Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers

Eric D. Young and Murray B. Sachs

Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205
(Received 13 June 1979; accepted for publication 26 July 1979)

This paper is concerned with the representation of the spectra of synthesized steady-state vowels in the temporal aspects of the discharges of auditory-nerve fibers. The results are based on a study of the responses of large numbers of single auditory-nerve fibers in anesthetized cats. By presenting the same set of stimuli to all the fibers encountered in each cat, we can directly estimate the population response to those stimuli. Period histograms of the responses of each unit to the vowels were constructed. The temporal response of a fiber to each harmonic component of the stimulus is taken to be the amplitude of the corresponding component in the Fourier transform of the unit's period histogram. At low sound levels, the temporal response to each stimulus component is maximal among units with CFs near the frequency of the component (i.e., near its place). Responses to formant components are larger than responses to other stimulus components. As sound level is increased, the responses to the formants, particularly the first formant, increase near their places and spread to adjacent regions, particularly toward higher CFs. Responses to nonformant components, except for harmonics and intermodulation products of the formants (2F, 2F,F + F, etc.), are suppressed; at the highest sound levels used (approximately 80 dB SPL), temporal responses occur almost exclusively at the first two or three formants and their harmonics and intermodulation products. We describe a simple calculation which combines rate, place, and temporal information to provide a good representation of the vowels' spectra, including a clear indication of at least the first two formant frequencies. This representation is stable with changes in sound level at least up to 80 dB SPL; its stability is in sharp contrast to the behavior of the representation of the vowels' spectra in terms of discharge rate which degenerates at stimulus levels within the conversational range.

PACS numbers: 43.63.Pd, 43.70.Dn, 43.63.Bq, 43.63.Rf

INTRODUCTION

The means by which the spectra of complex stimuli such as speech sounds are encoded at the level of the auditory nerve presents an interesting challenge to studies of auditory neurophysiology. The most straightforward representation which can be advanced is for each fiber to respond at a discharge rate proportional to the energy within its tuning curve (Evans, 1978; Kiang and Moxon, 1974); fibers whose characteristic frequencies (CFs) are near peaks in the spectrum of the stimulus have more energy within their tuning curves than fibers with CFs between the peaks, and therefore respond at higher discharge rates (Hashimoto et al., 1973). This scheme, in which the stimulus spectrum is represented in terms of discharge rate as a function of place along the basilar membrane, is limited at high sound levels by the restricted dynamic range of the rate response of auditory-nerve fibers (Kiang et al., 1965; Kiang and Moxon, 1972, 1974; Sachs and Abbas, 1974; Evans, 1978). In a previous paper (Sachs and Young, 1979a) we have directly measured the discharge rate of large populations of single auditory-nerve fibers in cats to steady-state vowel stimuli; as expected, the profiles of discharge rate versus fiber CF showed peaks which corresponded to the formant energy peaks in the vowels when the stimuli were presented at low sound levels. At higher levels, the peaks in the rate profiles disappeared, principally because of the effects of rate saturation. Although it is not possible to rule out conclusively a rate/place representation of speech on the basis of these results (Sachs and Young, 1979a), the instability of the rate profiles with stimulus level would greatly complicate the problem of speech perception. In the present paper, we will discuss an alternative representation for speech which provides an internal representation of the spectrum of steady-state vowel stimuli which is stable over the range of sound levels from 20 to 80 dB SPL.

The classical alternative to representation of stimuli in terms of rate and place in the auditory system is representation in terms of the temporal patterns of response of single neurons. Rose and his collaborators have demonstrated that the instantaneous rate of auditory-nerve fibers responding to stimuli with frequencies below about 6 kHz is modulated by a rectified version of the stimulus waveform, as modified by cochlear filtering and nonlinearities (Rose et al., 1967, 1969, 1971, 1974; Brugge et al., 1969; Hind et al., 1967). Because the important frequencies for most speech sounds fall within this range, it is natural to expect that considerable information about the spectrum of speech signals could be carried in the temporal response patterns of auditory-nerve fibers. Studies of the responses of auditory-nerve fibers to speech masked by noise (Kiang and Moxon, 1974) and of two-tone approximations to vowels (Reale and Geisler, 1979) have been reported which support this viewpoint. In this paper, we will present an analysis of the temporal responses to steady-state vowels of the same populations of auditory-nerve fibers upon which our previous rate paper was based. We will show that responses to the formant frequencies dominate these temporal response patterns and that a good estimate of the stimulus spectrum, which is quite stable over a broad range of stimulus intensities, can be...
easily derived from the response patterns. This representation retains place as the basic frequency parameter, and substitutes the amplitude of temporal response for rate as the response measure. It is thus generally similar to a recent model for periodicity pitch perception (Goldstein, 1973; Goldstein and Srulovicz, 1977).

The results presented in this paper are based on the study of responses to perfectly periodic synthesized vowel stimuli. Such stimuli are considerable simplifications of natural speech in which both the formant frequencies and the pitch of vowels are time-varying quantities. By simplifying our stimuli in this way, we have been able to focus our attention on fundamental questions in the encoding of multicomponent stimuli which otherwise would have been obscured by complexities in the stimulus. The results of this study provide a basis for the study of more natural speech.

A. Methods

The conclusions presented in this paper are based on results from 13 successful experiments. These are the same experiments which provided the rate data described in our earlier paper (Sachs and Young, 1979a). Healthy adult cats, free from external ear obstructions and middle-ear infections, were anesthetized with sodium pentobarbital. The bulla was opened and the bony septum between the bulla and the middle ear was removed to eliminate acoustical resonance effects (Guinan and Peake, 1967). The auditory nerve on one side was exposed by retracting the cerebellum medially. Single-unit responses were recorded with glass micropipettes filled with 3M NaCl. Since our goal was to study large numbers of units in each cat, experiments typically lasted from 3 to 5 days. During this period, anesthesia was maintained by pentobarbital as needed; 10 cc of lactated Ringers were injected subcutaneously every 12 hours. The physiological state of the cochlea was assessed by monitoring the threshold of the click-evoked N1 potential at the round window and the thresholds of single auditory-nerve fibers. The experiment was terminated if a significant threshold shift occurred.

Sound stimuli were delivered via a closed acoustic system utilizing an electrostatic earspeaker (Sokolich, 1977). Sound pressure was monitored near the eardrum through a calibrated high impedance probe tube (Sokolich, 1977). The transfer function for the acoustic delivery system measured in one cat is shown at the top of Fig. 1. For the frequencies and sound levels used in these experiments, all distortion products were at least 65 dB below the fundamental. Stimuli were synthesized digitally and presented via a digital-to-analog converter. Tones and clicks were synthesized on line. Steady-state vowels were synthesized digitally with a terminal-analog synthesizer prior to the experiment and recalled from a disk memory for presentation. All vowels had a fundamental frequency of 128 Hz. The acoustic spectra, measured near the cat's eardrum, of the three vowels for which results will be presented in this paper are shown in the lower three graphs of Fig. 1. The formant frequencies for the vowels are near the average adult male values in the data of Peterson and Barney (1952). Formant frequencies were adjusted slightly to make them exactly equal to a harmonic of the vowel's fundamental frequency. In the remainder of this paper, we will refer to aspects of these spectra in terms of harmonic numbers. The formant frequencies, in terms of harmonic numbers, for these three vowels are given in Table I. In each experiment, only one of these vowels was used as a stimulus. The vowels were presented once per second in 400 ms bursts with rise and fall times of 10 ms. Sound levels specified in this paper are given in terms of rms pressure levels expressed as decibels re 0.0002 dyn/cm².

The following experimental paradigm was followed for all units. As the microelectrode was advanced through the auditory nerve, broadband noise bursts were presented as a search stimulus. In the experiments in which noise thresholds were determined, the level of the search noise was at least 20 dB above the noise threshold for all units studied. When an auditory-nerve fiber was isolated, its CF was determined using audio-visual cues.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>/I/</td>
<td>3</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>/O/</td>
<td>6</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>/E/</td>
<td>4</td>
<td>14</td>
<td>19</td>
</tr>
</tbody>
</table>

TABLE I. Formant frequencies of vowels used in these experiments: in terms of harmonic number re 128 Hz fundamental.
For units with CF less than 3 kHz, the CF was verified by measuring the peak spacing in the PST histograms of the units’ responses to clicks (Kiang et al., 1965). Responses of the unit to the vowel being studied were then recorded; the vowel was presented at sound levels ranging in 10-dB steps from the lowest level to which the unit would respond to about 85 dB SPL. The same sound levels were used for all units studied in any one animal.

