THE ANALOG COCHLEA AS A TOOL IN PHONETIC RESEARCH

JOHN J. GODFREY

In the search for the distinctive features of a language's sound system, it remains true that "the closer we are in our investigation to the destination of the message (i.e., its perception by the receiver), the more accurately can we gage the information conveyed by its sound shape" (Jakobson, Fant and Halle 1967:12). Incomplete and conflicting information on auditory processing, however, has prevented us from extracting the 'auditory features' of speech sounds in a principled way. At the Aerospace Medical Research Laboratories, extensive research has been directed at the mechanical transform of auditory signals effected by the inner ear. A major instrument in this research has been the Analog Cochlea, a twenty-four channel electronic model of inner ear function based on von Bekesy's physiological measurements. This device models the basilar membrane as a leaky transmission line, in which the high frequencies dissipate first and low frequencies last. The cochlea thus transforms a two-dimensional signal (time vs. amplitude) into three dimensions: time, position along the surface of the basilar membrane, and displacement at each point.

In Figure 1, we see the Analog Cochlea response to four cycles of the vowel /i/ from a male speaker. In the foreground (bottom) is the basilar membrane's basal or high-frequency end; in the background, the apical or low-frequency termination. Time is on the horizontal axis, covering in this case 36 msec., while displacement is represented vertically. The curvature of the traveling wave, due to the delay in propagation time along the sense organ, can be clearly seen. One notices

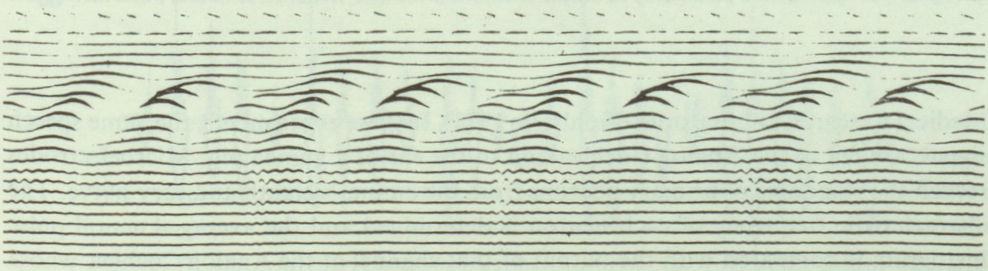

Fig. 1. Analog Cochlea Surface — For four cycles of the vowel /i/.

the repetition of the pattern in its entirety from glottal pulse to glottal pulse; but note especially that the two most prominent features of this pattern, namely the broad peak-to-peak intervals in the background and the narrow ones in the foreground, are each spread over a large area of the basilar membrane. Features appear not just in one channel (or one point on the membrane), but spread over several channels. Such poor spatial resolution is a result of the low Q of the basilar membrane response —Bekesy measured a Q of 2, or a half-power bandwidth of 500 Hz at 1000 Hz. Obviously, the length of the basilar membrane should not be regarded as the dimension of frequency, except in the crudest sense. The transform, because of the low Q, does have the effect of distributing the signal's components over a significant length of the membrane, and thus presenting each of its features to a large number of neural channels, since the receptor cells are arrayed along the length of the membrane. To illustrate graphically that this is in fact the way the auditory nervous system sees speech sounds, Figure 2 reproduces results obtained by Dr. Goldstein of the Aerospace

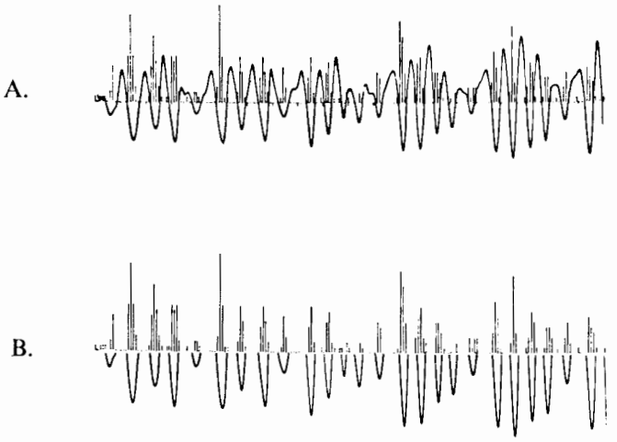

Fig. 2. Alignment of Basilar Membrane displacement with Pulse Occurrence Histogram from a single neuron. A. Basilar membrane motion predicted by model, superimposed on neural pulse histogram. B. Rarefaction phase only of basilar membrane motion, matched to neural pulse histogram.

