Speech coding in the auditory nerve: I. Vowel-like sounds

Bartrand Delgutte* and Nelson Y. S. Kiang
Eaton-Peabody Laboratory of Auditory Physiology, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts 02114, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. Departments of Otolaryngology and Physiology, Harvard Medical School, Boston, Massachusetts 02115, and Neurology Service, Massachusetts General Hospital, Boston, Massachusetts 02114

(Received 11 November 1982; accepted for publication 30 September 1983)

Discharge patterns of auditory-nerve fibers in anesthetized cats were recorded in response to a set of nine steady-state, two-formant vowels presented at 60 and 75 dB SPL. The largest components in the discrete Fourier transforms of period histograms were almost always harmonics of the vowel fundamental frequency that were close to one of the formant frequencies, the fundamental frequency or the fiber characteristic frequency (CF). For any fiber, the position of its CF relative to the formant frequencies (F1 and F2) appears to determine which of these components dominates the response. Specifically, the response characteristics of the tonotopically arranged array of fibers can be described in terms of five CF regions: (1) a low-CF region below F1 in which the largest response components are the harmonics of the fundamental frequency closest to CF; (2) a region centered around CF = F1 in which the first formant and its harmonics are the largest components; (3) an intermediate region between F1 and F2 with prominent components at both the fiber CF and the fundamental frequency; (4) a region centered around CF = F2 in which harmonics close to the second formant are the largest for frequencies above the fundamental; and (5) a high-CF region in which response spectra tend to show broad, multiple peaks at the formant and fundamental frequencies. These CF regions are related to the phonetic descriptions of vowels. For example, the extent of the low-CF region is largest for "open" vowels (which have a high F1), and the intermediate region is distinct only for "spread" vowels for which F1 and F2 are more than 1.5–2 octaves apart. For all vowels, response activity for the majority of fibers is concentrated near the formant frequencies, in contrast to responses to broadband noise for which components near CF are dominant.

PACS numbers: 43.63.Pd, 43.70.Dn, 43.70.Ve

INTRODUCTION

This paper is the first of a series reporting responses of auditory-nerve fibers to speechlike sounds. While the methods used to record activity in anesthetized cats are standard for auditory physiologists, the choice of stimuli is dictated by concepts developed by phoneticians. The first paper concerns steady-state, vowel-like sounds, the second discusses theoretical processing schemes that can extract phonetically important information from responses to vowel stimuli, the third describes responses to steady-state voiceless fricatives, the fourth examines responses to consonants characterized by rapid changes in amplitude and spectral characteristics, and the final paper introduces effects of background noise.

Vowels belong to a more general class of speech sounds known as sonorants, which are characterized by a quasiperiodic waveform with low-frequency energy concentrated around the resonant frequencies (formants) of the vocal tract. Other sonorants are glides, diphthongs, and nasal consonants. Strictly speaking, natural vowels are neither steady state nor periodic, but they can be approximated by steady-state, periodic stimuli that are readily recognized as vowels by listeners. Although natural vowels have, in general, three or more formants, there is considerable evidence that the first two (the lowest in frequency) formants suffice for the identification of most vowels (Carlson et al., 1975; Bladon and Fant, 1978). Because of the simplicity in specifying two-formant vowels, they were chosen as stimuli for this study.

A previous series of observations using a smaller set of four-formant, steady-state vowels (Sachs and Young, 1979; Young and Sach, 1979) suggested that the profile of average discharge rate across the tonotopically arranged array of auditory-nerve fibers was a poor candidate for representing formant frequencies because virtually all the fibers would be discharging at high rates for stimulus levels well within those normally used in conversations. Although the question of the usefulness of average discharge rates for vowel coding is by no means closed, the present study concentrates on how formant pattern and fundamental frequency (corresponding to voice pitch) could be represented in the fine time patterns of discharge of the most sensitive auditory-nerve fibers. This choice precludes examination of issues such as the role of the efferent system, the middle-ear muscles, or the higher-threshold fibers in vowel discrimination.

*Present address: Centre National d'Etudes des Télécommunications, Lannion, France.
I. METHODS

A. Stimuli

The two-formant vowel stimuli were generated by a computer program that simulates the summed outputs of two bandpass filters excited by a periodic pulse train. The repetition rate of the pulses, which represents the fundamental frequency, is 125 Hz for all stimuli. The bandpass filters represent the resonances of the vocal tract associated with the formant frequencies $F_1$ and $F_2$. The details of stimulus generation are given in the Appendix.

Figure 1 shows the waveforms and spectra of the stimuli that were used in these experiments. These nine stimuli resemble the "cardinal vowels" used by phoneticians as reference sounds in identifying vowel-like sounds in unknown languages (Jones, 1956). Although in all cases the spectra have peaks near the formant frequencies, the relative heights of the components in the vicinity of these peaks depend on the relationship between the formant frequency and the fundamental frequency. When a formant frequency coincides with a harmonic of the 125 Hz fundamental, this harmonic is considerably more intense than the adjacent harmonics. This occurs, for instance, for both formants of /a/ which correspond to harmonics 4 and 12. More generally however, the formant frequencies do not coincide with harmonics of the fundamental frequency. For instance, the 700 Hz first formant of /a/ is about halfway between the 625 Hz fifth harmonic and the 750 Hz sixth harmonic, so that the two harmonics have nearly equal amplitudes.