Stimuli were usually repeated until at least 1000 spikes were collected during the analysis period. The time of occurrence of each spike was recorded with 10-μs resolution; the spike timer was reset at the start of each fundamental stimulus period so the histograms of responses synchronized to the pitch period of the vowel could be computed. To minimize the effects of transients, the analysis was confined to the last 380 ms of the 400-ms stimulus. Period histograms of the responses were computed using 64 or 128 bins per cycle of the vowel (as needed to avoid aliasing). Four examples of period histograms are shown in the left column of Fig. 2. These histograms are estimates of the instantaneous discharge rate as a function of time through one fundamental period (7.81 ms) of the vowel, averaged over the analysis period. The amplitude of the unit’s response at the frequency of each of the harmonics of the vowel was estimated by computing the discrete Fourier transform of the period histogram, i.e., expressing the N bin histogram \( r(n) \) in the form

\[
r(n) = R_0 + 2 \sum_{k=1}^{N/2-1} R_k \cos\left(2\pi k n/N + \theta_k\right),
\]

where \( n = 0, 1, 2, \ldots, N-1 \).

The magnitudes of the Fourier transforms (the \( R_k \) of the histograms in Fig. 2) are shown in the right column. (These Fourier transforms were computed from histograms containing two cycles of the vowel on the abscissa, so there is a “noise point between each pair of vowel harmonics.) Note that the units of \( r(n) \) and of \( R_k \) are spikes/s.

We have chosen to measure a unit’s response at a particular harmonic by the synchronization index, defined as the magnitude of that harmonic in the Fourier transform divided by the number of spikes in the record (i.e., as \( R_k/R_0 \)). This measure has been commonly used (Goldberg and Brownell, 1973; Johnson, 1974; Pfeiffer and Kim, 1975); its appropriateness has recently been questioned by Kim and Molnar (1979) who point out that the synchronization index saturates at moderate stimulus levels, while the total response at the frequency of the harmonic, as measured by \( R_k \), is still increasing. They suggest, instead, using the magnitude of the harmonic component divided by spontaneous rate, i.e., \( R_k/SP \). Division by spontaneous rate reduces some of the scatter among units; it was justified in terms of the exponential model for auditory-nerve fiber response (Siebert, 1970; Goldstein, 1972; Littlefield, 1973; Johnson, 1974) in which spontaneous rate appears as a multiplicative factor [see Eq. (A1) in the Appendix]. However, for the purpose of initial analysis of our data, we are more interested in which frequency components dominate the response of a fiber than in the absolute magnitude of the response. Equation (1) can be rewritten as below,

\[
r(n) = R_0 \left(1 + 2 \sum_{k=1}^{N/2-1} (R_k/R_0) \cos\left(2\pi k n/N + \theta_k\right)\right),
\]

The response has now been divided into two components: The term outside the brackets (\( R_0 \)) is the average discharge rate of the unit and the terms inside the brackets express the shape of the period histogram, normalized to an average rate of 1. The coefficients of this expansion are the synchronization indices, \( R_k/R_0 \). Thus the synchronization index reflects the relative response at each frequency when discharge rate is factored out.
For auditory-nerve fibers $R_a/R_o$ saturates while $R_a$ is still increasing, as Kim and Molnar have pointed out. This saturation means that further growth of the response ($R_a$) at a particular harmonic occurs at the same rate as growth of average discharge rate ($R_o$). The relative contribution of that harmonic to the response is not growing; rather the response at that harmonic is increasing because the whole response is getting stronger. Therefore, in Figs. 3–5 and 7–9 we will express the response at each harmonic in terms of the relative contribution of that harmonic to the response rate as growth of average discharge rate ($R_o$). The response ($R$) at a particular harmonic occurs at the same frequency as the corresponding component of the vowel’s spectrum in the overall response of the population of interest. Therefore, in these figures we will abbreviate this function as ALSR in the remainder of the paper.

In Figs. 6, 10, and 11 it is the representation of the vowel’s spectrum in the overall response of the population that is of interest. Therefore, in these figures we have plotted the average value of the response at each harmonic, expressed in units of discharge rate; the average for each harmonic was computed over all fibers whose CFs were within 0.5 octave of the frequency of the harmonic. If $R_M$ is the magnitude of the $k$th component [as defined in Eq. (1)] of the response of the $i$th fiber to the vowel, then we define the “average localized synchronized rate” at the $4$th harmonic to be

$$\text{ALSR}(k) = \frac{1}{M_k} \sum_{i=0}^{M_k} R_{2i}$$

where $C_k$ is the set of units with CF between $0.707 f_o$ and $1.414 f_o$; $M_k$ is the number of units in $C_k$; and $f_o$ is the fundamental frequency of the vowel. For convenience we will abbreviate this function as ALSR in the remainder of the paper.

B. Results

Figure 2 displays examples of the temporal patterns of response which we have observed. The left column shows period histograms of responses to $/a/$ of four single units studied on 11/28/77. The right column shows the amplitude spectrum of the discrete Fourier transform of the period histogram next to it. Each period histogram and Fourier transform pair was obtained from a different unit; the CFs of the units are shown in the center column. One period of the voltage waveform of the vowel is shown above the column of histograms.

The top unit in Fig. 2 responded to the vowel with three phases of increased discharge rate of approximately equal amplitude; this unit seems to be responding principally to the third harmonic of the vowel at this level. The frequency of the third harmonic (384 Hz) is near the unit’s CF. The second unit in Fig. 2 has a CF near the sixth harmonic of the vowel and its response is dominated by the sixth harmonic. Responses such as these, in which one frequency component dominates, are seen most often (but not exclusively) in units whose CFs are near a formant frequency. The Fourier transforms of these two period histograms (right column in Fig. 2) have a large component at the frequency of the dominant harmonic; they also have significant energy at integer multiples (harmonics) of the dominant harmonic. The components at integer multiples of the dominant harmonic arise, at least in part, from the rectification inherent in the fact that auditory neurons cannot discharge at negative rates; the inhibitory half cycles of period histograms such as these are therefore clipped at zero rate. This rectification can be modeled by an exponential function which relates the instantaneous rate in an auditory-nerve fiber to the cochlear excitation pattern driving it. This model is discussed in detail in the Appendix along with a method of fitting it to period histograms of responses of auditory-nerve fibers. The filled circles on the Fourier transform plots of the first two units in Fig. 2 show the spectrum of model period histograms obtained by assuming that a component at the dominant harmonic (the 3rd harmonic for the top unit and the 6th harmonic for the second unit) is present at the input of an exponential nonlinearity of the form of Eq. (A1) (along with a much smaller component at the first or second harmonic, which is not necessary in these two cases and contributes insignificantly to the responses). It is apparent that a rectifying nonlinearity of this type is sufficient to account for much of the distortion observed. We will refer to response components arising as a consequence of the rectification process as rectifier distortion products and will refer to components at the “input” of the rectifier as primary components.3

The third and fourth histograms in Fig. 2 illustrate responses to which more than one stimulus component contributes significantly. The third histogram in Fig. 2 has nine peaks, suggesting that this unit is responding principally to the ninth harmonic of the $/a/$. This response is strongly modulated, however, indicating that other components also contribute. The Fourier transform shows a large component at the ninth harmonic, as expected; the next largest component is the sixth harmonic. Three of the remaining significant components, the 3rd, 15th, and 18th, can be interpreted as rectifier distortion products arising from a response to 6th and 9th harmonic primaries. The filled circles on the Fourier transform plot for this unit show the amplitude spectrum of a model period histogram computed by assuming that primary components at the 6th and 9th harmonics are present at the input of an exponential rectifier. Once again, it is apparent that many of the frequency components in the response can be accounted for by rectifier distortion. The presence of distortion arising in this way must be considered when evaluating the data to be shown in this paper.

The bottom histogram in Fig. 2 displays a pattern of response which is frequently observed in units with CFs above 1.5–2.0 kHz. The period histograms have a large peak which may be followed by damped oscillations; period histograms of this type result from broad spectrum responses to several components of the stimulus simultaneously.

