SRC-TR-87-93

The Acoustic Features of Speech Phonemes in a Model of Auditory Processing: Vowels and Unvoiced Fricatives

by

S. Shamma

The acoustic features of speech phonemes in a model of auditory processing: vowels and unvoiced fricatives
J. of Phonetics, 1987 (in press)

Abstract

The acoustic features of three types of stimuli (a harmonic series, naturally spoken vowels, and unvoiced fricatives) are analyzed based on the response patterns they evoke in a model of auditory processing. The model consists of a peripheral cochlear stage, followed by two central neural networks. At the peripheral stage, the asymmetrical nature of the cochlear filters, combined with the preservation of the fine temporal structure of their outputs, provide for a robust and level-tolerant spatiotemporal representation of the speech signals. At the subsequent central stages, the cochlear patterns are processed by two layers of lateral inhibitory networks (LIN) to extract the perceptually important parameters of the stimuli. For the harmonic series, an in-phase and an out-of-phase version (one harmonic inverted) are used to illustrate the role of the spatiotemporal cues in encoding the spectral and temporal features of the stimuli. With the more complex vowel sounds, the primary acoustic features encoded by the LIN outputs are the few largest harmonic components of the stimuli, i.e. those closest to the formant frequencies. The output patterns computed for different (male and female) speakers display moderate variability, especially in the *locations* of the output peaks. However, the results also suggest that the *relative* levels of the LIN peaks (or the weight distribution of the patterns) is a more stable and characteristic feature of the different vowel groups. The results for the unvoiced fricatives indicate that the most invariant and distinctive acoustic feature the auditory model extracts, is the location of the high frequency edge of each stimulus spectrum.

Author:
S. Shamma, Electrical Eng. Dept. & Systems Res Ctr.
Univ. of Md., College Park 20742

1. Introduction

The detailed analysis of the acoustic features of speech phonemes and their relationship to perception is a difficult goal that has significant implications in many fields. Over the decades, the description of these features has been primarily based on linear Fourier spectral analysis methods and representations (e.g. spectrograms) which are sometimes augmented by various perceptually and physiologically based transformations to accomplish specific tasks, e.g. the mel scale for inter-speaker normalization (Bladon [1986]) and vowel categorization (Fant [1973]), or critical band filtering for speach recognition systems applications (Hermansky et al. [1986]). However, it has become apparent in recent years that these spectral representations differ in varying degrees from those generated by the mammalian cochlea and subsequently processed by the central auditory system (Delgutte and Kiang [1984]; Sinex and Geisler [1983]; Young and Sachs [1979]). Some of the unique and important features of the latter representations arise from the particular shapes of the cochlear filters (highly overlapping and asymmetrical), and the preservation of the fine temporal structure of the filters' outputs (up to 3-4 kHz) (Pickles [1986]).

In speech research, the primary motivations for using accurate biophysical models of the auditory system are two fold: First, to determine the perceptual relevance of the various acoustic features as represented in their output patterns, and the relationship of these features to the phonemic classes of the speech signal; Second, to discover the functional principles that underlie the remarkable abilities of the biological system to resist the detrimental effects of high noise environments and wide input dynamic ranges. Recently, pioneering experimental recordings from large populations of auditory nerve fibers have made it possible to examine closely for the first time the cochlear response patterns to selected speech segments (Miller and Sachs [1983]; Young and Sachs [1979]). These data, together with other important discoveries in basilar membrane mechanics and hair cell function, have facilitated the construction of cochlear models that can adequately replicate the primary response features at the level of the auditory nerve (Shamma et al. [1986]). With such models, it is now possible to analyze the response patterns associated with a wide variety of speech sounds, and under many signal conditions. Beyond the peripheral auditory stages, however, little is known about the central neural networks and the processing they perform on the cochlear outputs. This adds a significant element of uncertainty to the analysis since apparently useful cues and response features at the auditory nerve level may be inconsequential for the phonemic perception and classification problem if the central nervous system ignores, or is incapable of processing them. Nevertheless, a few criterea maybe helpful in the critical assessment of these processing strategies and in utilizing the parameters thus extracted; these include the plausibility of their biological implementations and the degree to which the isolated parameters reflect and explain the psychophysical measurements (Shamma [1985a]).

In this report, we shall utilize a multistage model of auditory processing to examine the identity and the neural expression of the perceptually significant parameters of some sustained speech sounds. The outlines of the model are discussed in section 2. Illustrative response patterns of a simple harmonic series stimulus, several naturally spoken vowels of both sexes and a sequence of unvoiced fricatives, are discussed in detail in section 3.

2. Models for Auditory Processing

Speech signals are processed in a model of three stages mimicking the peripheral and central auditory analysis of sounds. The three stages are: the cochlear processing stage, the nonrecurrent lateral inhibitory network (LIN.I), and the recurrent lateral inhibitory network (LIN.II). The cochlear stage is based on a composite mathematical model of the biophysics of the cochlea (Shamma et al. [1986]). It incorporates a linear formulation of basilar membrane mechanics, a fluid-cilia coupling stage which transforms membrane vibrations into hair cell cilia displacements, and a simplified description of the inner hair cell nonlinear transduction of cilia displacements into intracellular electrical potentials. The potentials at each hair cell along the cochlear partition is then taken as a measure of the probability of firing of the nerve fiber innervating it. Many more details of cochlear function can be incorporated in such models, e.g. adaptation at the hair cell/ nerve synapse (Westerman and Smith [1984]), active mechanisms of basilar membrane motion (Neely and Kim [1982]), the effects of the middle ear muscles and of the efferent system (Winslow [1985]). Nevertheless, the simplified model used in this study reproduces the major response properties observed experimentally, especially with relatively steady and broad-band stimuli like the vowel and fricative phonemes considered here. The outputs of the cochlear model are computed at 128 equally spaced locations along the cochlear partition, and are all displayed together as a 2-dimensional spatiotemporal pattern representing the ensamble activity of the tonotopically organized array of auditory nerve fibers (Shamma [1985a]). The spatial axis is labeled by the characteristic frequency (CF) of each output channel in a manner described at the end of this section. More details regarding the computer implementation of the model are available in the appendix (A1).

In response to speech stimuli, the cochlea generates intricate spatiotemporal patterns which encode the perceptually significant parameters of the signal. Neural networks in the auditory system may extract these parameters by processing specific features created by the interplay of time and space in the response patterns. An example of such features are the discontinuities produced by the rapid change (at specific positions along the spatial axis) of the amplitude and phase of the responses to specific components in the signal (Shamma [1985b]). These features reflect the highly asymmetrical nature of the cochlear filters, and can be used as cues to the identity of the underlying stimulus frequency. For the extraction of these cues, we have analyzed and implemented models of spatially distributed LIN's (Shamma [1985a]; Shamma [1986b]). Examples of such networks abound in biological sensory systems (e.g. the visual system, and the somatosensory system), where they serve to sharpen a spatial input pattern by highlighting its edges and peaks, and in some cases they also sharpen its temporal course (Hartline [1974]). These networks are also capable of more complex functions (e.g temporary storage and oscillations) when the nonlinearities typical of neuronal function (e.g. threshold and saturation) are incorporated in their description (Morishita and Yajima

[1972]).

