SRC TR 86-31

The Auditory Processing of Speech

by

S.A. Shamma

The Auditory Processing of Speech

Shihab A. Shamma

Abstract

The processing of speech in the mammalian auditory periphery is discussed in terms of the spatio-temporal nature of the distribution of the cochlear response and the novel encoding schemes this permits. Algorithms to detect specific morphological features of the response patterns are also considered for the extraction of stimulus spectral parameters.

THE AUDITORY PROCESSING OF SPEECH

SHIHAB A. SHAMMA

Electrical Engineering Dept & Systems Research Ctr
University of Maryland, College Park, MD. 20742.
Mathematical Research Branch, NIH. Bethesda, MD

abstract

The processing of speech in the mammalian auditory periphery is discussed in terms of the spatio-temporal nature of the distribution of the cochlear response and the novel encoding schemes this permits. Algorithms to detect specific morphological features of the response patterns are also considered for the extraction of stimulus spectral parameters.

The remarkable abilities of the human auditory system to detect, separate, and recognize speech and environmental sounds has been the subject of extensive physiological and psychological research for several decades. The results of this research have strongly influenced developments in various fields ranging from auditory prostheses to the encoding, analysis, and automatic recognition of speech. In recent years, improved experimental techniques have precipitated major advances in our understanding of sound processing in the auditory periphery. Most important among these is the introduction of nerve-fiber population recordings which made possible the reconstruction of both the temporal *and* spatial distribution of activity on the auditory-nerve in response to acoustic stimuli [1, 2]. Sachs et al. utilized such data to demonstrate the existence of a highly accurate temporal structure that is capable of providing a faithful and robust representation of speech spectra over a wide dynamic range and under relatively low signal-to-noise conditions [3, 4]. Their work has since motivated further research into the various algorithms that the central nervous system (CNS) might employ to detect and extract these and other response features, and the possible neural structures that underly them [5, 6].

In pursuit of these goals, we have constructed and analyzed the spatio-temporal response patterns of cat's auditory-nerve to synthesized speech sounds [4, 5]. These patterns are formed by spatially organizing the temporal response waveforms (or PST histograms) of the auditory-nerve-fibers according to their characteristic frequency (CF) [4]. The resulting display highlights the interplay of temporal and spatial cues across the fiber array and suggest novel ways of viewing cochlear processing and encoding of complex sounds [7, 5]. The availability of such experimental data, however, is at present limited by technical constraints and the massive amount of processing required to handle them. Thus, in order to analyze new speech tokens, and to facilitate the necessary manipulation of stimulus and/or processing conditions and parameters, we have developed detailed biophysical and computational models of the auditory periphery and used them to generate spatio-temporal response patterns to natural and synthesized speech stimuli. Various CNS schemes for the estimation of stimulus spectral parameters are then investigated based on these patterns.

The Cochlear Model :

Computational algorithms for the cochlear processing of speech are developed that are based on detailed biophysical formulations of linear basilar membrane mechanics and nonlinear hair cell transduction characteristics [8]. Basilar membrane analysis is based on detailed 3-D hydroelastic models that are quite efficient to compute [8, 9]. These models are used to generate the transfer functions at points along the cochlear length, which are then employed directly in all subsequent processing of speech sounds. The output (membrane displacement) at each point is transduced into hair cell intracellular potentials through two stages representing the velocity fluid-cilia coupling and the nonlinear hair cell. The latter stage can be approximated in most cases by a cascade of a compressive nonlinearity (of the form: $V = z.\exp(au)/(1+\exp(au))$ where (z,a,x) are constants with definite biophysical interpretations) followed by a low pass filter (time constant=0.1 ms). The final outputs then approximately represent the instantaneous probability of firing of the auditory-nerve fiber array. Many more detailed refinements have often been included in this model (e.g. synaptic adaptation mechanisms, middle and outer ear transfer functions, and some form of automatic gain control) to reproduce the finer details of the

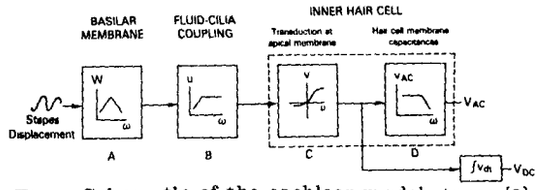


Fig.1: Schematic of the cochlear model stages [8].

responses. Nevertheless, the simpler model described above captures the major features of the experimental responses.

