Auditory Neural Processing of Speech

Bertrand Delgutte

Eaton-Peabody Laboratory, Massachusetts Eye and Ear Infirmary,
Research Laboratory of Electronics, Massachusetts Institute of Technology
Harvard-MIT Division of Health Sciences and Technology,
Department of Otology and Laryngology, Harvard Medical School

Chapter in **The Handbook of Phonetic Sciences,**
edited by W.J. Hardcastle and J. Laver
Oxford: Blackwell (1997), pp. 507-538.

Telephone: (617)573-3876
Fax: (617)720-4408
Email: bard@epl.meei.harvard.edu

October 17, 2002

Abstract

Recordings from single auditory neurons in response to speech and speech-like stimuli provide detailed descriptions of neural processing in the earliest stages of speech perception. These studies show that a great deal of information about both the formant pattern and voice pitch is available in both the average discharge rates and the temporal discharge patterns of auditory-nerve fibers and cochlear nucleus cells. Moreover, many auditory neurons show prominent responses to rapid changes in amplitude and spectrum that are phonetically important. Evidence for neural correlates of speech perceptual phenomena such as categorical perception and trading relations is emerging. Overall, phonetic features are prominently and robustly encoded in neural responses, suggesting that the auditory system shows predispositions for the particular set of acoustic features used for phonetic contrasts.

1. Introduction

In speech communication, the listener identifies the phonetic structure of the utterance produced by the speaker on the basis of the waveform of the acoustic signal. This complex decoding act involves both general auditory mechanisms as well as specialized language processing mechanisms. This chapter outlines what is known about the earliest stages of speech perception based on recordings from single auditory neurons in response to speech and speech-like stimuli. Such single-unit data make it possible to trace neural representations of the speech signal through subsequent stages of auditory processing. This survey updates and extends previous reviews of the neural processing of speech (Sachs, 1984; Sachs et al., 1988; Delgutte, 1982; Smoorenburg, 1987; Greenberg, 1988). It incorporates recent results on speech processing by central auditory neurons, and emphasizes the neural representation of dynamic features of speech.

Pioneered in the early 1970's (Kiang and Moxon, 1974; Kiang, 1975; Hashimoto et al., 1975), studies of speech coding in the auditory nerve have considerably matured, and a great deal of information is now available on responses to most phonetic categories, including vowels (Sachs and Young, 1979; Young and Sachs, 1979; Delgutte and Kiang, 1984a), stops (Miller and Sachs, 1983; Sinex and Geisler, 1983; Carney and Geisler, 1986), nasals (Deng and Geisler, 1987), and fricatives (Delgutte and Kiang, 1984b). These physiological studies have motivated the development of peripheral auditory models for speech processing, many examples of which can be found in recent volumes (Carlson and Granström, 1982; Schouten, 1987, 1992; Ainsworth, 1992; Cooke et al., 1993). The last decade has witnessed the first detailed studies of the encoding of speech at the next stage of auditory processing, the cochlear nucleus (Palmer et al., 1986; Kim et al., 1986; Blackburn and Sachs, 1990; Winter and Palmer, 1990). Some information is also available on the responses of midbrain and cortical auditory neurons (Watanabe and Sakai, 1973, 1978; Palmer et al., 1990; Steinschneider et al., 1994, 1995). This survey necessarily reflects these limitations in our knowledge: it emphasizes speech coding in the auditory nerve and cochlear nucleus, even though the most important processing in speech perception may well occur at more central stages.

Because speech has many properties in common with other acoustic stimuli, the neural processing of speech can be studied using both speech stimuli and nonspeech analogs. For example, voiced speech shares periodic excitation with music and animal vocalizations. Speech, music and sounds produced by animals in motion all show pronounced amplitude modulations at very low frequencies (2-16 Hz). In many cases, very simplified stimuli can be understood as speech. For example, sine-wave speech (a sum of frequency-modulated sine waves that match the first 2-3 formants of an utterance) is intelligible under certain circumstances (Remez et al.,

1981). Sine-wave speech has been valuable for understanding the encoding of vowels in the auditory nerve (e.g. Reale and Geisler, 1980; Sachs and Young, 1980). The frequency-modulated (FM) sounds used by bats for echolocation resemble the formant transitions of speech (Suga, 1992). FM sounds have long been used in studies of central auditory neurons (Suga, 1964; Whitfield and Evans, 1965). Among these nonspeech analogs, species-specific vocalizations are particularly germane because there may exist general neural mechanisms for the perception of conspecific communication sounds, including human speech (Ehret, 1992).