In this study, two stages of LIN's are used to process the cochlear spatiotemporal patterns, resulting in the extraction of parameters that are potentially significant in the auditory perception of speech phonemes. The first network (LIN.I) is linear and nonrecurrent, ... subsequent thresholding and time window averaging (details of this network's construction and operation are discussed in (Shamma [1985a]) and briefly in appendix A2). The second network (LIN.II) is a nonlinear recurrent LIN which further sharpens the LIN.I outputs. Its mode of operation is quite different - the strengths and profiles of the recurrent inhibitory connections are such that, a *locally* large peak in the input pattern dominates completely the output activity in its neigberhood, thus leaving in the final pattern only the prominent features extracted by the LIN.I (further details of the network topology and its inhibitory weights are discussed in appendix A2). Using the output of this network to various single tone stimuli, a frequency-to-position map was generated and used to label the spatial axis of the cochlear generated responses. The resulting scale is approximately linear at low frequencies ($\leq$ .4 kHz) and becomes logarithmic beyond 1 kHz (Shamma [1986b]).

3. The Auditory Responses

The auditory processing of complex sounds will be illustrated in the case of three types of sustained stimuli: (1) A synthetic harmonic series, (2) natural voiced vowels, (3) a series of unvoiced fricatives.

3.1 The harmonic series

Fig.1a illustrates the cochlear responses to a stimulus consisting of 15, closely spaced (100 Hz fundamental), in-phase harmonics (300-1700 Hz). The different harmonics excite travelling waves along the basilar membrane which are reflected in the fine temporal structure and spatial spread of the cochlear responses. An important property of the responses to the lower, partially resolved, components ($\leq$ .5 kHz) is that the travelling waves they initiate decay in amplitude, and begin to accumulate phase, rapidly at different locations along the array, each depending on the frequency of the underlying harmonic. The expression of both these features progressively deteriorates as the harmonics become spatially less segregated (less resolved) and begin to interfere (e.g. the responses at the CF's of the higher harmonics in Fig.1a, .7 $\leq$ CF $\leq$ 1.7 kHz).

Conversely, these features become much more pronounced in the responses to a given harmonic if it is distinguished from its neighbors by a larger frequency separation (e.g. a larger fundamental frequency), a higher amplitude (as under vowel formants), or even by a phase shift. For example, in Fig.1b, the 11th harmonic of the stimulus is inverted relative to the ensemble. Both features of the travelling wave due to this harmonic now stand out and combine to create the illusion of an edge or discontinuity in the response patterns around CF=1.1 kHz. It is precisely these kinds of edges that the LIN's can detect and signify at their outputs. Examples of LIN.I outputs for the in-phase and out-of-phase stimuli are shown in Fig.1c. In the in-phase condition, only small peaks corresponding to the partially resolved, low order harmonics are visible. In the out-of-phase condition, however, the peak due to the inverted harmonic is quite prominent relative to the background of the harmonic complex[1],

[1]Note that the difference between the in-phase and out-of-phase outputs would not exist for a resolved

a fact which may account for the audibility of such tones in analogous psychophysical experiments (Duifhuis [1970]). Since the cues upon which the LIN extraction here is based are temporal (the edges being created by disparities in the temporal course of the response waveform at adjacent locations of the channel array), the LIN outputs are relatively insensitive to the saturation of the cochlear channels resulting from large stimulus amplitudes (see also (Shamma [1986a])).

The LIN could equally well detect such peaks and edges if they exist in the spatial profiles of the average cochlear outputs, i.e. ignoring the fine temporal structure and sensing only a measure of the short-time average power in each channel (as, for instance, is done in the usual spectrograms). At moderate stimulus levels, however, such a representation deteriorates significantly because of the saturation of the cochlear outputs, leaving only residual cues (Sachs and Young [1979]). In the normal behaving animal, there are many ways in which this saturation might be circumvented (Pickles [1986]). Furthermore, it can be shown that, with a complex stimulus as that of Figs.1, the saturation level of the different channels across the array is not uniform, but rather varies slightly depending on the exact waveform of the underlying basilar membrane motion at each location. Thus, even with complete saturation of the average rates across the array, small potentially useful cues (discontinuities) do persist.

3.2 The vowel stimulus

We examine here three aspects of the auditory responses to vowels: The acoustic features, the variability of the responses across different speakers, and the representation of the pitch.

3.2.1 The acoustic features

The auditory response patterns of a naturally spoken vowels were analyzed, both from male and female speakers, and in many noise and signal amplitude conditions. The spatiotemporal outputs for four representative vowels (/i/,/⊃/,/U/,/u/) (Zue [1985]) are shown in Figs.2a. Besides the general features discussed earlier (Figs.1), the vowel responses possess a typical structure, also observed in all experimental data (Miller and Sachs [1983]; Shamma [1985b]; Sinex and Geisler [1983]) - that is the dominance of the entire response pattern by a few stimulus harmonics. These harmonics correspond to the largest components in the stimulus spectrum. Their frequency and amplitude depend on the formant structure of the vowel (usually F_1 and F_2) and its fundamental frequency (F_0). Thus, for the vowel /i/ (Fig.2a), there are two response domains: the first is at approximately CF $\leq$ 2 kHz, corresponding to the F_1 harmonics (2-4); an abrupt transition in the response patterns occurs at this point as the harmonics associated with the higher formants become dominant. These trends are seen again in the /⊃/ vowel responses, where F_1 and F_2 occupy intermediate locations (F_1 at CF $\approx$ 650 Hz; F_2 at CF $\approx$ 1 kHz). For /u/, the amplitude of F_1 relative to other formants is such that the low harmonics dominate the entire pattern. The voicing and the fundamental period of the vowels are reflected in the periodic structure of the responses, especially at the higher CF's since the channel bandwidths here are relatively broad. There are many more

harmonic. Thus, in a cochlear model with narrower filters, the 11th harmonic becomes more resolved, hence reducing the distinction between the outputs to the two stimuli.

detailed features in these outputs that presumably relate more specifically to characteristics of this speaker's voice and articulation (and upon which listeners' identification of different individuals is based).

For the central processing of these patterns, the LIN detects and generates output peaks at the locations of the edges produced by the dominant stimulus harmonics. These peaks usually encode the strongest high frequency formants, one or two resolved harmonics near F_1, and the fundamental F_0 if it is relatively high (e.g. a female voice). For instance, the LIN.I and LIN.II outputs of the 7 vowel series shown in Figs.2b-c (same male speaker as Fig.2a responses) reflect well the overall formant patterns of the vowels as derived from spectrograms (Zue [1985]). An additional important aspect of these outputs is the systematic change in the relative *amplitudes* of the low- and high-CF peaks along the sequence (or equivelantly, the downward CF shift of the 'center of gravity' of the output patterns). Thus, for the high vowels at either end (/i,u/), the high-CF peak is relatively large when the constriction is fronted (as in /i/), and vice versa in the back vowel /u/[2]. The open vowels /æ/ and /ɔ/ occupy an intermediate position in that the two peaks are comparable[3].