Figure 2 shows the positions of the nine vowel stimuli in a plot of $F_1$ vs $F_2$. The second formant frequencies were set to the values that are obtained in perceptual matching experiments with natural vowels (Carlson et al., 1975; Bladon and Fant, 1978). These matched values are usually close to the second formant of natural vowels, except for vowels with a high $F_2$ such as /i/ and /e/ when they are closer to the third formant. In Fig. 2, the vowel stimuli, except /æ/, are distributed along a quadrilateral where “peripheral” vowels in many languages tend to be found. The vowel /æ/, located roughly at the center of this quadrilateral, is often called "neutral vowel."

Dimensions that are traditionally used in phonetics to describe contrasts between vowels can be related to the diagram of Fig. 2 (Peterson and Barney, 1952). The vowels /i/, /e/, and /æ/, which have a low first formant, are called “close” vowels because of the position of the tongue during their production. The contrasting “open” vowels /æ/ and /ʌ/ are characterized by a high $F_1$. There is also a contrast between the “back” vowels /u/, /o/, and /ʌ/, which have a low second formant, and the “front” vowels /i/, /e/, /e/, and /æ/, which are characterized by a high $F_2$. Another dimension of possible perceptual importance is the “spread” dimension, defined by the ratio $F_2/F_1$ (Fant, 1973; Chisto- vich and Lublinskaya, 1979). Figure 2 shows that /i/, /e/, and /ʌ/ are the most spread, whereas /æ/, /e/, and /ʌ/ are the least.

B. Experimental procedures

The preparation of the animals and techniques to record from single auditory-nerve fibers are essentially as described in Kiang et al. (1965). Twenty-seven healthy adult cats weighting 1.6 to 3.5 kg were injected peritoneally with dailly barbituric acid in urethane solution. A cannula was
inserted into the trachea, the cartilaginous external auditory meatus was cut near the tympanic membrane, and the bulla opened to expose the round window. The posterior fossa of the cranial cavity was opened dorsally and the cerebellum retracted medially to allow visualization of the auditory nerve. The animal was placed in a sound-proof, vibration-insulated, electrically shielded chamber (Ver et al., 1974). An acoustic cavity containing the sound transducer (B & K 1-in. condenser microphone) and a calibrated probe microphone was sealed into the meatus.

Transfer characteristics of the acoustic system in individual animals were always flat within ±4 dB for frequencies below 5 kHz, and the maximum sound pressure in the flat region was approximately 115 dB SPL. For each animal, sound levels of the vowel stimuli were set relative to the average of the transfer ratio over the frequency range 0.1–5 kHz. Although this procedure results in variations in stimulus levels from one experiment to another, the standard deviation of interanimal variations in the transfer ratio did not exceed 2 dB for frequencies below 8 kHz. Distortion components generated by the acoustic system did not exceed 25 dB SPL for the most intense (75 dB) vowels.

The state of the preparation was assessed by measuring single-fiber thresholds and the visual detection level (VDL) of the click-evoked N1 potential recorded near the round window. The VDL was always stable within 5 dB during an experiment, and varied over a range of 15 dB between animals. Single-fiber responses were recorded with KCl-filled micropipettes. For each fiber, the characteristic frequency (CF) and threshold at CF were estimated from a tuning curve obtained using the same method as Kiang et al. (1970) and Liberman (1978). To obtain a relatively homogeneous population of fibers whose threshold was more than 15–20 dB above Liberman’s (1978) “best threshold curve” were discarded. This fiber selection procedure restricts recordings to the most sensitive auditory-nerve fibers relative to the 40-dB spread of thresholds that is found in normal animals (Liberman, 1978), so that detailed statements concerning the less sensitive populations of fibers cannot be made from the present data. When the selected fibers were pooled for the 27 animals, the standard deviations of the fiber thresholds in 1/4 oct bands of CFs were about 5 dB on the average.

The 184 ms vowel stimuli were repeated at a rate of 100/min to obtain post-stimulus time (PST) histograms (Gerstein and Kiang, 1960). The histograms were computed with a bin width of 0.05 ms on the basis of 250 to 500 presentations of the stimulus. This bin width enables estimation of the frequency components of the instantaneous discharge rate up to roughly 4 kHz (Johnson, 1978).

Period histograms were computed from the PST histograms by adding the histogram waveforms for each period of the sound burst. The responses to the first two and the last period of each burst were discarded because the responses to transient parts of the stimulus could differ significantly from responses to the central portion of the burst. Discrete Fourier transforms of the period histograms were used to estimate the synchronization indices for each harmonic of the 125 Hz fundamental up to 5 kHz (Johnson, 1978; Young and Sachs, 1979). The synchronization index, which varies between 0 and 1, indicates how well fiber discharges are synchronized to a particular frequency component. It is defined as the magnitude of the Fourier component at that frequency divided by the DC component, which is the average discharge rate. Plots of synchronization index against harmonic frequency will be called “normalized harmonic spectra.”

