Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds

Bertrand Delgutte

Eaton–Peabody Laboratory of Auditory Physiology, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts 02114 and Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

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

Several processing schemes by which phonetically important information for vowels can be extracted from responses of auditory-nerve fibers are analyzed. The schemes are based on power spectra of period histograms obtained in response to a set of nine two-formant, steady-state, vowel-like stimuli presented at 60 and 75 dB SPL. One class of “local filtering” schemes, which was originally proposed by Young and Sachs [J. Acoust. Soc. Am. 66, 1381–1403 (1979)], consists of analyzing response patterns by filters centered at the characteristic frequencies (CF) of the fibers, so that a tonotopically arranged measure of synchronized response can be obtained. Various schemes in this class differ in the characteristics of the filter. For a wide range of filter bandwidths, formant frequencies correspond approximately to the CFs for which the response measure is maximal. If in addition, the bandwidths of the analyzing filters are made compatible with psychophysical measures of frequency selectivity, low-frequency harmonics of the stimulus fundamental are resolved in the output profile, so that fundamental frequency can also be estimated. In a second class of processing schemes, a dominant response component is defined for each fiber from a 1/6 octave spectral representation of the response pattern, and the formant frequencies are estimated from the most frequent values of the dominant component in the ensemble of auditory-nerve fibers. The local filtering schemes and the dominant component schemes can be related to “place” and “periodicity” models of auditory processing, respectively.

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

INTRODUCTION

Speech intelligibility for normal human subjects is essentially constant over a broad range of stimulus levels, and is remarkably resistant to background noise and other degradations in the signal (Fletcher, 1953; Pollack and Pickett, 1958). Many aspects of auditory-nerve fiber responses to acoustic stimuli are strongly dependent on stimulus intensity and signal-to-noise ratio (Kiang et al., 1965; Kiang and Moxon, 1972, 1974; Sachs and Young, 1979; Voigt et al., 1981). Thus one needs to examine schemes for processing auditory-nerve data in order to obtain response measures that would be relatively invariant with respect to stimulus level and signal-to-noise ratio, while retaining information necessary for distinctions between speech sounds (Young and Sachs, 1979; Sachs and Young, 1980). In this paper, two classes of processing schemes that can be related to psychophysical models of auditory processing will be analyzed for steady-state, two-formant stimuli that correspond roughly to the cardinal vowels (Delgutte and Kiang, 1984a).

Perceptual data suggest that phonaemic distinctions among vowels are based primarily on the formant frequencies (Peterson and Barney, 1952; Pols et al., 1969; Shepard, 1972; Fant, 1973; Carlson and Granstrom, 1980). Another important vowel parameter is the fundamental frequency (corresponding to voice pitch), which plays a role in prosody, in the distinction between voiced and voiceless sounds, and, for tone languages, in word identification. The goal of the proposed processing schemes for vowels will be to estimate the formant frequencies and the fundamental frequency from auditory-nerve activity.

The class of processing schemes originally proposed by Young and Sachs (1979) consisted essentially of analyzing the response patterns through bandpass filters centered at the fiber CFs in order to obtain a tonotopically arranged representation of the stimulus. The formant frequencies of vowels were estimated from peaks in the profiles of the filter outputs against characteristic frequency. This class of “local filtering” schemes is further studied in this paper by examining the consequences of changing the characteristics of the analyzing filters in a manner consistent with psychophysical measures of frequency selectivity. Another class of schemes is possible in which a dominant response component is defined for each fiber, and the formant frequencies are identified with the dominant components that occur most frequently in the ensemble of auditory-nerve fibers. Unlike the “local filtering” schemes, the “dominant component” scheme does not involve the fiber CFs in the computation, so that it would not depend critically on a precise tonotopic organization.

I. METHODS

The proposed processing schemes are based on a set of auditory-nerve fiber data from anesthetized cats that has been described previously (Delgutte and Kiang, 1984a). Although these schemes were applied to all nine vowel stimuli used by Delgutte and Kiang (1984a), the illustrative figures

---

*Present address: Centre National d’Etudes des Télécommunications, Lannion, France.*
will concentrate on /i/, /æ/, and /u/, which have the most extreme formant frequencies, and /a/ whose formant frequencies are about average within the range of variations found in language. The figures are also limited to stimuli presented at 75 dB SPL, although the processing schemes were analyzed for 60 dB SPL as well. The general conclusions are applicable to all nine vowel stimuli at both stimulus levels.