Medical Research Laboratories (Goldstein 1971). In these experiments, the same speech signals are fed to the Analog Cochlea and to the ear of a guinea pig. Microelectrodes record the responses of primary neurons from the animal's eighth auditory nerve. The characteristic frequency of each neuron is determined and the averaged neural pulse histogram is compared with the output of that channel of the Analog Cochlea whose resonant frequency most closely matches the characteristic frequency of the neuron.

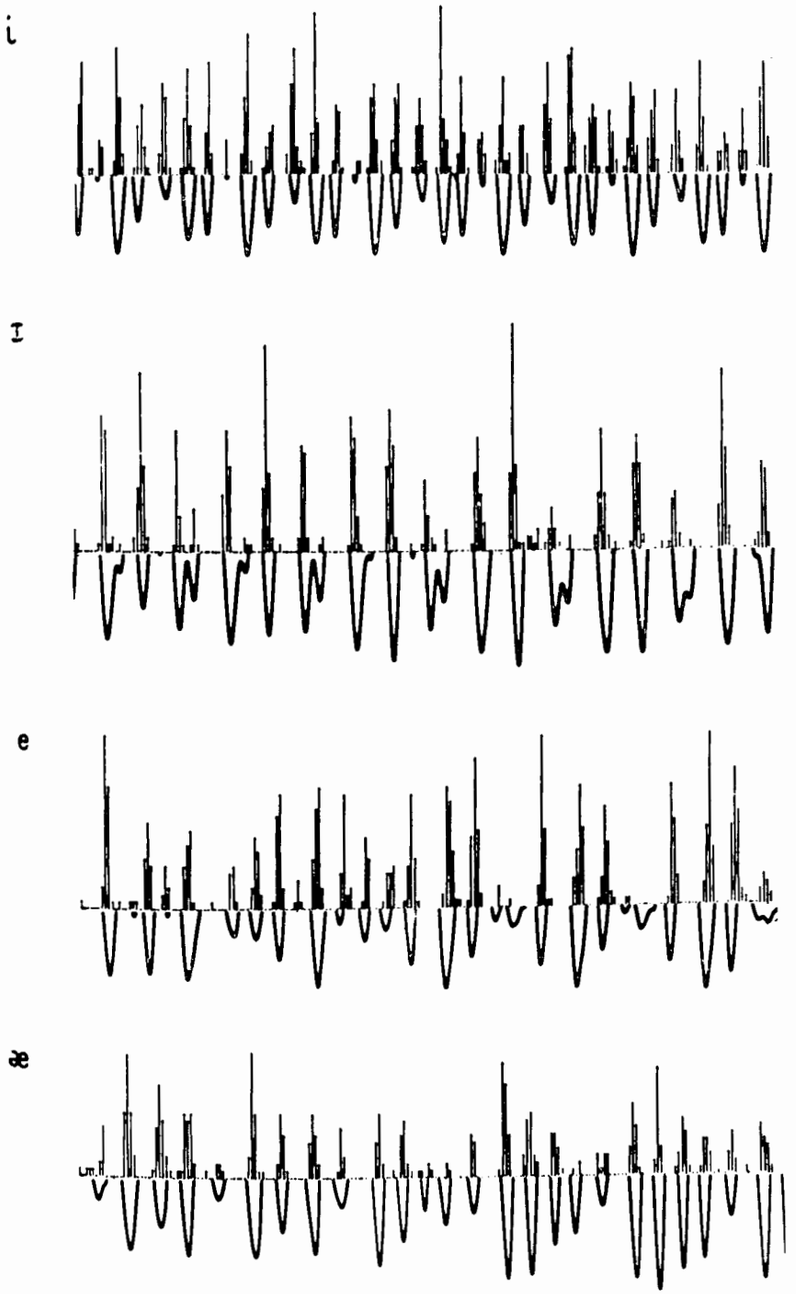

Fig. 3. Single neuron responses and Basilar Membrane displacement for four vowels. (Neuron: CF772 Hz, Speaker: ♂).

Figure 2 illustrates the correlation between the downward (excitatory) deflections of a single point on the basilar membrane as modeled by the Analog Cochlea and the actual average pulse output of a neuron which innervates that area of the cochlea, with the same speech sound as stimulus. One sees immediately that the neural firings are faithfully tracking the displacement of the membrane as shown by the model. The signal is five cycles of the vowel /æ/ by a male speaker. Figure 3 shows the responses of neuron and cochlear model to four sounds: /i/, /ɪ/, /e/ and /æ/.