Examples of the model outputs are shown in Figs 2a,3 in response to a naturally spoken (female) /bɛt/ and a synthesized vowel /a/, respectively. In Fig.2a the response is to the onset of the vowel portion of the stimulus (whose spectrogram is shown in Fig.2b(right)). The periodic nature of the response is evident at regular intervals corresponding to the fundamental period of the stimulus. Strong harmonics, located near the formants of the vowel, dominate the response patterns over relatively broad segments of the channel array. Within each segment (e.g.$0.4<CF<1.6$ KHz) the travelling waves exhibit two important characteristics observed earlier in the experimental data: (1) Rapid apical decay due to the asymmetrical tuning of the basilar membrane amplitude. (2) phase shifts or delays in the response waveforms near the CF of the underlying harmonic, due to the rapid accumulation of phase-lag in the travelling wave near its point of resonance. The response to the plosive /t/ in /bɛt/ is also shown in Fig.2a, with its noisy character and high frequency content evident in the response patterns.

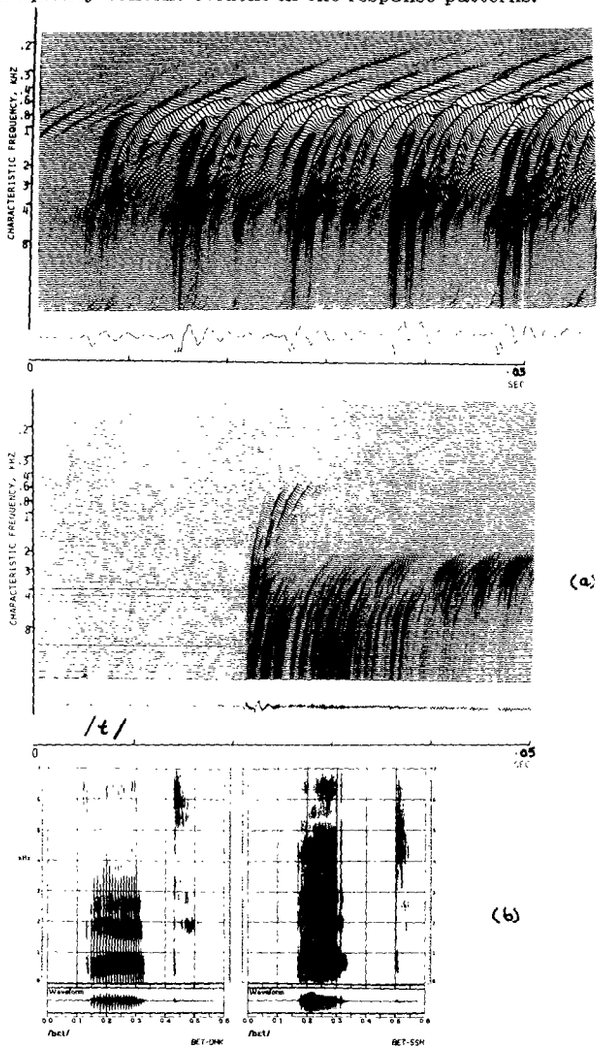


Fig.2: (a) Spatio-temporal responses of the cochlear n to selected portions of /bɛt/ spoken by a female. (b) trograms of /bɛt/ spoken by a male (left) and a (right) [12].

ENCODING THE ACOUSTIC SPECTRUM IN THE SPATIO-TEMPORAL RESPONSES OF THE AUDITORY-NERVE

Shihab A. Shamma

Department of Electrical Engineering and Systems Research Center, College Park, MD 20742
and
Mathematical Research Branch, NIH, Bethesda MD 20982

INTRODUCTION

Acoustical spectral features play an important role in the auditory perception and recognition of speech and complex sounds (Fant, 1973). These features are initially derived in the peripheral auditory system through the fine frequency selectivity of the basilar membrane (BM). Given the remarkable stability and robustness of the acoustic percepts, the underlying spectral cues must be preserved in the responses of the auditory-nerve over large ranges of intensity and noise conditions. Experimental evidence from single unit studies, however, has not been easy to interpret because several nonlinear phenomena of cochlear function conspire to alter drastically the apparent nature of nerve responses for different stimulus conditions (Sachs and Young, 1980). Consequently, several response measures have been proposed to detect and extract reliably the encoded spectral parameters. They may be organized along a continuum between two extremes: (1) Purely spatial measures that utilize only the spatial profile of the average rate of response along the tonotopically organized nerve-fiber array (e.g. place-code theory) (Sachs and Young, 1979). (2) Temporal periodicity measures that dispose of the tonotopic axis, using instead the periodicities in the response (phase locking) as measures of the stimulus spectral content (e.g. Dominant Frequency algorithm) (Sinex and Geisler, 1983). Other measures, such as the Average Localized Synchronous Rate (ALSR) and the Generalized Synchronous Rate (GSR), are intermediate in that they combine both aspects of the response (Young and Sachs, 1979; Seneff, 1984). The results of these studies have demonstrated that, bearing in mind such experimental constraints as anasthesia, processing the temporal cues is essential in providing a faithful representation of the stimulus spectra over wide dynamic ranges and low signal-to-noise ratios (Sinex and Geisler, 1983; Young and Sachs, 1979). These findings, however, raise a further dilemma: Unlike the average rate measure which can be derived using a simple counting or integrating neural network, CNS extraction of the temporal measures requires complex and precise neural network topologies that are capable of performing accurate periodicity analysis (Delgutte, 1984).