II. RESULTS

A. Responses of individual fibers

Figure 3 shows period histograms obtained for fibers with different CFs in response to the /i/, /æ/, and /u/ stimuli presented at 75 dB SPL. For each vowel, the response patterns change considerably as a function of CF, so that a description of responses for the entire tonotopically arranged array of auditory-nerve fibers is required. For each CF, the response patterns to the three vowels are usually quite different, indicating that responses of individual fibers can obtain information about the identity of the stimulus. Figure 4 shows the normalized harmonic spectra of the period histograms shown in Fig. 3. Together, these two figures illustrate how fiber responses vary as a function of both CF and the formant frequencies of the vowel stimuli.

The first formant of /i/ at 0.25 kHz is close to the CF of the 0.23 kHz fiber. The largest component in the response spectrum of this fiber is the first formant frequency, and the period histogram consists of two peaks separated by intervals of 1/CF. In the response spectrum for the 0.39 kHz fiber, the first formant is also the largest component, but there are large components at 2F1 = 0.5 and 3F1 = 0.75 kHz. Accordingly, the period histogram shows two pairs of peaks separated by intervals of about 1/CF. For the other two vowels, frequency components near 1/CF are also the largest in the response spectra of the fibers whose CFs are within one octave of the first formant (the 0.65 and 1.1 kHz fibers for /æ/, and the 0.23 and 0.39 kHz fibers for /u/). However, unlike the first formant of /i/ which coincides with the second harmonic, the 0.3 kHz first formant of /u/ is almost midway between the 0.25 kHz second and the 0.375 kHz third harmonic. The largest component in the spectrum of the 0.23 kHz fiber for /u/ is the second harmonic, whereas the third harmonic component is the largest for the 0.39 kHz fiber. Accordingly, the number of peaks in the period histogram is two for the 0.23 kHz fiber and three for the 0.39 kHz fiber. A similar result is obtained for /æ/ in which the 0.8 kHz first formant is between harmonics 6 and 7: the period histogram of the 0.65 kHz fiber has six peaks, whereas the histogram of the 1.1 kHz fiber has seven peaks.

The largest components in the response spectra of low-CF fibers are not always harmonics near 1/CF. The CF of the 0.23 kHz fiber is nearly two octaves below the first formant of /æ/. The period histogram for this fiber shows two peaks at intervals of about 1/CF, and the largest spectral component is harmonic 2 which is the closest to CF. The response spectrum of the 0.39 kHz fiber for /æ/ also shows intense low-frequency components but in this case there are components near 1/CF as well.

Fibers whose CFs are close to the second formant of a vowel stimulus (the 3.1 kHz fiber for /i/, the 1.8 kHz for
FIG. 3. Normalized period histograms for 12 auditory-nerve fibers with seven different CFs in response to the /i/, /æ/, and /u/ stimuli presented at 75 dB SPL. Each period histogram is normalized by the maximum discharge rate. The approximate CFs of the fibers (shown at the left) are spaced 3/4 oct apart. Numbers identifying the animal and the fiber are listed above each histogram. The horizontal markers at the top correspond to time intervals of 1/F1 and 1/F2, and the markers above each histogram correspond to intervals of 1/CF. There is an arbitrary delay between the onset of the stimulus period and zero time of the period histogram. This delay (in ms) is listed as follows for each row in increasing order of CF: 3.7 (for /u/), 3.0 (for /u/, 2.0 for /i/), 2.4, 1.9, 1.7, 1.4, and 1.2.

FIG. 4. Normalized harmonic spectra of the period histograms shown in Fig. 3. The DC component (equal to 1 for all spectra) is omitted from the plots for clarity. Dotted lines show the positions of the formant frequencies F1 and F2, and arrows below the frequency axes mark the CFs of the fibers.
/æ/, and the 0.65 and 1.1 kHz fibers for /u/ have intense response components near F2, and their period histograms show peaks separated by intervals of about 1/F2. For the /i/ stimulus, the response spectrum of the 3.1 kHz fiber has a large fundamental component in addition to the peak near F2. Because the first and second formants of /i/ are widely separated, the largest response components for the 0.65, 1.1, and 1.8 kHz fibers are not close to either F1 or F2. The spectra for the 1.1 and 1.8 kHz fibers show a broad peak near the CF and a peak at the fundamental frequency. Accordingly, the period histograms show peaks at intervals of 1/CF during the first half of the period. The period histogram of the 0.65 kHz fiber has peaks at irregular intervals and its spectrum shows many prominent components including those near CF.

Fibers whose CFs are more than about one octave above the second formant tend to have response spectra with broad peaks at several frequencies. The spectra of the 3.1 and 5.2 kHz fibers show peaks near both F1 and F2 in response to the /æ/ stimulus, with the F1 peak being somewhat larger. In contrast, the largest components for the 1.8 and 5.2 kHz fibers in response to /u/ are near F2, while the largest component for the 5.2 kHz fiber in response to /i/ is the fundamental.

In summary, the largest spectral components in the response patterns of auditory-nerve fibers for the two-formant vowel stimuli usually are harmonics that are close in frequency to a formant, the fundamental, or the CF. The principal factor that determines which of these components will be the largest is the relation of the fiber CF to the formant frequencies.