While these analogies are useful, speech differs in many important respects from nonspeech analogs. For example, while sine-wave speech may be intelligible, it is not spontaneously identified as speech, in part because it lacks the periodic amplitude modulation of voiced sounds. While bat FM sounds may resemble formant transitions on spectrographic displays, FM rates for bat echolocation sounds often exceed 10 MHz/sec, which is three orders of magnitude greater than rates for typical formant transitions. Moreover, formant transitions are not true FM sounds in that the actual component frequencies do not vary, only the peak of the spectral envelope does. Thus, nonspeech analogs should be closely scrutinized for their similarities and differences with speech when using these analogs as models for investigating the neural processing of speech.

While no single acoustic property is unique to speech, speech is characterized by particular combinations of acoustic properties occurring in specific frequency ranges that set it apart from other acoustic stimuli (Stevens, 1980). First, speech shows an alternation between relatively intense segments corresponding to vowels and weaker segments corresponding to consonants. This more or less regular amplitude modulation occurring at a 3-4 Hz rate is essential for speech understanding (Houtgast and Steeneken, 1973). Second, the spectral envelope of speech shows pronounced maxima (corresponding to formants) interleaved with minima at intervals of 1000-1200 Hz. Third, speech shows both nearly periodic segments corresponding to sonorants such as vowels, and irregular (noise-like) segments corresponding to obstruents such as fricative consonants and bursts of stop consonants. Thus, speech is characterized by a triple alternation in amplitude envelope, spectral envelope, and fine spectral structure. In order to get a realistic picture of the neural encoding of speech, it is necessary to use stimuli possessing all of these characteristics.

A methodological issue in studying the neural substrates of speech perception is the degree to which the mechanisms involved are specific to humans. Perceptual studies generally use human subjects, while single-unit studies must use animal models (usually cats or small rodents, more rarely monkeys). There is a long-standing debate between proponents of specialized speech perception mechanisms and those who favor general auditory mechanisms (Liberman et al., 1967; Stevens and Blumstein, 1978; Kuhl and Miller, 1978; Delgutte, 1982; Bregman, 1990;

Liberman and Mattingly, 1989; Miller and Jusczyk, 1990; Kluender, 1994). The evidence is broadly consistent with the view that speech perception *at the phonetic level* requires no more than an adaptation of general auditory mechanisms to a particular class of stimuli rather than specialized neural mechanisms. Cats (Dewson, 1964), chinchillas (Kuhl and Miller, 1978; Kuhl, 1981), monkeys (Sinnott et al., 1976; Kuhl and Padden, 1982), and certain birds (Kluender et al., 1987; Dooling et al., 1989) can discriminate speech sounds in ways that can be strikingly similar to humans. When processing at the more peripheral stages of the auditory nervous system is considered more specifically, the general layout of the cochlea and brainstem auditory pathways is largely similar in most species of mammals including humans (Moore, 1987). Many physiological properties of auditory neurons are similar in unrelated species such as rodents, bats, and cats, suggesting that they may represent very general neural mechanisms common to most mammals including humans. Thus, it is likely that a great deal can be learned about the neural processing of human speech by studying responses of single units in non-human mammals. At the very least, physiological studies using animal models are relevant to the issue of whether speech perception at the phonetic level requires specialized neural mechanisms.

Speech is made up of a wide variety of phonetic elements (speech sounds and features) that differ in both temporal and spectral characteristics. Such diversity is essential for speech to function effectively as a communication signal. Distinctions among major phonetic categories such as vowels, stops, nasals and fricatives are based primarily on dynamic features such as temporal envelope characteristics (onsets and offsets, silent intervals, durations) and changes in the gross distribution of spectral energy. Distinctions among vowels and place-of-articulation distinctions among consonants depend on more detailed spectral features such as the formant pattern, and how this pattern changes with time (formant transitions). Voicing and pitch information depend on waveform periodicity or, equivalently, harmonicity of the spectral pattern. This chapter follows this tripartite classification of the acoustic characteristics of speech: Section 2 discusses the neural encoding of dynamic features, Section 3 the spectral pattern, and Section 4 pitch and voicing. Section 5 describes specific attempts to identify neural mechanisms underlying categorical perception, context dependence and trading relations. In each section, the responses of auditory-nerve fibers are described first, followed by those of central auditory neurons.

