# 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

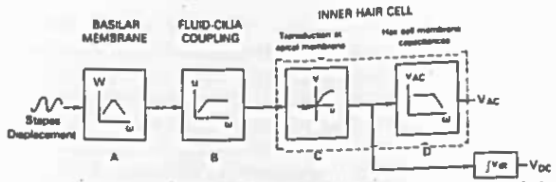


Fig.1: Schematic of the cochlear model stages [8].

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.

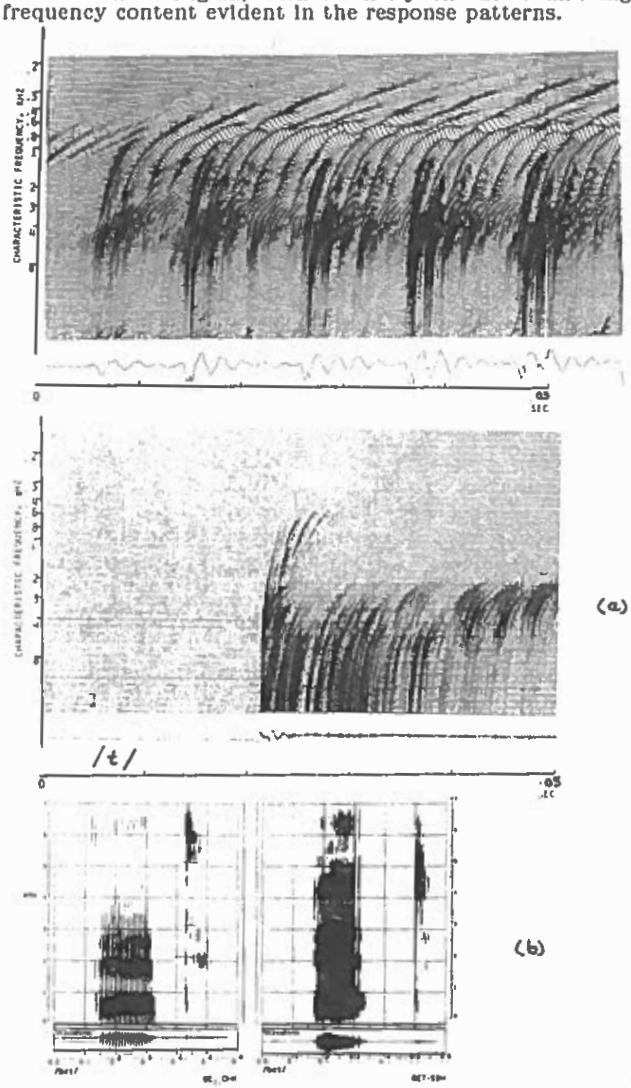


Fig.2: (a) Spatio-temporal responses of the cochlear model to selected portions of /bɛt/ spoken by a female. (b) Spectrograms of /bɛt/ spoken by a male (left) and a female (right) [12].

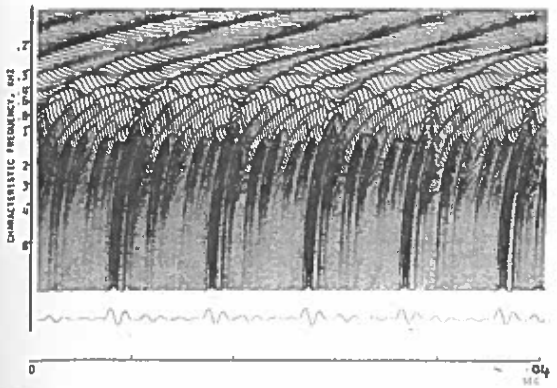

Fig.3: Spatio-temporal responses to synthesized vowel /a/. F0=130 Hz; F1=730 Hz; F2=1090 Hz; F3=2440 Hz.

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.

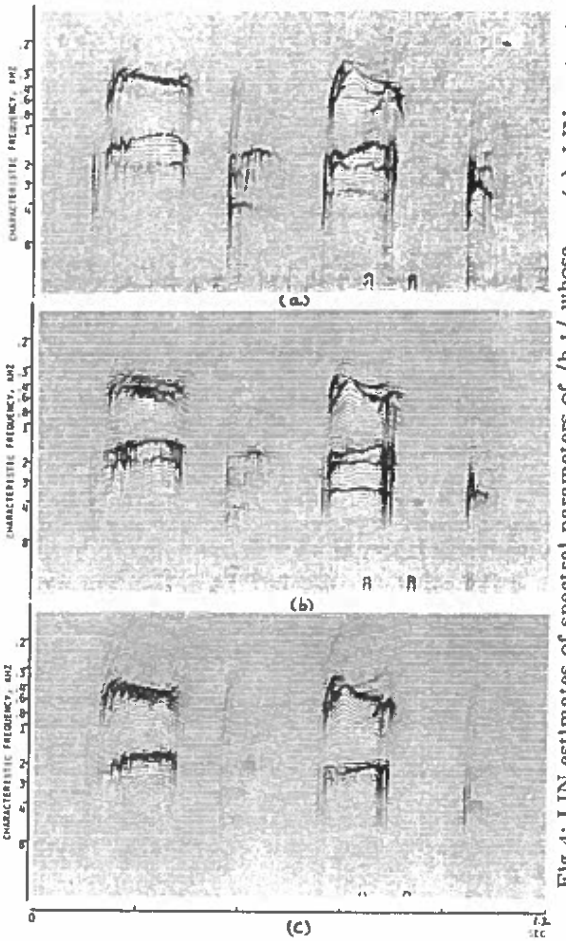

Fig.4: LIN estimates of spectral parameters of /b ɪ/ whose spectrograms are shown in Fig.2. Parameters of the LIN network are published elsewhere [5]. (a) LIN outputs for moderate stimulus levels. (b) LIN outputs for high stimulus levels. (c) LIN outputs for severly saturated channels.

[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).

[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).