Figures 3–5 show the distribution of temporal response, as measured by the synchronization index, for a population of 269 units studied in one cat with $/e/$ as the stimulus. Each figure shows the distribution of synchronization at one stimulus level; 38 dB SPL in Fig. 3, 58 dB SPL in Fig. 4, and 78 dB SPL in Fig. 5. In each figure there are 12 plots; the upper left-hand
FIG. 3. Distribution of response as a function of CF for 269 units studied on 11/13/78 using /æ/ at 38 dB as stimulus. Plots at upper left show distribution of normalized rate. Each of remaining plots shows distribution of synchrony to one harmonic; harmonic indicated by number at upper left. Calculations based on period histograms with 128 bins/cycle. Arrows point to CF which is equal to frequency of harmonic (i.e., to place of harmonic). Each point derived from response of one unit. Points plotted with open square symbols are from units with spontaneous activity less than 1/s; points plotted as Xs are from higher spontaneous units. Lines are averages of the points computed with a triangular weighting function whose base was 0.25 octaves wide (see footnote 4).
FIG. 4. Same as Fig. 3 except showing responses to /ɛ/ at 58 dB from the same experiment.
FIG. 5. Same as Fig. 3 except showing responses to /ɛ/ at 78 dB from the same experiment.
plot shows the distribution of normalized rate to the
vowel as a function of CF. Each of the remaining 11
plots shows the synchronization index for one harmonic
of the stimulus (indicated by the number at upper left
in each plot), plotted as a function of CF. Data for units
with spontaneous rate less than 1/s are plotted with
open square symbols; data for units with higher sponta-
nous rates are plotted with xs. The arrow in each
plot points to the CF corresponding to the harmonic for
which responses are shown in that plot (referred to be-
low as the “place” for that harmonic). The lines are
averages of the data points computed with a triangular
window whose base was 0.25 octave wide. The har-
monics for which data are shown were selected to be
typical of the response distributions observed. Response
components fell into three groups: those corresponding
to the formant frequency harmonics (indicated by a
block letter F in the upper left corner of the plot); those
related to harmonics and intermodulation products of
the formant frequencies; and those at other harmonics,
not related to the formants.

Figure 3 shows the distribution of synchronization to
the components of the /e/ presented at 38 dB SPL.
Notice that for each of the harmonics shown (with the
exception of harmonic 1), there is a peak in the syn-
chronization distribution in the vicinity of the arrow;
at this level there is a tendency for the response to a
particular harmonic to be strongest among fibers with
CFs near the harmonic’s frequency. Synchronization
to the first formant component, harmonic 4, is par-
ticularly strong and widespread, reflecting the fact that
this is the largest component in the stimulus (see Fig.
1). This is a general finding in our results: The largest
amount of temporal response is observed to the first
formant frequency and this response tends to spread
widely toward units with CFs above its frequency. The
synchronizations to the second (14th harmonic) and third
(19th harmonic) formants are also slightly stronger than
responses to nearby components (compare them with
synchronization to the 13th and 17th harmonics), but
the effect is not nearly as great as for the first formant.
Notice the large synchronization to the fundamental
component (harmonic 1) among fibers with CFs above
1.0 kHz. This is associated with the occurrence of
broad spectrum responses as in the bottom histogram
of Fig. 2.

Figure 4 shows the distribution of synchronization to
the same stimulus at a 20 dB higher level, 58 dB SPL.
Several important changes can be noted. Synchroniza-
tion to the first formant has spread toward lower CFs,
and increased greatly at CFs above 3 kHz. Much of
this spread is due to recruitment of units which did not
respond at 38 dB. Synchronizations to the first two
harmonics of the first formant (harmonics 8 and 12)
spread to the same regions as responses to harmonic 4.
Notice that growth of these harmonics does not occur
close to their places, where the amplitude of response
stays approximately constant or decreases. Synchroni-
sations to the second and third formants (harmonics 4
and 19) grow slightly, especially in units with CFs
above their frequencies. Notice that the only CF region
not dominated by the first formant at this level is the
region around 2 kHz where the second formant response
is large. Similarly, there is a notch in the distribution
of second formant synchronization at the place of the
third formant. Synchronization to harmonics not closely
related to the formants (i.e., 1, 6, 9, 13, and 17) is
smaller at this level than at 38 dB (except for the broad
spectrum responses of units with CF above about 2
kHz). This reduction is particularly noticeable in the
vicinity of the first formant, e.g., at harmonics 1 and
6.

Figure 5 shows the continuation of these processes
after an additional 20-dB increase in stimulus level. Syn-
chronization to the first formant (harmonic 4) has con-
tinued to increase, especially in the region of the place
of the second formant. In another experiment in which
/æ/ was used as the stimulus, the notch in the first
formant response at the second formant place was com-
pletely gone at this level. Synchronization to the first
two harmonics of the first formant, harmonics 8 and 12,
have grown considerably over a broad region including
their places. Synchronization to the second and third
formants has become smaller in the vicinity of their
places. Synchronization to nonformant-related har-
monics, such as the 1st, 9th, 13th, and 17th, is now
almost gone, especially at the place of these harmon-
ics. This effect is striking; in Fig. 3, there were sig-
nificant peaks in the distributions of these nonformant-
related harmonics near the harmonics’ places. These
peaks have disappeared in Fig. 5. Temporal response
to the 78 dB /æ/ in the population of units is almost en-
tirely dominated by responses to the first formant and
its harmonics. The principal exceptions to this rule are
the responses to the second and third formant frequen-
cies remaining in the vicinity of their places and the
response at the various difference frequencies associ-
ated with the formants and their harmonics (e.g., har-
monic 6 = 14 – 8, see Discussion; harmonic 10 = 14 – 4;
harmonic 12 = 8 + 4 or 3 × 4, see Discussion).

The population responses shown in Figs. 3–5 are
summarized in Fig. 6. The points in this figure show
the temporal response to each harmonic expressed as
discharge rate and averaged over units with CFs near
the harmonic’s frequency [ASIR in Eq. (3)], plotted as
a function of harmonic frequency. There is one plot for
each stimulus level used in experiment 11/13/78. The
results are plotted on a logarithmic ordinate and each
plot is shifted vertically from the previous one by an
order of magnitude. The scale shown on the ordinate
is for the top plot (78 dB SPL); the scales for successive
plots below the top one are simply shifted downward by
one order of magnitude for each plot (the average re-
sponse rate at harmonic 4 is about 100 spikes/s in all
cases). The lines drawn through some of the points in
this figure will be explained later. The choice of ±0.5
octave as the averaging bandwidth is, of course, arbi-
trary; this particular value was chosen because it en-
compases the peaks near each harmonic’s place in
Fig. 3. The conclusions to be drawn from this plot are
not strongly dependent upon the bandwidth.

The plots in Fig. 6 show the average temporal re-
sponse to a particular harmonic of the stimulus in the
The points show average localized synchronized rate computed according to Eq. (3) for responses to /e/ at all sound levels used in experiment 11/13/78. There is one point for each harmonic up to the 32nd. Points corresponding to formant frequencies are plotted with filled circles. Ordinate is scaled logarithmically. Plots are shifted vertically from one another by one order of magnitude for clarity. Maximum response in each plot is about 100 spikes/s. The lines are drawn through all points except those corresponding to the 2nd and 3rd harmonics and the sum and difference tones of the first two formants. See rules given in text.

region of the cochlea where response to that frequency is expected to be maximal. As such, they reflect a combination of rate, place, and periodicity information. To the extent that they occur at places inappropriate to the harmonic’s frequency, the effects of rectifier distortion (but not combination tones) are minimized in these plots. The points at the first three formant frequencies (4th, 14th, and 19th harmonics) have been plotted with filled circles. Notice that the ALSR is always largest at the first formant. At lower sound levels the profile of the ALSR is a good reflection of the spectrum of the /e/ (see Fig. 1), with local maxima at the first three formant frequencies. The only serious deviation is the slightly elevated response at harmonic 8. At higher levels, the large responses locked to harmonics of the first formant become more evident, particularly at harmonics 8 and 12. Except for these two harmonics, however, the largest peak above the first formant continues to be at the second formant. The salience of the third formant is considerably reduced at higher levels, in part due to responses at the harmonics of the first formant and the intermodulation frequencies of the first and second formants.

Figures 7–10 show the same analysis for experiment 11/28/77 in which /a/ was the stimulus. The basic pattern of synchronization described above for the /e/ can also be seen in these figures; in particular, note that the first and second formant frequencies (6th and 9th harmonics) and their distortion products (particularly the 3rd, 12th, and 15th harmonics) dominate the synchronization, especially at the highest level (Fig. 9). As was the case for the /e/, the first formant synchronization increases and spreads as sound level increases. The responses to the /a/ in this experiment differ from those for the /e/ shown in Figs. 3–5 in that less synchronization (in the sense of a peak at a harmonic’s place) is observed to harmonics not related to formant frequencies, even at the lowest sound level (shown in Fig. 7; compare Fig. 3). Suppression of synchronization to harmonics above the first formant is considerable at the highest level; even the second formant response (9th harmonic) is reduced, and the small third formant (19th harmonic) response observed near its place at 57 dB (Fig. 8) is eliminated entirely at 77 dB (Fig. 9). Synchronization to harmonics above the first formant is replaced by synchronization to the first formant (6th harmonic) and its harmonics (e.g., the 12th harmonic response shown in Fig. 9). The suppression of synchronization is less for harmonics below the first formant as is shown by harmonic 3. The distribution of the 12th and 15th harmonics in Figs. 7 and 8 is illustrative of the problem of rectifier distortion products. Notice that the distribution of the 12th harmonic follows that of the first formant (6th harmonic) very closely (the dotted lines on the 12th harmonic plots reproduce the distributions of the 6th harmonic at the same level). Similarly, the 15th harmonic distributions are large where both the 6th and 9th harmonic distributions are large (9th harmonic distribution is repeated with dotted lines). In particular, note that there are no peaks at the place of the 12th harmonic in Figs. 7 and 8. This is the behavior to be expected from rectifier distortion components (see Appendix A for a detailed analysis of these distributions). At the highest level (77 dB) shown in Fig. 9 the distributions of the 12th, and perhaps also the 15th, harmonics differ from those expected of a rectifier distortion product by having peaks in the vicinity of their places.