All the processing schemes are based on period histograms and their discrete Fourier transforms, which were obtained as described by Delgutte and Kiang (1984a). The computation of the processing schemes is described in the Appendix. Because knowledge of the fundamental frequency is necessary for computing period histograms, the processing schemes presented here make use of a priori information about the stimulus. In principle, these schemes could also be based on segments of post-stimulus (PST) time histograms or their short-time spectra, so that the processing would not depend on the fundamental frequency. In a subsequent paper (Delgutte and Kiang, 1984b), it will be shown that the vowel processing schemes are also effective when based on short-time processing of PST histograms, using the 10–20 ms windows that are commonly used in speech analysis (Flanagan, 1972; Rabiner and Schafer, 1978).

II. RESULTS

A. Local filtering schemes

The basis for the class of processing schemes proposed by Young and Sachs (1979) is seen in Figs. 6 and 8 of Delgutte and Kiang (1984a), where the response amplitude along the curved line representing \( f = CF \) varies considerably with CF, and usually shows local maxima near the places of the formant frequencies. In essence, these schemes introduce a bank of filters that select the response components in the vicinity of the fiber CFs. The filter outputs are averaged across all fibers in a narrow band of CFs centered at the fiber center frequency, and the average is expressed as a function of the filter center frequency. Because each filter analyzes only the responses of fibers whose CFs are near its center frequency, these schemes provide a tonotopically arranged representation of the stimulus. In the original formulation of Young and Sachs (1979), the Average Localized Synchronized Rate (ALSR), was evaluated only for center frequencies of the analyzing filter corresponding to harmonics of the vowel fundamental frequency, and the filter bandwidths were sufficiently narrow to resolve individual harmonics throughout the range of frequencies. In more recent work with stop consonants (Sachs et al., 1982), the ALSR has been evaluated for both harmonic and nonharmonic frequencies, but the filter bandwidths were still chosen to be near 50 Hz. In the present analysis, a response measure resembling the ALSR was evaluated using filters whose center frequencies were finely spaced on a logarithmic scale, and the filter bandwidths were chosen in a manner more consistent with psychophysical measures of frequency selectivity.

Different processing schemes arise by choosing the analyzing filters differently. With these modifications, local filtering schemes can be applied to both periodic and nonperiodic stimuli, and can be tested for how well they resolve partials of a complex tone. A further modification of the Young and Sachs scheme is that the filter outputs were normalized by the mean square discharge rate. This normalization is intended to compensate for the considerable scatter of discharge rates in our data. In addition, it seems to improve the behavior of the processing schemes in certain cases, particularly when the filters are broadly tuned. To avoid confusion, the response measures obtained in the modified local filtering schemes will be called Average Localized Synchronized Measures (ALSM).

The first filter used to compute ALSMs is a 1/6 oct Gaussian bandpass filter whose transfer function is illustrated in Fig. 1(a). This choice of bandwidths corresponds roughly to the sharpest psychophysical measures of frequency selectivity (Houtgast, 1974; Vogten, 1978; Moore, 1978). Figure 2(a) shows plots of ALSM against the center frequency \( f_c \) of this narrow bandpass filter for the /i/, /æ/, /u/, and /a/ stimuli. The profile shows clear peaks at the formant frequencies above 1 kHz (F2 of /i/, /æ/, and /a/). For lower-frequency formants, a peak appears if the formant frequency coincides with one of the harmonics of the fundamental frequency \( F1 \) of /i/ and /æ/, while there are two peaks flanking the formant frequency if it falls between two harmonics \( F1 \) of /æ/ and /u/, F2 of /u/. For /æ/ and /a/, there are also large peaks at the low-frequency harmonics (second and third harmonics for /æ/, second harmonic for /a/). Such low-frequency peaks are found for all the relatively open vowels (/æ/, /a/, /e/, /s/, and /a/). For /i/ and the other spread vowels /e/ and /æ/ there are large peaks at 2F1 and in the CF region between F1 and F2 where response spectra have intense components near CF (Delgutte and Kiang, 1984a). If the formant frequencies were estimated simply from the positions of the two largest peaks in the

