THE AUDITORY FEATURES OF VOWEL AND FRICATIVE PHONEMES

SHIHAB A. SHAMMA

Electrical Engineering Department & Systems Research Center, University of Maryland, College Park, Maryland 20742

Mathematical Research Branch (NIDDK), National Institutes of Health, Bethesda, Maryland 20982

ABSTRACT

The acoustic features of vowels and fricatives are examined in the response patterns of a model of auditory processing. For the vowels, a few harmonics dominate the peaks of the internal representation, reflecting the *formant structure* by their spatial locations, and the *front cavity resonance* by their relative amplitudes. For the fricatives, the most prominant feature extracted is the location of the cut-off frequency in their highpass-like spectra.

Understanding the nature of sound processing in the auditory system is an essential step in determining the acoustic elements of speech sounds and their relevance to perception and articuation. In recent years, important discoveries in peripheral auditory function (both at the basilar membrane/hair cell level [1–3], and from auditory-nerve recordings [4], have facilitated the construction of cochlear models that can adequately replicate the primary response features in the auditory nerve [5]. With such models, it is relatively easy to analyze the response patterns associated with a wide variety of speech sounds, and under many signal conditions. Specifically, it is now possible to generate internal auditory representations of the acoustic spectra, and hence to examine closely the expression of such acoustic features as vowel formant locations, amplitudes, and transitions, and fricative spectral shapes. It is important to note, however, that beyond the peripheral auditory stages, little is known about the central neural networks and the processing they perform on the cochlear outputs. This adds an element of uncertainty to the analysis since apparently useful cues and response features at the auditory nerve level may be irrelevant for phonemic perception and classification if the central nervous system ignores, or is incapable of processing them. We shall address this point further after first illustrating the response patterns to the stimulus /position/ (Fig.1), as generated by a cochlear model.

The peripheral auditory model consists of 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 [6], active mechanisms of basilar membrane motion [7], the effects of the middle ear muscles and of the efferent system [8]. This simplified model reproduces the major response properties observed experimentally, especially with relatively steady and broad-band stimuli like vowels and fricatives. 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 [9]. The spatial axis is labeled by the characteristic frequency (CF) of each output channel, i.e. the frequency of the tone which produces its maximum final output at that location (see below for further details of the central processing of the cochlear outputs).

The responses to the vowel portions of the stimulus (/I/,/u/) posses a typical structure that is observed in all experimental data [10,11] - that is the dominance of the entire pattern by a few stimulus harmonics. These harmonics correspond to the largest components located near the formants of the stimulus spectrum. They excite travelling waves along the basilar membrane which are evident in the fine temporal structure and spatial spread of the responses. Because of the unique asymmetrical shape of the cochlear filters, the waves decay in amplitude, and begin to acumulate phase rapidly at locations along the array, depending on the frequency of the underlying harmonic [9]. The expression of 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). For each of the vowel responses, the identity of the underlying dominant harmonics can be deduced from two sources: (1) The temporal course of the response (e.g. by measuring the frequency of the synchronized response), or (2) by the location of the above

described features along the spatial axis (i.e. a tonotopic axis). Thus for the vowel /I/ (Fig.1), there are two response domains, the first is the apical region CF $\leq$ 2 kHz, corresponding to the F_1 harmonics (2-3), each of which decaying and experiencing phase shifts at its appropriate CF location. An abrupt transition in the response patterns occurs at approximately 2 kHz as the harmonics associated with the higher formants become dominant. These trends are seen again in the /u/ vowel responses, where F_1 is at a lower CF location ($\approx$ 250 Hz) and F_2 is considerably weaker.

The auditory responses of the fricative portions /z/,/š/ differ considerably from those of voiced vowels. To start with, there is a random component in the excitation that is quite evident in the cochlear responses as randomly initiated travelling waves. Another distinctive aspect is the predominance of the high frequency components and their sudden decay at a different location for each of the fricatives. For the voiced fricative /z/, there is an additional voiced component in the excitation waveform.