Figure 10 shows the ALSR for the responses to /a/ at all levels used in experiment 11/28/77. The prominence of the first two formants (6th and 9th harmonics, plotted with filled circles) is evident. At higher levels, harmonics (12th, 18th) and intermodulation products (3rd, 15th) of the formants are increasingly prominent. The third formant (19th harmonic) is not prominent in this case except at 47, 57, and 67 dB. Note that the general features of the spectrum of the stimulus (see Fig. 1) are once again well preserved in this plot, except for the responses at harmonics and intermodulation products of the first two formant harmonics.

Figure 11 shows a similar ALSR for the responses to /I/ at all levels used in experiment 10/18/77. The behavior of the distributions of synchrony observed in this experiment is very similar to that shown for /e/ in Figs. 3–5. The first three formants in this case were the 3rd, 16th, and 20th harmonics. Once again, the for-
FIG. 7. Distribution of response as a function of CF for 261 units studied on 11/28/77 using /a/ at 37 dB SPL as the stimulus. Layout of plots and symbols same as in Fig. 3. Dotted curve on 12th harmonic plot is average curve for 6th harmonic synchrony; dotted curve on 15th harmonic plot is average curve for 9th harmonic synchrony.
FIG. 8. Same as Fig. 7 except showing responses to /a/ at 57 dB from the same experiment.
FIG. 9. Same as Fig. 7 except showing responses to /a/ at 77 dB from the same experiment.
The lines drawn through some of the points in Figs. 6, 10, and 11 are meant to show the extent to which the ALSR resembles the stimulus spectrum if presumed distortion components are ignored. These lines were drawn through all the points remaining after eliminating points according to the following two rules:

1. The harmonic at which the largest amplitude response is observed is the first formant. Eliminate its second and third harmonics.

2. The harmonic above the first formant at which the most prominent peak is observed, discounting points eliminated by rule 1, is the second formant (most prominent means largest compared to adjacent points, not necessarily largest overall). Eliminate its second and third harmonics; eliminate the first sum and difference harmonics of the first two formants [i.e., $(F_1 + F_2)$ and $(F_2 - F_1)$].

The similarity of these plots to the spectra of the stimuli is clear. Notice particularly that changes in stimulus level over a 50–60 dB range do not alter the qualitative aspects of these plots; they continue to reflect the first two formants of the stimulus. This stability is in marked contrast to the behavior of discharge rate profiles in the same circumstances [compare the rate profiles shown at upper left in Figs. 3–5, 7–9; Sachs and Young, (1979a)].

The importance of suppression of response to harmonics not related to the formants can be seen by inspection of Fig. 11. ALSR profiles (with distortion components removed) have been reproduced here from Figs. 6, 10, and 11. Each set of three plots in Fig. 12 shows data from one vowel at three sound levels, with the plots superimposed for comparison. The average response to the first formant grows or stays constant as level increases for all three vowels. The response to the second formant may grow monotonically (for /I/) with level or may be suppressed at the highest level (for /a/ and /e/). However, responses to harmonics between the first and second formants are always suppressed at the highest level and this suppression is sufficient to maintain a local maximum of response at the second formant frequency. Similarly, responses to harmonics above the second formant do not increase with stimulus level beyond moderate levels.

I. DISCUSSION

A. Summary of results

The patterns of synchrony in the responses of auditory-nerve fibers to steady-state vowels described in this paper can be summarized as follows: At low stimulus levels, synchronized responses are usually observed at the frequency of each harmonic in the stimulus; these responses are maximal in the vicinity of the place corresponding to their frequency. Larger amounts of synchrony are observed for the larger stimulus components at the formant frequencies. As sound level is increased, the responses to the first formant frequency grows in amplitude and spreads throughout the population of auditory-nerve fibers, but most rapidly in the direction of higher CFs. This growth is accompanied by increased synchrony at harmonics of the first for-
B. Nature of response components

Responses synchronized to the components of a complex stimulus can be generated in three ways: They can be responses to energy present in the stimulus; they can be responses to propagating combination tones generated in the cochlea (Goldstein and Kiang, 1961; Smoorenburg et al., 1976; Siegel, Kim, and Molnar, 1977; Kim, Siegel, and Molnar, 1979); or they can be generated as rectifier distortion products from primary responses to stimulus or combination tones (Pfeiffer and Molnar, 1974; Goldstein, 1972). Although it is not generally possible to ascertain the extent to which each of these contributes to the responses of a particular unit, it is possible to make some inferences about the nature of responses when knowledge of the distribution of synchrony through a large population of fibers is available. Responses arising from stimulus components or from propagating combination tones are expected to be distributed in a manner similar to that of the responses to a pure tone of the same frequency, taking into account the effects of synchrony suppression (Rose et al., 1974). In particular, these responses should be maximal in the vicinity of the place in the population appropriate to the frequency of the component (Kim, Siegel, and Molnar, 1979; Pfeiffer and Kim, 1973). The distribution of rectifier distortion products, on the other hand, should follow the distribution of the primary responses from which they are generated and therefore have a distribution more appropriate to the primaries’ frequencies than to their own (see Appendix). Based on the exponential model for the production of rectifier distortion, the amplitudes of rectifier distortion products are expected to be smaller than the amplitudes of the primaries producing them (Goldstein, 1972; Johnson, 1974; but see Appendix for some special problems in the case of vowels). These two properties of rectifier distortion allow it to be easily differentiated from primary responses in most cases.

The problem of differentiating responses to stimulus components and combination tones is more difficult in the case of multicomponent stimuli like our vowels, because there is a stimulus component at every frequency at which a combination tone might be expected. Smoorenburg and collaborators (1976) have shown that the amplitudes of the $f_2-f_1$ and $2f_2-f_1$ combination tones in the responses of cells in the anteroventral cochlear nucleus are maximum when the sound levels of the primaries are approximately equal; their amplitudes are decreasing functions of the frequency spacing of the primary components, and are always less than the amplitudes of the primaries except at values of $f_2-f_1$ of 100 Hz or less. Products like $f_1+f_2$ and $2f_1+f_2$ have not been reported to be significant so that combination tones occur primarily at frequencies below those of the primaries. Because our vowels have generally low-pass spectra and because combination tones are expected to occur at frequencies below those of the primaries and be smaller in amplitude than the primaries, combination tones should be smaller than stimulus components in most cases in our study. The exceptions to this conclusion are the first and second harmonics of all vowels and the third harmonic in the case of /æ/ which might arise as both $f_3-f_1$ and $2f_3-f_1$ combination products of the first and second formants. Another apparent exception is the second harmonic of the first formant discussed below.

Several of the significant responses observed in our data can be unequivocally identified as rectifier distortion products. The clearest examples are provided by the responses at the 13th and 15th harmonics of the /æ/ at 37 and 57 dB SPL. As is shown in Figs. 7 and 8, the distributions of these responses follow qualitatively those of the primaries which would produce them as rectifier distortion products (the 12th harmonic is the second harmonic of the 6th harmonic and the 15th harmonic is the sum tone of the 6th and 9th harmonics). The top two plots of Fig. 11 in the Appendix show the
distributions of the 15th and 12th harmonic responses which are expected in this experiment on the basis of the exponential model, assuming that these two components arose as rectifier distortion products of primary responses at other frequencies. These distributions were computed by fitting a two-component exponential model of the form of Eq. (A1) to all the period histograms of responses to the 57 dB SPL /a/. The primary components for each fit were taken to be the two harmonics at which the largest responses were observed (subject to certain limitations, see the Appendix for details of the procedure). No units had primaries at the 12th harmonic and only four units had 15th harmonic primaries. The amplitudes of the 12th and 15th harmonic components in the resulting model histograms are the points plotted in Fig. A1. The smooth curves are the averages of the model points and the broken lines are the averages of the corresponding distributions of these harmonics in the data, taken from Fig. 8. The agreement of the expected and observed distributions is quite good in these two cases. Other cases in which responses appear to result largely from rectifier distortion include the 3rd harmonic response in units with CFs above 1 kHz in Figs. 7 and 8 (6th and 9th harmonic primaries) and the 10th harmonic response in Fig. 5 (4th and 14th harmonic primaries).