![FIG. 1. Transfer functions of four different filters with center frequency \( f_c \) that were used to analyze period histograms for the computation of average localized synchronized measures (ALSM). The details of the filtering operation are given in the Appendix. (a) Gaussian bandpass filter with a bandwidth of 1/6 oct. These functions correspond to \( H_f(f) \) in Eqs. (A2) and (A3). (b) Gaussian bandpass filter with a bandwidth of 2/3 oct. (c) Comb filter with a bandwidth of 1/6 oct. The implementation of this filter is shown in Fig. 3(b). (d) Comb filter with cosinusoidal transfer function. The implementation of this filter is shown in Fig. 3(a).](image-url)
FIG. 2. Four average localized synchronized measures plotted against center frequency of the analyzing filter for the /I/, /ae/, /u/, and /a/ stimuli presented at 75 dSPL. For each measure, the places of the formant frequencies along the f\* dimension are marked by dashed lines. (a) ALSM obtained using the 1/6 oct Gaussian bandpass filter of Fig. 1(a). (b) ALSM obtained using the 2/3 oct Gaussian bandpass filter of Fig. 1(b). (c) ALSM obtained using the 1/6 oct comb filter of Fig. 1(c). (d) ALSM obtained using the comsinooidal comb filter of Fig. 1(d).

ALSM profile, the nonformant peaks could lead to incorrect results.

Figure 1(b) shows the transfer function of a Gaussian filter with a 2/3 oct bandwidth that was used in place of the 1/6 oct filter. This choice of bandwidth is at least twice greater than the psychophysical critical bandwidth (Zwicker and Eldert, 1967; Scharf, 1970). Figure 2(b) shows profiles of ALSM against f\* for this broadband filtering scheme. In contrast to the results for the narrow-band scheme, there are peaks centered at the first formant of /ae/ and the second formant of /u/. However, the peak at the second formant of /I/ is less prominent, and the amplitude of the nonformant peak at 2 kHz is increased. This loss of second-formant information is typical for spread vowels (/I/, /ae/, and /u/) processed by broadband schemes. A combination of a broadband scheme for the low frequencies and a narrow-band scheme for the high frequencies would be preferable to either scheme alone for formant-frequency estimation.

Because cochlear transduction mechanisms are nonlinear, the response of an auditory nerve fiber to a single tone includes a DC component and components at harmonics of the stimulus frequency in addition to the fundamental component (Rose et al., 1967; Johnson, 1974). Thus, a comb filter, which would extract the response components near DC as well as components near the harmonics of its center frequency, may be better adapted to the response signal than a bandpass filter. As shown in the Appendix, a simple comb filter can be realized by evaluating the autocorrelation function of the period histogram for a time delay of 1/f\*. This computation requires only a delay, a multiplication, and a low-pass filter [Fig. 3(a)]. The transfer functions of the resulting filters have a sinusoidal shape, as shown in Fig. 1(d). Figure 2(d) shows ALSM profiles for this sinusoidal comb filtering scheme. For all stimuli, there are clear peaks at the formant frequencies except for the first formant of /ae/, where peaks occur at the two harmonics flanking the formant frequency. The nonformant peaks between F1 and F2 for /I/ are less prominent than for the bandpass filtering schemes. For /ae/ and /u/, open vowels in general, there are also peaks at the low-frequency harmonics. For all vowels, the ALSM is large at high frequencies as 1/f\*, approaches zero where the auto-correlation functions have their maximum.

To estimate the formant frequencies, it is desirable to smooth the spectrum in order to eliminate the fine spectral structure due to voice excitation. On the other hand, to estimate the fundamental frequency it is necessary to use a spectral representation in which individual harmonics of the fundamental are resolved. This is the case for the narrow-band filtering scheme of Fig. 2(a), as there are clear peaks at many of the harmonics below 1 kHz. Specifically, for /I/ there are peaks at harmonics 2, 4, 5, and 7, for /ae/ at harmonics 2, 3, 4, 6, and 7, for /u/ at harmonics 2, 3, 4, 5, and 6, and for /a/ at harmonics 2, 4, and 8. With the possible exception of /a/, these harmonic sequences would suffice to estimate the fundamental frequency.

In order to obtain a narrow-band comb filter, the implementation of Fig. 3(a) was modified by placing the 1/f\* time delay in a feedback loop [Fig. 3(b)]. The transfer function of the resulting filter is derived in the Appendix and shown in Fig. 1(c). Profiles of ALSM against f\* for this narrow comb filtering scheme are shown in Fig. 2(c). There are peaks at essentially the same low-frequency harmonics as for the narrow-band scheme of Fig. 2(a), so that this scheme would be useful in fundamental-frequency estimation. However, small peaks at nonharmonic frequencies are occasionally introduced (e.g., at 0.19 kHz for /a/ and at 0.33 kHz for /ae/). This scheme could also be used in formant-frequency esti-
mation since there are prominent peaks at the positions of high-frequency formants and at harmonics close to low-frequency formants.