The CNS derives its auditory percepts from the cues available in the spatiotemporal outputs of the cochlea. The identity of these cues and the way they may be extracted and processed are two issues that are essentially inseperable. In the cochlear patterns, there is an abundance of spatial and temporal cues to the physical parameters of the stimulus [12]. However, given the complexity of the extraction algorithms involved, only a subset of of these cues are probably relevant in that the CNS is actually capable of utilizing them. Since little is known about the anatomical and functional role of the neural networks of the central auditory system, little can be said in support of any processing algorithm aside from the general plausibility arguments regarding its biological implementation and the degree to which the isolated parameters explain the psychophysical measurements [13].

In viewing the cochlear outputs as spatiotemporal images, a set of cues emerge that are robust and particularly easy to extract. These are the spatial edges due to the asymmetrical shape of the cochlear filters [9]. As noted earlier, such edges occur at the regions separating the responses to the strong, resolved components of the stimulus. While the expression of these edges is dependent on the integrity of the phase locked reponses (i.e. the ability to visualize regions of different temporal character), they can also appear in the high frequency regions (where phase locking is minimal) as the peaks and valleys of the spatial average rate profiles. In all cases, the *location* of these edges along the tonotopically organized spatial axis, and their saliency, are reliable indicators of the stimulus frequency and amplitude [13]. As with normal visual images, such spatial discontinuities can be detected and highlited by relatively common and simple neural networks as the lateral inhibitory networks (LIN) [14].

We have processed the cochlear patterns of a wide variety of sounds with models of recurrent and nonrecurrent LIN's [13,15]. The results shown in Fig.2 for part of the word /position/ and in Fig.3 for a vowel series, are generated with a two layer LIN: the first nonrecurrent and performs the initial edge detection and extraction, the second is a recurrent version which further sharpens the outputs of the first layer and preserves only locally large peaks [15]. For the vowel portions of the stimulus, the LIN's typically extract two or three peaks corresponding to the components near the nominal formant frequencies of the vowels; an additional low frequency peak sometimes appears corresponding to the fundamental or second harmonic components of voiced sounds (especially for females). The variability in the locations of these peaks for different speakers and sexes seems to be similar to that observed in traditional spectrogram outputs [15], though this remains to be confirmed with much larger data samples. An interesting aspect of these and other vowel outputs [15] is the systematic change in the *relative amplitudes* of the high-CF and low-CF peaks (or equivilantly, the location of *center of gravity* of the pattern) for different vowels (Fig.3). Thus, for high vowels (e.g. /i,u/), the high-CF peak is always relatively large when the constriction is fronted (as in /i/), and vice versa in the back vowel /u/. In all close vowels (e.g. the frontal /i,y/ and back /u/), the place of the constriction seems to be 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. The open vowels /æ/ and /ɔ/ occupy an intermediate position in that the two peaks are comparable.

These relations are summarized schematicaly in Fig.4. 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. 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' [16], which also move in the same direction for this sequence of vowels [17]. Finally, the effects of lip-rounding in this schematic are viewed only as local modulations (in the direction of the arrows) 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 [17].

Correlates of the pitch percept associated with voiced

Fig.1
The spatiotemporal response patterns of a cochlear model to the word /position/. The spatial axis represents the basal-to-apical (bottom-to-top) spread of the cochlear partition; It is labeled by the Characteristic Frequency (CF) of each output channel (see text). The scale marks on the time axis = 20 msec.