2. Coding of rapid changes in amplitude and spectrum

Much information in speech is available in rapid changes in amplitude and spectral characteristics that are apparent in spectrographic displays (Fant, 1973). The temporal relationships among these discrete events, as well as their spectral distribution provide phonetic information. For example, stop consonants are characterized by two events occurring in rapid

succession: a rapid increase in amplitude primarily in the high frequency region corresponding to consonantal release (burst onset), and an increase in amplitude in the low frequency region corresponding to the onset of voicing. The time interval between these two events is the voice onset time (VOT), which is an important cue for voicing distinctions among stop consonants (Lisker and Abramson, 1964). Not only the spectro-temporal distribution of these events, but also their amplitude envelope is phonetically important. For example, affricate consonants (as in "chop") have abrupt increases in amplitude of the frication noise, while fricative consonants (as in "shop") have a more gradual onset. More generally, rapid changes in amplitude and spectrum point to time intervals that are rich in information about the identity of the neighboring phonetic segments, including vowels and consonants (Delattre et al., 1955; Blumstein and Stevens, 1980; Stevens, 1980). These events are particularly important in continuous speech, where they may help segment utterances into perceptually-manageable chunks.

2.1 Representation in the auditory nerve. Adaptation.

Rapid changes in amplitude and spectrum are prominently represented in the discharge patterns of auditory-nerve fibers (ANFs). Figure 1A shows the activity of the auditory nerve in response to a speech utterance whose spectrogram is shown in Fig. 1B. In this *neurogram* display, each trace shows the average response of a small number of ANFs displayed as a *post-stimulus time histogram* (PSTH). PSTHs represent the rate of discharge averaged over short intervals, or "bins" as a function of time following stimulus onset. The bin width in Fig. 1A is 1 msec. Fibers are arranged according to their *characteristic frequency (CF),* the frequency to which they are most sensitive. For pure-tone stimuli, each fiber responds to a limited range of frequencies at a given sound level (Kiang et al., 1965). This frequency selectivity is due to the mechanical tuning of the basilar membrane and hair cells in the cochlea. There is a precise mapping between the CF of an ANF and its place of innervation along the cochlea (Liberman, 1982b). This mapping between CF and spatial position is replicated at every major station in the auditory pathway up to the auditory cortex (Irvine, 1986; Brugge and Reale, 1985). Thus, the CF dimension is a fundamental organizing principle of the auditory nervous system that must be taken into account in any model for the neural processing of acoustic stimuli.

The rapid spectral and amplitude changes pointed to by arrows in the spectrogram of Fig. 1B are also apparent in the neurogram of Fig. 1A. Low-CF fibers show an abrupt increase in discharge rate, associated with a prominent peak whenever the spectrogram shows an abrupt increase in energy in the low-frequency region (filled arrows in Fig. 1). These events occur at the transitions from obstruent to sonorant segments, and at the onset of voicing for stop consonants. On the other hand, high-CF fibers show a rise/peak in discharge rate when the spectrogram shows a rapid increase in energy in the high-frequency region (open arrows). This

occurs at the transitions from sonorant to obstruent segments, and at the onset of the release burst for stop consonants. Thus, the spatio-temporal pattern of auditory-nerve activity contains pointers to regions of rapid changes that contain important phonetic information.

These neural pointers differ from those seen in the spectrogram in that the rapid rise is often followed by a prominent peak and then a gradual decay in instantaneous discharge rate. Such peaks in discharge rate are seen for any stimuli that has an abrupt onset such as a tone burst. The decay in discharge rate following an abrupt onset is called *adaptation*, and may be caused in part by the depletion of neurotransmitter at the synapses between hair cells and ANFs in the cochlea (Smith, 1979). Following an adapting stimulus, the responses to subsequent stimuli are depressed (Smith, 1979; Harris and Dallos, 1979). Adaptation occurs on different time scales, ranging from a few milliseconds to several seconds and even minutes (Smith, 1979; Kiang et al., 1965).

Adaptation plays several roles in the encoding of speech in the auditory nerve (Delgutte, 1980; Delgutte and Kiang, 1984c; Delgutte, 1986). First, peaks in discharge rate resulting from adaptation point to spectro-temporal regions that are rich in phonetic information, as shown in Fig. 1. Second, adaptation increases the temporal precision with which onsets are represented. Third, adaptation enhances spectral contrast between successive speech segments. This enhancement arises because a fiber adapted by stimulus components close to its CF is less responsive to subsequent stimuli that share spectral components with the adapting sound. On the other hand, stimuli with novel spectral components stimulate "fresh", unadapted fibers, thereby producing an enhanced response. A fourth role of adaptation is to encode phonetic contrasts based on characteristics of the amplitude envelope. For example, for a majority of auditory-nerve fibers, the abrupt onset of affricate consonants results in a more prominent adaptation peak than the more gradual onset of fricative consonants having the same spectral characteristics (Delgutte, 1980; Delgutte and Kiang, 1984c).