B. Band averaging

In order to construct a standardized representation of responses in the entire array of auditory-nerve fibers arranged tonotopically, one needs data entries at regular intervals along the CF dimension. For this purpose, we averaged the magnitude spectra of period histograms for all fibers whose CF lie in a narrow band of frequencies. Specifically, the averaging was carried out with a trapezoidal weighting window having a central width of 0.25 oct and a total width of 0.55 oct (effective width 0.4 oct). Whenever data were available, the center frequencies of the CF bands were sampled at intervals of 0.25 oct between 0.1 and 10 kHz. Fibers were selected on the basis of their thresholds at CF, as described in Sec. I B. The number of fibers used in the computation of each “band-average” spectrum was about four on the average. Overall, more than 300 fibers from 27 animals were used in this way.

![Figure 5](image-url)  
**FIG. 5.** Normalized harmonic spectra for individual units and band-average spectra in response to the vowel /i/ presented at 75 dB SPL. Response spectra for three units in each of five CF bands are shown in the first three columns, and the corresponding band-average spectra are shown in the last column. The center frequencies of the 0.35 oct bands are spaced 1 oct apart. The number of units involved in the computation of each band average is, from low to high frequencies: 3, 5, 7, 5, and 7. The CF of each unit is listed above the spectrum, and the center frequency of each band is listed above the band average. The positions of the formant frequencies F1 and F2 are shown by dotted lines, and the fiber CFs are marked by arrows.
were involved in the band averaging.

The band-averaging procedure is meaningful only if the response spectra of fibers within each CF band are similar. Figure 5 compares response spectra for three fibers in each of five CF bands with the corresponding band-average spectra. As far as possible, fibers in the first column are near the low-frequency limit of each band, fibers in the second column near the center, and fibers in the third column near the high-frequency end. For the bands centered at 0.23, 0.46, and 1.83 kHz, the spectra of all three fibers share the same largest component, so that the band-average spectrum is also dominated by this component. This pattern is typical when the band center frequency is close to one of the formant frequencies. For the band centered at 3.67 kHz, the band-average spectrum also resembles the spectra of the individual fibers, except for the fiber in column 2. The spectra for the fibers in the band centered at 0.92 kHz show multiple peaks whose locations vary among the fibers. The band-average spectrum has intense components over the broad frequency region where spectral peaks of the individual fibers are found. Large variations between fibers are usually restricted to narrow CF regions between the formant frequencies. On the whole, the band-average spectrum preserves the major features of the response spectra of individual fibers and is a fair estimate of the spectrum for a fiber whose CF is near the center frequency of the band.

C. Response of the array of fibers

Figure 6 shows pseudo-perspective displays of the distribution of spectral response components across the array of auditory-nerve fibers for the nine vowel stimuli presented at 75 dB SPL. For most vowels, the strongest activity is distributed along the horizontal dashed lines corresponding to the formant frequencies, indicating that the response patterns of auditory-nerve fibers are dominated overall by frequency components near the formant frequencies. Prominent activity distributed along the curved line representing $f = CF$ occurs only in the CF region between $F_1$ and $F_2$ for the spread vowels /ɜ/, /ɹ/, and /ɹ/, and in the low-CF region for the open vowels /æ/, /a/, /ɛ/, and /ɜ/.

For each vowel, the array of auditory-nerve fibers can be divided into five CF regions on the basis of which components are the largest in the response spectra of Fig. 6. The five regions, which are shown schematically in Fig. 7, consist of a low-CF region below $F_1$, a region centered around CF = $F_1$, an intermediate ("mid") region for CFs between $F_1$ and $F_2$, a region centered around CF = $F_2$, and finally a high-CF region above $F_2$. The positions and extents of these CF regions are

![Diagram](image_url)

**FIG. 6.** Pseudo-perspective representation of normalized band-average power spectra for 0.55 oct CF bands in response to the nine vowel stimuli presented at 75 dB SPL. The normalized power spectrum is the square of the normalized harmonic spectrum. Each band-average power spectrum is plotted with frequency along the oblique axis, and amplitude along the vertical axis. Spectrum points with an amplitude lower than 0.05 are omitted for clarity. The "envelopes" of the vertical lines have been drawn for each power spectrum in order to improve visual continuity. The center frequencies of the CF bands are sampled every quarter octave. Horizontal dashed lines show the positions of the fundamental frequency $F_0$ and the formant frequencies $F_1$ and $F_2$ along the frequency axis. Oblique dashed lines mark the places of the formant frequencies along the CF dimension. The curved dashed line is the locus of points for which frequency is equal to CF.
more than 1 oct for the back vowels /a/ and /o/ in less than 1.2 oct for /i/. Thus the front/back dimension of phonetics would seem to correlate with the position and extent of the F2 region along the CF dimension.

For vowels in which the two formant frequencies are spread more than 2 oct apart (/i/, /e/, and /I/), there is a distinct intermediate CF region in which response spectra show a broad peak near CF and a prominent fundamental component. For vowels in which F1 and F2 are closer together, there is an abrupt jump from the F1 region to the F2 region without an intermediate region. For these vowels, the boundary between the two regions is usually about 2/3 of the distance between F1 and F2, on a logarithmic scale. Thus the spread dimension of phonetics correlates with the extent of the intermediate region and of the regions dominated by F1 and F2.