In order to address this issue, we have examined the spatio-temporal response patterns derived from cat's auditory-nerve-fiber population responses to synthesized speech stimuli (Miller and Sachs, 1983; Shamma, 1985a,b). We have also developed biophysical and computational models of cochlear function to extend the analysis where data are not available, and to examine the effects of manipulating stimulus and processing parameters (Fig.1) (Shamma et al., 1986). The main result that will be illustrated here is, that the stimulus spectral parameters are encoded as specific and stable morphological features in the spatio-temporal response patterns of the auditory-nerve. And, while temporal phase-locking may be important for the expression of these features,

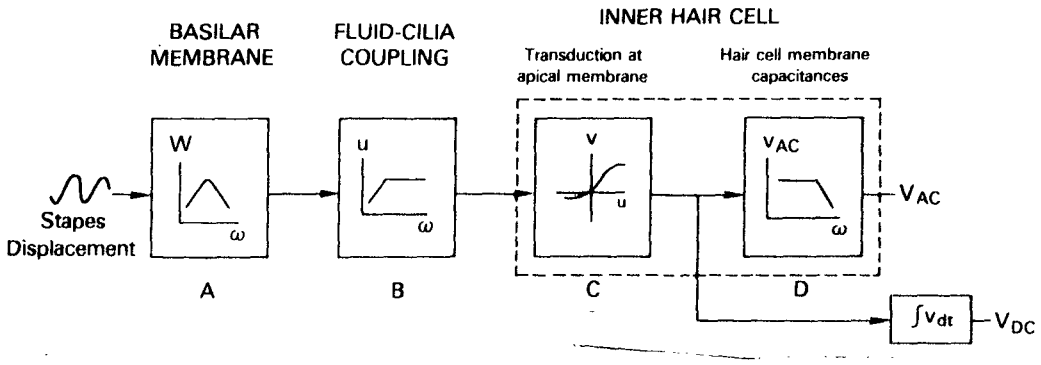


Fig. 1 Schematic of the simplified biophysical model of cochlear processing.

periodicity analysis need not be performed to extract them, but rather, simple edge detection algorithms suffice (e.g. lateral inhibitory networks (LIN)).