The roles of adaptation in speech coding can be placed into the broader context of a functional model for the auditory processing of speech proposed by Chistovich et al. (1982). Chistovich et al. hypothesized that two separate systems operate in parallel: a tonic system that continuously delivers a running spectral representation of the stimulus, and a phasic system that detects acoustic transients (onsets and offsets) in individual frequency bands. The phasic system has two functions: by itself, it provides important temporal cues to the identity of phonetic segments, and it also provides pointers for sampling the output of the tonic system at times that are particularly important.

Delgutte (1986) showed how adaptation in auditory-nerve fibers leads to a simple and robust implementation of the phasic system proposed by Chistovich et al. (1982). Using a model of the peripheral auditory system incorporating adaptation, he showed that peaks in discharge

rate such as those visible in Fig. 1 can be reliably detected in the response patterns of model ANFs by means of a template matching technique. For a corpus of French stop-vowel syllables produced by both male and female speakers, peaks in discharge rate were consistently detected at the burst onset for high-CF (> 1 kHz) model fibers, while peaks at voicing onset were detected for low-CF fibers. Thus, the model was able to reliably measure VOT by computing the time interval between these high-CF and low-CF events. This line of research was further developed by Wu et al. (1992) and Schwartz et al. (1982), who introduced many refinements in the model and showed that the phasic system encodes other phonetic events besides VOT.

In summary, adaptation produces prominent features in the response of the auditory nerve to the rapid changes in amplitude and spectrum in speech. Adaptation should not be seen as an epiphenomenon reflecting the inability of neurons to sustain high discharge rates for long periods of time, but as the first stage in a phasic neural system specialized for the processing of acoustic transients.

2.2 Representation in central auditory nuclei

Little is known about the encoding of dynamic features of speech in the central nervous system. The available data on the response of the central auditory neurons to speech stimuli (Watanabe and Sakai, 1973, 1978; Steinschneider et al., 1994, 1995), as well as more extensive data on their responses to dynamic nonspeech stimuli (reviewed by Langner, 1992) suggest that the prominent representation of acoustic transients initiated at the level of the auditory nerve is further enhanced in the central nervous system, consistent with the ideas of Chistovich et al. (1982).

Many central auditory neurons respond primarily at the onset of tone-burst stimuli, giving little or no sustained response (for reviews, see Rhode and Greenberg, 1992; Irvine, 1986; Brugge and Reale, 1985). Such "onset" neurons are found at virtually every major stage of processing in the auditory pathway, beginning with the cochlear nucleus (CN). At first sight, onset neurons might appear to provide appropriate processing for the phasic system postulated by Chistovich et al. (1982). However, onset cells in the CN tend to discharge to every pitch period of vowel stimuli (Kim et al., 1986; Palmer and Winter, 1992). Thus, these cells signal many more events in addition to the major changes in amplitude and spectrum that are apparent in Fig. 1. The response of CN onset cells can be understood if we consider their sensitivity to amplitude-modulated (AM) tones. In response to these stimuli, neural discharges tend to occur at a particular phase within the modulation cycle, a phenomenon known as *phase locking.* For CN onset cells, phase locking is most precise for modulation frequencies in the 100-400 Hz range (Frisina et al., 1990). These best modulation frequencies closely coincide with the range of fundamental frequencies of human voice. These observations suggest that the precise phase-

locking of CN onset cells to the pitch period may result from the pronounced AM that voiced speech shows at the fundamental frequency.

Many cells at more central locations than the CN in the auditory pathway also show preferred sensitivity to a particular range of AM frequencies, but their best modulation frequencies are generally lower than for CN cells. For example, a majority of cells in the inferior colliculus (IC), the principal auditory nucleus in the midbrain, have best modulation frequencies in the 10-300 Hz range (see review by Langner, 1992). Most cortical neurons have best modulation frequencies in the 3-100 Hz range. Cells with best modulation frequencies in the 10-50 Hz range would not be expected to phase lock to the pitch period of speech stimuli. On the other hand, these cells are likely to respond vigorously to the major changes in amplitude and spectrum associated with phonetic events. The multiple-unit recordings of Steinschneider et al. (1994) from the primary auditory cortex of the awake macaque are consistent with this hypothesis. Steinschneider et al. found a population of units that showed a "double onset" pattern in response to [da] and [ta] syllables: These units respond with a transient burst of activity at the consonantal release, and a second burst at the onset of voicing. A similar response pattern has been observed by Watanabe and Sakai (1973) for one IC neuron in response to a [ta] syllable. Thus, cells may exist that directly encode the VOT of stop consonants in their response patterns, consistent with the phasic cells postulated by Chistovich et al. (1982). Such cells are found in the primary auditory cortex, and possibly in the auditory midbrain as well.