The local filtering schemes remain effective when the vowels are at a 15 dB lower level (60 dB SPL), although nonformant peaks in the low-CF region for open vowels, and between F1 and F2 for spread vowels tend to be more prominent.

B. Dominant component scheme

The processing scheme that can be proposed as an alternative to local filtering schemes is based on the observation that, for the majority of auditory-nerve fibers, the largest components in the spectra of period histograms are close to the formant frequencies of the vowel stimuli (Young and Sachs, 1979; Delgutte and Kiang, 1984a). This is shown clearly in Figs. 6 and 8 of Delgutte and Kiang (1984a), where the major zones of response activity are located along the horizontal lines at the formant frequencies. In order to obtain spectral representations of period histograms that would be more closely related to psychophysical frequency resolution than discrete Fourier transforms, the histograms were analyzed by a bank of Gaussian bandpass filters whose center frequencies were spaced on a logarithmic scale, and whose bandwidths (1/6 oct) increased with frequency. This frequency analysis of period histograms was carried out in the same manner as for the narrow-band local filtering scheme, except that each histogram was analyzed by the entire filter bank instead of only by the filters with center frequencies close to the fiber CF. A dominant component is defined as the largest component over a certain frequency range in the 1/6 octave spectral representation of the period histogram. In order to separate the formant frequencies from the fundamental, it is useful to define a low-frequency dominant component and a high-frequency dominant component, the cutoff being at 0.2 kHz. In addition, to eliminate cases in which all spectral components would have nearly equal magnitudes, the amplitude of a dominant component has to exceed 8% of the mean square discharge rate.

The circles and crosses in Fig. 4, respectively, show the low-frequency and high-frequency dominant components plotted against fiber CF for the /i/, /æ/, /u/, and /a/ stimuli presented at 75 dB SPL. Predictably, for all vowels, the low-frequency dominant component is always equal to the fundamental, which is the only stimulus component between 0.1 and 0.2 kHz. The important point is that it exceeds detection threshold for many fibers, particularly in the high-CF region for all vowels, and below 0.5 kHz for /æ/. For all vowels except /i/, the high-frequency dominant component is close to one of the formant frequencies over nearly the entire range of CFs, and jumps from F1 and F2 without passing through intermediate frequencies. More precisely, for formants below 1 kHz, it is close to one of the harmonics of the 125 Hz fundamental frequency flanking a formant frequency. For the /i/ stimulus, the high-frequency dominant component is equal to F1 or 2F1 for many fibers, but it coincides with F2 only over a narrow strip, and is near CF for many fibers. With the exception of the second formant of

\[
\text{FIG. 4. Low-frequency and high-frequency dominant components plotted against the characteristic frequency of individual fibers for the /i/, /æ/, /u/, and /a/ stimuli presented at 75 dB SPL. To compute the dominant components, the period histogram for each fiber was analyzed by a bank of 1/6 oct Gaussian bandpass filters whose center frequencies were swept from 0.1 to 4 kHz in steps of 1/12 oct. For each filter, the analysis was implemented as described in the Appendix for the narrow bandpass filtering scheme (Eqs. (A2) and (A3)). Open circles and crosses show the largest component in the 1/6 oct spectral representation of the period histogram for frequencies above and below 0.2 kHz, respectively. A symbol is plotted only if the amplitude of the corresponding component exceeds 8% of the mean square discharge rate. Dashed vertical lines mark the places of the formant frequencies along the CF dimension, and dashed horizontal lines show the dominant components corresponding to the formant frequencies and fundamental frequency. The oblique dashed line is the locus of points for which the dominant component equals CF.}
\]

/i/, the formant frequencies (or the harmonics of the fundamental closest to the formant frequencies) could in principle be estimated by constructing a histogram of the distribution of the high-frequency dominant component for the entire range of CFs, and picking the two largest modes in the histogram. This procedure is similar in some respects to the one proposed by Carlson et al. (1975) and Carlson and Granström (1982) for the processing of vowels. The dominant-component scheme remains effective at 60 dB SPL, although the CF region where F1 dominates is somewhat narrower and the contribution of fibers with CFs above 5 kHz becomes minimal.