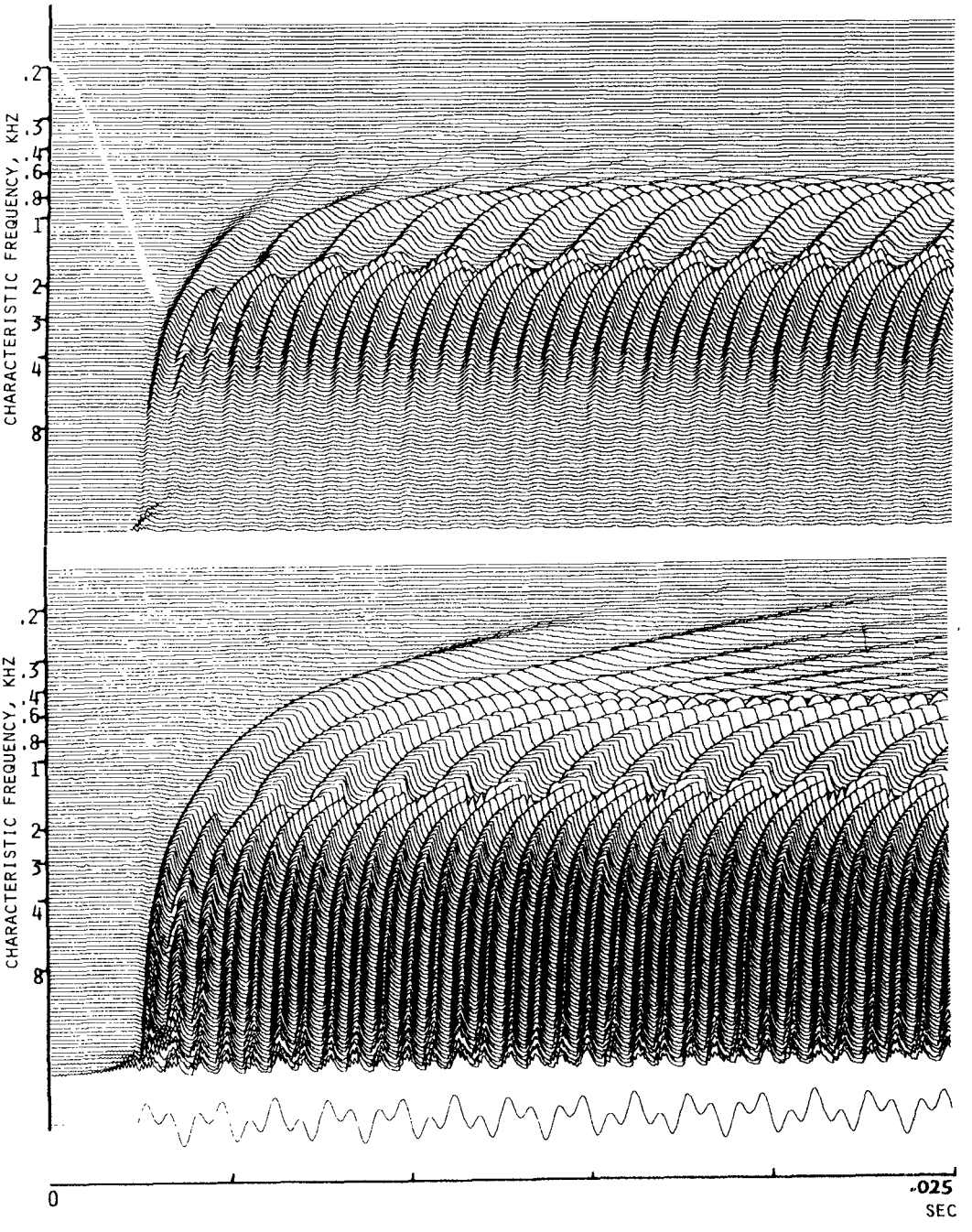
THE COCHLEAR RESPONSES

Before discussing the details of the cochlear model, we illustrate in Figs. 2a and 3a its spatio-temporal responses to two stimuli: A two-tone complex (600/1400 Hz), and a naturally spoken /dot/ (Zue, 1985). The patterns are constructed by spatially ordering the response waveforms from uniformly spaced locations along the cochlear partition model. They display the major response features observed earlier in the experimental data (Shamma, 1985). Thus in Fig.2a, the two tones recruit relatively broad regions of synchronous responses across the channel array. Near the characteristic frequency (CF) locations corresponding to the tones, two important features occur: (1) The travelling waves diminish rapidly (reflecting the steep apical slopes of the amplitude of the BM filters), and (2) the travelling waves slow down abruptly, causing the synchronous responses near CF to appear phase shifted or delayed relative to their neighbors. Both these amplitude and phase characteristics combine to form regions of rapid change (w.r.t. the spatial axis) that appear as edges extended parallel to the time axis, and are localized near the CF places of the stimulus components.

A simplified schematic of the biophysical cochlear model used to generate these patterns is shown in Fig.1 (Holmes and Cole, 1984; Shamma et al., 1986). It consists of a linear 3-D hydroelastic basilar membrane model, a velocity fluid-cilia coupling stage, and a biophysical model of the inner hair cell transduction mechanism. More detailed formulations have also been used (e.g. including outer and middle ear transfer functions, synaptic adaptation, and some form of automatic gain control), but the simplified model retains the major response features of interest here. The model contains one nonlinearity, due to threshold and saturation of the hair cell transduction mechanism. Its effects on the response waveform of each channel are rather simple to describe: It rectifies and compresses the waveform, allowing for only a limited dynamic range. This nonlinearity leaves invariant major spatial features of the spatio-temporal patterns over wide ranges of stimulus intensity. In contrast, when viewed in the frequency domain, its numerous manifestations are often misleading because of the apparent variability of the experimentally used synchrony response measures.

what does this mean

For instance, the V_{AC} tuning curves (Fig.1) (or the frequency transfer functions of the phase-locked responses) (Moller, 1983) of a given channel exhibit nonlinear dependence on stimulus intensity, such as bandwidth broadening and downward shifts of the frequency of peak response (BF) (Shamma, 1986; Moller, 1983). The shifts in the model occur because of the combined effects of a saturating nonlinearity, followed by a lowpass filter. It can be shown, however, that these shifts do not contradict a stable tonotopic map since the latter is a spatial map, and its definition at each location involves a single frequency, rather than the many frequencies of the tuning curve paradigm. With two tone stimuli, more complex phenomena can be simulated by the model, resembling those of synchrony suppression observed in hair cell and auditory-nerve-fiber responses (Javel et al., 1983). Here, the synchronous responses due to one tone are reduced by the presence of a second larger tone in the channel. This behavior is usually accompanied by a saturation of the channel, and hence



where is the 600/1400 response

Fig. 2a-b Spatio-temporal response patterns of the model to a two-tone stimulus (600/1400 Hz). (a) Moderate levels. (b) High levels (+40dB). Stimulus waveform is shown below each pattern. The CF axis is derived independently using single tone stimuli.