Finally, there is a high-CF region above F2 in which the fibers respond to the intense, low-frequency components of the vowel stimuli. Response spectra in that region show broad peaks at the formant frequencies and the fundamental. For /a/, components near CF can also be prominent in the response of fibers with CFs near 1.5 kHz. In general, formants in the frequency range 0.5 to 1 kHz are the most intense components. Specifically, the first formants of /e/, /æ/, /a/, and /o/ and the second formants of the close back vowels /u/ and /o/ are the largest components. When there is no formant within that range (as for the spread vowels /i/, /e/, and /I/), the fundamental component tends to be the largest. For the back vowels /a/, /a/, and /o/, the spectral peaks associated with the formants are particularly broad. For the 75 dB vowels, the high-CF region extends at least up to 10 kHz, although discharge rates become progressively lower, particularly for the vowels /u/ and /o/ which have the lowest second formants.

Figure 8 shows the spatial distribution of response components for the vowel stimuli at a 15 dB lower sound pressure level (60 dB SPL). The division of the CF dimension into five regions remains valid at this level, although the F1 region is somewhat more restricted around the F1 place, so that the low-CF region for open vowels, and the intermediate region for spread vowels are more prominent than at the higher stimulus level. Components near F1 become smaller in the F2 region for /e/, /e/, /æ/, and /a/. In addition, the extent of the high-CF region above F2 is greatly decreased, as responses to many of the vowels drop below detection level for CFs above 5 kHz.

The fundamental component of the response spectrum is of particular interest since it is obviously a possible cue for voice pitch. An intense fundamental component usually indicates the existence of prominent fluctuations at the fundamental frequency in the envelope of period histograms, whereas a weak fundamental usually indicates a flatter envelope (Figs. 3 and 4). Because the fundamental component of speech is usually weak, and evokes little response in most fibers when presented alone (Hashimoto et al., 1975), it seems likely that these fluctuations originate largely through nonlinear behavior in the cochlea. Specifically, so long as the frequency-selective elements do not resolve individual harmonics of the fundamental frequency, their outputs should...
show considerable envelope modulation at the fundamental frequency. Rectification of this modulation at a later stage of processing would produce a prominent fundamental response component. Figure 9 shows the distribution of the fundamental component across CF for the vowels /i/, /æ/, and /u/ presented at both 60 and 75 dB SPL. At both levels, the fundamental component is large in the high-CF region for all vowels, in the CF region between F1 and F2 for /i/, and near 0.3–0.5 kHz for /æ/. Generally speaking, a large fundamental component is found in the CF region below F1 for open vowels, and between F1 and F2 for spread vowels. In the few cases when a fiber's CF was near the fundamental frequency, the fundamental component was predictably large. In contrast, the fundamental component is small for fibers with CFs near the formant frequencies (except in the case of the second formant for /i/). There is some tendency for the fundamental response component to be smaller at the higher stimulus level, as the other response components grow larger. These results are consistent with those of Delgutte (1980) for single-formant stimuli.

FIG. 8. Same as Fig. 6 for the 60-dB vowel stimuli.

FIG. 9. Band-average synchronization index at the fundamental frequency plotted against center frequency of 0.55 oct CF bands in response to the /i/, /æ/, and /u/ vowels at 60 and 75 dB SPL. The band-average synchronization index is the fundamental component of the normalized band-average harmonic spectrum. The CF bands are the same as in Figs. 6 and 8. The places of the formant frequencies F1 and F2 are marked by dashed lines.
III. DISCUSSION

A. Comparison with previous studies of vowel coding in the auditory nerve

The present results on responses to two-formant, steady-state vowels are generally consistent with those of Young and Sachs (1979) for four-formant vowels, although there are differences in the relative amplitudes of the response components associated with F1 and F2. In our data, the first formant was never found to dominate the response patterns of fibers whose CF is near F2, whereas that situation arises at high stimulus levels in the data of Young and Sachs. The second formant was sometimes the largest component in the high-CF region (e.g., for /u/ and /ɛ/) in our data, whereas F1 was always dominant in that region in the data of Young and Sachs. In our data, the amplitudes of the distortion components 2F1 and 3F1 at their places were usually lower than the F2 amplitude at its place, whereas the opposite seems to be the rule at high stimulus levels in the data of Young and Sachs. These differences are probably due in part to the fact that the amplitude of the second formant relative to the first formant was 3 to 10 dB higher in our stimuli than in those of Young and Sachs. In addition, the formant frequencies always coincided with harmonics of the fundamental frequency in the stimuli of Young and Sachs. For our stimuli, when F1 coincided with a harmonic, response components near 2F1 tended to be larger than for a similar vowel in which F1 was about halfway between two harmonics.

The generally lower amplitudes of components near F2 in the data of Young and Sachs may also be due in part to suppression by higher-order formants. As expected, considerable differences between the two sets of data are found in the high-CF region where fiber responses can show significant components at the third formant frequency for the four-formant stimuli used by Young and Sachs. For certain vowels (like /u/ and /ɛ/), our data tend to show more prominent components along the f = CF line than those of Young and Sachs, probably because the spread between the formants is increased for the two-formant stimuli. Some of these differences must be perceptually relevant since listeners can distinguish two-formant vowels from full vowels. However, vowel identity can usually be maintained in spite of these differences provided F2 of the two-formant vowel is adjusted to a higher value for certain vowels (Carlson et al., 1975).