Our contention, then, which I can only assert here, is that this model preserves all and only that information which is the input available to the peripheral auditory system. It thus provides a unique and promising means of analyzing speech sounds to extract the auditory correlates of phonetic and phonological features. To illustrate this, we may examine—though necessarily only in a very discursive way here—the cochlear surface features for some English sounds.

In order to relate the surfaces to more familiar transforms of the speech signal, we begin with features which appear to separate some English vowels. Figure 4

	/i/		/a/		/u/	
Speaker	Int. 1	Int. 2	Int. 1	Int. 2	Int. 1	Int. 2
A. G.	2.8	.42	1.5	.88	2.6	.72
J. G.	3.2	.41	1.9	.73	3.1	.86
T. H.	3.0	.50	1.6	.73	2.8	.72

Fig. 4. Distinctive Cochlear Surface Intervals (in milliseconds) for three vowels.

shows the dominant peak-to-peak intervals, in milliseconds, which occur at the beginning of each glottal pulse in the cochlear patterns for the vowels /i/, /a/, and /u/. Three male speakers of one metropolitan New York dialect pronounced six CVC words containing each vowel, thus providing a variety of consonantal contexts. Measurements were taken from a single glottal pulse pattern one-third to one-half way through each vowel, then averaged. Not surprisingly, these intervals can be directly correlated with formants—the reciprocal of each interval would yield an appropriate formant frequency in Hz. One advantage of this display in the study of vowels is that one sees clearly when two 'formants' manifested by the Sonagraph do not in fact constitute two distinct sensory inputs to the ear, but rather merge or interact. Further studies with a larger population with all vowels, and across two dialects, should be completed this year. For the moment I simply point out that we can separate the vowels of this dialect quite well, and in the same manner that the ear must accomplish this task.

However, a method of analysis which can take into account each glottal pulse in succession becomes especially valuable for the phonetician, as it is for the auditory system, in detecting and tracking changes in the speech signal—the well-known transitional cues produced by rapidly changing articulatory gestures.

Figure 5 shows a complete cochlear surface display for the word 'dodge' pronounced by a male speaker. Let us look briefly at some auditory cues which can be tentatively identified as distinctive.

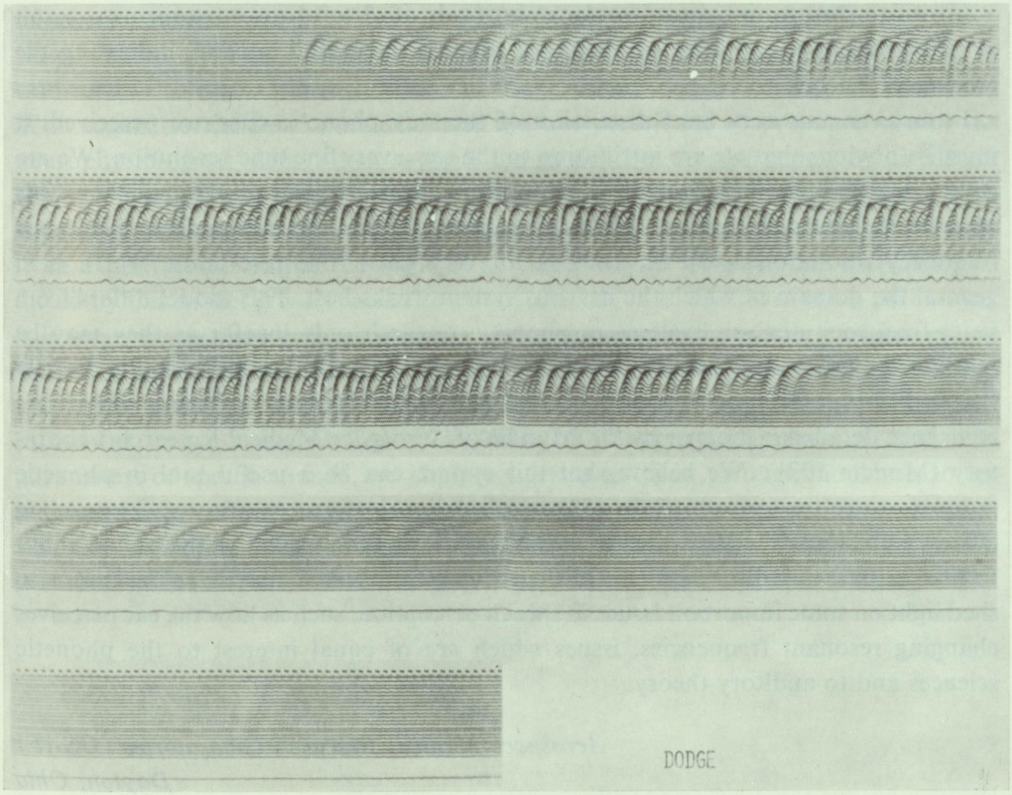

Fig. 5. Analog Cochlea Surface for the word 'dodge'.