the loss of the information regarding the absolute levels of the stimuli. Nevertheless, the relative levels of the primary synchronous responses continue to reflect approximately the relative levels of the input tones over large ranges of absolute levels. Simulations with the cochlear model show that, compared to the linear case, the output relative levels of the highly saturated channels are generally enhanced by at most 6 dB in favor of the stronger tone. This apparently complex behavior again simply reflects the relative stability of the shape of the compound response waveform, since the instantaneous saturating nonlinearity does not significantly alter the gross features of the response pattern (e.g. the positions of the peaks and troughs in the waveform). Similar interpretation can be applied to other recently observed nonlinearities of the synchronous responses to multi-tonal stimuli (Horst et al., 1986). There are, however, other response nonlinearities that apparently arise from different cochlear sources and which are thus not reproduced by this model (e.g. rate suppression and propagating distortion products).

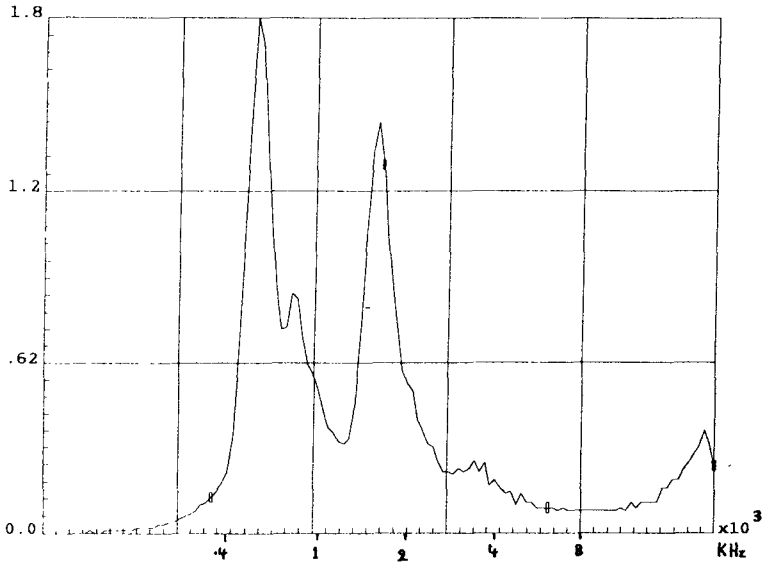


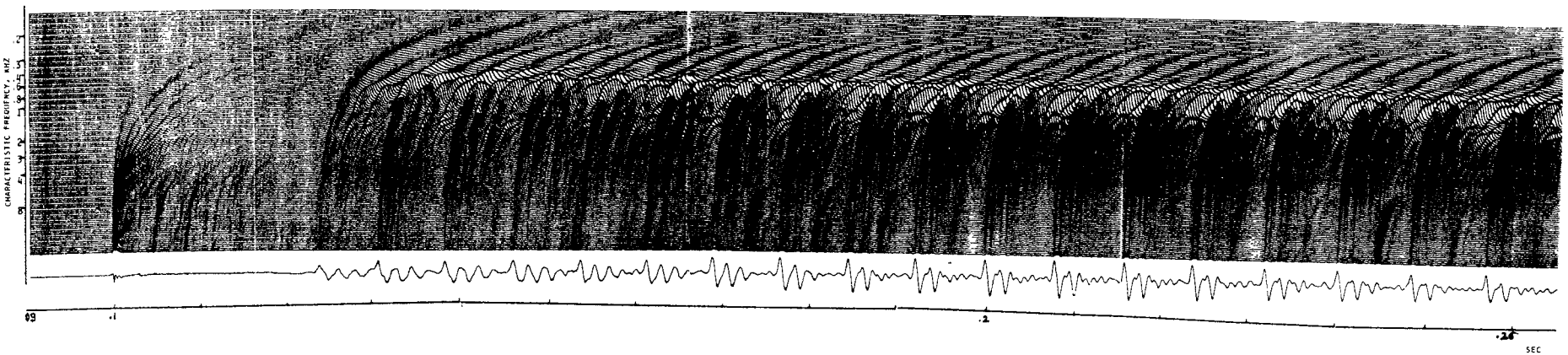
Fig. 2c Output of LIN processing of Fig. 2b patterns. The scale on the ordinate is linear (arbitrary units). The abscissa is the same CF scale of Figs. 2a-b.