III. DISCUSSION

A. Effectiveness of the processing schemes

The effectiveness of the local filtering schemes in extracting the fundamental frequency and formant frequencies of vowel stimuli depends on the choice of the analyzing filters. Narrow-band or narrow-comb filtering schemes are good at fundamental frequency estimation, and are also useful for extracting formant frequencies. In contrast, broadband and broad-comb filtering schemes are useful only for formant frequency estimation, and the broadband scheme does poorly at detecting the second formant of spread vow-
els. The broad-comb scheme seems more accurate than the narrow-band and narrow-comb schemes for estimating low-frequency formants because the peaks at the harmonics flanking the formant frequency that are found for narrow filters are merged into one peak centered at the formant frequency. However, it might be more parsimonious to use the same set of narrow filters for estimating fundamental frequency and formant frequencies, and follow this analysis by a stage of spectral smoothing to improve the estimates of low-frequency formants.

The importance of filter bandwidths for the local filtering schemes is confirmed by a recent study of consonant–vowel syllables (Sachs et al., 1982). An ALSR measure was computed from short segments of the PST histograms, using analyzing bandwidths of about 50 Hz for all frequencies. The ALSR profiles showed peaks at harmonics of the fundamental frequency at least up to 3 kHz, while such peaks were found only up to 1 kHz in the present narrow-band schemes.

Among the local filtering schemes, those based on comb filters were implemented with only delays, multiplications, additions, and low-pass filters (Fig. 3). In Licklider’s (1951) view, these operations could be realized by known physiological mechanisms: Delays could correspond to chains of neurons, addition to spatial summation at a synapse, low-pass filtering to temporal summation, and, for low probabilities of discharge, multiplication could be approximated by a coincidence detector. In this manner, timing cues might be converted to place cues at an early stage of neural processing. A possible problem might be preserving precise synchrony of discharges for the 0.3 to 5 ms delays that would be required for estimating the formant frequencies of speech.

A possible alternative to the local filtering schemes is the dominant component scheme based on a 1/6 oct spectral representation of period histograms. This scheme would successfully estimate the formant frequencies of the vowel stimuli (except for the second formant of /i/). It might not be so effective for the vowels used by Young and Sachs (1979), in which the second formant amplitude was somewhat lower than in our stimuli, because the F1 response components were usually larger than the F2 components at high stimulus levels in those data. On the other hand, the increase in the bandwidths of the analyzing filters with center frequency would tend to emphasize the higher frequencies in the 1/6 oct representation of PST histograms. In general, the optimization of the dominant-component scheme might require a fixed weighting of the frequency components. The fundamental frequency of the vowel stimuli could also be estimated from the distribution of the dominant component in very low frequencies (<0.2 kHz), although this cutoff frequency would presumably have to be adjusted for each speaker. Alternatively, because low-frequency harmonics of the fundamental frequency are resolved in the 1/6 oct representation of period histograms, the fundamental frequency might be estimated from the pattern of most frequent low-frequency components in the ensemble of fibers.

Both the dominant-component scheme and certain local filtering schemes have difficulties in estimating the 3.2 kHz second formant of /i/. The F2 of this two-formant vowel is actually located in the third formant region for natural vowels (Carlson et al., 1975). Although the frequencies of the first two formants are the most important for speech perception, the third and possibly higher formants also contribute, particularly for front vowels (Carlson et al., 1975), and for place-of-articulation distinctions among glide, nasal, and stop consonants (Cooper et al., 1952). In the ALSR scheme of Young and Sachs (1979), the third formant of steady-state vowels was poorly represented at high stimulus levels, although its prominence improved somewhat when the resonance of the human ear canal was taken into account. The difficulties of the processing schemes in estimating the third formant frequency may be due in part to the decrease in synchronization of discharges with frequency (Rose et al., 1967; Johnson, 1980).

The proposed processing schemes have been tested for steady-state vowels. The main frequency components of sonorant sounds (vowels, diphthongs, glides, and nasals) are distributed in the same regions of the spectrum as the steady-state stimuli used in this paper, but their formant frequencies and fundamental frequency change over time. The processing schemes are in principle compatible with the short-time analyses that are commonly used for speech (Flanagan, 1972; Rabiner and Schafer, 1978). In particular, the bandwidths of the analyzing filters do not require longer time windows than the typically used values of 10–20 ms. Available results suggest that the features of the response patterns that are essential for the effectiveness of the processing schemes are also present for speechlike sounds with changing formant frequencies (Sachs et al., 1982, Sinex and Geisler, 1983). Specifically, a scheme resembling the narrow-band ALSM was shown to be effective for estimating the formant frequencies and fundamental frequency during the formant transitions of /da/ syllable (Sachs et al., 1982), while the largest components in the discrete Fourier transform of segments of PST histograms generally coincided with formant frequencies of /ba/ syllable, so long as the formants were clearly apparent in the stimulus spectrum (Sinex and Geisler, 1983).