Both the present data and those of Young and Sachs (1979) are based on steady-state, synthetic vowels. In real speech, essential information for vowel identification, can also be provided by changes in formant frequencies (Lindblom and Studdert-Kennedy, 1957; Strange et al., 1976). If formant frequencies of speech were to change slowly enough, they would remain within the bandwidths of the frequency selective elements in the cochlea for time intervals of about 10 ms. Under such conditions, the response patterns of auditory-nerve fibers for speech stimuli with changing spectral characteristics might not be much different from responses to steady-state vowels. The available data on responses to stimuli that have changing formants are consistent with this view (Sachs et al., 1982; Sinex and Geisler, 1983).

D. Average rates and fine time patterns of fiber discharges

Descriptions of the responses of auditory-nerve fibers to acoustic stimuli can be made in terms of either the average rates of discharge, or the fine time patterns of spike activity, or some combination of the two. The average discharge rates of a fiber over short times can be regarded as being related to the distribution of stimulus energy relative to the fiber tuning characteristics, so that any representation of the stimulus in terms of average rates must involve responses from fibers with different CFs. In contrast, considerable information about the relative amplitudes of different stimulus components is in principle available in the fine time pattern of discharges of even a single fiber, so long as the fiber is capable of responding to the components at all. Obviously, timing cues from fibers with different CFs could also be integrated by more central stations. Since the present description of vowel coding focuses on fine time patterns of discharge, the role of average rates cannot be assessed except through arguments at a general level.

Sachs and Young's (1979) study of steady-state vowels showed that, at stimulus levels within the normal range for conversations, profiles of average discharge rates against CF for the most sensitive fibers did not always show clear peaks at the formant frequencies, although some cues were available for distinguishing certain classes of vowels (for instance vowels with "spread" versus "compact" formant patterns). The limited data for the least sensitive population of fibers suggested that formant peaks could remain present in rate profiles for levels as high as 85 dB SPL. Rate profiles for the first few tens of ms following the onset of the vowel stimuli showed formant peaks up to higher levels than the profiles for the steady-state portions of the stimuli (Young et al., 1981). Any conclusions to be drawn from these results are limited by the experimental procedures. These limitations, which also apply to the present study of speech coding, will now be examined.

A first limitation is that nothing is known about the activity of the unmyelinated auditory-nerve fibers, which constitute 5%–10% of the entire population (Arnesen and Olsen, 1978; Liberman, 1982). Another limitation is that the use of barbiturates could have affected the auditory nerve activity, so that the results would be invalid for awake, behaving animals. The effects of barbiturates on the compound action potential (CAP) of the auditory nerve do not seem to be large (Peake et al., 1962), but this measure may not be particularly sensitive. The effects on the efferent system are unknown. Experiments on electrical stimulation of the crossed olivocochlear bundle (COCB) demonstrated that the discharge rates of auditory-nerve fibers can be significantly reduced (Fex, 1962; Wiederhold and Kiang, 1970). Thus fibers that would have been discharging at their maximum rates in the absence of COCB stimulation could have their discharge rates drop below saturation through efferent activity. The effects of the other components of the efferent input to the cochlea (Warr and Guinan, 1979) are totally unknown. Behavioral experiments with animals in which the COCB has been cut have not yet established the function of
the efferent projections to the cochlea (Dewson, 1968; Traholtis and Elliot, 1970). Barbitalates also suppress the the activity of middle-ear muscles. Although these muscles should not be a large factor at the stimulus levels used in the published studies of speech coding (Møller, 1965), they could conceivably have effects at high intensities where some of the arguments become critical.

The use of cats as the experimental animal raises the usual questions about the applicability of these data to the coding of speech in humans. The anatomy of the cochlea and auditory nerve shows general similarities between the two species (Schuknecht, 1974), although some peculiarities in the spiral ganglion cells of humans have been described (Ota and Kimura, 1980; Kiang et al., in press). Single-unit recordings from the auditory nerve appear, on the whole to be comparable for many species of mammals, including primates (Evans, 1975; Rose et al., 1967). The compound action potentials (CAP) of the auditory nerve evoked by click and tone-burst stimuli behave similarly in humans and other mammals (Kiang and Peake, 1960; Stephens et al., 1973; Montandon et al., 1975; Eggermont, 1976). Comparison of CAP tuning curves in cats and humans show no great differences in tuning, although the data are limited to frequencies above 2 kHz (Gorga and Abbas, 1981; Harrison et al., 1981; Van Heusden and Smoorenburg, 1981). The general shapes of the masking functions ("psychophysical tuning curves") in cats and humans, and fiber tuning curves in cats appear to be similar (Zwicker, 1974; Bauer, 1978; Pickles, 1979). One clear difference between the two species is the range of frequencies to which the ear is responsive (Miller et al., 1963). In cats, the behavioral threshold to tones roughly follows the threshold at CF of the most sensitive auditory-nerve fibers (Kiang et al., 1965). If a similar relationship holds in humans, one would expect human fiber thresholds to be considerably higher than cat thresholds for frequencies above 8–10 kHz. These differences in thresholds should not have a great effect for the coding of vowels, which have mostly low-frequency components. They may also be differences between cats and humans at the very low frequencies where the human should be more sensitive. Results on speech discrimination by animals are also relevant. Although cats have been shown to discriminate /s/ from /a/ (Dewson, 1964), no data are available for vowels with closely spaced formant frequencies. Indeed, Miller (1977) reports that chinchillas have difficulty in discriminating certain vowels although they easily distinguish /a/ from /s/ (Burdick and Miller, 1975).

