A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers

Rick L. Jenison
Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706

Steven Greenberg
Department of Neurophysiology, University of Wisconsin, Madison, Wisconsin 53706

Keith R. Kluender
Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706

William S. Rhode
Department of Neurophysiology, University of Wisconsin, Madison, Wisconsin 53706

(Received 21 November 1990; accepted for publication 28 March 1991)

A composite model of the auditory periphery, based upon a unique analysis technique for deriving filter response characteristics from cat auditory-nerve fibers, is presented. The model is distinctive in its ability to capture a significant broadening of auditory-nerve fiber frequency selectivity as a function of increasing sound-pressure level within a computationally tractable time-invariant structure. The output of the model shows the tonotopic distribution of synchrony activity of single fibers in response to the steady-state vowel [ε] presented over a 40-dB range of sound-pressure levels and is compared with the population-response data of Young and Sachs (1979). The model, while limited by its time invariance, accurately captures most of the place-synchrony response patterns reported by the Johns Hopkins group. In both the physiology and in the model, auditory-nerve fibers spanning a broad tonotopic range synchronize to the first formant (F₁), with the proportion of units phase-locked to F₁ increasing appreciably at moderate to high sound-pressure levels. A smaller proportion of fibers maintain phase locking to the second and third formants across the same intensity range. At sound-pressure levels of 60 dB and above, the vast majority of fibers with characteristic frequencies greater than 3 kHz synchronize to F₁ (512 Hz), rather than to frequencies in the most sensitive portion of their response range. On the basis of these response patterns it is suggested that neural synchrony is the dominant auditory-nerve representation of formant information under "normal" listening conditions in which speech signals occur across a wide range of intensities and against a background of unpredictable and frequently intense acoustic interference.

PACS numbers: 43.71.An, 43.71.Cq, 43.71.Es, 43.64.Bt

INTRODUCTION

Over the past decade there has been increasing interest in computational models of the auditory periphery (e.g., Allen, 1985; Lyon, 1982; Seneff, 1986; Deng et al., 1988; Shamma, 1988; Ghitza, 1988). The majority of these models treat cochlear frequency analysis as relatively time and intensity invariant. They are consequently based on a single bank of linear time-invariant filters to model the dispersion of frequency components across the cochlear partition and auditory-nerve fiber array (e.g., Seneff, 1986; Shamma, 1988; Ghitza, 1988; Meddis and Hewitt, 1991; Patterson and Holdsworth, 1991).

Experimental studies by Rhode (1971) and others (e.g., Johnstone et al., 1986; Robles et al., 1986) have demonstrated that the motion of the basilar membrane (BM) is highly nonlinear. The input–output function for basilar membrane motion becomes highly compressive at moderate-to-high sound-pressure levels (SPLs) typical of conversational speech (60–70 dB SPL). One consequence of such compression is a broadening of the membrane’s frequency selectivity, which is reflected in the filtering characteristics of auditory-nerve (AN) fibers, particularly in those most sensitive to frequencies above 3 kHz. For high SPLs, the frequency response of both the BM and high-CF AN fibers approximates a low-pass filter. Such broadening in frequency selectivity is observed in the "tail" component of the frequency threshold curve of high-characteristic-frequency (CF) AN fibers and in the basalward spread of activity evoked by low-frequency signals at high SPLs. For example, at 70 dB SPL, a 1-kHz sinusoid will produce a response across a very large proportion of AN fibers whose CFs range between 0.5 and 8 kHz (Pfeiffer and Kim, 1975; Kim et al., 1980). We believe that such upward spread of activity is especially important for coding certain features of the speech signal, such as the first formant.

Because most models of the auditory periphery assume
that cochlear filtering is linear, and hence does not change as a function of sound-pressure level or center frequency, they do not capture the intricate pattern of excitation distributed across the auditory nerve in response to complex, broadband spectra such as speech. For this reason, these models are essentially "place" representations in that they reflect only the response of AN fibers whose CFs are close to that of the stimulus frequencies. In general, their specific focus has not been on the spread of excitation associated with the dominant spectral regions of the input signal (i.e., formant peaks).

Distinguishing the current work from other models are filter functions that have been designed to accurately simulate the change in frequency selectivity of single AN fibers over a broad range of intensities. These filter functions are derived from average-rate iso-intensity functions of 97 cat high spontaneous-rate (SR) AN fibers, and are thus based directly on physiological filter properties, insofar as they are reflected in the rate response of single fibers. Given the diversity of thresholds, spontaneous rates and filtering for any given CF range, the filter functions are derived from ensembles of iso-intensity curves of fibers distributed over a delimited CF range, and confined to units that are highly spontaneously active. Low-SR AN fibers \( (n = 45) \) were omitted from the current analysis due to the variability of their tuning. However, they are included in the database for future analysis.