These processing schemes are only some of the possible ways of extracting phonetically important information from auditory-nerve responses to vowel stimuli. For instance, being based on the power spectra or autocorrelation functions of period histograms, they systematically ignore the response phase. Another restriction is that no use was made of average discharge rates because the response spectra and autocorrelation functions were normalized. In the local filtering schemes of Young and Sachs (1979) and Sachs and Young (1980), average rates enter as multiplicative factors. This should have little effect at high stimulus levels because most fibers would be discharging at high rates in the CF range of interest (Sachs and Young, 1979). At lower levels, use of average rates might improve the behavior of the local filtering schemes because the rate profiles show maxima near the places of the formant frequencies. Of course, there is no proof that the fine timing cues on which the proposed processing schemes are based are actually used by the central processor, and one could conceive of schemes based solely on average discharge rates (Delgutte, 1982; Delgutte and Kiang, 1984a).
B. Comparison with psychophysical models

The local filtering schemes can be related to the concept of "excitation pattern," or "central spectrum" that has been proposed to organize a wide variety of psychophysical results (Fletcher, 1953; Zwicker and Feltkeller, 1967; Scharf, 1970; Plomp, 1976; Goldstein, 1978). A central spectrum is thought to be produced by running spectral analysis of acoustic stimuli in the peripheral stages of the auditory system. Estimates of the frequency selectivity of this analysis are usually between 1/6 and 1/3 oct for the 2 kHz region in the human (Zwicker et al., 1957; Chistovich, 1957; Plomp, 1964; Houtsmul and Goldstein, 1972; Houtgast, 1974; Zwicker, 1974; Patterson, 1976; Moore, 1978; Vogten, 1978). According to one view, the perception of the pitch of complex tones and vowel quality would essentially be pattern recognition operations on the central spectrum (Plomp, 1976; Goldstein, 1978; Srulowicz and Goldstein, 1983). Because auditory-nerve fiber responses are frequency selective, the central spectrum could in principle be obtained from the profile of average discharge rates against CF. The results of Sachs and Young (1979) suggest that it is not readily apparent how the formant frequencies and fundamental frequency of vowel stimuli could be estimated from rate profiles at high stimulus levels, at least for the most sensitive auditory-nerve fibers. The present results suggest that this information could be extracted from ALSM profiles provided that the bandwidths of the analyzing filters are sufficiently narrow. According to this interpretation (Srulowicz and Goldstein, 1983; Young and Sachs, 1979; Sachs and Young, 1980), the central spectrum would be produced by a series of two filter banks, a peripheral filter bank located in the cochlea, and a hypothetical central filter bank operating on the discharge patterns of auditory-nerve fibers. For each channel, the two filters have the same center frequency, which is the characteristic frequency of auditory-nerve fibers. In our results, harmonics of the fundamental frequency up to about the eighth are often associated with peaks in the ALSM profiles for local filtering schemes that have bandwidths corresponding to psychophysical measures of frequency selectivity. In contrast, when the filter bandwidths are increased above the values suggested by psychophysics, some of these low-frequency harmonics are no longer resolved in ALSM profiles. In the ALSM profiles of Sachs et al. (1982), which were obtained with considerably narrower bandwidths than those of psychophysical measures, resolved harmonics were found up to higher frequencies than suggested by psychophysical data. These results imply that psychophysical frequency selectivity would be primarily determined by the hypothetical central filters rather than by the peripheral filters. This is consistent with ideas expressed by Kiang and Moscov (1974) on the nature of critical bands.

While the local filtering schemes are related to "place" models of auditory processing, the dominant component scheme resembles certain "periodicity" models (Moore, 1982). In this scheme, dominant periodicities in the response patterns of auditory nerve fibers are detected by Fourier analysis with a limited frequency resolution. Formant frequencies and fundamental frequency are estimated from the most frequent dominant periodicities in the ensemble of auditory-nerve fibers. In this scheme as in the local filtering scheme, the peripheral auditory processing is followed by a hypothetical central Fourier analysis. If the bandwidths of the central analyzing filters are increased beyond 1/6 oct, the identities of the dominant components are not in general greatly affected, but the ability to resolve closely spaced response components is decreased (not shown). Thus, the dominant component schemes would imply that the psychoacoustic model of masking and frequency resolution could be based on different mechanisms.