In summary, physiological studies of central auditory neurons generally support the Chistovich et al. (1982) notion of a phasic system that encodes rapid changes in amplitude and spectrum. Systematic studies are needed to determine the neural mechanisms leading to these phasic responses, as well as the exact nature of the phonetic events encoded by these cells at different stages in the central auditory pathway. The possibility of higher-order neural circuits that would respond to combinations of onsets occurring at different times in different frequency bands also needs to be investigated. In the next section, we turn to the other component of the Chistovich et al. (1982) model, the tonic system that delivers a running spectral representation.

3. Coding of spectral patterns

Techniques for the physical analysis of speech such as spectrograms or linear prediction provide spectral representations for successive time frames of the waveform. Experience with speech synthesis and recognition suggests that such short-time spectral representations contain sufficient information for speech understanding (e.g. Flanagan, 1972). For certain classes of speech sounds such as vowels or fricative consonants, the spectrum can be approximately constant over many time frames. Such steady-state stimuli provide an appropriate starting point for studies of the neural representation of the spectral patterns of speech. The most important

features for vowel perception are the frequencies of the spectral maxima associated with the first two or three formants (Peterson and Barney, 1952; Carlson et al., 1975).

3.1 Rate-place representation in the auditory nerve

The simplest neural codes for the representation of speech spectra are *rate-place* schemes, which display the amount of neural activity (average discharge rate) as a function of CF. Rate-place schemes, which constitute modern formulations of Helmholz's (1863) place theory of hearing, are based on the frequency selectivity of the cochlea and the tonotopic organization of the auditory nervous system. Because virtually every major station in the auditory pathway from the auditory nerve to the cortex is tonotopically-organized (Irvine, 1986; Brugge and Reale, 1985), rate-place schemes provide very general neural representations of the short-time spectrum.

The ability of rate-place schemes to represent the spectra of steady-state vowels was investigated in a classic paper by Sachs and Young (1979). Sachs and Young recorded the activity of a large number of ANFs in the same cat in response to a vowel stimulus, and analyzed how the average discharge rate varies as a function of fiber CF. Their results, reproduced in Fig. 2 for the vowel [ε], show that, for low stimulus levels, discharge rate is maximum for fibers whose CFs are close to the frequency of one of the first three formants. Thus, for these low levels, rate-place schemes provide a good representation of the formant pattern. As stimulus level increases, the representation of the formants degrades. For higher levels still well within the conversational range, the rate-place pattern takes on a lowpass shape with little or no information about the positions of the formants. Similar degradations in the representation of the formant frequencies occur when moderate-level background noise is introduced, even if such noise does not impair intelligibility (Sachs et al., 1983; Delgutte and Kiang, 1984d).

The results of Sachs and Young (1979) illustrate a very general "dynamic range" problem in auditory neuroscience (Evans, 1981): The dynamic range of single ANFs is much smaller than the wide (> 100 dB) range over which listeners can understand speech and discriminate sound intensities. Specifically, the discharge rates of ANFs only increase over a 20-40 dB range of stimulus levels between threshold and a level where discharge rate reaches a maximum (Kiang et al., 1965; Sachs and Abbas, 1974). In response to vowels, the discharge rates of fibers with CFs close to a formant reach their maximum at moderate sound levels. With further increases in level the rates of fibers with CFs between formants also reach their maximum, so that the dips between formants in the rate-place profiles are eliminated. While the Sachs and Young results pose a serious challenge to rate-place schemes, several factors need to be considered before reaching any conclusion as to the viability of these schemes for speech encoding.

The first factor is that the data of Fig. 2A-C represent only the responses of the subset of auditory-nerve fibers that discharge vigorously in the absence of intentionally-applied acoustic stimuli. These high spontaneous rate (SR) fibers form a majority of ANFs, and always have the lowest thresholds for pure tones at the CF (Liberman, 1978). There is a minority of fibers that have low SRs, and thresholds that can be as much as 40-60 dB higher than those of high-SR fibers with the same CF (Liberman, 1978). Low-SR and high-SR fibers differ not only in threshold but also in many other characteristics such as dynamic range (Schalk and Sachs, 1980; Winter et al., 1990), morphology and position of their synaptic terminals on inner hair cells (Liberman, 1982a), and patterns of projections to the cochlear nucleus (Liberman, 1991). These variations in SR and threshold among ANFs constitute an organizing principle possibly as important as tonotopic organization, which must be taken into account in descriptions of responses to speech stimuli.