The present model, while incorporating intensity specific frequency selectivity, is essentially a time-invariant system. The filter functions employed are based upon steady-state rms levels of the input signal over a 25.6-ms interval. Although time invariance is a compromise solution, the temporal fine structure of neural activity can still be reasonably well portrayed as a function of intensity.

The model was developed in MATLAB™, an interactive matrix computation, and graphics program, and has been run on a wide range of computer platforms, including a 386 PC, a Macintosh IIci and a DECstation 5000. A 25.6-ms segment of speech requires approximately 60 s of processing time on a DECstation 5000 (2300 times real time), which is approximately six times faster than the Mac IIci for this task.

I. AUDITORY PERIPHERY

As a consequence of the mechanical properties of the organ of Corti, complex sounds are decomposed into a series of constituent signals distributed along the cochlear partition. In essence, the cochlea acts as a limited-resolution spectrum analyzer. At low SPLs, the BM responds most vigorously to low frequencies at its apex and to high frequencies at the base. This place coding of frequency is translated into discrete populations of AN fibers responding to individual frequencies. At higher SPLs, this place coding breaks down for low frequencies as detailed below.

A distinctive property of AN fibers is their ability to temporally synchronize their discharge to a restricted portion of the stimulus cycle (Rose et al., 1967). Fibers are capable of encoding information concerning stimulus frequency in terms of the temporal properties of the discharge, in addition to the place identity of the fiber. In response to a complex signal, with a broad range of frequencies the temporal information could, in principle, provide better resolution of the stimulus spectrum than possible through place information alone (Srulovicz and Goldstein, 1983).

AN fibers vary in terms of their spontaneous discharge activity. Approximately 60% are highly active (i.e., spontaneous rates \( > 18 \) spikes/s) in the absence of sound. About a quarter show moderate rates of background activity \( (> 0.5 < 18 \) spikes/s), while the remainder (15%) are spontaneously inactive or nearly so \( (< 0.5 \) spikes/s) (Liberman, 1978).

Iso-intensity functions (response areas) were recorded from 142 single AN fibers in the sodium pentobarbitol anesthetized cat. Response areas were obtained by recording the extracellular potential of single auditory-nerve fibers either in response to 25- or 50-ms sinusoidal signals of variable frequency and sound-pressure level. Each point in the response area (for a single frequency by SPL) is based on the average firing rate to five stimulus presentations. Details of the surgical and recording procedures are described in Rhode and Greenberg (1991).

Response areas were mapped by measuring the magnitude of the discharge output as a sinusoidal signal was stepped through the spectrally responsive region of the fiber at constant SPL (Rose et al., 1967). Response areas from a cat high-SR AN fiber with a characteristic frequency of approximately 9 kHz is shown in Fig. 1. The fiber serves as an example of the properties discussed earlier. The iso-intensity curves evidence a broadening of frequency selectivity as the input stimulus increases in SPL from 10 to 80 dB.

The average-rate iso-intensity function of single fibers is assumed to reflect the summed filtering effects of mechanical BM tuning, inner hair cell (IHC) transduction, IHC neurotransmitter release, and auditory-nerve spike generation. Thus all varieties of nonlinearities such as BM compression, IHC receptor potential saturation, spike rate saturation of the AN, thresholding, etc., are reflected in the output response of AN fibers.

![Figure 1](image_url)

**FIG. 1.** Iso-intensity curves (response areas) for a cat auditory-nerve fiber with a characteristic frequency of approximately 9 kHz, as derived from the presentation of sinusoidal stimuli of variable frequency and sound-pressure level (10-80 dB, in 10-dB steps) (see Greenberg and Rhode, 1991), for details of stimulus presentation and data collection.)
II. DERIVATION OF FILTERS

The AN fiber data were cast in matrix form in MATLAB™ and could be freely organized along any number of dimensions. The database of nerve-fiber response areas is organized as stacked matrices of 142 AN fibers by frequency (0.05–20 kHz), ordered by sound-pressure level (10–80 dB in 10-dB steps). Of the 142 fibers, 97 high-SR fibers were extracted from the database for analysis.

A. Characteristic frequency axis

Synthetic auditory-nerve fiber channels were derived from ensembles of AN response areas within a range centered about fixed intervals along the length of the basilar membrane partition. Channels were mapped uniformly along a theoretical frequency-position axis. The channel “CFs” were sampled based on the frequency-position function derived by Greenwood (1961, 1989). The equation for computing frequency in Hz as a function of x expressed as a proportion of basilar length is

\[ F = A(10^a - k). \]  

The constants \( A = 456, a = 2.1, \) and \( k = 0.8 \) were obtained from Liberman’s (1982) best fit of the frequency-position function to AN data from the cat.