(1) Transient marking the release of /d/: characteristic of postdental stops are both the area of the cochlea occupied and the intervals of .2 and .6 msec;

(2) Voice onset time: here the onset of voicing is coincident with stop release;

(3) Transition: intervals in the middle channels are seen to increase from about .6 to about .85 msec. as the vowel achieves steady state, marking the apical stop transition. A similar interval transition in reverse is seen in the final vocalic patterns, as the closure for /j/ is being effected. This undoubtedly corresponds to the second formant transitions of spectral analysis.

(4) Vowel intervals: each glottal pulse of the vowel /a/ begins with an interval, extending over seven or more channels of the cochlear surface, of about 1.6 msec. (which would appear in Sonagrams as a resonance of about 630 Hz.).

(5) A second interval, extending over two or more channels of the mid-frequency area, of about .8 msec. (the Sonagram's second formant at about 1250 Hz.). Note

that these last two features, corresponding to Formants 1 and 2, interact on the cochlear surface. This is to be expected, given the bandwidth of the basilar membrane at these frequencies, a fact which the Sonagraph does not show. Such an interaction might, in fact, form the basis of a definition of 'compactness' for the cochlea.

(6) After closure, groups of peaks at intervals of .3 to .4 msec. occur as a result of palatal friction, and they are spaced at roughly the glottal interval, indicating the voicing of the /j̈/.

From even this very brief discussion of the way phonetic cues are perceived, it must be obvious that we are attributing to the ear a very fine time resolution. We are rather sure, in fact, that populations of neurons act probabilistically to mark intervals of a fraction of a millisecond. In essence, we find that the ear trades off its poor frequency resolution along the cochlea for very good time resolution, which is in general the domain in which the nervous system works best. This model differs from strict frequency analysis systems in phonetic research only insofar as they may be motivated by place-frequency theories of hearing or conceived as representing directly the auditory correlates of speech sounds and features. Some of these differences have been detailed in a paper by Dr. Mundie of Aerospace Medical Research Laboratory (Mundie 1971). We believe that this system can be a useful tool in phonetic research, since our investigations of the auditory cues of speech proceed from strong evidence that our model preserves just that information which is the actual input to the peripheral auditory system. In the Analog Cochlea we may have the means to shed light on some important issues in speech perception, such as how the ear perceives changing resonant frequencies, issues which are of equal interest to the phonetic sciences and to auditory theory.

Aerospace Medical Research Laboratories (USAF)
Dayton, Ohio

REFERENCES

Goldstein, A.
1971 Paper presented at the Meeting of the Acoustical Society of America, Washington, D.C.
Jakobson, R., G. Fant, and M. Halle
1967 *Preliminaries to Speech Analysis* (Cambridge, Mass., M.I.T. Press).
Mundie, J.R.
1971 Paper presented at the Meeting of the Acoustical Society of America, Washinton, D.C.

DISCUSSION

GUIRAO (Buenos Aires)
Mr. Godfrey's paper is very inspiring though I wonder if it is not too premature to attempt the type of electronic device he just described. There are still many problems unsolved regarding the information encoded through the auditory nerve. Then, it

is not clear yet whether the basilar membrane by itself could analyze complex sounds. On the other hand we know that the auditory system can make extremely fine discriminations of temporal changes with relation to localization of stimuli (from the source), but we do not know if it is so regarding temporal sequence of acoustical parameters of speech sounds.

GODFREY
With regard to the first comment, time did not permit me to present in detail the neurophysiological data which support the model, but see the papers by Goldstein (1971) and Mundie (1971). The electronic analog of the cochlea is based on Békésy's experimental measurements in the first place; and the fact that neural firing patterns from experimental animals correspond so closely to the displacement patterns predicted by this model greatly increases our confidence in it. Far from being premature, it is more timely than ever, as evidence (from research at the Institute of Perceptual Research in the Netherlands, for example, and the work of Nordmark and others) continues to mount in favor of a time-domain analysis theory of hearing. Thus, in answer to your second doubt, regarding the temporal sequence of acoustical parameters of speech sounds, we aim precisely to test this, through modeling, to see if the ear's capabilities as predicted by the model, and supplemented by neurophysiological evidence from the auditory nervous system, explain the facts of speech perception adequately.