These relations are summarized schematicaly in Fig.2d. On the left, the vowels are organized along a continuum in the plane of A_1,A_2 - the relative amplitudes of the low and high-CF peaks respectively. The small arrows indicate the effects of lip-rounding (see below, and footnote 2). The figure on the right illustrates the organization of the same vowels on the plane of two articulatory features: The open-close axis reflecting tongue height, and the front-back axis indicating the position of the constriction. These two figures are closely related, in that the vowel continuum in the A_1,A_2 plane (left) can be thought of as the continuum that would result if we project the vowels in the articulatory plane unto the front-back axis. Since movement along the latter axis correlates well with the length of the front cavity, the organization of the vowels in the A_1,A_2 plane (i.e. the relative height of the LIN peaks) may also reflect the effects of the position (frequency) of the 'front cavity resonance' and the so-called F_2' (Carlson, Grantstorm and Fant [1970]), which also move in the same direction for this sequence of vowels (Kuhn [1975]). Finally, the effects of lip-rounding in this schematic are viewed only as local modulations (in the direction of the arrows; see footnote 2) of the parameters already established by the articulatory features. Therefore, it is possible to reach the same point along the vowel continuum of the left figure with different combinations of lip-rounding and front-back articulations (Kuhn [1975]).

3.2.2. The response veriability

It has been suggested that the long observed variability in vowel spectra across speakers, and particularly between the sexes (Fant [1973]), may decrease significantly when physio-

[2]LIN.II outputs (not shown) of the vowels /y,ɔ/ (latter vowel pronounced as in the initial segment of the Russian vowel /bl/) present more extreme opposing situations, with /y/ having a very small low-CF peak, and /ɔ/ a very small high-CF peak. Note that for both groups of close vowels (the frontal /i,y/ and back /u,ɔ/), the place of the constriction is the primary factor in determining the overall weight distribution of their outputs. Lip rounding seems to have only a secondary effect, increasing slightly the relative size of the higher CF peaks.

[3]The absolute and relative amplitudes of the LIN peaks are clearly influenced by the type of filtering and other processing stages of the auditory model. However, given a fixed set of conditions (or model parameters), the change in the relative amplitudes among the different vowels is primarily a property of the signal parameters and not of the analysis system.

logically based amplitude and frequency scaling of these spectra is applied (Bladon [1986]). Fig.3 illustrates some of the variability in the LIN.II outputs to the (nominal) vowels /i/, /ɔ/, and /u/, with two male and two female speakers[4]. The male outputs are reasonably consistent among themselves and with those of Fig.2c in terms of the positions of peaks. Female outputs, however, differ in many details. For instance, both female output peaks are typically located at slightly higher CF's than corresponding male peaks. One of the females (Fem2) had an additional peak for /i/, and a missing peak in /ɔ/. Nevertheless, despite this variability, the earlier observations regarding the relative levels of the peaks of the vowel sequence largely hold for all outputs here. Three factors may contribute to the observed variability: (1) Subtle differences in vowel colors among the different speakers (this in fact may account for (Fem2) differences, as she had a distinct pronounciation); (2) Different F_0 frequencies between males and females which affect mostly the positions of the low-CF peaks (females $F_0 \approx 250$ Hz;males $F_0 \approx 125$ Hz); (3) The effects of the differing pharangial cavity length between males and females (Fant [1973]). It is apparent, therefore, that the LIN.II outputs are influenced at this stage by many of the subtle qualities of the tested vowels, and that the categorical organization of different vowel systems must therefore occur at higher level networks in the brain. However, more definitive statements regarding the degree of variability in the LIN.II outputs as compared to the usual spectra must clearly await the analysis of a much larger sample of speakers and vowels.

3.2.3 Pitch representation

There is considerable evidence in the literature that the perception of the pitch of voiced speech or a harmonic series is derived from the lower CF, relatively resolved, harmonic components (Moore and Glassberg [1986]). It is still unclear, however, whether purely spatial cues, purely temporal cues, or both are involved in encoding these harmonics, or what the nature of the neural central mechanisms that process these cues is. In the LIN outputs, the beating in the time course of the central peaks emerges as a likely correlate of the pitch percept (Fig.3). For instance, since this beating occurs at the LIN peaks, it is effectively derived from the edge or discontinuous regions of the cochlear spatiotemporal outputs. At these regions, the LIN combines locally dissimilar traces arising from different *resolved* harmonics, producing output peaks that are proportional to the sharpness of these edges. When the harmonics are well resolved, only harmonic pairs may interact locally at the LIN inputs, and hence the output beating is phase-insensitive (Greenberg [1980]). However, the beating waveform becomes phase-sensitive when more partially resolved harmonics interact, and disappears entirely when no local harmonics are resolved (and therefore no LIN peaks exist). Similarly, no beating occurs where phase-locking deteriorates (e.g. the cochlear outputs of resolved high frequency harmonics). All these properties of the LIN outputs and many more, find their correlates in the pitch perception literature (Moore and Glassberg [1986]) and in neurophysiological data (Greenberg [1980]). Finally, the saliency of the pitch percept can also be correlated with the number and consistency of the beating outputs across the channel array. It is unclear, however, if or how the more central neural networks may further process

[4]The LIN outputs in these figures are averaged with a relatively narrow window ($\approx$ 4 msec) in order to bring out the temporal character of the response (see discussion of pitch correlates next subsection). The display of the temporal modulation of the female outputs can be further improved with a narrower window ($\leq$ 2 msec).

the LIN output correlates to produce the pitch percepts.

3.3 The unvoiced fricatives

The auditory responses of unvoiced fricatives differ considerably from [illegible] vowels. To start with, the random nature of the excitation is quite evident in the cochlear responses (Fig.4a) where the travelling waves are now initiated at random points along the time axis. Another distinctive aspect is the predominance of the high frequency components and their sudden decay at a different location for each of the fricatives. The edges created by this decay in the spatiotemporal patterns are directly related to the steep roll-offs at the high cut-off frequencies of the frictive spectra. The LIN detects these discontinuities, producing large peaks at the corresponding output locations.

Fig.4b illustrates the LIN.II outputs of the fricative sequence /s/,/./,/s/,/X/,/h/, whose articulation involves a progressive backward shift of the point of constriction. The most prominent feature in each of the LIN.II outputs is the largest and lowest CF peak, which is extracted from the CF regions described above. Along the fricative sequence, the progressive downward shift of the CF of this peak reflects the lengthening of the frontal cavity which largely determines the low frequency extent and overall shape of the spectra of these fricatives (Fant [1973]). In our small sample, the existence and the CF range of these 'edge' peaks are reliable cues to the identity of the fricative - in agreement with findings of speaker normalization studies (Bladon [1986]), and of psychoacoustical experiments using high-pass filtered white noise (with variable cut-off frequencies) to model these stimuli (Fant [1973]). The other peaks in the LIN output reflect the fine structure of the the fricative spectrum, and are often indicative of the vowel context. Finally, note that the absence of the voicing in these stimuli is paralleled by the randomness and absence of beating in the LIN outputs.