One hundred twenty-eight-channel CFs were sampled from 0.10 to 0.70 of the frequency-position axis (2.5 to 17.5 mm assuming a mean basilar length of 25 mm), providing a frequency range of channel CFs corresponding to 375–13 093 Hz. The ensemble window centered at each channel CF was chosen to span 0.13 (3.25 mm) of the cochlea length, ensuring sufficient ensemble membership despite fewer AN fibers above 4 kHz and of slightly greater extent, 0.15 (3.75 mm), of the frequency range of channel CFs corresponding to 375–13 093 Hz. The ensemble window length at higher CFs followed by inverse Fourier transformation. The resulting windowing method was used to design the matrix of FIR filters from the desired frequency response functions. A Hanning window was applied to each ensemble eigenvector, followed by inverse Fourier transformation. The resulting

rows of observed AN fibers. Filter response functions were derived by extracting the principal response of the AN fibers within each matrix by analysis of principal components. Principal component analysis is a mathematical technique where the principal axes or eigenvectors are a linear combination of the original variables. The method produces a unique set of orthogonal basis vectors with the first principal component accounting for the highest possible amount of variance of any linear combination of the variables. Eigenvectors and eigenvalues of the cross-product matrix were derived by direct decomposition of the data matrix using singular value decomposition (SVD) (Golub and von Loan, 1983). SVD yields a fundamental representation of the basic properties of a particular matrix, even in cases where the matrix is singular or not of full rank. Assuming the aligned iso-intensity matrix \( \mathbf{R} \) to be a \( m \times n \) matrix (\( m \) = number of fibers and \( n \) = number of frequency samples),

\[ \mathbf{R} = \sum_{k=1}^{r} s_k u_k v_k^T. \]  

where \( r \) is the rank of \( \mathbf{R} \), \( u_k \) and \( v_k \) are the left and right \( k \) th singular vectors, \( s_k \) is the \( k \) th singular value of \( \mathbf{R} \), and \( T \) denotes the transpose. The representation may also be shown in matrix form as

\[ \mathbf{R} = \mathbf{USV}^T. \]  

The columns of \( \mathbf{V} \) are the orthonormal eigenvectors \( \mathbf{R}^T \mathbf{R} \) and the columns of \( \mathbf{U} \) are the orthonormal eigenvectors of \( \mathbf{RR}^T \). The first column vector of \( \mathbf{V} \) was used to describe the shape of the filter function that captures the maximum variance of matrix \( \mathbf{R} \) subject to the constraint of mutual orthogonal axes. Note that the mean was retained in \( \mathbf{R} \) rather than transforming to a mean normalized covariance matrix. Retaining the mean is desirable from a statistical standpoint (Rummel, 1970), as well as physiologically motivated (i.e., we are interested in the contribution of the magnitude of the response). If the ensemble of iso-intensity curves are highly correlated, such that the percentage of total variance accounted for by the first eigenvector is large, then the first eigenvector closely scales to the column mean of \( \mathbf{R} \). In our analysis, the first eigenvector accounted for at least 90% of the total variance for each of the selected ensembles of aligned iso-intensity curves.

C. Finite impulse response (FIR) filter design

One method of instantiating the frequency response of each channel is to design finite impulse response filters (FIR) from the first eigenvectors accounting for the maximum variance within a channel ensemble. An FIR filter takes a sequence of input samples \( x \) and produces a sequence of output samples \( y \) by the formula

\[ y_n = \sum_{k=-M}^{M} c_k x_{n-k}. \]  

A windowing method was used to design the matrix of FIR filters from the desired frequency response functions. A Hanning window was applied to each ensemble eigenvector, followed by inverse Fourier transformation. The resulting
vector of $2M + 1$ (129) coefficients $c_k$ corresponds to the impulse response of the channel. These FIR filters, in contrast with infinite impulse response (IIR) filters, have linear phase.

All filters were normalized to unity gain at the peak response of each channel at each intensity level as shown in Fig. 2. This results in characterizing the linear spectral shape, but not the nonlinear level-dependent gain.

The resulting filter structure is one whereby the impulse response for each fiber channel can be selected or indexed according to the estimated signal level. For purposes of this paper, an entire filter bank is selected based upon the rms level of the input signal. The equivalent sound-pressure levels are expressed in decibels re: 1 least significant bit (lsb), which provides an arbitrary reference to the smallest integer quantization of the input signal.

D. Derivation of rate-intensity functions

Conversion to fiber discharge rates is accomplished by applying a derived rate-intensity function that maps a partic-
ular input level to the resulting instantaneous rate of each channel. The rate-intensity functions used in the model were derived from AN response areas using a window equal in width to that used in deriving the filter response functions. The peak discharge rate for each iso-intensity function at each intensity level defined the nonlinear average rate-intensity function for each AN fiber in the database. The instantaneous rate-intensity function is best predicted by the response modulation (RM) function, which generally reflects greater dynamic range at higher intensities than predicted by average rate-intensity functions, particularly for fibers with CFs below 2 kHz (Smith and Brachman, 1980a). However, as an approximation for steady-state signals, a transformation based upon average rate intensities was adopted with the understanding that the mapping to discharge rate slightly underestimates the dynamic range of fibers with lower CFs.