In several other cases responses appear to be a mixture of rectifier distortion and primary response. A good example of this situation is provided by the responses to the 12th harmonic of the /a/ at 77 dB SPL. As is shown in Fig. 9, this response distribution resembles that of the 6th harmonic over part of the CF range, but contains a significant peak just above the place of the 12th harmonic which does not appear in the 6th harmonic distribution. The bottom plot in Fig. A1 of the Appendix shows the expected distribution of 12th harmonic response based on the exponential model. The deviation of the expected and the observed response amplitudes in the vicinity of the 12th harmonic place is clear. Note also that in this region there were eight units in which the 12th harmonic was apparently a primary component (triangular symbols), because it was one of the largest two components of the response (in particular, larger than the 6th harmonic component). Another situation in which the response is most likely a mixture of rectifier distortion and a primary response is the 9th harmonic of responses to /e/ (Figs. 3-5). The distribution of this harmonic follows that predicted for a second harmonic rectifier distortion component of the first formant below 0.9 kHz and above 3 kHz but deviates in the vicinity of its place.

In all of the vowels which we have used, we have observed large responses to the second harmonic of the first formant at high sound levels (around 80 dB) which could not be explained as rectifier distortion. The 8th harmonic responses to /e/ of units with CFs above 0.9 kHz in Fig. 5 and the 12th harmonic responses to /a/ of units with CFs near 2.0 kHz in Fig. 9 are examples. Although these response distributions are consistent with rectifier distortion over part of the CF range, they are significantly larger than expected over a range of CFs in the vicinity of their place (demonstrated for the 12th harmonic of /a/ in Fig. A1). The 8th harmonic of the responses to /e/ is more typical of the behavior of the first formant’s second harmonic in the vowels we have studied in that a large response is observed over a wide area including and above this harmonic’s place. If it is accepted that these responses do not result from rectifier distortion, then their amplitudes must be regarded as unusual. The stimuli contain components at these frequencies, but they are small, roughly the same size as adjacent harmonics to which little or no response is observed at these levels (e.g., harmonic 9 in Fig. 5 and harmonics 10 and 14 in Fig. 9). There is no apparent reason why the stimulus energy at the first formant’s second harmonic should induce responses significantly larger than those at adjacent frequencies. In cases where these second harmonics are large, significant synchrony is observed to frequencies which can be interpreted as rectifier distortion involving the second harmonic as a primary. Good examples can be seen in Fig. 5: The distribution of harmonic 13 is approximately that of a sum product of harmonics 4 and 8; and the response to harmonic 6 is consistent with a difference product of harmonics 14 and 8 (except for the frequency range from 1.5 to 2.3 kHz, where harmonic 6 is larger than predicted for rectifier distortion from this source; harmonic 6 must arise from other sources as well). Thus it seems likely that the large responses to the second harmonics of the first formants are primary responses, but unlikely that they are responses to stimulus energy. One possibility is that they arise from propagating second harmonic combination tones generated by nonlinearities in the cochlea. These second harmonics responses may be related to the “peak splitting” reported in responses to single tones at high sound levels (Kiang and Moxon, 1972; Johnson, 1974).

C. Synchrony suppression

It is well known that the response of an auditory-nerve fiber to one stimulus component can be suppressed by stimulus energy at other frequencies (Sachs and Kiang, 1968; Hind et al., 1967). When a two-tone stimulus is used, synchrony to either or both tones can be observed in the responses of auditory-nerve fibers, depending upon the relative amplitudes and frequencies of the tones; in general, as the amplitude of one of the stimulus tones is increased, the temporal response of a unit will be increasingly dominated by synchrony to it and synchrony to the other tone will be suppressed (Rose et al., 1974; Arthur, 1976; Johnson, 1974). A two-tone stimulus in which one tone is at the unit’s CF and a second, more intense tone is located below CF is most relevant to our vowel results for units with CFs above the first formant. If the relative level of the two components in such a complex is held constant while the overall level is increased, synchrony to the CF tone will behave nonmonotonically; it will increase at low levels but then decrease at higher levels. At the same time, synchrony to the more intense low-frequency tone will increase, so the unit’s responses will go from domination by the CF tone at low levels toward domination by the low-frequency tone at higher levels (Rose et al., 1974; Reale and Geisler,
1979; Bernardin, unpublished results). This is the behavior which we have observed in responses to vowels; the small but significant synchrony to harmonics not related to formants which is observed near their places at low sound levels is gone at higher levels, replaced mostly by synchrony to the first formant frequency. Since synchrony grows monotonically with sound level for single-tone stimuli (Arthur, 1976; Johnson, 1974; Rose et al., 1974), the failure of nonformant-related harmonics to grow with vowel level must reflect suppression by first formant energy. The effects of suppression on the responses to the second and third formants are not as severe, but can be clearly seen in the nonmonotonic behavior of synchrony to these frequencies as sound level is increased. The only case in which synchrony to a second or third formant did not decrease at higher sound levels was the second formant of /I/; in this case the response saturated at a level slightly below the maximum synchrony expected at its frequency (Johnson, 1974). Even though the first formant dominates the responses in most cases, it is probably also suppressed under certain circumstances, such as in the region of the place of the second formant in the case of /e/ (Figs. 4 and 5). Similar effects have been reported for two-tone stimuli by Kim and collaborators (1979).

In our previous paper (Sachs and Young, 1979a), we have argued that the rate response of units with CFs near the second or third formants could be approximated by a two-tone stimulus situation in which one tone (second or third formant) is at the unit's CF and a second, stronger tone (first formant) is located below CF. The similarity of rate versus level functions for responses to vowels and for responses to two-tone stimuli with constant amplitude difference between the levels of the tones led us to conclude that the same suppressive mechanism was operating in both cases. Inspection of the distributions of synchrony in Figs. 3-5 and 7-9 suggest that the two-tone approximation is not a bad one for units with CFs near the second or third formants. It is therefore not surprising that the rate-level functions to the vowel should resemble those for two-tone stimuli.

Upon reflection, it is clear that synchrony suppression is vital to the maintenance of the shape of the ALSR functions over such a wide range of sound levels, and therefore to the preservation of information about the stimulus spectrum in the temporal patterns of response of auditory-nerve fibers. If the responses to harmonics between the formants were not strongly suppressed, they would increase with sound level and saturation of synchronized rate at high stimulus levels would flatten any measure of the internal representation of the stimulus spectrum, such as the ALSR. Synchrony suppression thus potentially plays an extremely important role in the perception of stimuli with complex spectra such as speech or music. In discussing cochlear filtering of broadband noise signals, several authors have emphasized the linear component of the system as reflected by the reverse correlation function (de Boer and de Jongh, 1978; Möller, 1977; Evans, 1978). The Fourier transform of the reverse correlation function provides an estimate of the bandwidth of the linear component of cochlear filtering; this bandwidth increases only slightly at high stimulus levels (Möller, 1977; Evans, 1978). Furthermore, the reverse correlation function can be used as an estimate of the impulse response of the system to predict the waveform of the firing probability of auditory-nerve fibers to pseudorandom noise. The predicted waveforms agreed reasonably well with measured values (de Boer and de Jongh, 1978). Results such as these have led to the suggestion that the linear component of cochlear filtering is predominant in responses to broadband noise. In the case of periodic vowels, however, our results show that cochlear frequency analysis at high sound levels is highly nonlinear, as evidenced by the strong synchrony suppression effects. The contribution made by the linear filtering process isolated by reverse correlation is unclear, especially for units with CFs between the formants whose responses at high levels are dominated by the first formant and display no synchrony to harmonics near their CF.