4.0 Summary

We have examined the acoustic features of various complex stimuli as derived from the response patterns of a model of auditory processing. The details and nature of the spectral and temporal cues detected by the models are illustrated by the reponses to a synthesized harmonic series in two cases: An in-phase and an out-of-phase version. Preserving and making use of the temporal structure of the cochlear outputs, emerges as a powerful means for encoding the stimulus parameters and may also explain the psychophysical percepts usually associated with the above phase manipulations. Extending the analysis to speech phonemes reveals a wide disparity in the relative contributions of the different stimulus components to the models outputs. Thus, with natural vowels, a few large harmonics (near the formants) dominate the entire cochlear patterns; Furthermore, the relative amplitudes of the corresponding LIN output peaks appears to be a characteristic feature of each vowel. In the case of unvoiced fricatives, the acoustic feature that the auditory model most reliably detects is the parameters (location and slope) of the high frequency roll-off (edge) in the stimulus spectra. group.

Appendix

A1. The cochlear model:

The cochlear spatiotemporal patterns are computed using digital algorithms based on a detailed biophysical model of the cochlea (Holmes and Cole [1984]; Shamma et al. [1986]). At each of 128 location along the basilar membrane, the transfer function is computed and used in an FFT-based overlap-and-add method to generate the membrane's response to the stimulus. This output is then highpass filtered (y_n=x_n-.8x_{n-1}; modelling both outer ear and fluid-cilia coupling stages) and compressed by a sigmoidal function of the form: $y = M \cdot 1/(1 + b \cdot e^{-a \cdot x})$, where a, b, and M are parameters of the nonlinearity, and y, x are the output and input respectively. Finally, a lowpass filter smooths the output (time constant = .1 msec). The parameters of the compressive nonlinearity should be such that approximately 30 dB of linear gain is available between threshold and saturation (defined as .1 - .9 of maximum output level, M) and that the output is saturated at moderate sound levels (approximately 60 dB SPL).

A2. The lateral inhibitory networks:

The cochlear outputs are processed by two stages of LIN's:

LIN.I - The first network is implemented in a nonrecurrent topology for computational efficiency. Its mode of operation is discussed in detail in (Shamma [1985a]). This network is modelled by a highpass, linear phase, FIR filter (symmetric coefficients: -.02, -.08, .3, .6, .3, -.08, -.02). Each of the 128 resulting outputs is then half-wave rectified, and time-window averaged (window width=10 msec, computed at 4 msec intervals) to generate the final traces shown in Fig.2b.

LIN.II - The second network is a recurrent network whose purpose is to sharpen further the LIN.I outputs. At its steady state, it is described by the set of equations:

$y_i = g(x_i - \sum_{ij} w_{ij}\, y_j)$, for i = 0 .. 127,

where g(u) = max(u,0), x_i is the input to the network at the i^{th} location, and y_i is the corresponding i^{th} output which satisfies the mapping above. At each time instant, the LIN.I input is applied (the **x** vector) and a set of outputs is computed (the **y** vector) by iterating the mapping from zero initial conditions. The w_{ij} connectivity is chosen *wide and strong* enough to preserve in the LIN.II outputs only the large and moderately separated peaks (approximately 6 traces or more) (Morishita and Yajima [1972]). The symmetric w_{ij} profile of inhibitory connection used is given by the set: 0 (midpoint), .02, .05, .1, .15, .2, .25, .25, .2, .15, .1, .0 and its reflection.

References

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

R. CARLSON, B. GRANTSTORM AND G. FANT, *Some studies concerning perception isolated vowels*, STL-QPSR (1970).

B. DELGUTTE AND N. Y. KIANG, *Speech coding in the auditory nerve: I. Vowel-like sounds*, J. Acoust. Soc. Am. , 75 (1984), pp. 866–878.

H. DUIFHUIS, *Audibility of high harmonics in a periodic pulse*, J. Acoust. Soc. Am., 48 (1970), pp. 888–893.

C. G. FANT, *Acoustic description and classification of phonetic units*, in Speech Sounds and Features, MIT, Cambridge, MA (1973).

S. GREENBERG, *Temporal neural encoding of pitch and vowel quality*, UCLA working papers on phonetics (1980).

H. K. HARTLINE, *Studies on excitation and inhibition in the retina*, Rockefeller University Press, New York (1974).

H. HERMANSKY, K. TSUGA, S. MAKINO AND H. WAKITA, *Perceptually based processing in automatic speech recognition*, Proc. IEEE-ICASP (37.5), Tokyo (1986).

M. H. HOLMES AND J. D. COLE, *Cochlear mechanics: analysis for a pure tone*, J. Acoust. Soc. Am., 76 (1984), pp. 767–778.

G. M. KUHN, *On the front cavity resonance and its possible role in speech perception*, J. Acoust. Soc. Am., 58(2) (1975), pp. 428–433.

M. I. MILLER AND M. B. SACHS, *Representation of stop consonants in the discharge patterns of auditory-nerve fibers.*, J. Acoust. Soc. Am., 74 (1983), pp. 502–517.

B. C. J. MOORE AND B. R. GLASSBERG, *The role of frequency selectivity in the perception of loudness, pitch and time*, in Frequency Selectivity in Hearing, B. C. J. Moore, ed., Acadamic Press, London (1986), pp. 251–302.

I. MORISHITA AND A. YAJIMA, *Analysis and simulation of networks of mutually inhibiting neurons*, Kybern., 11 (1972), pp. 154–165.

S. T. NEELY AND D. O. KIM, *An active cochlear model shows sharp tuning and high sensitivity*, Hearing Res., 9 (1982), pp. 123–130.

J. O. PICKLES, *The neurophysiological basis of frequency selectivity*, in Frequency Selectivity in Hearing, B. C. J. Moore, ed., Acadamic Press, London (1986), pp. 51–122.

M. B. SACHS AND E. D. YOUNG, *Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate*, J. Acoust. Soc. Am., 66 (1979), pp. 470–479.

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

S. A. SHAMMA, *Speech processing in the auditory system. I: Representation of speech sounds in the responses of the auditory-nerve*, J. Acoust. Soc. Am., 78 (1985b), pp. 1612–1621.

S. A. SHAMMA, *The auditory processing of speech*, Proceedings of the Symposium on Speech Recognition, Montreal (1986a).

S. A. SHAMMA, *Encoding the acoustic spectrum in the spatio-temporal responses of the auditory-nerve*, in Auditory Frequency Selectivity, B. C. J. Moore and R. Patterson, eds., Plenum Press, Cambridge (1986b), pp. 289–298.

S. A. SHAMMA, R. CHADWICK, J. WILBUR, J. RINZEL AND K. MOORISH, *A biophysical model of cochlear processing: intensity dependence of pure tone responses*, J. Acoust. Soc. Am. (1986).