The ensemble of rate-intensity data for each channel of the model was fit by a weighted sigmoidal function

\[ y_n = b e^{-\alpha/x_n} \]

where the free parameters \( a \) and \( b \) were estimated using the simplex method for multidimensional minimization. The discharge rate \( y_n \) saturates at the value of \( b \) for large values of \( x_n \) relative to \( a \). Several other sigmoidal functions where considered such as the logistic, hyperbolic tangent, and arc tangent; however, all functions imposed an unrealistic degree of nonlinear expansion for small input values of \( x_n \), the resulting 128 rate-intensity functions are shown in Fig. 3.

### III. COMPOSITE MODEL OF THE AUDITORY PERIPHERY

The composite model is composed initially of a matrix (CF by intensity) of linear FIR filters. Following the initial filtering stage, the resulting signal is half-wave rectified, reflecting the directional transduction of the basilar membrane motion by the hair-cell mechanism that results in neurotransmitter release of the inner hair cells and depolarization of auditory-nerve fibers. Next, a saturating nonlinearity captures the refractory nature of the nerve fibers based upon the estimated limiting values from rate-intensity functions discussed in Sec. II.

Synchrony reduction serves as the final processing stage in the composite model. The degree to which AN fibers exhibit synchrony decreases systematically in response to frequencies greater than 800 Hz (Johnson, 1980). A low-pass, FIR filter derived from Javel et al.'s (1988) fit to Johnson's (1980) data is shown in Fig. 4. This function has approximately a 3.5-dB/oct rolloff between 1 and 2.5 kHz, and a rolloff of 16 dB/oct above 2.5 kHz.

Finally, to visualize the magnitude of synchrony as a function of frequency, a periodogram (Kay and Marple, 1981) was computed over a 25.6-ms interval for each channel by

\[ P_s(f) = \frac{1}{N\Delta f} \left| \sum_{n=0}^{N-1} x_n e^{-2\pi i n f \Delta t} \right|^2. \]

A normalized periodogram \( S_s(f) \) analogous to the synchronization index was computed by

\[ S_s(f) = \frac{P_s(f)}{P_s(0)}, \]

where \( P_s(0) \) corresponds to the average discharge rate (dc component) over the analysis interval. The resulting normalized periodogram \( S_s(f) \) therefore reflects the relative response at each frequency having factored out the discharge rate.

To capture the relevant details of the channel filter characteristics, the model operates at a sampling rate of 40 kHz, which allows encoding of the signal spectrum up to 20 kHz.

### IV. EVALUATION OF THE MODEL WITH SPEECH

A synthetic steady-state vowel [ɛ] (as in the word pet) was used to evaluate the model. The vowel [ɛ] was synthesized using a parallel formant-wave-function (FWF) synthesizer based upon the formulas developed by Rodet (1980). The stimulus was initially generated at a sampling frequency of 10 kHz, and was subsequently upsampled to 40 kHz for input to the model. Figure 5 shows the magnitude of the harmonics of the stimulus [ɛ], which was synthesized to have formant frequencies and amplitudes similar to those of the vowel [ɛ] used by Sachs and Young (1979) in their physiological study. Frequencies corresponding to the perceptually significant peaks in the vowel's spectral envelope
overall sound-pressure level of 80 dB. The top panel shows a representation (three-dimensional response surface) that captures the long-run characteristics resulting from multiple presentations. The equivalency of representation is based upon the assumption that a channel represents spatial integration that can be considered statistically isomorphic to temporal integration for single units.

The synchronization index, as noted previously, provides a convenient way of representing synchrony in each channel. As shown in the bottom panel of Fig. 6, this particular channel is dominated by $F_1$ (512 Hz).

Figure 7(a)−(c) shows a normalized synchrony-place representation (three-dimensional response surface) that includes spectral components above dc for the output of high-SR fiber channels over a 40-dB (40–80 dB) range. It is evident from these plots that formant-relevant information is well encoded in the synchrony-place patterns, even at the highest sound-pressure level (80 dB). While all three formants are preserved across the entire range of intensities, synchrony to $F_1$ clearly becomes dominant at higher sound-pressure levels as a consequence of the broadening of the fiber channel filters.

The model's synchrony-place representation was corroborated by comparing its output with the single-unit data of Young and Sachs (1979). Their results are based on recording from a population of 269 AN fibers from a single cat in response to the synthetic vowel [ε]. They computed the magnitude of synchronization for each harmonic of the stimulus and expressed the result in terms of the synchronization index. Our analysis differs in that our model computed the synchronization index of the filtered, half-wave-rectified, and compressed waveform over a 25.6-ms Hanning-windowed segment of 40-kHz sampled speech. In contrast, Young and Sachs computed the synchronization index directly from the period histogram (64 or 128 bins over 7.81 ms) of single AN fibers. These period histograms were derived from responses over a 400-ms interval. Given that the probabilistic synchrony patterns are simulated by the time-varying output of the model, it is unnecessary to perform any sort of period averaging on the output. To ensure sufficient definition of the harmonics, a 1024-point fast Fourier transform (FFT) was used in our analysis.