D. Temporal encoding of speech

It is obvious that the first step in speech perception must be obtaining an internal representation of the acoustic speech signal which contains sufficient information about its spectrum to allow the identification of phonemes. In the case of vowels, estimates of the formant frequencies would suffice (Peterson and Barney, 1952; Pols, Tromp, and Plomp, 1973; Klein, Plomp, and Pols, 1970; Broad and Wakita, 1977); models of vowel perception have frequently sought to estimate the formant frequencies of parameters closely related to them (e.g., Carlson, Fant, and Granstrom, 1975; Karnickaya et al., 1975; Pols, Tromp, and Plomp, 1973). The data in this paper relate to the internal representation of vowel sounds at the level of the auditory nerve. One of the challenges in considering the auditory representation of speech is the fact that speech intelligibility is essentially constant over a broad range of levels, extending well above 100 dB SPL (Pollack and Pickett, 1959; to about 100 dB SPL in the absence of the stapedius muscle reflex, Borg and Zakrisson, 1973). It seems likely that the internal representation of speech should also be relatively insensitive to changes in stimulus level. In our previous paper we have shown that peaks corresponding to formant peaks in the spectra of vowel stimuli exist in the profile of discharge rate over populations of auditory-nerve fibers at low sound levels, but as rate saturation limits the discharge of fibers with CFs in the vicinity of the formant frequencies, the peaks in the rate profiles are lost (Sachs and Young, 1979a). This effect can be seen from the rate profiles plotted in Figs. 3-5 and 7-9 of this paper. Effects related to two-tone suppression contribute to the loss of formant representation, rather than sharpening it, as might be expected for a lateral suppressive mechanism. We have pointed out that the loss of a sharp representation of the stimulus spectrum at high sound levels does not conclusively rule out the rate and place representation because there are still differences in the gross shape of the rate profiles for different vowels and there is a subpopulation of low spontaneous rate, high
threshold auditory-nerve fibers (Liberman, 1978) which extends the range over which formant peaks can be seen in at least some part of the population of auditory neurons (Sachs and Young, 1979a). By utilizing all the available cues, it is conceivable that speech recognition could be carried out on the basis of rate alone, but this would require that stimulus intensity be estimated separately and used to control which set of cues is examined and which rules are used to identify particular patterns of activity. This does not seem to be a particularly appealing or parsimonious approach, particularly in view of the data in this paper.

Although a representation of speech in terms of rate and place does not seem to be adequate, there are compelling reasons to retain a representation based on place and seek a replacement for rate as the response measure. One argument for a place representation is that every auditory region of the central nervous system is tonotopically organized (Clopton, Winfield, and Flammingo, 1974). It is unlikely that place along the basilar membrane would be preserved through successive levels of central processing if it were not an important parameter of the internal representation of sound. Another argument for representation in terms of place comes from the study of periodicity pitch (Schouten, Ritsma, and Cardodo, 1962) which refers to a pitch sensation evoked by stimuli such as tonal complexes consisting of several successive harmonics of the same fundamental. The pitch of such stimuli is at the fundamental frequency of the complex whether there is energy in the stimulus at that frequency or not. This pitch was originally attributed to the periodicity at the frequency of the fundamental which appears in the envelope of such a stimulus. Subsequent investigations have made this hypothesis untenable (reviewed in Goldstein, 1978). The importance of a place representation of the spectrum of the stimulus to periodicity pitch was made apparent by Houtsma and Goldstein, (1972) who showed that periodicity pitch could be evoked if the frequency components of the complex stimulus were separated and presented to different ears. Goldstein (1973) subsequently formulated an optimum processor model for pitch recognition which worked by first estimating the frequencies of the components of the stimulus and then finding the harmonic template which provided the best fit of its components to the estimated frequencies from the stimulus. Thus the pitch sensation was related to the separation of frequency components (place) in the (possibly dichotic) stimulus rather than to temporal features of the stimulus waveform (periodicities). This model successfully explains many of the properties of periodicity pitch which a strictly periodicity-based analysis cannot (Wightman, 1973; Goldstein et al., 1978; Gerson and Goldstein, 1978). The pitch sensations evoked in a number of situations by noise stimuli having flat long-term spectra can also be explained in terms of a similar place mechanism (Bilsen, 1977; Pierce, Lipes, and Cheetham, 1977).

The similarity of periodicity pitch and vowel identification is evident; in both cases the problem reduces to specifying the spectrum of a complex signal. For periodicity pitch, it is the frequencies of the components of the signal which must be estimated; in the case of speech, it is the frequencies at which the stimulus has peaks of energy which are of interest. In both cases profiles of discharge rate as a function of place on the basilar membrane seem inadequate to supply the needed information at high stimulus levels (Sachs and Young, 1979a; Smoorenburg and Linschoten, 1977; Evans, 1978). It is clear from the results presented in this paper that the patterns of temporal response of auditory-nerve fibers are quite adequate to supply information about the spectrum of a vowel stimulus at sound levels up to at least 80 dB SPL. This can be most clearly seen in the ALSR plots of Figs. 6, 10, and 11. These plots reflect a combination of rate (the higher the average rate, the larger the temporal response), place (only those units with CF's near the harmonic are averaged) and synchrony information. Responses to the first two and sometimes the third formant frequencies stand out in these plots at all stimulus levels. The most important feature of these results is the stability of the plots over a wide range of sound levels. The major change as sound level is raised is an increase in the response at the harmonics and intermodulation frequencies of the first two formants. This phenomenon is not a serious problem for a synchrony-based response measure because it is quite predictable; in fact, the existence of such harmonics and intermodulation products might even be useful in providing additional information about the identity of the formants because each set of formant frequencies would generate its own unique set of distortion products.

Utilization of the information present in the temporal patterns which we have described requires that the CNS perform a limited spectral analysis of the responses of auditory neurons; the analysis is limited in the sense that it is only necessary to extract one frequency component from the response of each fiber (or ensemble of fibers with similar CF's), and that is the component at or near the fiber's CF. This is what we have done in constructing the ALSR functions. Whether the CNS is capable of such a calculation can be questioned, but the fact that central neurons are sensitive to the small interaural time delays necessary for binaural hearing (Rose et al., 1966; Goldberg and Brown, 1969) suggests that calculations of the kind needed are at least feasible.

We have used the ALSR computation only to demonstrate the existence of a stable representation of the spectrum of vowels in the temporal patterns of auditory-nerve fibers. The ALSR is based on the computation of period histograms. This form of analysis requires a phase reference for the stimulus signal which the central nervous system lacks; even if such a phase reference could be derived from the responses of auditory neurons, the period histogram is of limited usefulness for signals such as real speech which are not perfectly periodic. There are many other calculations which are feasible for the CNS, however, and which yield essentially the same information as the ALSR. Carlson and his collaborators (1975) have presented a model for vowel perception based on counting zero-crossings in the outputs of a bank of bandpass filters. Because the formant harmonics tend to dominate responses to vow-
els, a computation of this type applied to some estimate of the instantaneous rate of auditory-nerve fibers (derived from an ensemble average instead of a time average) would yield results with prominent formant representation. Another general class of computations which are, in principal, feasible for the CNS are those based on the interval statistics of auditory-nerve fiber discharge. Using interval histograms as a temporal response measure gives roughly the same information as the ALSR (Sachs and Young, 1979b) because the Fourier transform of the interval histogram of an auditory-nerve fiber's discharge is approximately the power spectrum of the period histogram of the same data (Johnson, 1974). Goldstein and Srulovicz (1971) have considered the use of the interval histogram in estimating the frequency of pure tones and of components of harmonic series. The theoretical variance of the estimate of frequency derived in this way depends upon stimulus frequency and duration in the same way as the JND for frequency measured psychophysically; the decay of precision of this measure as frequency increases is exactly the behavior required of the estimator of the frequencies of the components of harmonic series in Goldstein's optimum processor model for periodicity pitch.

In this paper we have used only steady-state stimuli; real speech, of course, contains few, if any, portions with steady spectral characteristics. In particular, the vowels of connected speech need not have steady formant frequencies. Their formants undergo dynamic transitions depending upon the consonantal context in which they are found (Liberman et al., 1967). In rapid speech a vowel may not even achieve its canonical formant configuration (Lindblom, 1963; Stevens, House, and Paul, 1966; Lindblom and Studdert-Kennedy, 1967). Studies of vowel identification have shown that the transitional part of the vowel where the short-term spectrum is changing from the formant configuration appropriate to the consonant preceding or following the vowel strongly affects perceived vowel identity (Lindblom and Studdert-Kennedy, 1967) and may even be a better cue to vowel quality than the quasi-steady portion of the syllable (Strange et al., 1978). Regardless of their importance for vowel quality, the short-term spectra of these transitional regions clearly carry information on the identity of consonants (Stevens and Blumstein, 1978; Halle, Hughes, and Radley, 1957). The problem of estimating the spectra of nonsteady stimuli from the discharges of auditory neurons should be subject to the same basic considerations which have been brought to bear on steady-state vowels in this and our previous paper. Encoding of vowel stimuli in terms of discharges was limited by rate saturation and two-tone suppression (Sachs and Young, 1979a). Studies of the responses of auditory neurons at the onset of stimulation have shown that there is no measurable latency for two-tone suppression (Arthur, Pfeiffer, and Suga, 1971) and that the rate saturation characteristics of auditory neurons are basically the same at stimulus onset as they are after a short period of adaptation (Smith, 1979). In the case of vowels, the profiles of discharge rate versus CF computed over the first 50 ms of the vowel stimulus are simply scaled versions of the steady-state profiles (Sachs and Young, 1979a). Encoding of the spectra of transitional stimuli in terms of temporal patterns of response presents several interesting problems and promises to be a fruitful area for future research.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institute of Neurological and Communicative Disorders and Stroke. M. B. Sachs and E. D. Young held N.I.H. Research Career Development Awards.