D. G. SINEX AND C. D. GEISLER, *Responses of auditory-nerve fibers to consonent-vowel syllables*, J. Acoust. Soc. Am., 73 (1983), pp. 602–615.

L. A. WESTERMAN AND R. L. SMITH, *Rapid and short term adaptation in auditory nerve responses*, Hear. Res., 15 (1984), pp. 249–260.

R. WINSLOW, *A quantitative analysis of rate-coding in the auditory nerve*, Ph.D. Dissertation, Johns Hopkins University (1985).

E. D. YOUNG AND M. B. SACHS, *Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers*, J. Acoust. Soc. Am., 66 (1979), pp. 1381–1403.

V. ZUE, *Speech spectrogram reading*, MIT, Lecture Notes and Spectrograms (1985).

Figure Legends

Fig.1

(1a) The spatiotemporal response patterns of a cochlear model. The spatial axis represents the basal-to-apical (bottom-to-top) spread of the cochlear partition; it is labeled by Characteristic Frequency (CF) of each output channel (see text). The waveform of the in-phase harmonic series stimulus is also shown. The intensity of the stimulus here is such that the cochlear outputs are mostly saturated (see appendix A1). The scale marks on the time axis = 5 msec.

(1b) Same as Fig.1a legend but with the out-of-phase stimulus.

(1c) Cross sections of the LIN.I outputs of the harmonic series stimuli for the in-phase (solid) and out-of-phase conditions (dashed). Ordinate scale is in arbitrary units.

Fig.2

(2a) The cochlear spatiotemporal responses to the vowel portions of the stimuli: b/i/t, b/⊃/t, and b/u/t (Zue [1985]). The scale marks on the time axis = 20 msec.

(2b) The LIN.I outputs of the sequence of vowels /i,I,ϵ,æ, ⊃,U,u/ in the context of b..t (Zue [1985]).

(2c) The LIN.II outputs of the vowel sequence as in Fig.2b legend.

(2d) A schematic of the relationship among the vowel stimuli in terms of the LIN.II output parameters (left) and two articulatory features (right). The small arrows (left) point to the direction of the effect of lip-rounding on the vowels. The formant axes are also shown for orientation purposes.

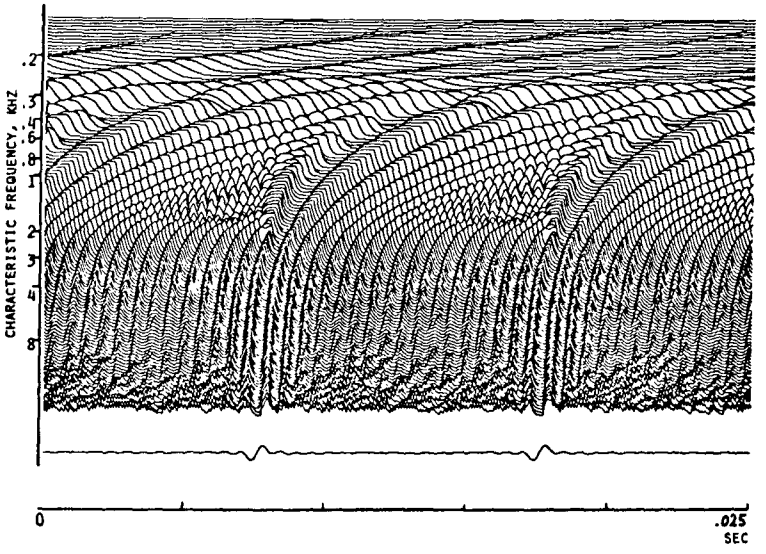
Fig.3

The LIN.II outputs corresponding to a series of vowels (indicated) in the /b..t/ context. M = male speaker, Fem = female speaker. Results from two male and two female speakers are shown (each uttering vowels /i,⊃,u/). The averaging window of the LIN outputs is 4 msec long.

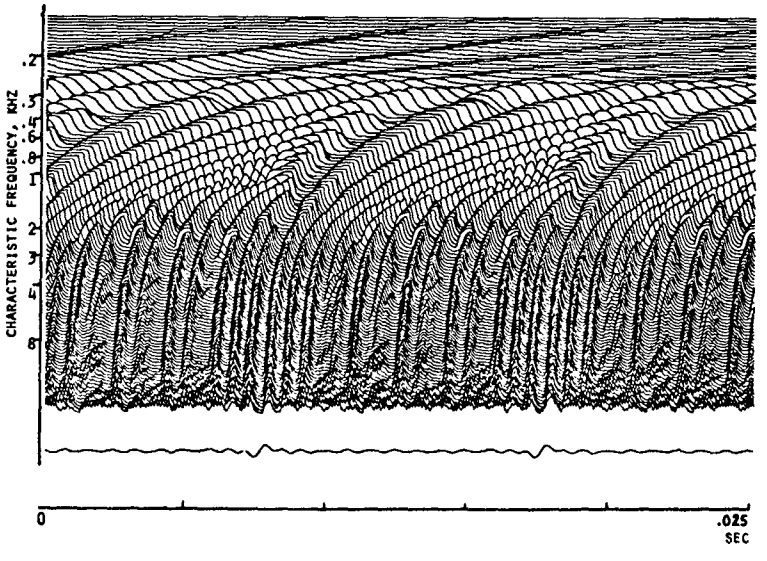
Fig.4

(4a) The spatiotemporal response patterns to the fricatives /s,X,h/. Each scale mark on the time axis = 20 msec.

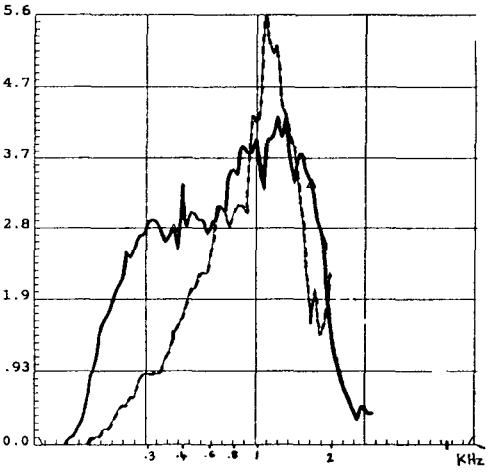
(4b) The LIN.II outputs of the unvoiced fricatives /s,.,s,X,h/. All stimuli are processed in the context /..oat/ (as in 'boat'). LIN averaging window width = 10 msec.