Slices of the response surfaces derived from the composite model of the auditory periphery can be directly compared to the respective harmonics of Young and Sachs' synchronization index plots. Figure 7(a)−(c) includes both Young and Sachs' representation for harmonics corresponding to $F_1$ (4), $F_2$ (14), and $F_3$ (19) as well as slices from the composite model response surface at three intensities.

Figure 7(a) compares the model's output with Young and Sachs' data for an SPL of 38 dB. One of the more striking aspects of the physiological data is the large proportion of fibers synchronizing to the first formant frequency and the spread of this response toward higher CF units. Similar, but less extensive, synchronization is also observed to the second and third formant frequencies. With respect to the Young and Sachs data, synchronization to the first formant spreads significantly toward CFs above 3 kHz. There is a notch in the distribution of the first formant synchronization at the CF locations corresponding to $F_2$ and $F_3$. Again, the output of the model at 58 dB is similar to the Young and Sachs' data. At this moderate SPL, much of the auditory nerve is synchronized to the first formant. Figure 7(c) shows another 20-dB increase in stimulus level to 78 dB. The synchronization to the first formant proliferates in the physiological data to cover most of the AN fiber population. The only CF region not dominated by the first formant at this level is the region of the first formant notch between 2 and 4 kHz, where the second formant response is large. The synchronized response to $F_n$, in both physiology and the model, is significantly reduced relative to the patterns generated at the lower sound-pressure levels. At higher stimulus levels, the peak frequency synchronization computed by the model at harmonics corresponding to $F_2$ and $F_3$ is shifted to higher frequencies relative to that evidenced by the physiology. This overestimation of frequency is a result of basing the channel CF primarily on the high frequency tail of the ensemble iso-intensity functions, but is probably not of significant magnitude.

The significance of this tonotopic distribution of synchrony information remains controversial. In Young and Sachs' (1979) average localized synchronized rate (ALSR) model, the distribution of synchrony is irrelevant for fre-
frequency coding, except insofar as it affects the synchronization magnitude in any given channel to that channel's center frequency range. Thus, in the ALSR model, the system effectively computes the synchronization solely to those frequencies in the vicinity of the fiber CF. Synchronization to other frequencies is ignored. In this sense, the ALSR is a strict "place" model of spectral coding, in that the cue for frequency is the region of a fiber's projection from the cochlear partition. Synchronization is used only to estimate the magnitude of response for the unit CF. However, it is clear from Young and Sachs' data in Fig. 7(b) and (c) that significant spread of synchronization to $F_1$ does occur at higher SPLs and that the extent of such dispersion could also serve to convey response magnitude in a manner analogous to Zwicker's (1970) excitation pattern model for loudness.

V. DISCUSSION

A. Comparison with other models

Although it is now well established that the frequency selectivity of the auditory periphery is highly nonlinear, most models of the auditory periphery continue to use linear filters (e.g., Shamma, 1985; Seneff, 1988; Ghitza, 1988; Patterson and Holdsworth, 1991). In these models, the band-
[c] 38 dB SPL

[c] 58 dB SPL

(a) 