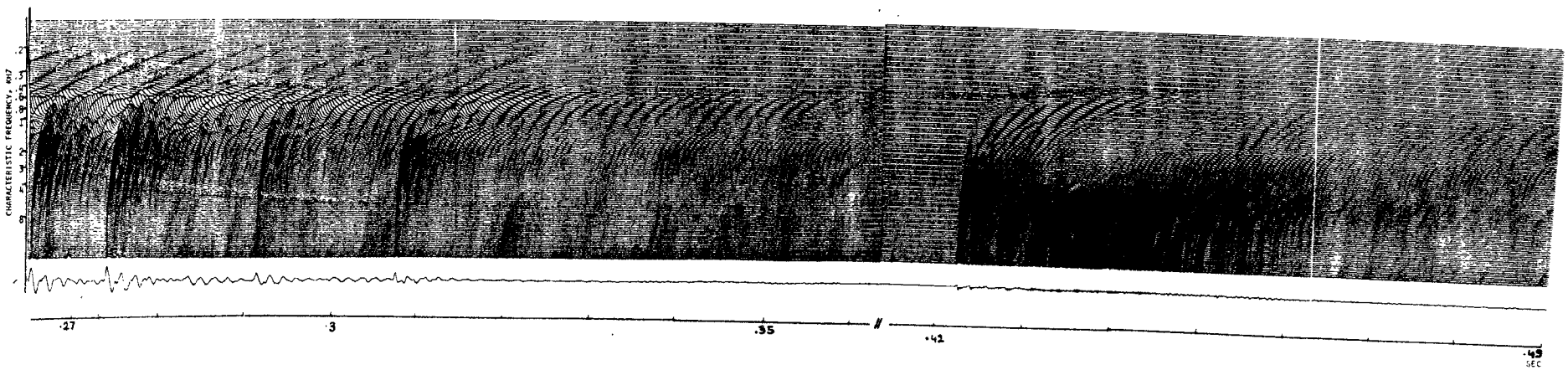
The stability of the spatio-temporal response features is illustrated in the two-tone patterns of Fig.2. The synchronous activity due to the two tones (600/ 1400 Hz) is spatially segregated due to the frequency analysis of the BM. If we move along the spatial axis (across the channel array) we observe the two regions of rapid transitions in the proximity of the CF locations 600 Hz and 1400 Hz. The first (CF=600 Hz) is due to the decay of the travelling wave amplitude and the rapid accumulation of phase-lag; The other (CF=1400 Hz) is marked by a rapid synchrony change from one tone to the other, again reflecting directly the tuning of the BM filters, and hence the change in the relative outputs across this border. The location of these transition regions (or edges) is quite stable with increasing sound intensity. This is demonstrated in Fig.2b where the stimulus level is 40 dB higher, resulting in more saturated output waveforms and a basal spread of the response. The response pattern remain relatively unaltered, and the edges stationary. Note that the fine temporal structure of the channel responses is important for the expression of these features. For instance, without synchrony, the border at CF=1400 Hz would simply disappear with the saturation of the channels in Fig.2b. For much higher frequencies, however, where phase locking diminishes in the basal CF regions, only the amplitudes of the BM filters contribute to forming the peaks and edges of the response spatial profile.

THE CENTRAL PROCESSING OF THE COCHLEAR RESPONSES

The response edge features can be viewed as spatially localized markers of the spectral components of the stimulus through the tonotopic map. In order to determine their locations, simple edge detection algorithms can be employed along the spatial axis. In the central auditory system, this may be accomplished by a spatially distributed lateral inhibitory network that receives its input from the tonotopically organized auditory-nerve-fiber array. These principles are demonstrated in Fig.2c which illustrates the reliable detection of the edges of the spatio-temporal patterns using a specific realization of a nonrecurrent LIN. This algorithm essentially subtracts from each trace a weighted sum of its close neighbors. In the particular case shown here, the weighting profile extends up to three traces on either side (which corresponds to an inhibitory field width of approximately 1/3 octave). The LIN output is then rectified and its rms value computed at 2 ms intervals using a 10 ms window. The plot in Fig.2c represents the LIN output, 20 ms after the onset of the saturated responses of Fig.2b. The locations of the two peaks correspond to the frequencies of the stimulus components.

More complex and dynamic patterns are seen in the responses (Fig.3a) to a naturally spoken /dot/ (whose spectrogram is shown in Fig.3d). During the voiced portions of the stimulus, the harmonics nearest to the formant frequen-