CHARACTERISTIC FREQUENCY, KHZ
.2
.3
.5
.6
.8
1
2
3
4
8
0
.025
SEC

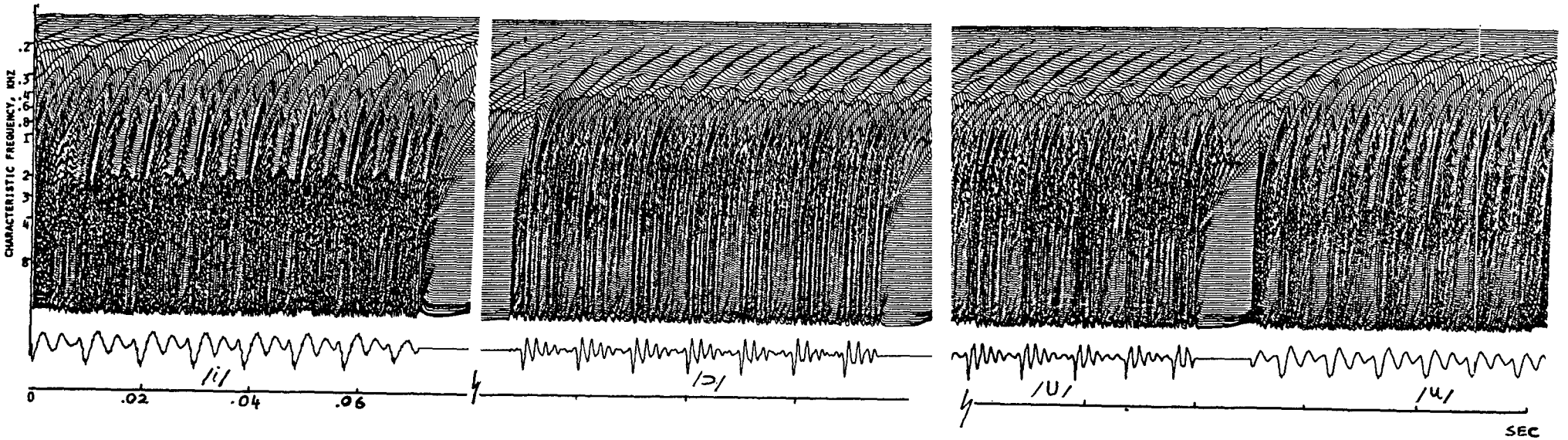
1a


CHARACTERISTIC FREQUENCY, KHZ
.2
.3
.5
.6
.8
1
2
3
4
8
0
.025
SEC

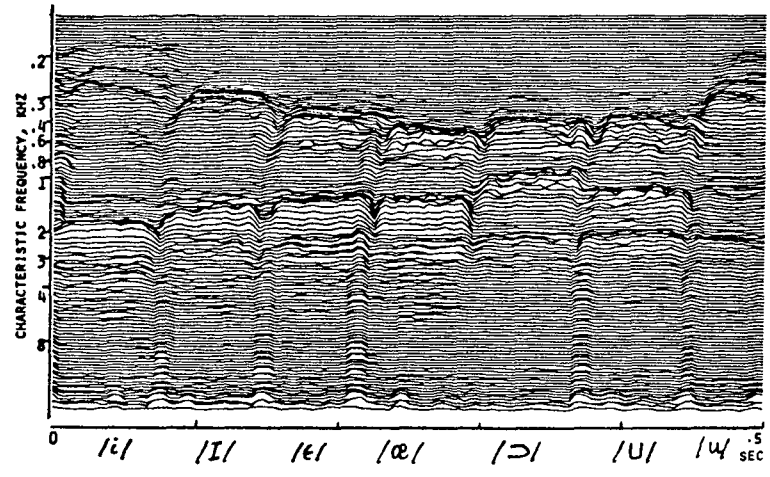
1b



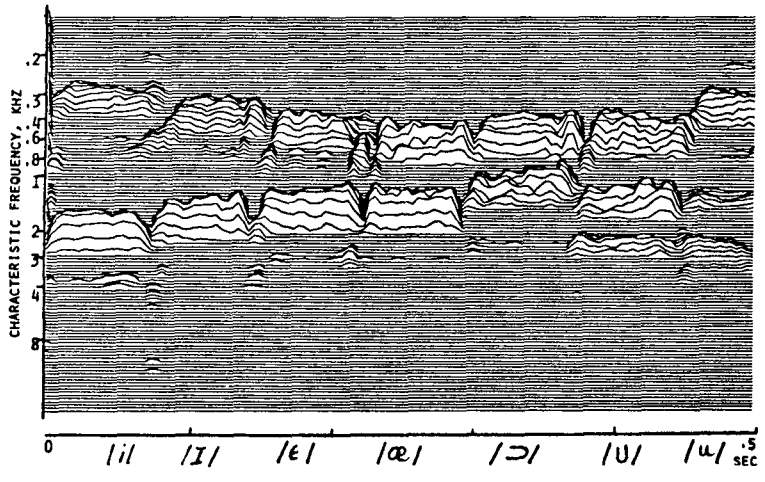
1C


CHARACTERISTIC FREQUENCY, KHZ
.2
.3
.5
.65
.8
1
2
3
5
8
/i/
/ɔ/
/u/
/ɑ/
0
.02
.04
.06
SEC

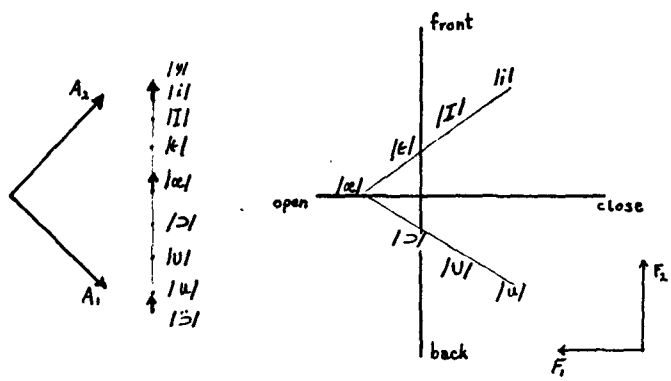
2a


CHARACTERISTIC FREQUENCY, KHZ
.2
.3
.4
.5
.6
.8
1
2
3
4
6
8
0
/i/
/I/
/ɛ/
/æ/
/ɔ/
/U/
/u/
.5
SEC

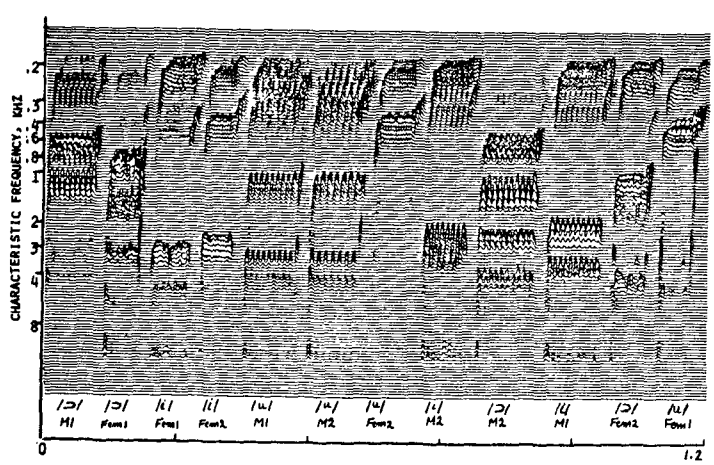
2b


CHARACTERISTIC FREQUENCY, KHZ
.2
.3
.4
.5
.6
.8
1
2
3
4
6
8
0
/i/
/I/
/ɛ/
/æ/
/ɔ/
/U/
/u/
.5
SEC

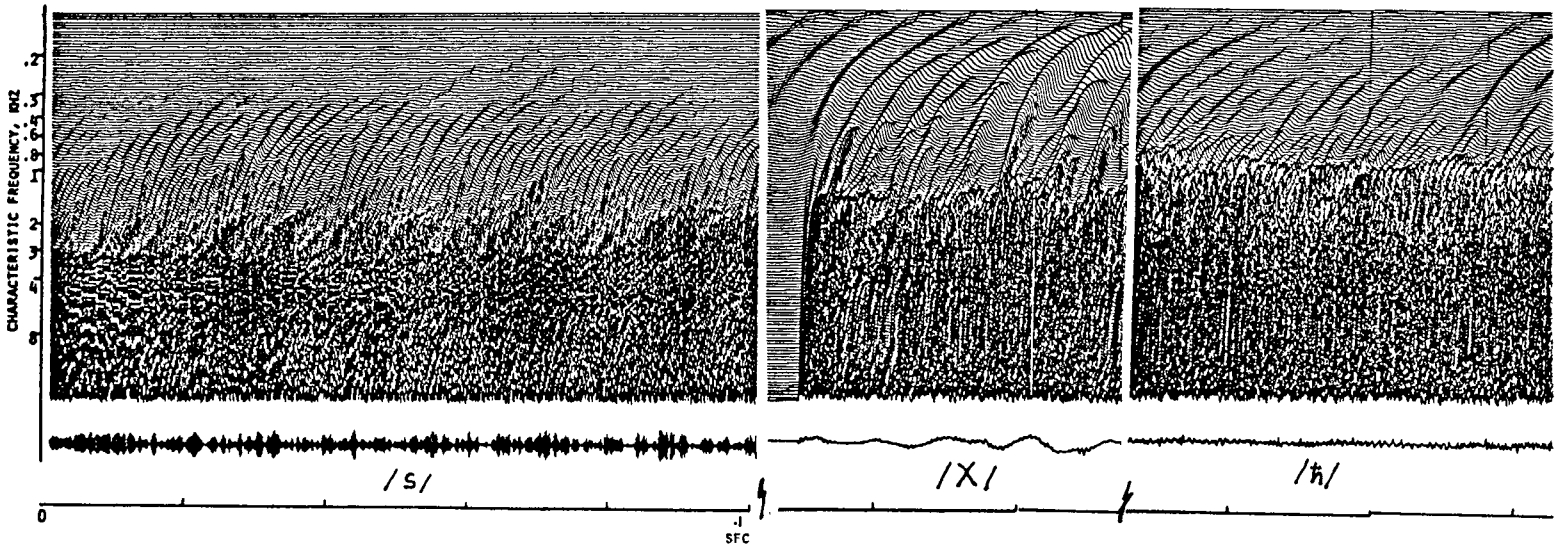
2c


A2
A1
/ɯ/
/i/
/ɪ/
/ɤ/
/ɑ/
/ɔ/
/ʊ/
/u/
/ɜ/
front
back
open
close
/i/
/ɪ/
/e/
/ɑ/