(b)
FIG. 7. Comparison of physiological population responses with the composite model simulation for vocalic stimuli presented at three different sound-pressure levels [38 (40), 58 (60), 78 (80) dB]. In the top panel, the distribution of synchronization magnitude as a function of characteristic frequency for cat auditory-nerve fibers responding to the vowel [ɛ] is displayed. Arrows point to the cochlear frequency position for each harmonic (corresponding to $F_1$, $F_2$, and $F_3$) for which the synchronization coefficient (ordinate) was computed. Each point represents the synchronization magnitude for a single fiber whose CF corresponds to the frequency shown on the abscissa. Data are from Young and Sachs (1979) and are published with the permission of the authors and the Acoustical Society of America. Displayed below the physiological responses are the corresponding harmonic slices across CF derived from the composite model. The bottom three-dimensional surface illustrates the concatenation of all harmonics above dc of the composite model output. The abscissa is the cochlea position expressed in terms of the characteristic (center) frequency for each of the 128 high-SR fiber channels. The frequency of the spectral analysis of the channel outputs is indicated on the ordinate. Note that the upper frequency ranges of the ordinate and abscissa differ by approximately 1 oct. The magnitude (synchronization index) for each frequency by fiber channel coordinate is indicated by the surface height of the contour. The magnitude scale is the same in (a)–(c). (a) Physiological population responses and corresponding composite model simulation for the vowel [ɛ] presented at 38 and 40 dB SPL, respectively. The three formant peaks are clearly delineated in the temporal responses of auditory-nerve fibers with CFs close to these spectral maxima. The distribution of activity synchronized to these formant peaks is relatively localized. (b) Physiological population responses and corresponding composite model simulation for the vowel [ɛ] presented at 58 and 60 dB SPL, respectively. The distribution of formant-synchronized activity becomes much broader, particularly for phase locking to $F_1$. Note that at this level $F_1$ synchrony is observed among both low- and higher-CF fibers (channels). There is a gap in the $F_1$ distribution corresponding to synchronization to $F_2$ and $F_3$. (c) Physiological population responses and corresponding composite model simulation for the vowel [ɛ] presented at 78 and 80 dB SPL, respectively. At this sound-pressure level, characteristic of conversational speech, the distribution of $F_1$ synchrony, as well as its magnitude, increases substantially, dominating the activity of the entire auditory-nerve array, except for a circumscribed CF region surrounding $F_2$ and possibly $F_3$. Although there is still an "island" of synchrony to $F_1$, phase locking to $F_1$ has been markedly reduced.
The result of these simplifications is that the distribution of function of sound-pressure level, and, in some instances, do not even vary as a function of cochlear frequency position. The result of these simplifications is that the distribution of temporal activity across the auditory-nerve-fiber array is not explicitly simulated. Linear, bandpass filters restrict the distribution of synchrony pertaining to the formant peaks to channels whose center frequencies are closest to the spectral maxima. Little or no spread of formant-related synchrony to the high-CF channels is observed in these linear models, in marked contrast to the physiological data of Young and Sachs (1979), Delgutte and Kiang (1984a), Miller and Sachs (1983), and Deng and Geisler (1987). Secker-Walker and Searle (1990) have recently pointed out that the population-response data of Miller and Sachs (1983) is inconsistent with narrowly tuned filters, though they made no attempt to characterize the shape of the filters that would describe the synchrony distribution.

The filter characteristics of the auditory models developed by Sinex and Geisler (1984) and Ghizza (1988) have also been based on auditory-nerve-fiber filtering. Unfortunately, in both instances, the filter characteristics were derived from single-fiber frequency-threshold curves (FTCs). FTCs provide a poor measure of cochlear filtering. Because the transduction is highly nonlinear (compressive), particularly at higher SPLs, the filtering observed near threshold, at the tip of the FTC is usually not a good estimate of the filtering at moderate or high SPLs (Lyon, 1991). This is particularly true for channels above 4 kHz in which the filtering broadens appreciably and becomes nearly low pass above 60 dB SPL. FTC-based models typically set the filter bandwidths considerably too narrow at SPLs common to speech (> 60 dB), and as a result the tonotopic distribution of low-frequency synchrony information is unrealistically narrow.

Iso-intensity functions, such as those upon which the present model is based, portray more realistically the extent to which the compressive and saturating nonlinearities are stimulus level and frequency dependent. With this information, it is possible to “sculpt” the filters as a function of SPL and cochlear place to conform to physiologically established parameters. One can then evaluate the accuracy of these filters by comparing the model's output with the population response profile of Young and Sachs (1979), as was done in the present study. Because of the relatively close fit between their physiological response profiles and the present simulation, we conclude that the shape of the filters used are a reasonable approximation to those in the auditory nerve. It would be of interest to compare the output of our simulation with those of other models in terms of the tonotopic distribution of neural synchrony.

Figure 8 illustrates the spatio-temporal output of 32 channels evenly distributed across the tonotopic range in response to the vowel [e] at intensities between 40 and 80 dB. This alternative representation affords a temporal view of the synchrony pattern prior to computation of the periodogram, allowing for a more direct comparison with other auditory models. In addition, intensities intermediate to those used in the comparison with the Young and Sachs data are shown. Several general characteristics of this representa-
FIG. 8. Spatio-temporal output of 32 channels evenly distributed across the tonotopic array, in response to the vowel [ε] at intensities ranging between 40 and 80 dB. The spatio-temporal representation, complementing the synchrony-place distribution surface, illustrates the progressive recruitment of synchrony to $F_1$ with increasing intensity. At lower intensities synchronization to the formant frequencies at high CFs is amplitude modulated at the frequency of the fundamental. Synchronization to the fundamental can also be observed in lower CF channels in the 50- to 70-dB range. A band of $F_2/F_1$ synchrony is clearly delineated in the mid-CF channels corresponding to the notch in the synchrony-place distribution.

cies is to reduce (and in many instances to abolish) the fiber's synchrony to frequencies close to its CF (for CFs below 4 kHz), a phenomenon variously described as synchrony recruitment (Sinex and Geisler, 1983) or synchrony suppression (Young and Sachs, 1979). The ALSR metric can be considered a place model because it only measures the amount of synchrony to frequencies appropriate to the fiber's cochlear position. However, the mechanism by which a fiber would "know" its appropriate cochlear-frequency position and be able to evaluate the temporal activity in its passband is assumed, but not motivated, by the ALSR model. In rate-based place models, this problem is circumvented because the neural activity metric (i.e., average rate) is the same for each fiber. Consequently, the spectral profile could be derived from the relational pattern of excitation without matching neural activity to threshold filter properties. For rate-place models, every fiber (or at least every frequency channel) "counts." The spectral representation is based, in principle, on the entire tonotopic distribution of rate activity. In contrast, the ALSR representation is largely based on
the activity of a comparatively small proportion of fibers, those with CFs close to the peaks of spectral energy. The output of other fibers is essentially ignored since their temporal activity does not conform to the unit CF, so it does not matter that these fibers may be highly synchronized to some other frequency.