Because low-SR fibers have higher thresholds and wider dynamic ranges than high-SR fibers, their discharge rates continue to grow at stimulus levels for which the responses of high-SR fibers have reached their maximum. Figure 2E-F, from Sachs and Young (1979) shows rate-place patterns for low-SR fibers in response to the vowel [ɛ] presented at conversational speech levels. Despite the smaller number of these high-threshold fibers, the rate-place patterns show local maxima at the first two formant frequencies up to the highest level that was investigated (75 dB SPL). Thus, high-threshold fibers provide a rate-place representation of the formant pattern for moderate and high stimulus levels, while low-threshold fibers provide this information for low stimulus levels. Delgutte (1982) proposed a rate-place scheme that gave a good representation of the formants over a broad range of levels by adaptively weighting information from low- and high-threshold fibers depending on stimulus level. Delgutte (1987) further showed that a very similar scheme arises when modeling psychophysical performance in intensity discrimination based on statistical descriptions of auditory-nerve activity. Thus, the same rate-place model that accounts for basic psychophysical tasks such as masking and intensity discrimination can also be applied to speech encoding over a wide dynamic range (Delgutte, 1995). Similar ideas have been expressed by Winslow et al. (1987), who further proposed a neural circuit based on patterns of synaptic connections in the cochlear nucleus that could implement level-dependent adaptive weighting.

Feedback is another factor that needs to be considered in assessing the viability of rate-place schemes for speech encoding. The activity of ANFs is modulated by feedback pathways from the brainstem. The best understood of these feedback systems is the medial olivocochlear (MOC) pathway, which consists of neurons whose cell bodies are located in the superior olivary complex, and whose axons terminate on outer hair cells in the cochlea (Warr, 1992). Stimulation of MOC neurons shifts the dynamic range of ANFs by as much as 15-30 dB towards higher

intensities for stimulus frequencies near the CF (Wiederhold and Kiang, 1970; Guinan and Gifford, 1988). Thus, fibers that would discharge at their maximum rate in the absence of efferent stimulation become capable of encoding stimulus level when MOC neurons are stimulated. These effects are particularly striking for transient signals in continuous background noise, where stimulation of the MOC pathway can exert a strong anti-masking effect (Winslow and Sachs, 1987; Kawase et al., 1993). These physiological experiments describe MOC effects in an open-loop condition. In natural conditions, MOC effects would be induced through reflex action, possibly modulated by central control (Warren and Liberman, 1989). This mode of activation might lead to further signal processing capabilities. For example, MOC neurons innervating cochlear regions in which the signal-to-noise ratio is particularly low might be selectively activated. Although the role of MOC feedback in speech encoding has not been directly investigated, results with tonal stimuli strongly suggest that this role is likely to be important, particularly in the presence of background noise.

To summarize, while rate-place profiles for low-threshold (high-SR) fibers provide a poor representation of the formant frequencies of vowels at conversational speech levels, rate-place coding cannot be discounted as a general scheme for spectral representation because high-threshold fibers and feedback systems are likely to provide the necessary information for high intensities and low signal-to-noise ratios. A major advantage of rate-place schemes is that they are equally effective for obstruent sounds such as fricative consonants as for vowels (Delgutte, 1980; Delgutte and Kiang, 1984b).

3.2 Temporal representation in the auditory nerve

An alternative to rate-place schemes for the encoding of speech spectra are temporal schemes, which can be seen as modern formulations of Wever and Bray's (1930) volley principle. Because ANF discharges are phase-locked to low-frequency (< 5 kHz) pure tones (Rose et al., 1967; Johnson, 1980), intervals between these discharges tend to occur at integer multiples of the stimulus period. Such interspike interval information can in principle be used to derive very precise estimates of the tone frequency (Siebert, 1970; Goldstein and Srulovicz, 1977). Phase locking is not limited to pure tones, but also occurs for complex periodic tones, including steady-state vowels (Young and Sachs, 1979; Reale and Geisler, 1980; Delgutte and Kiang, 1984a). In this case, the temporal patterns of discharge of single fibers contain information about the frequency content of the stimulus.

When discussing temporal schemes for auditory processing, it is important to be specific about time resolution. Any viable scheme must be able to track short-time variations in the spectrum of speech. For example, rate-place schemes typically average the instantaneous rate of discharge over 5-40 msec moving frames similar to those used for speech analysis and synthesis

(Flanagan, 1972). These time frames are consistent with psychophysical estimates of temporal resolution (e.g. Moore et al., 1988). In contrast, detection of phase locking of ANF discharges requires an analysis with a much finer time resolution. For example, in order to clearly demonstrate phase locking to a 5-kHz stimulus, temporal resolution finer than 50 μsec is required. In the following, the term *temporal scheme* will be restricted to models of auditory processing that make use of such fine time information. Thus, both rate-place and temporal schemes make use of information distributed in time, albeit on different scales.