CHARACTERISTIC FREQUENCY, KHZ
09
.1
.2
.26
SEC


CHARACTERISTIC FREQUENCY, KHZ
.27
.3
.35
.42
.45
SEC

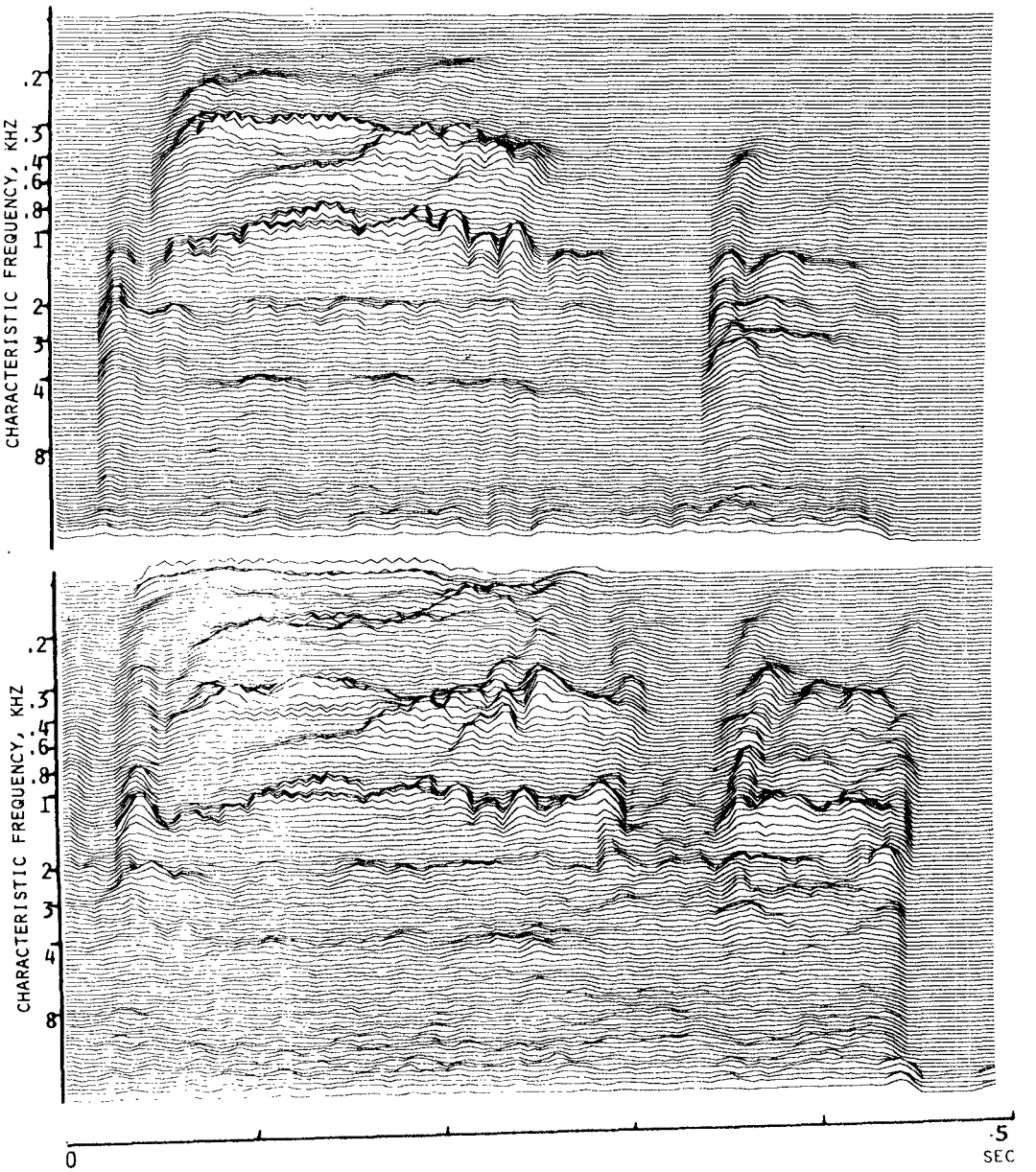


Fig.3b-c (b) Outputs of LIN processing of Fig. 3a patterns. Note that the time origin and scale here are different. Outputs can be aligned with Fig. 3a patterns based on stimulus onset. (c) Output of LIN processing of a highly saturated /dot/ response patterns (+50dB).

cies dominate the synchronized activity, being the largest and best resolved stimulus components. For example, during the vowel portion /o/ (t= 0.2-0.26 s) the response is dominated by the 5th harmonic (5 peaks/fundamental period of the response) in the .5 < CF < 1 kHz region, and by the 7th and 8th harmonics near the 1 < CF < 2 kHz region. There is also a clear disruption of the patterns near CF = 3 kHz. At a later interval (t = 0.27-0.31 s) the fundamental frequency of the stimulus decreases; since the formants remain approximately stationary (Fig.3d), the same CF regions are now dominated by higher harmonics. As before, edges occur at the CF's of these components, which can be extracted by LIN processing as shown in Fig.3b. The simulations and parameter extraction are repeated in Fig.3c under severe channel saturation (+50 dB) in order to illustrate the robustness of this approach. The LIN outputs trace the

Fig. 3a Spatio-temporal responses to naturally spoken /dot/. The time axis has been adjusted to correspond to that of Fig. 3d spectrogram.