The present model addresses these concerns in that the representation of certain key components of the speech spectrum, such as $F_1$ and $f_o$, are distributed over a broad tonotopic range of fibers. At high SPLs, synchrony to $F_1$ spans a range encompassing several octaves of cochlear frequency. Distributed representations, as evidenced by the use of coarse coding by various sensory systems (Heiligenberg, 1987; Baldi and Heiligenberg, 1988), yield extremely accurate encoding of information despite broadly tuned receptive fields. Central auditory processes may well have developed to efficiently integrate synchrony response properties across the range of tonotopic activity.

The speech spectrum is likely to be encoded in a number of different ways in order to ensure sufficient redundancy as to maximize the signal's intelligibility under a wide range of acoustic conditions. At low sound-pressure levels, both rate-place and synchrony-place information may very well provide an adequate representation of the formant pattern. At higher sound-pressure levels, the ability of rate-place information to provide a clearly delineated representation of the spectral envelope is severely compromised, among all but the lowest-SR fibers (Sachs et al., 1988). However, the amplitude of most consonants is sufficiently low as to preclude rate saturation across the auditory-nerve array (Delgutte and Kiang, 1984b). These low-intensity segments could thus effectively be encoded by rate-place information, at least in quiet.

B. Comparison of the simulation with physiological responses

Evidence in support of the composite model is provided by the close correspondence with the physiological observations of Young and Sachs (1979). It is clear from both empirical data and the simulation that the encoding of speech in the auditory periphery appears to differ in important ways from the conventional spectrum analysis or linear prediction coding. While the formant structure of voiced speech is well preserved in the synchrony-place representation across sound-pressure level, a type of distributed representation emerges across auditory-nerve fibers at higher intensities. A distributed representation of information is naturally more robust to interference relative to coding schemes dependent solely on rate-place information (Greenberg, 1988, 1991a,b).

It is rather remarkable that the output of the composite model is as similar to that obtained through direct physiological recording given that the analyzing filters were derived from pure-tone response areas. The average-rate iso-intensity curves serve as an estimate of the filtering characteristics of the entire cochlear partition, despite the fact that the function reflects many different components of cochlear transduction including mechanical tuning, the response of the inner hair cells and the transmission of this information to the auditory nerve. It is important to note that our approach has been functional rather than theoretical, whereby the linear structure has been statistically described by principal components and the nonlinear structure captured by nonlinear fitting of the rate-intensity functions.

C. Limitations of the model

Because the filters were derived from pure-tone response areas such nonlinearities as lateral suppression are not included in the model. Sokolowski et al. (1989) have suggested that such suppression may play an important role in the ability of certain AN fibers to maintain a rate-place representation of complex stimuli at higher SPLs. These fibers, typically of the low-SR variety, show considerably more suppression than the other fibers. The addition of lateral suppression contours to the filters would be expected to improve the rate-place representation of vocastic spectra.

Low-SR fibers, omitted from the current model, tend to be more sharply tuned than higher SR fibers. Future development of the model will include the increased dynamic range and frequency selectivity of low-SR fibers in the form of a separate population of filter functions. This addition would also be expected to enhance the rate-place representation.

The filters were derived from the average-rate behavior of AN fibers, and yet the simulation is focused on the temporal aspects of neural activity. Over most of the range of the filters, the disparity between rate-alone derived filters and synchronized-rate filters will be small. However, near rate threshold of the high-SR units the difference could be significant. In this range, the fiber will be highly synchronized to low-frequency signals even though the magnitude of the driven rate (firing rate—spontaneous rate) will be small or even zero (Rose et al., 1967). As a consequence, the filter shapes derived exclusively from average rate are somewhat more sharply tuned, particularly on the low-frequency slope, than would be filters derived from synchronized rate. For this reason, the upper-CF range of phase-locking to low-frequency signals in the present simulation is probably lower than would be the case with synch-rate filters. This disparity may become slightly more significant in the case of background noise, where the upper-CF phase locking may be particularly affected.

Both temporal and spectral features of speech are continuously varying—yet the present implementation of the composite model is restricted to the simulation of neural responses to steady-state signals for reasons of computational practicality. Thus the model, in its present form, does not accurately simulate auditory-nerve fiber rapid adaptation (Smith and Brachman, 1980a,b), likely to play an important role in encoding many consonantal features (Delgutte and Kiang, 1984b).