In the face of these limitations, it is not possible to present to assess the role of average discharge rates for vowel coding. In particular, one can conceive of schemes that would involve both the action of the efferent system, perhaps in selected CF regions, and a central combination of the rate profiles from populations of fibers with different sensitivities. Such schemes might effectively represent the formant frequencies of vowels even at high stimulus levels and low signal-to-noise ratios. On the other hand, fine time patterns of discharge certainly contain sufficient cues for discriminating vowels, even if one considers only the most sensitive population of auditory-nerve fibers. There is no proof that these cues are actually used by the central processor, but it would be surprising if they were totally ignored, particularly at low frequencies. It is of course possible that a combination of average discharge rates and fine time patterns of discharge is used, and that the central processor weights the cues according to temporary conditions. Such a strategy would be of survival value for organisms living in varied environments. In this view, it would be important to describe the conditions under which specific ways of representing the stimulus would be most effective.

C. Comparison with psychophysical and phonetic data

The presence of peaks in the low-frequency (< 3 kHz) region of the spectrum of a steady-state sound stimulus seems to be an essential cue for listener's identification of this sound as speech (Remez, 1979). All the available physiological data show that the responses of most auditory-nerve fibers to vowels are dominated by components near the formant frequencies. In contrast, responses to broadband noise stimuli are dominated by components near the fiber CF (De Boer and Kuypers, 1965; Ruggero, 1973; Møller, 1977; De Boer and De Jongh, 1978). This contrast in the distribution of response components between broadband noise and vowel-like sounds may help the central processor in distinguishing speech from other environmental sounds, although dynamic cues would also play a role in continuous speech.

The identification of vowels appears to be primarily determined by the frequencies of the local spectral maxima associated with the formant frequencies. In general, vowel identity is not sensitive to the relative amplitudes of the formant peaks and to the general tilt of the spectrum (Carlson and Granstrom, 1980; Chistovich and Lublinskaya, 1979). In spite of differences in formant amplitudes between our stimuli and those of Young and Sachs (1979), the responses of the vast majority of auditory-nerve fibers in both sets of data are indeed dominated by the formant frequencies.

Although the frequencies of the local spectral maxima associated with formants are important in vowel perception, the exact properties of the stimulus that the auditory system uses in estimating these frequencies are unclear. Some psychophysical results (Chistovich, 1971) can be interpreted as if the auditory system detects the most intense harmonic near the formant frequency. Other results (Carlson et al., 1975) suggest that the auditory system computes a weighted average of the two or three most intense harmonics. This average would more nearly coincide with the "true" formant frequency (i.e., the resonant frequency of the vocal tract) than with the most intense harmonic. In our data, when a harmonic was sufficiently close to a formant frequency so that its amplitude exceeded that of its neighbors by about 6 dB, the components synchronized to the neighboring harmonics were considerably suppressed. When the two largest harmonics near a formant frequency had amplitudes within 6 dB, considerable response components were found at both harmonics, so that a central averaging scheme would in principle be possible.

Deletion of several nonformant harmonics from an /æ/-like stimulus has little perceptual effect when the harmonics are between F1 and F2, or in the high frequencies, but is easily heard when the harmonics are below F1 (Carl-
son et al., 1979). Our results show that fibers whose CFs are close to a low-frequency harmonic of /æ/ have large response components at that harmonic. There was usually little synchronization to harmonics above F2, unless they coincided with distortion or intermodulation products of the formant frequencies. This was also the case for harmonics between F1 and F2 when the formant frequencies are less than 1.5 oct apart. However, for spread vowels, like /æ/, /e/, or /i/, large response components at harmonics between F1 and F2 were found for fibers whose CFs were between F1 and F2. One might expect that deletion of harmonics between the formant frequencies would be more easily detected psychophysically for such vowels. Absence of synchronization of discharges to components near the fiber CFs may be a correlate of psychophysical masking of these components.