APPENDIX

In the main body of this paper we have discussed the possibility that many of the response components observed in our data arise as distortion products produced by the equivalent of rectification in the cochlear transduction process. In this Appendix we will consider some of the special difficulties which arise in considering the problem of rectifier distortion for stimuli like our vowels. We will then compare the distribution of presumed rectifier distortion observed in three cases with the predictions of the exponential model for the instantaneous rate of auditory-nerve fibers (Siebert, 1970; Goldstein, 1972; Littlefield, 1973; Johnson, 1974 Molnar, 1974). The distribution of the 12th and 15th harmonic responses to the 57 dB $\alpha$/ (Fig. 8) will be shown to be consistent with the predictions of this model, whereas the distribution of the 12th harmonic responses to the 77 dB $\alpha$/ (Fig. 9) will be shown to deviate significantly for CFs just above this harmonic's place.

For the case of a fiber responding to two tones, the exponential model expresses the instantaneous rate of discharge $r(t)$ as,

$$r(t) = SR \cdot \exp \left[ a_1 \cos (\omega_1 t + \phi_1) + a_2 \cos (\omega_2 t + \phi_2) \right]. \quad (A1)$$

SR is the spontaneous rate; $a_1$, $a_2$, $\phi_1$, and $\phi_2$ are the amplitudes and phases of the primary components at frequencies $\omega_1$ and $\omega_2$ and $a_0$ is a stimulus-dependent constant term (i.e., $a_0 = 0$ for no stimulus). The primary components are not limited to frequency components actually present in the acoustic stimulus; they also include propagating combination tones generated in the cochlea which apparently set up a traveling wave on the basilar membrane and otherwise act like actual stimulus components (Seigel, Kim, and Molnar, 1977; Kim, Siegel, and Molnar, 1979). Equation (A1) is introduced only to model the distortion produced by the rectification involved in the conversion of the cochlear excitation pattern to discharge rate. Cochlear filtering and the effects of cochlear nonlinearities such as synchrony suppression are expressed as modifications of the primary components.

Equation (A1) may be rewritten in an expansion in terms of sinusoids as below (Olver, 1864):

$$r(t) = SR \cdot \exp \left[ \left( I_0(a_0) + 2 \sum \_n I_n(a_0) \cos (n\phi_0) \right) \cdot \left( I_0(a_2) + 2 \sum \_m I_m(a_2) \cos (m\phi_2) \right) \right]. \quad (A2)$$
where $I_n(a)$ is the modified Bessel function of order $n$. If the multiplication of the series in Eq. (A2) is carried out, response components at all frequencies $(\omega_1 \pm m\omega_2)$ will result. We will term all the components of the response except those at the primary frequencies **rectifier distortion products**. For the special case of $\omega_1$ and $\omega_2$ relatively prime, the amplitude of the component of frequency $(\omega_1 \pm m\omega_2)$ will be (Johnson, 1974; Molnar, 1974),

$$2R_{\omega_1 \pm m\omega_2} = 2SR e^{i\omega_1 t} I_n(a_1)I_n(a_2), \quad (A3)$$

where $R$ is the magnitude of the Fourier transform, as defined in Eq. (1) of the main body of the paper. The response amplitudes at the primaries are given by Eq. (A3) with $n = 1$ and $m = 0$ (or vice versa). Note that the rectifier distortion component at frequency $(\omega_1 \pm m\omega_2)$ has the same amplitude as the component at $(\omega_1 \mp m\omega_2)$. The Bessel functions $I_n(a)$ are monotonic decreasing functions of their order $n$, so that the amplitudes of response components will be a decreasing function of the order of the distortion (Johnson, 1974; Goldstein, 1972). That is,

$$R_{\omega_2} > R_{\omega_1 \pm \omega_2} > R_{2\omega_1 \pm \omega_2} > R_{3\omega_1 \pm 2\omega_2} > \cdots \quad (A4)$$

Notice particularly that the amplitude of the response at each of the primaries will be larger than the amplitude of any distortion component involving that primary, although not necessarily larger than all distortion components. For example, $R_{\omega_2}$ may be larger or smaller than $R_{\omega_1}$. 

Examination of Eq. (A3) shows that the amplitude of a rectifier distortion product depends on the amplitudes of the primary components at the input of the rectifier, i.e., upon $a_0$, $a_1$, and $a_2$. Thus the distribution of a rectifier distortion component will be determined by the distribution of $a_0$, $a_1$, and $a_2$ and therefore by cochlear filtering appropriate to the primary frequencies $\omega_1$ and $\omega_2$, rather than the frequency of the distortion product. The response to a distortion product will be large at places where the responses to both primary components are large and therefore the distribution of the rectifier distortion product will resemble the product of the distributions of response to the primaries rather than the distribution expected for a tone of the same frequency as the distortion product. In particular, peak response will not necessarily occur at the place appropriate to the frequency of the distortion product, but at a place determined by the distribution of $a_0$, $a_1$, and $a_2$. This property differentiates rectifier distortion products from response components resulting from stimulus components or propagating combination tones, which peak at the place appropriate to their frequency (Pfeiffer and Kim, 1975; Kim and Molnar, 1979; Kim, Siegel, and Molnar, 1979).

The expectations discussed above concerning the distribution of response to primary components and rectifier distortion components have been used in this paper to differentiate the various sources of response. Response components which peaked at or near their place were considered to be responses to stimulus components (or possibly combination tones); whereas components which peaked at a place inappropriate to their frequency and for which suitable primary response distributions were observed were considered to be rectifier distortion. However, in the case of responses to vowels, there are two uncertainties in this argument. First, the frequencies of the primary components are rationally related by ratios of relatively small integers (such as $9/6$) for the case of periodic tonal complexes like our vowels. In this circumstance, the amplitudes of response components are not given by simple equations like (A3) because carrying out the multiplication in Eq. (A1) will lead to several terms at each frequency. If $\omega_1$ and $\omega_2$ are the $K_1$ and $K_2$th harmonics of a fundamental frequency $\omega_0$ then the response component at harmonic $K$ (frequency $K\omega_0$) is given by Eqs. (A5) and (A6). These two equations represent the vectorial sum of all the terms from Eq. (A2) at frequency $K\omega_0$, $R_K$ and $\theta_K$ are the amplitude and phase, respectively, of the component of $r(t)$ at frequency $K\omega_0$.

$$2R_K \cos(\theta_K) = 2SR e^{i\omega_0 t} \sum_{p=q}^{q+r} \sum_{p-q=Q(p,K)} I_p(a_1)I_q(a_2) \times \cos(p\phi_1 + q\phi_2), \quad (A5)$$

$$2R_K \sin(\theta_K) = 2SR e^{i\omega_0 t} \sum_{p=q}^{q+r} \sum_{p-q=Q(p,K)} I_p(a_1)I_q(a_2) \times \sin(p\phi_1 + q\phi_2), \quad (A6)$$

where $M$ is large enough to include all significant components of the response and $Q(p,K) = \{q | -M, M \}$ such that $(pK_1 + qK_2) = K_0$. (A7)

Since there are now several terms with differing phase contributing to each response component, including those at the frequencies of the primaries, it is possible that these terms will partially cancel one another. Thus the simple rank ordering of response components discussed above in Eq. (A4) does not necessarily hold. It is also not necessarily true that the distribution of rectifier distortion components is simply or monotonically related to the amplitude of primary components. In many cases, the terms in Eqs. (A5) and (A6) with the lowest order of distortion (smallest $p$ and $q$) is large compared to the other terms in the series and Eqs. (A3) and (A4) are reasonable approximations. The deviation of Eqs. (A5) and (A6) from Eq. (A3) is worst for cases like /a/ where the primaries are frequently at 6 and 3 or 6 and 9, producing many terms in Eqs. (A5) and (A6) with small orders of distortion and comparable magnitudes. For example, the response component at harmonic 12 resulting from primaries 6 and 9 contains a term proportional to $I_1(a_1)I_1(a_2)$ and a second term proportional to $I_1(a_1)I_2(a_2)$ which may be approximately equal. The deviation of Eqs. (A5) and (A6) from Eq. (A3) is negligible for cases like /e/ with primaries 4 and 14. The response component at harmonic 10 with primaries 4 and 14 is dominated by a term proportional to $I_1(a_1)I_1(a_2)$; the next term in the series is proportional to $I_1(a_1)I_1(a_2)$ and is negligible. One important aspect of the rank ordering of response components in Eq. (A4) is usually not affected by cancellation: The re-
response components are still largest at the primary frequencies, in the sense of Eq. (A4). This rule is violated only for extreme values of the $a_i$, infrequently encountered in our data. For example, for primaries at harmonics 6 and 9, empirical computation of model histograms show that the largest response components are at harmonics 6 and 9 except that the response at harmonic 3 may be slightly larger than the response at harmonic 9 if $a_6 > 2.8$ and $\phi_6$ is within about 30° of $3\phi_6/2$.