The local filtering and the dominant component schemes differ considerably in the way that they process auditory information. The local filtering schemes require only one filtering element in each band of CFs, but the center frequency of the filter has to correspond precisely with the fiber CF. In contrast, the dominant component schemes initially require a full spectral analysis in every CF band, but they do not necessitate a precise tonotopic organization beyond this stage of analysis. Consequently, these schemes should perform better than the local filtering schemes in cases when the normal tonotopic organization of central auditory pathways is disrupted. Another difference is that the high-CF fibers contribute to the estimation of the formant frequencies in the dominant-component schemes, while they play little role in the local filtering schemes because the response components at their CFs are weak. Thus the dominant component schemes should be the most effective when there is considerable loss of low-CF fibers. On the other hand, their effectiveness should be reduced at low stimulus levels or in high-pass background noise, when the high-CF fibers no longer respond to the vowel stimuli.

In summary, several processing schemes have been shown to be effective in estimating formant frequencies and fundamental frequency from responses of auditory-nerve fibers to a set of stimuli resembling the cardinal vowels. Among these schemes, the narrow-band and narrow-comb local filtering schemes seem to be the most attractive because they are useful for both fundamental frequency and formant frequencies, can be related to the psychophysical concept of central spectrum, and require only one filtering element in each band of CFs. Clearly, these schemes are not the only possible ones, and further evaluation of different processing schemes is needed. One evaluation procedure consists of comparing the effectiveness of a processing scheme under conditions of signal degradation with the performance of human listeners. Schemes that lose their effectiveness in conditions for which subjects' behavior is essentially unaffected, or schemes that remain effective when subjects' performance is poor would not be good candidates as models of the central processor, at least for those conditions.

ACKNOWLEDGMENTS

This paper is based in part on a doctoral dissertation submitted to the Department of Electrical Engineering and Computer Science at M.I.T. I thank N. Y. S. Kiang and K. N. Stevens for guidance, encouragement and support at all stages of the work, and P. M. McGaffigan for drawing the figures. Some of the ideas presented here originated from discussions with H. S. Colburn and S. Senneff. This work was


Bertrand Delgutte: Speech coding in the auditory nerve. II
supported by N.L.H. Grants NS 13126 and NS04332, and by a C. J. Lebel fellowship.

APPENDIX: REALIZATION OF THE LOCAL FILTERING SCHEMES

In the local filtering schemes, an average localized synchronized measure (ALSM) is computed as a function of the center frequency of a filter in a bank. For each value of the filter center frequency \( f_c \), the ALSM was computed by (1) analyzing the period histogram of each fiber whose CF was in a narrow frequency band centered at \( f_c \), and (2) averaging the results of the filtering operations for all fibers in that band. Specifically, if \( \text{LSM}_i(f_c) \) denotes the result of the filtering operation for fiber \( i \) with characteristic frequency \( CF_i \), the value of the ALSM at \( f_c \) was obtained by averaging the \( \text{LSM}_i(f_c) \) using a trapezoidal weighting function \( W_{f_c}(CF) \) centered at \( CF = f_c \), with a central width of 1/6 oct and a total width of 1/2 oct:

\[
\text{ALSM}(f_c) = \frac{\sum \text{LSM}_i(f_c) W_{f_c}(CF)}{\sum W_{f_c}(CF)}. \quad (A1)
\]

The value of \( f_c \) was varied from 0.1 to 5 kHz in steps of 1/12 oct. To complete the description of the local filtering schemes, one needs to specify how the \( \text{LSM}_i(f_c) \) were obtained for each one of the four filtering schemes in Fig. 2.

A. Bandpass schemes

For both the narrow-band and the broadband filtering schemes, \( \text{LSM}_i(f_c) \) was the mean power at the filter output normalized by the mean square discharge rate. The mean output power was evaluated by integrating over frequency the product of the magnitude squared of the filter transfer function \( H_{f_c}(f) \) by the power spectrum of the period histogram \( P_i(f) \). Because the power spectrum only contains energy at harmonics \( f_k \) of the vowel fundamental frequency, this integration reduces to a sum over harmonic number \( k \):

\[
\text{LSM}_i(f_c) = \sum_{k=0}^{\infty} P_i(kf_c) H_{f_c}(kf_c) / \sum_{k=0}^{\infty} P_i(kf_c). \quad (A2)
\]

The denominator corresponds to the normalization by the mean square discharge rate. For both bandpass schemes, the filter transfer function was Gaussian:

\[
H_{f_c}(f) = \exp \left[ -\left( \frac{f-f_c}{b_c} \right)^2 \right]. \quad (A3)
\]

The bandwidth \( b_c \) was set to 0.116 \( f_c \) (1/6 oct) for the narrow-band scheme, and to 0.47 \( f_c \) (2/3 oct) for the broadband scheme. The corresponding transfer functions are shown in Fig. 1(a) and (b), respectively. In the ALSR scheme of Young and Sachs (1979), the normalization by the mean square discharge rate was omitted, and the filter transfer function was in effect set to 1 if \( f = f_c \), to 0 elsewhere, so that \( \text{LSM}_i(f_c) \) was simply the magnitude of the Fourier component of the period histogram at frequency \( f_c \).