Phonetic descriptions of vowels are traditionally organized according to the degree of vocal-tract opening (tongue height) and place of major constriction. The physiological results show that these phonetic dimensions have clear correlates at the level of the auditory nerve as variations in the position and extent of different CF regions within which response properties are essentially homogeneous. Specifically, the open–close dimension is correlated with the position of a CF region around the F1 place over which discharges are primarily synchronized to F1 and its harmonics, and with the extent of a low-CF region in which harmonics near CF are the largest components. The front–back dimension is correlated with the position and extent of a region around the F2 place over which F2 is the largest response component for frequencies above the fundamental, and with the extent of a high-CF region in which response spectra have broad peaks at F1, F2 and the fundamental frequency. The spread dimension, proposed by Fant (1973), is correlated with the extent of an intermediate region between F1 and F2 over which large response components near CF and F0 are found, and with the extent of the regions centered at F1 and F2. These correlations between phonological dimensions and properties of auditory-nerve activity are of a continuous nature, and there is no basis in our data for dividing, for instance, the tongue-height dimension into discrete “open” and “close” vowel categories. These results suggest that the basis for phonetic categories may depend in part on properties of the auditory processing apparatus, even its most peripheral stages. This view is not inconsistent with those that emphasize speech production mechanisms in defining these categories (Liberman et al., 1957; Repp, 1981). The identification of certain sounds may require integrating many acoustic cues that normally occur together in speech and would be treated as a perceptual unit at some higher level of processing. These various cues would have to be compatible with both the integrating abilities of the auditory system and the constraints imposed by the speech production apparatus.

Acknowledgments

This work is based in part on a doctoral dissertation submitted to the Department of Electrical Engineering and Computer Science at M.I.T., for which the first author thanks the second one and K. N. Stevens for guidance, encouragement and support. Dr. H. S. Colburn, Dr. D. H. Klatt, Dr. W. T. Peake, and Dr. V. W. Zue made valuable comments on the manuscript. We thank P. M. McGaffigan for drawing the figures, and M. M. Jazak, J. W. Larrabee, E. M. Marr, L. P. Miller, and J. M. Rho for assistance in the experiments. This work was supported by NIH grants NS 13126 and NS 04332, and by a C. J. Lebel fellowship.

APPENDIX: STIMULUS GENERATION

The two-formant vowel synthesizer consists of a voicing source in cascade with two bandpass filters connected in parallel. The voicing source generates impulses at a rate equal to the fundamental frequency. The impulse train is sent to a first-order digital low-pass filter with a cutoff frequency of 0.5 kHz. This filter represents the combined effects of glottal source spectrum and radiation characteristic. Each bandpass filter is a second-order digital filter whose transfer function is specified by the formant frequency and bandwidth. To make the low-frequency components more realistic, the second bandpass filter was followed by a high-pass filter whose transfer function increases linearly with frequency up to 1.5 kHz, and is flat thereafter (Klatt, 1980). The outputs of the two branches are added outside of phase to avoid the generation of a spectral zero between the two-formants frequencies. This parallel configuration of the filters was chosen over a cascade configuration so that the relative amplitude of the two formants A2/A1 would be under independent control. The formant parameters for the nine vowel stimuli are listed in Table A1.

The synthesizer was implemented by the impulse-invariance method (Oppenheim and Schater, 1975) using a sampling rate of 20 kHz. A period of the steady-state output waveform was replicated 2.3 times to generate a burst of sound with a duration of 184 ms. A 2 ms linear rise–fall time was then applied to the burst to suppress discontinuities at the onset and offset. The synthesizer generated 12-bit integer sequences that were sent to a digital-to-analog converter and low-pass filtered at 10 kHz by a fifth-order elliptic filter. All stimuli were delivered to the earphone so that positive voltages in Fig. 1 represent condensation.

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>F1 (kHz)</th>
<th>BW1 (Hz)</th>
<th>F2 (kHz)</th>
<th>BW2 (Hz)</th>
<th>A2/A1 *</th>
</tr>
</thead>
<tbody>
<tr>
<td>/æ/</td>
<td>0.25</td>
<td>50</td>
<td>3.2</td>
<td>250</td>
<td>-5</td>
</tr>
<tr>
<td>/e/</td>
<td>0.40</td>
<td>60</td>
<td>2.4</td>
<td>120</td>
<td>-6</td>
</tr>
<tr>
<td>/i/</td>
<td>0.60</td>
<td>60</td>
<td>2.0</td>
<td>120</td>
<td>-5</td>
</tr>
<tr>
<td>/a/</td>
<td>0.80</td>
<td>70</td>
<td>1.8</td>
<td>120</td>
<td>-6</td>
</tr>
<tr>
<td>/o/</td>
<td>0.70</td>
<td>130</td>
<td>1.2</td>
<td>70</td>
<td>-6</td>
</tr>
<tr>
<td>/u/</td>
<td>0.55</td>
<td>90</td>
<td>1.0</td>
<td>100</td>
<td>-7</td>
</tr>
<tr>
<td>/ɪ/</td>
<td>0.30</td>
<td>60</td>
<td>0.7</td>
<td>110</td>
<td>-9</td>
</tr>
<tr>
<td>/ɪ/</td>
<td>0.30</td>
<td>60</td>
<td>1.6</td>
<td>150</td>
<td>-9</td>
</tr>
<tr>
<td>/æ/</td>
<td>0.50</td>
<td>60</td>
<td>1.5</td>
<td>100</td>
<td>-9</td>
</tr>
</tbody>
</table>

*A2/A1 is the ratio of the magnitudes of the transfer function of the entire synthesizer at the formant frequencies. It corresponds to the actual ratio of stimulus components only when the formant frequencies coincide with harmonics of the fundamental frequency.