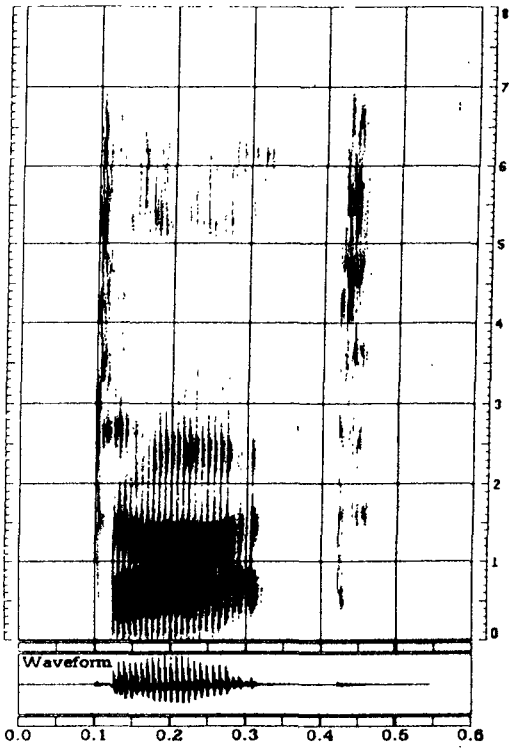


Fig. 3d Spectrogram of the /dot/ stimulus.

trajectories of the edges corresponding to the four stimulus formant regions (F1-F4) as they evolve in time. Thus, near the onset of the stimulus, F1 is represented by the 4th harmonic of the stimulus. At later times, the F1 shifts to higher frequencies and the 5th harmonic peak emerges and later dominates. Similarly, the F2 movements can be seen in the succession of harmonic peaks that are extracted at different time intervals. These trends can be readily detected in the response patterns themselves (Fig.3a).

SUMMARY

The spectral encoding scheme outlined above exploits the properties of the spatio-temporal distribution of responses in the auditory-nerve. It is a place-code in that the spectral parameter estimation proceeds by localizing specific spatial features along the ordered tonotopic axis. It is a temporal-code in that these features are formed by the amplitude and/or phase characteristics of the basilar membrane fine temporal responses across the auditory-fiber array. Simple edge detection networks are sufficient to isolate these cues in the CNS.

ACKNOWLEDGEMENTS

This work is partially funded by grants from NSF, the Minta Martin Foundation, and the Graduate Research Board at the University of Maryland.

REFERENCES

Delgutte, B. (1984). Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds, J. Acoust. Soc. Am., 75, 3, 879-886.

Fant, C. G. (1973). Speech Sounds and Features, MIT, Cambridge, MA.

Holmes, M. H. and Cole, J. D. (Sept. 1984). Cochlear mechanics: analysis for a pure tone, J. Acoust. Soc. Am., 76, 3, 767-778.

Horst, J. W., Javel, E., and Glenn, R. F., (1986). Coding of spectral fine structure in the auditory-nerve, I, Fourier analysis of period and interspike interval histograms, J. Acoust. Soc. Am., 79 (2), 398-416.

Javel, E., McGee, J., Walsh, E. J., and Farley, G. R., (1983). Studies of syn chrony suppression in normal and hearing-impaired cats, in Mechanisms of Hearing, ed. W. Webster, Monash University Press, Clayton.

Miller, M. I. and Sachs, M. B. (1983). Representation of Stop Consonants in the Discharge patterns of Auditory-Nerve Fibers, J. Acoust. Soc. Am., 74, 502-517.

Moller, A. R. (1983). Frequency selectivity of phase-locking of complex sounds in the auditory nerve of the rat, Hearing Res., 11, 267-284.

Sachs, M. B., and Young, E. D. (1980). Effects of nonlinearities on speech en coding in the auditory-nerve, J. Acoust. Soc. Am., 68, 858-875.

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

Seneff, S. (1984). Pitch and spectral estimation of speech based on auditory synchrony model, Working Papers on Linguistics, MIT.

Sinex, D. G., and Geisler, C. D. (1983). Responses of Auditory-Nerve Fibers to Consonent-Vowel Syllables, J. Acoust. Soc. Am., 73, 602-615.

Shamma, S. A. (1985a). Speech Processing in the auditory System I: Represen tation of speech Sounds in the responses of the auditory-nerve, J. Acoust. Soc. Am., 78, 1612-1621.

Shamma, S. (1985b). Speech processing in the auditory system. II: Lateral inhibition and the processing of speech evoked activity in the auditory-nerve, J. Acoust. Soc. Am., 78, 1622-1632.

Shamma, S. A., Chadwick, R., Wilbur, J., Rinzel, J. and Moorish, K., (1986). A biophysical model of cochlear processing: intensity dependence of pure tone responses, submitted to the J. Acoust. Soc. Am.

Young, E. D., and Sachs, C. D. (1983). Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers, J. Acoust. Soc. AM., 66, 1381-1403.

Zue, V. (1985). Speech Spectrogram Reading, Lecture Notes and Spectrograms, MIT.