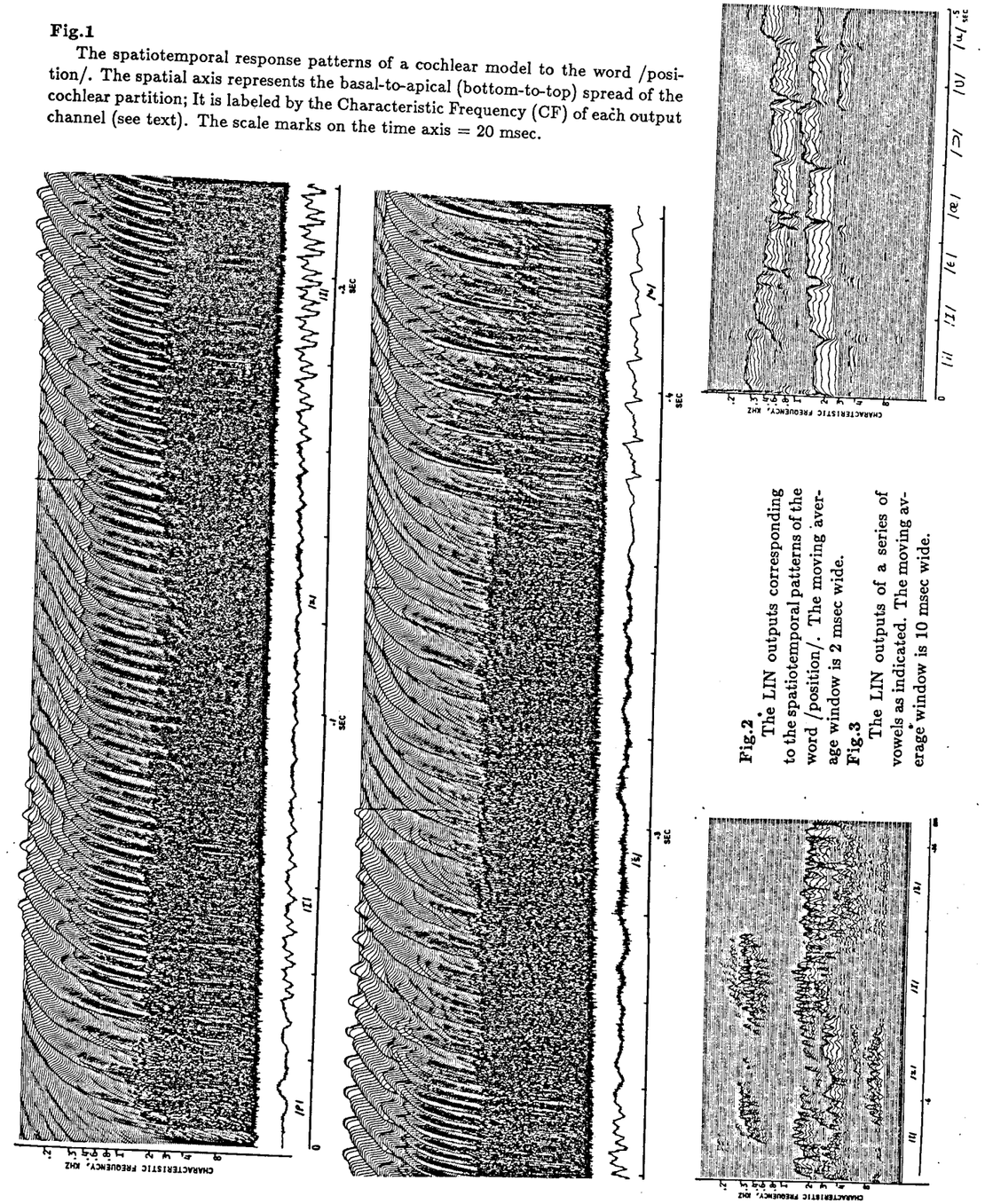

Fig.2
The LIN outputs corresponding to the spatiotemporal patterns of the word /position/. The moving average window is 2 msec wide.

Fig.3
The LIN outputs of a series of vowels as indicated. The moving average window is 10 msec wide.

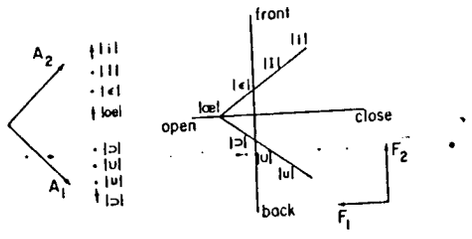

Fig.4

A schematic of the relationship among vowel model parameters (left) and articulatory features (right).

vowels can also be descerned in the LIN responses. In Fig.2 outputs, this is seen in the *beating* of the LIN peaks at the voiced portions of the stimulus[1]. The origin of this temporal character is the combining by the LIN of locally dissimilar waveforms at the regions of discontinuities in the cochlear patterns. These responses are due to different *resolved* harmonics of the same fundamental, and hence beat at this frequancy [15]. As expected, in the LIN outputs of unvoiced fricated speech (e.g. /š/, and the high CF region of /z/) the regular beating is absent.