/ɔ/
/ʊ/
/u/
F2
F1

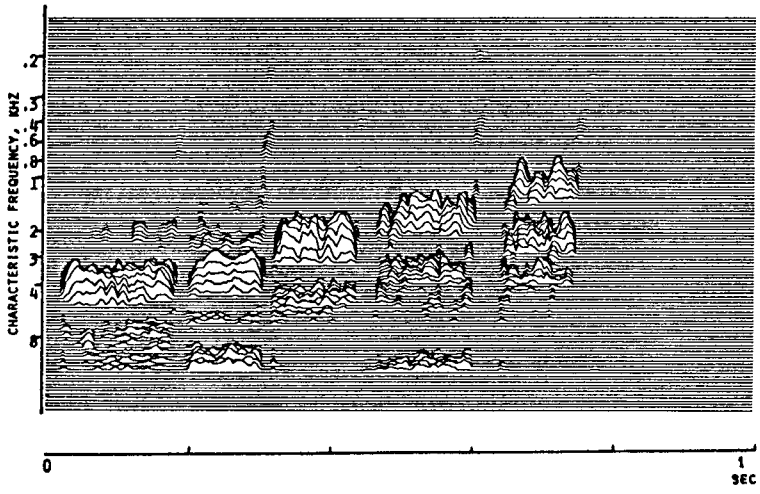
2d


CHARACTERISTIC FREQUENCY, KHZ
.2
.3
.4
.5
.6
.8
1
2
3
4
5
8
/ɔ/ M1
/ɔ/ Form1
/k/ Form1
/i/ Form2
/u/ M1
/ɤ/ M2
/ɤ/ Form2
/i/ M2
/ɔ/ M2
/ɤ/ M1
/ɔ/ Form2
/u/ Form1
0
1.2
SEC

3


CHARACTERISTIC FREQUENCY, KHZ
/s/
/X/
/ħ/
0
.1
SEC

4a


CHARACTERISTIC FREQUENCY, KHZ
0
1
SEC

4b

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

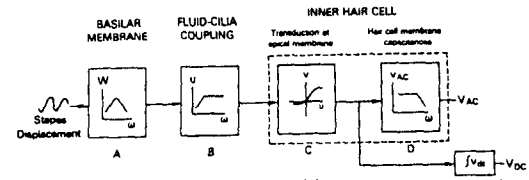


Fig.1: Schematic of the cochlear model stages [8].

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.

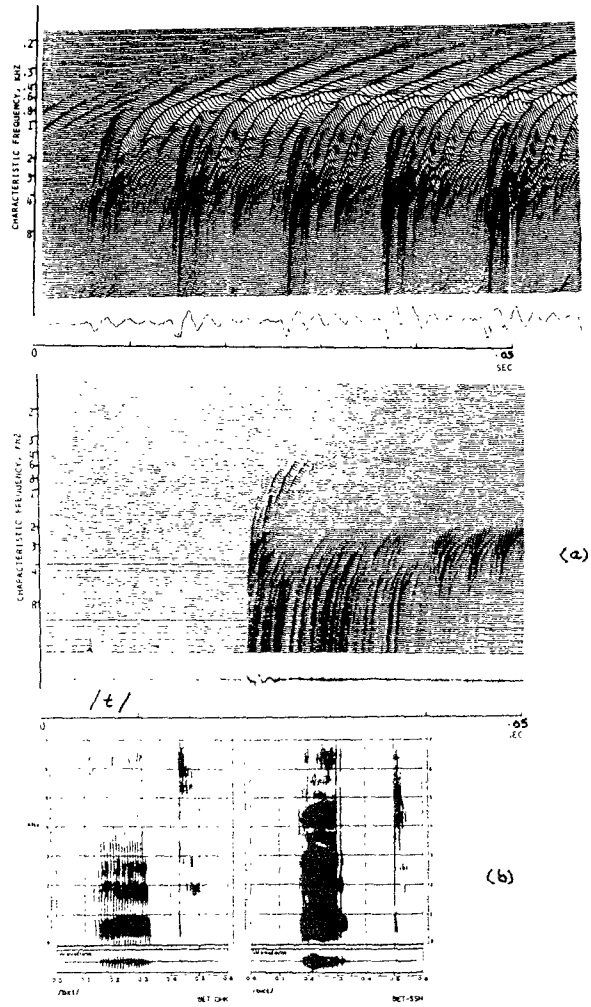


Fig.2: (a) Spatio-temporal responses of the cochlear model to selected portions of /bɛt/ spoken by a female. (b) Spectrograms of /bɛt/ spoken by a male (left) and a female (right) [12].