Figure 3 shows the spatio-temporal patterns of discharge of the auditory nerve for the steady-state vowel [ae] presented at 60 dB SPL. This neurogram shows PSTHs arranged by CF as in Fig. 1A, but differs in that it has a much finer time resolution, thereby revealing the phase locking of neural discharges to the stimulus waveform. The duration of the time axis is 20 msec, corresponding to two periods of the stimulus. Figure 3 demonstrates two important points. First, the fine time patterns of discharge depend systematically on CF. Thus, the frequency selectivity of the cochlea, which is the basis for rate-place coding, also plays an important role in shaping temporal patterns of response. The second point is that a majority of fibers convey formant information in their discharge patterns. Specifically, for fibers with CFs between 500 Hz and 1300 Hz, peaks in the response patterns are separated by intervals of approximately 1.2 msec, which is the reciprocal of the first formant frequency F1 = 750 Hz. For CFs between 1300 Hz and 1800 Hz, response patterns show peaks at intervals of 0.7 msec, which is the reciprocal of F2 = 1450 Hz. For higher CFs, more complex response patterns are observed, with some periodicities related to F1, others to F2, and yet others to the fundamental frequency F0. Thus, the temporal response patterns of the majority of ANFs provide information about the formant frequencies and the fundamental frequency.

The results shown in Fig. 3 are typical for vowels at moderate to high stimulus levels (Young and Sachs, 1979; Delgutte and Kiang, 1984a): The general rule is that fibers tend to phase lock to the formant frequency that is closest to their CF. There are however considerable variations in the extent and positions of the CF regions in which ANFs phase lock to a particular formant frequency depending on the formant pattern of each vowel (Delgutte and Kiang, 1984a). For example, for *low* vowels (e.g. [a] or [ae]), which have a high F1, low-CF fibers (< 500 Hz) phase lock to the low-frequency harmonic of F0 closest to the CF rather than to the first formant. No such low-CF region is found for *high* vowels such as [i] and [u]. For *diffuse* vowels such as [i], for which F1 and F2 are widely separated, fibers with CFs between F1 and F2 primarily phase lock to the CF or to F0. This intermediate CF region is lacking in *compact* vowels such as [a]. Thus, there exist correlates of phonetic features in the spatio-temporal patterns of discharge of ANFs.

While Figure 3 demonstrates that fine temporal patterns of discharge contain a great deal of information about the stimulus spectrum, generation of this display requires an independent time reference that precisely indicates the onset of each pitch period. Such a time reference would not be directly available to the central nervous system during speech perception. Figure 4A shows an alternative display based on interspike intervals, which does not require such a time reference. Each trace shows the all-order interspike interval distribution (also known as *autocorrelation histogram*) for one ANF. As in Fig. 3, the fibers are arranged by CF. This display is an instance of the interval-place representation proposed by Licklider (1951) in his duplex theory of hearing. Licklider pointed out that the all-order distribution of neural interspike intervals is formally identical to an autocorrelation function if the neural discharges are modeled as a train of impulses. The autocorrelation function can be implemented using coincidence detectors and delay lines, elements which are known to exist in the central nervous system. If the autocorrelation function is evaluated for different lags using separate neural circuits differing in the length of their delay lines, a scheme for transforming a temporal code into a place code is obtained. Licklider proposed that an array of such neural circuits computes the autocorrelation function of the spike train in every CF region, forming a two-dimensional ("duplex") representation of neural activity as in Fig. 4A along both CF (or cochlear place), and autocorrelation lag (or interspike interval). Much of the information about formant frequencies available in the neurogram of Fig. 3 is also apparent in the interval-place representation of Fig. 4A. Specifically, for fibers with CFs between 500 and 1300 Hz, the first 3 peaks in the interspike interval distribution are approximately at 1/F1 and its multiples, while for CFs between 1300 Hz and 1800 Hz, the first 6 peaks are approximately at 1/F2 and its multiples. Thus, short interspike intervals (< 5 msec) may provide sufficient information for vowel identification (Palmer, 1990; Cariani and Delgutte, 1993). Longer intervals provide information about the fundamental frequency, a point to which we return in Section 4.1.