The LIN outputs (Fig.2) of the fricatives show major peaks that correspond to the most important discontinuity in the spatiotemporal patterns, i.e. the edge created by the rapid cut-off of their high frequencies (Fig.1). The downward CF shift of this peak from that of /z/ to /š/ reflects the lengthening of the frontal cavity which largely determines the high frequency extent and overall spectral shape of the fricative [18].

In summary, auditory processing of speech phonemes isolates specific features that may play an important role in the perception and recognition of these sounds. These auditory features can be related to articulatory aspects such as the formant resonances of vowels and the front cavity resonances of fricatives and vowels. They also contain cues to other attributes of the speech signal, e.g. pitch.

Acknowledgment: This work is supported in part by an NSF initiation award NSFD ECE-85-05581 and NSF Grant CDR-85-00108.

REFERENCES

[1] A. J. Hudspeth & D. P. Corey, "Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli," *Proc. Nat. Acad. Sci. U.S.A.* 74(6) (1977), 2407–2411.

[2] I. J. Russell, "Origin of the receptor potential in inner hair cells of the mammalian cochlea - evidence for Davis's theory," *Nature* 301(27) (1983), 334–336.

[3] R. Patuzzi & P. M. Sellick, "Basilar membrane motion and inner hair cell output," *J. Acoust. Soc. Am.* 74(6) (1983), 1734–1741.

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

[5] S. A. Shamma, "Encoding the acoustic spectrum in the spatio-temporal responses of the auditory-nerve," in *Auditory Frequency Selectivity*, B. C. J. Moore & R. Patterson, eds., Plenum Press, Cambridge, 1986, 289–298.

[6] L. A. Westerman & R. L. Smith, "Rapid and short term adaptation in auditory nerve responses," *Hear. Res.* 15 (1984), 249–260.

[7] S. T. Neely & D. O. Kim, "An active cochlear model shows sharp tuning and high sensitivity," *Hearing Res.* 9 (1982), 123–130.

[8] R. Winslow, "A quantitative analysis of rate-coding in the auditory nerve," Ph.D. Dissertation, Johns Hopkins University , 1985.

[9] 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 (1985), 1612–1621.

[10] E. D. Young & 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), 1381–1403.

[11] D. G. Sinex & C. D. Geisler, "Responses of auditory-nerve fibers to consonent-vowel syllables," *J. Acoust. Soc. Am.* 73 (1983), 602–615.

[12] J. O. Pickles, "The neurophysiological basis of frequency selectivity," in *Frequency Selectivity in Hearing*, B. C. J. Moore, ed., Acadamic Press, London, 1986, 51–122.

[13] 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 (1985), 1622–1632.

[14] H. K. Hartline, *Studies on excitation and inhibition in the retina*, Rockefeller University Press, New York, 1974.

[15] S. A. Shamma, "The acoustic features of speech phonemes in a model of auditory processing: Vowels and unvoiced fricatives," *J. Phonetics* (1987 (in press)).

[16] R. Carlson, B. Grantstorm & G. Fant, "Some studies concerning perception of isolated vowels," STL-QPSR, 1970.

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

[18] C. G. Fant, "Acoustic description and classification of phonetic units," in *Speech Sounds and Features*, MIT, Cambridge, MA, 1973.

[1] The voicing in the /z/ segment is clearly visible in the Fig.1 responses to the 3rd harmonic component of the fundamental (same as the F_1 harmonic of the preceeding and succeeding vowel) and will be clearer in the LIN output of a slightly louder stimulus. In Fig.3 the voiced vowel outputs also beat, but the LIN moving average window is set at 10 msec in order to clarify the display of the relative amplitudes; This in turn averages out the beating.