Rob Kassel was very helpful in constructing some of the figures. The paper is much improved due to the careful reading of earlier versions by Katy Kline, Don Johnson, Ken Stevens, Steven Greenberg, Quentin Summerfield and an anonymous reviewer.

This research was supported by DARPA under Contract N00039-85-C-0254, monitored through Naval Electronic Systems Command.

References

Delgutte, B. (1980) Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers, *Journal of the Acoustical Society of America*, **68**, 843–857.

Delgutte, B. & Kiang, N. Y. S. (1984*a*) Speech coding in the auditory nerve: I. Vowel-like sounds, *Journal of the Acoustical Society of America*, **75**, 866–878.

Delgutte, B. & Kiang, N. Y. S. (1984*b*) Speech coding in the auditory nerve: III. Voiceless fricative consonants, *Journal of the Acoustical Society of America*, **75**, 887–896.

Delgutte, B. & Kiang, N. Y. S. (1984*c*) Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics, *Journal of the Acoustical Society of America*, **75**, 897–907.

Eggermont, J. J. (1973) Analog modelling of cochlear adaptation, *Kybernetik*, **14**, 117–126.

Glass, J. R. & Zue, V. W. (1986) Signal representation for acoustic segmentation, *Proceedings of the First Australian Conference on Speech Science and Technology*, 124–129.

Geisler, C. D. (1988) Representation of speech sounds in the auditory nerve, *Journal of Phonetics*, **16**, 19–35.

Goldhor, R. S. (1985) *Representation of consonants in the peripheral auditory system: a modeling study of the correspondence between response properties and phonetic features*. Technical Report 505. Cambridge, MA: M.I.T.

Harris, D. M. & Dallos, P. (1979) Forward masking of auditory nerve fiber responses, *Journal of Neurophysiology*, **42**, 1083–1107.

Hudspeth, A. J. & Corey, D. P. (1977) Sensitivity, polarity and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli, *Proceedings of the National Academy of Science, U.S.A.*, **74**, 2407–2411.

Javel, E., Geisler, C. D. & Ravindran, A. (1978) Two-tone suppression in the auditory nerve of the cat, *Journal of the Acoustical Society of America*, **63**, 1157–1163.

Johnson, D. H. (1974) The response of single auditory-nerve fibers in the cat to single tones: synchrony and average discharge rate. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.
Johnson, D. H. (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones, *Journal of the Acoustical Society of America,* **68**, 1115–1122.
Johnson, D. H. & Swami, A. (1983) The transmission of signals by auditory-nerve fiber discharge patterns, *Journal of the Acoustical Society of America,* **74**, 493–501.
Kiang, N. Y. S., Watanabe, T., Thomas, E. C. & Clark, L. F. (1965) *Discharge patterns of single fibers in the cat's auditory nerve*, Cambridge, MA: M.I.T. Press.
Miller, M. I. & Sachs, M. B. (1983) Representation of stop consonants in the discharge patterns of auditory-nerve fibers, *Journal of the Acoustical Society of America,* **74**, 502–517.
Rhode, W. S. & Smith, P. H. (1985) Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers. *Hearing Research,* **18**, 159–168.
Sachs, M. B. & Abbas, P. J. (1976) Phenomenological model for two-tone suppression, *Journal of the Acoustical Society of America,* **60**, 1157–1163.
Sachs, M. B., Blackburn, C. C. & Young, E. D. (1988) Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus, *Journal of Phonetics,* **16**, 37–53.
Sachs, M. B. & Young, E. D. (1980) Effects of nonlinearities on speech encoding in the auditory nerve, *Journal of the Acoustical Society of America,* **68**, 858–875.
Seneff, S. (1984) Pitch and spectral estimation of speech based on an auditory synchrony model, *Proceedings of ICASSP-84*, San Diego, CA.
Seneff, S. (1985) *Pitch and spectral analysis of speech based on an auditory synchrony model.* RLE Technical Report 504. Cambridge, MA: M.I.T. Press.
Seneff, S. (1987) Vowel recognition based on line-formants derived from an auditory-based spectral representation, *Proceedings of the Eleventh International Congress of Phonetic Sciences*, Tallinn, Estonia, U.S.S.R.
Sinex, D. G. & Geisler, C. D. (1983) Responses of auditory-nerve fibers to consonant-vowel syllables, *Journal of the Acoustical Society of America,* **73**, 602–615.
Smith, R. & Zwislocki, J. J. (1975) Short-term adaptation and incremental responses of single auditory-nerve fibers, *Biological Cybernetics,* **17**, 169–182.
Srulovicz, P. & Goldstein, J. L. (1983) A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of the frequency spectrum, *Journal of the Acoustical Society of America,* **73**, 1266–1276.
Yost, W. A. & Nielsen, D. W. (1977) *Fundamentals of hearing—an introduction.* New York: Holt, Rinehart and Winston.
Young, E. D. & Sachs, M. B. (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers, *Journal of the Acoustical Society of America,* **66**, 1381–1403.
Zwicker, E. (1961) Subdivision of the audible frequency range into critical bands (frequenzgruppen), *Journal of the Acoustical Society of America,* **33**, 248–249.