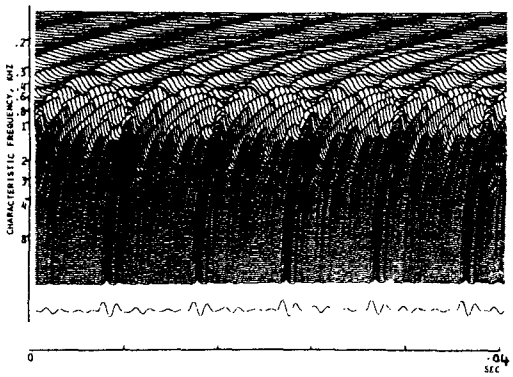


Fig.3: Spatio-temporal responses to synthesized vowel /a/. F0=130 Hz; F1=730 Hz; F2=1090 Hz; F3=2440 Hz.

The Central Processing of Auditory-Nerve Responses

This stage involves the extraction and utilization of the perceptually relevant cues from the response patterns of the cochlear nerve. Conceptually, it is a particularly difficult problem because the nerve patterns contain a rich variety of cues pertaining (in unknown ways) to a multitude of perceptual tasks. Thus, in studying a particular encoding scheme on the auditory nerve, or in implementing algorithms for automatic speech recognition applications, *a priori* decisions have to be made as to the appropriate response measures that need to be used and the ways these are to be combined. For instance, in the estimation of the spectral parameters of speech (e.g. formants) several measures have been proposed that range from purely spatial, i.e. discarding the fine temporal structure of the nerve responses (e.g. using the distribution of the *average rate* profiles across the tonotopically organized nerve-fiber array), to purely temporal, i.e. utilizing primarily the periodicities in the response as measures of the spectral content (e.g. the dominant frequency algorithm) [10]. Others in between include the Average Localized Synchronous Rate (ALSR) [3] and the Generalized Synchrony Detector [11].

An alternate approach is to view the response patterns essentially as 2-D spatio-temporal images with specific morphological features acting as spectral cues. One such feature, for instance, are the *edges* in the profiles of activity across the spatial axis created by one or both of the amplitude and phase changes eluded to earlier [5, 7]. The strength and position of the edges along the tonotopic axis are related to the signal spectral parameters through the dependence of the above two response characteristics on the frequency and amplitude of the stimulus (or its resolved harmonics in case of complex sounds). Edge detection algorithms, based on realistic biological lateral inhibitory network (LIN) topologies, can be used to extract these features and thus signify the spectrum of the underlying acoustic stimulus [5]. The LIN possesses several desirable properties which include: (1) A spatially distributed structure which is naturally suited for fast parallel processing implementations; (2) A robust performance in the presence of certain severe stimulus and/or channel distortions. The latter point is illustrated in the LIN outputs of Figs.4 under three conditions: (a) Moderate stimulus levels where few channels are saturated. (b) 40 dB higher stimulus levels where most channels are saturated: Despite channel saturation, the edges in the cochlear response patterns remain intact, and so do the LIN outputs near F1-F4 (These should be compared to the spectrograms of Fig.2.b). (c) Fig .4c simulates the case where the channel nonlinearity has a large slope [a], and the response waveforms become highly saturated. The outputs here are derived by a spatial first-difference operation evaluated *only* at the spatial zero crossings of the response pattern. The F1 and F2 are still extracted, though higher formants are now lost.

Acknowledgements

This work is supported in part by an Initiation grant from NSF, by the Mathematical Research Branch (NIH), and by a grant from the Minta Martin Foundation.

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

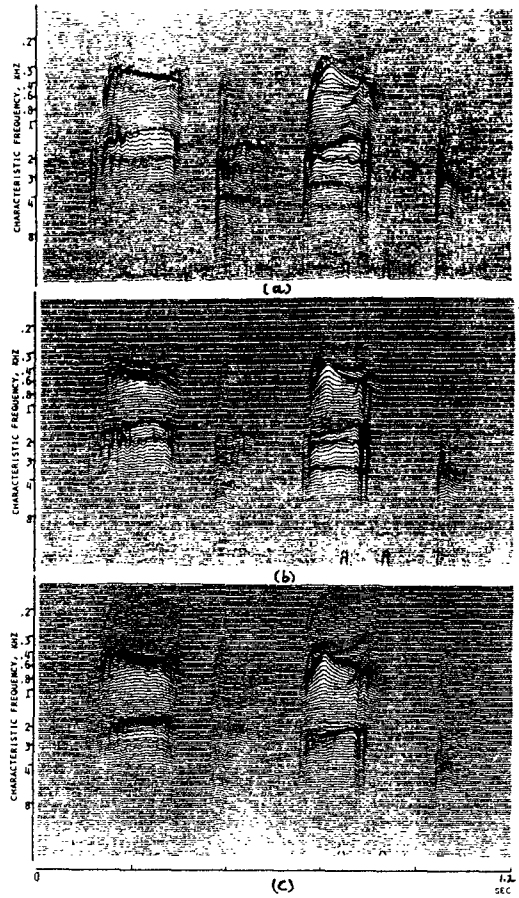


Fig.4: LIN estimates of spectral parameters of /b t/ whose spectrograms are shown in Fig.2. Parameters of the LIN network are published elsewhere [5]. (a) LIN outputs for moderate stimulus levels. (b) LIN outputs for high stimulus levels. (c) LIN outputs for severly saturated channels.

[2] R. R. Pfeifer and D. O. Kim, "Cochlear Nerve Fiber Responses: Distribution Along the Cochlear Partition," *J. Acoust. Soc. Am.* vol. 58, pp. 867-869 (1975).

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

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

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

[6] B. Delgutte, "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds," *J. Acoust. Soc. Am.* vol. 75, no. 3, pp. 879-886 (1984).

[7] S. A. Shamma, "Speech Processing in the auditory System. I: Representation of speech Sounds in the responses of the auditory-nerve," *J. Acoust. Soc. Am.* vol. 78, pp. 1612-1621 (1985).

[8] S. A. Shamma, R. Chadwick, J. Wilbur, and J. Rinzel, "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," *submitted to the J. Acoust. Soc. Am.*, (1986).

[9] M. H. Holmes and J. D. Cole, "Cochlear mechanics: analysis for a pure tone," *J. Acoust. Soc. Am.* vol. 76, no. 3, pp. 767-778 (Sept. 1984).

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

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

[12] V. Zue, "Speech Spectrogram Reading," Lecture Notes and Spectrograms, MIT (1985).