In addition to its static nature, the model is limited in several other respects. First, the saturating nonlinearity is based entirely on the rate-level function at a fiber CF, where the function is typically most compressive at higher SPLs. Incorporating frequency-dependent rate-level information may provide more realistic simulations of the rate-place profile since the model, at present, overestimates the degree of
cochlear compression and neural saturation. We believe that this limitation could be rectified by instantiating the frequency-dependent compressive characteristics into a dynamic representation of the response-area-based filters.

The crucial step toward a dynamic model is one of representing discharge rate as a continuous function of intensity and frequency. Hence, the response-area function can be visualized as a surface rather than a family of iso-intensity curves. The difficulty in deriving this function lies in choosing appropriate interpolation and approximation methods. Poor approximation of the function may lead to oscillation between interpolation points sampled from the original data. We are optimistic that at least one of several methods we are investigating will yield a well-behaved solution.

Each channel would then be represented by a unique function of intensity and frequency. The parameters of this function would be the acoustic signal expanded into its constituent frequency components on a time sample-by-sample basis. This construction requires an alternative to the basic idea of the short-time Fourier transform, such as a joint distribution describing the intensity of a signal simultaneously in time and frequency (Cohen, 1989).

We are currently investigating these incremental extensions to the current model. The additional computational demands to address the dynamics and frequency-dependent compressive characteristics are rather substantial. While the dynamic model will certainly improve the accuracy of the synchrony profiles, it will also be less tractable.

Conventional spectral analysis via the periodogram is not an ideal means of measuring synchrony. In this study, Fourier analysis has been used as a tool of convenience, as a means of quantifying the magnitude of periodicity passing through auditory-nerve-fiber channels, and as a means to compare the simulation's output with the data of Young and Sachs (1979), where the synchronization index was used as a quantitative measure. Fourier analysis is relatively insensitive to the modulation frequency corresponding to the fundamental \( f_0 \) of voiced speech. Consequently, synchronization to \( f_0 \) observed in high-CF fibers is underestimated in our composite model.

A physiologically plausible measure of synchrony, akin to auto- and cross-correlation, could be based on some form of coincidence detection, believed to occur at the level of the posteroverentral cochlear nucleus (Greenberg and Rhode, 1987; Kim et al., 1986; Rhode and Smith, 1986) and in more central regions of the auditory pathway (e.g., Langner and Schreiner, 1988; Abeles, 1982). However, until considerably more is known about such central temporal mechanisms, we believe that the synchronization index computed from the periodogram is an adequate method for quantifying synchrony if one is not focusing on \( f_0 \).

D. Potential applications of the composite model

We envision three potential applications of our composite model. First, it can serve as a means of visualizing the distribution of rate- and synchrony-based activity across the auditory nerve in response to complex sounds such as speech. Because of the nonlinearities in cochlear filtering, it would be difficult to visualize the rate-place and synchrony-place excitation patterns without such a simulation tool. Due to the limitations described above, the model is restricted, for the present, to signals with steady-state spectra such as synthetic vowels. Another version of the model is currently under development that will be capable of processing dynamic stimuli and plotting the evolution of the auditory periphery's response through time.

The simulation's output could be used, as well, to infer the nature of the afferent input into the cochlear nucleus. This sort of information would be of particular utility in understanding the processes governing the different physiological properties of cochlear nucleus neurons, particularly those, such as onset choppers, which receive a broad tonotopic projection from the auditory nerve.

Finally, the model could be used as a front-end, representative of spectral preprocessing in the auditory periphery, for studying broader psychological issues concerned with perceptual constraints on phonetic inventories, in a manner analogous to the simulations of perceptual contrast developed by Liljencrantz and Lindblom (1972) and Lindblom (1986). The output of the simulation could also serve as the input into a self-organizing neural network in order to determine features of the auditory-nerve fiber activity pattern that maximize transmitted information.

ACKNOWLEDGMENTS

This work was supported, in part, by a Dean of Letters and Science Predoctoral Fellowship awarded to the first author, as well as by NIDCD Grants DC 00719 (KRR), NS 26274 (SG), and NS 17590 (SG, WSR). We thank Dr. Eric Young, Dr. Shihab Shamma, and an anonymous reviewer for valuable comments on a draft of this paper. We would also like to thank Carol Dizack for help in preparing some of the figures and Terry Stewart for photography.

1 The rank of the matrix corresponds to the maximal number of independent rows or columns.

2 Linear, time-invariant filters are insufficient at accommodating the saturating nonlinearities characterized by the iso-intensity curves. In the current work, nonlinear gain takes the form of an instantaneous mapping to discharge rate. Ideally, filtering and nonlinear amplification would be integrated into a time-variant adaptive filter (see Sec. V).