A second complication, in the case of vowels, is that there is energy in the stimulus at a large number of frequencies, including the presumed distortion frequencies. Thus the potential number of primaries is large and although Eqs. (A5) and (A6) can be extended to cover this situation, they become very complex and the problem of estimating the $a_i$ and the $\phi_i$ is difficult. In what follows, we will apply the two-primary model of Eq. (A1) to responses to vowels; this is a good approximation for the case of /a/ at 57 and 77 dB, where little response is observed at frequencies other than those related to the formants.

The possibility that the distribution of rectifier distortion might differ significantly from that predicted from the distribution of the primaries makes it worthwhile to compare the distortion observed in our data with the predictions of the exponential model expressed in Eqs. (A5) and (A6). We have done this by estimating the parameters $(\alpha_0, a_1, a_2, \phi_1, \phi_2)$ for the best fit of Eq. (A1) to each period histogram of responses to a particular vowel, resulting in a model histogram for each period histogram. The primaries for each model histogram were chosen to be at the frequencies of the largest two components in the period histogram (disregarding the constant component), except that the frequency of the primary with the second largest response component was not allowed to be a harmonic of the primary with the largest response component. This restriction is necessary since large rectifier distortion products are expected at the harmonics of the primary components. This choice of primary components is based on the rank ordering of primaries and distortion components discussed above; suitable precautions were taken to eliminate the small number of units for which it was likely that some distortion components would be larger than one of their primaries. The best-fitting parameters were those which produced a model histogram having exactly the same components (amplitude and phase) at the two primaries and the same average value (constant component) as the period histogram.

The procedure for the estimation was as follows. Two equations, like (A5) and (A6), can be written relating the response at each primary frequency (amplitudes $R_{a_1}$ and $R_{a_2}$; phases $\theta_{a_1}$ and $\theta_{a_2}$) to the five unknown parameters of the fit, resulting in four equations. An additional equation can be written for the average value of the histogram:

$$R_0 = SR \cdot e^{a_0} I_0(a_1) \cdot I_0(a_2) + 2 \sum_{p=-1}^{1} \sum_{q=-1}^{1} I_p(a_1) \cdot I_q(a_2) \cdot \cos(p\phi_1 + q\phi_2).$$

These five nonlinear equations in five unknowns were solved by a modified Newton–Raphson technique. The procedure converged in the vast majority of cases (e.g., 174 out of 179 for the data in the top two plots of Fig. A1); most of the cases where it failed to converge were situations in which the identity of the primaries was in doubt.

Figure A1 shows three examples of the results of this analysis. The top plot shows the distribution of synchronization to the 15th harmonic for model histograms fit to the responses to /a/ at 57 dB SPL from the experiment of 11/28/77 (data in Fig. 8); the middle plot shows the distribution of synchronization to the 12th

![Figure A1](https://example.com/figureA1.png)

**FIG. A1.** Distributions of 12th or 15th harmonic in model histograms fit to period histograms of responses to /a/ at 57 or 77 dB from 11/28/77. Points are amplitudes of components in model histograms. Symbol with which point is drawn shows the primaries used in computation of model histogram from which point is taken; defined in the legend. Solid lines are averages of model points computed using triangular window with base width 0.25 octaves. Dashed lines are averages of same component in data histograms (as in Figs. 8 and 9). Points from histograms with 12th harmonic primaries (triangles) in bottom plot are not included in either data or model averages. Arrows point to CF which is equal to harmonic's frequency (i.e., to place of harmonic).
For the same model fits; and the bottom plot shows the distribution of synchronization to the 12th harmonic for model histograms fit to the responses to the /a/ at 77 dB SPL from the same experiment (Fig. 9). The shape of the symbol with which each point is plotted identifies the dominant primaries used to compute the model histogram for that unit. In most cases, the primaries were either the 6th and 9th harmonics (squares), the 6th harmonic and some harmonic other than the 9th (diamonds), or the 3rd harmonic and some harmonic other than the 6th or 12th (Xs). In a few cases, the 12th or 15th harmonics themselves were primaries (triangles or filled diamonds, respectively). The smooth curves in this figure show the average synchronization computed from the model points in the same way as for Figs. 3-5 and 7-9 in the main body of the paper. The dotted curves show the average synchronization curves for the data histograms. There is reasonable agreement between the average data and model distributions in the top two plots, for responses at 57 dB. Notice that there were no cases in which the 12th harmonic was a primary at this level and only four cases in which the 15th harmonic was a primary. Thus the 12th and 15th harmonic components of these model histograms are exclusively or almost exclusively rectifier distortion products at this level. The principal deviations between the average model and data curves are a tendency of the exponential model to underestimate the distortion among units with CFs less than 0.5 kHz and a tendency to overestimate the distortion in the regions just below the place corresponding to the frequency of the harmonic (indicated by the arrows). The responses of the low CF units for which significant deviation is observed are mostly dominated by the 3rd harmonic. The predicted 12th and 15th harmonic components in these cases are small compared to the noise in the data histograms, which partially accounts for the deviation. But the deviation also reflects the fact that the exponential model, when fit in this way, consistently underestimates the distortion at harmonics more than three standard deviations above its expected value if there were no response component at that frequency (using the Rayleigh distribution). Average lines are plotted only if the weighted number of spikes in the analysis period upon which the period histogram was based. With 600 spikes in the record, a synchronization index of 0.1 is about three standard deviations larger than its expected value if there were no response component at that frequency (using the Rayleigh distribution). Average lines are plotted only if the weighted number of spikes in a window was at least two. Both high and low spontaneous rates are included in the averages.

In contrast to the two distributions for 57-dB data, the bottom plot showing the model predictions for the 12th harmonic at 77 dB shows a large deviation between the model prediction and the average value observed in the data. This deviation occurs just above the place of the 12th harmonic. In this region, there are nine cases in which the 12th harmonic was the dominant response component (plotted with triangles). Values for these units are not included in either the data or model average curves. Thus, the curves show a comparison of the amplitude of the 12th harmonic only in those units in which the 12th harmonic resulted from rectifier distortion in the model histograms. The most likely interpretation of the large deviation of model and data near 2 kHz is that the data average reflects responses to a propagating combination tone at the second harmonic of the large stimulus component at the first formant. This combination tone produces significant primary response just above the 12th harmonic’s place which is not included in our model responses because of the rules adopted for choosing primary components for the model histograms. Presumably the deviation occurs above the 12th harmonic’s place because of the effects of synchrony suppression (compare the behavior of the 9th harmonic responses in the same experiment, Figs. 7, 8, and 9; it is displaced upward in a similar fashion; see also the results of Kim, Siegel, and Molnar, 1979).

The synthesizer used was provided by J. M. Heinz. It is essentially the same as the one described by Rubiner (1968). Impulses at the pitch rate are shaped by a conjugate pole pair low-pass filter to model the glottal excitation; this is followed by five conjugate pole pair resonators which model the formants. Finally, the output is differentiated to produce a +6 dB octave radiation impedance correction.

Of course, not all the distortion components in the responses of auditory-nerve fibers result from rectification. Combination tones can also be generated within the cochlea, presumably as part of cochlear mechanics, and propagate on the basilar membrane like real stimulus tones (Siegel et al., 1977; Kim et al., 1979). Components generated in this way would appear at the input of the rectification process described in the text, i.e., as primaries.

Normalized rate is defined as average discharge rate over the last 390 ms of the 400-ms vowel stimulus minus spontaneous rate, divided by the unit’s saturation rate minus spontaneous rate. Saturation rate was estimated for each unit by presenting a CF tone 50 dB above threshold. For details about the construction of the normalized rate plots, see our earlier paper (Sachs and Young, 1979a).

Points are not plotted in Fig. A1 if the primaries for the model histogram were at harmonic numbers which could not produce significant distortion at the 12th or 15th harmonics (such as 2 and 9) or if the amplitudes of the 12th or 15th harmonic component in both the data and model histograms were less than 2.276√N, at which point the synchronization index is three standard deviations above its expected value if there were no response component at that frequency (Rayleigh distribution).


of the cat which may be Relevant to Localization of a Sound Source," J. Neurophysiol. 29, 289-314.