B. Broad-comb scheme

The broad-comb filter was realized as shown in the block diagram of Fig. 3(a). The output \( A_i(t, f_c) \) is the short-time autocorrelation function of the period histogram \( r_i(t) \) evaluated for a time delay of 1/\( f_c \). Specifically, the low-pass filter corresponded to an integration over one period \( T = 8 \) ms of the vowel stimuli:

\[
A_i(t, f_c) = \frac{1}{T} \int_0^T r_i(t-u)r_i(t-u-1/f_c)du, \quad (A4)
\]

where \( u \) is a dummy time variable. Because the histogram also has period \( T \), the result of Eq. (A4) is the autocorrelation function of the period histogram evaluated for a delay of 1/\( f_c \), and does not depend on absolute time \( t \). In practice, the integration over time in Eq. (A4) was approximated by a summation over 160 bias of the period histogram. Using the Fourier transform relation between the autocorrelation function \( C_i(u) \) and the power spectrum \( P_i(f) \), one obtains:

\[
C_i(1/f_c) = \sum_{k=0}^{\infty} P_i(kf_c) \cos(2\pi f_c kf_c). \quad (A5)
\]

Except for the truncation of the Fourier series, this expression is identical to the numerator of Eq. (A2) for \( H_{f_c}(f) = \cos(2\pi f / f_c) \). This cosinusoidal function is shown in Fig. 1(d). The final result \( \text{LSM}_i(1/f_c) \) was obtained by normalizing \( C_i(1/f_c) \) by the mean square discharge rate \( C_i(0) \).

C. Narrow-comb scheme

The narrow-comb filtering scheme was realized according to the block diagram of Fig. 3(b). The output \( A_i(t, f_c) \) is the short-time cross correlation between the period histogram \( r_i(t) \) for fiber \( i \) and the intermediate variable \( y(t) \), which is obtained by sending \( r_i(t) \) through a delayed-feedback mechanism:

\[
y(t) - \sum_{m=0}^{\infty} (a_m) r_i(t-m/f_c). \quad \text{(A6)}
\]

As previously, multiplying by \( r_i(t) \) and integrating over one period of the vowel stimulus, the output signal becomes independent of time \( t \):

\[
A_i(t, f_c) = \sum_{m=0}^{\infty} (a_m) C_i(m/f_c), \quad \text{(A7)}
\]

where \( C_i(u) \) is the autocorrelation function of the period histogram. Again, using the Fourier transform relation between power spectra and autocorrelation functions, this expression can be made to resemble the numerator of Eq. (A2):

\[
A_i(t, f_c) = \sum_{k=0}^{\infty} P_i(kf_c) \sum_{m=0}^{\infty} (a_m) \cos(2\pi mf_c kf_c). \quad \text{(A8)}
\]

The function corresponding to the weighted sum of cosines in Eq. (A8) is shown in Fig. 1(e). The feedback gain \( a_m \), which controls the bandwidth of the comb filter, was set to 0.7 to obtain 1/6 oct filters. Again, the final result \( \text{LSM}_i(1/f_c) \) was obtained by normalizing \( A_i(t, f_c) \) by the mean square discharge rate.

1The mean square discharge rate is a relevant normalizing factor because the result of the filtering operation represents a mean power with units of (spike/s)^2, instead of an amplitude (in spike/s) in Young and Sachs' (1979) scheme. Because the power spectrum of the period histogram has about the same frequency dependence as the magnitude spectrum of the interval histogram (Johnson, 1978), this should make the result more similar to the "average localized interval rate," a measure based on interval histograms.
which was later proposed by Sachs and Young (1989).

2 Similar results are obtained if the width of the CF band over which filter outputs are averaged is increased to 1 oct.

3 Similar results are obtained with 1/6 oct rectangular filters.

* Another scheme, in which the output of the feedback loop $y(t, f)$ is squared and low-pass filtered, gives very similar results.


Zwicker, E., and Feldtkeller, R. (1967). Das Ohr als Nachrichtenempfänger [Hirzel, Stuttgart].