3.3 Temporal processing schemes

Because temporal patterns of discharge of ANFs provide rich, highly-redundant information about the spectra of sonorants, many different processing schemes have been proposed for extracting this temporal information. The goal of these schemes is to derive a compact representation that contains essential information for intelligibility for a wide range of stimulus conditions. The scheme that has received the most attention in the literature is the *Average Localized Synchronized Rate (ALSR)* proposed by Young and Sachs (1979). The ALSR is closely related to the central spectrum model of Srulovicz and Goldstein (1983). In these models, the temporal pattern of discharge of each ANF is processed by a central (neural) filter whose center frequency matches the fiber CF. The central filter selects the frequency

components of the response that are close to the CF and possibly its harmonics. The time-average output of each central filter is then displayed as a function of CF to form the ALSR or central spectrum. Because the ALSR for a particular CF is based on fine temporal information from ANFs innervating a specific cochlear place, it combines temporal and place information and constitutes a *temporal-place* model of auditory processing (Young and Sachs, 1979). Figure 5B shows the ALSR for the vowel [ae] derived from the neurogram of Fig. 3. The ALSR shows pronounced peaks near the frequencies of the first 2-3 formants. In general, the ALSR provides a good representation of the formant frequencies over a wide range of stimulus levels for many different classes of speech sounds, including steady-state vowels (Young and Sachs, 1979; Delgutte, 1984), whispered vowels (Voigt et al., 1982), vowels in background noise (Sachs at al., 1983; Delgutte and Kiang, 1984d), and formant transitions of stop consonants (Miller and Sachs, 1983; Delgutte and Kiang, 1984c). However, the ALSR poorly encodes the spectra of sounds such as fricative consonants that have intense frequency components above 3 kHz because there is little or no phase-locking at these high frequencies (Delgutte and Kiang, 1984b).

The ALSR is important because it demonstrates that a simple temporal scheme can provide a spectrum-like representation that contains essential information for speech intelligibility, at least in the low-frequency region. On the other hand, the ALSR makes use of only a small fraction of the temporal information available in the auditory nerve, and the particular form of information reduction performed by the ALSR may not be the most physiologically plausible. In particular, there is no physiological evidence for the existence of central filters matched to the CFs of auditory-nerve fibers.

One alternative to temporal-place schemes is the pooled (also known as "ensemble", "summary", or "aggregate") interspike interval distribution, which is obtained by summing interspike interval histograms across the tonotopically-arranged ensemble of ANFs (Ghitza, 1988; Palmer, 1990; Meddis and Hewitt, 1991, 1992; Delgutte and Cariani, 1992). Because this scheme eliminates explicit place information by integrating across all cochlear places, it is an example of a *purely temporal* scheme for spectral representation. Figure 4B shows the pooled interspike interval distribution for the vowel [ae] derived from the interval-place representation of Fig. 4A. The Fourier transform of the pooled distribution (Fig. 5C) shows peaks at the frequencies of the first 2-3 formants, indicating that essential information for vowel identification is available in this representation. In general, purely temporal schemes such as the pooled interval distribution do well for low-frequency sounds such as vowels and voiced formant transitions (Sinex and Geisler, 1983; Delgutte, 1984; Carney and Geisler, 1986; Geisler and Gamble, 1989; Palmer, 1990), but have difficulty for high-frequency sounds such as fricative consonants (Delgutte and Kiang, 1984b). It should also be noted that the vast majority of cells at

the most peripheral stages of the auditory nervous system show sharp frequency tuning, so that there is no evidence for wide-scale spatial integration in the brainstem.

A third alternative to temporal-place and purely temporal schemes is a class of *spatio-temporal coincidence* schemes that rely on differences in spike arrival times for ANFs innervating different places along the cochlea (Loeb et al., 1983; Shamma, 1985, 1988; Deng et al., 1988; Carney, 1994). These schemes differ in the detail of their assumptions about how sensitivity to coincidence is achieved. Although these schemes have not been as thoroughly tested against physiological data as the other two classes of schemes, it appears that some of them provide sufficient information for encoding at least sonorant spectra (Deng et al., 1988; Shamma, 1985). From the point of view of central processing, these schemes assume the existence of cells sensitive to the coincidence of spikes across their spatially-distributed inputs, a notion for which physiological evidence is mounting (Carney, 1990; Rothman et al., 1993; Joris et al., 1994; Palmer et al., 1994). Unlike schemes based on interspike intervals, spatio-temporal coincidence schemes do not require long neural delays, and make use of precise temporal relationships imposed by cochlear mechanics among the discharges of ANFs tuned to different CFs. This class of schemes appears to be a promising avenue for research into central auditory mechanisms for processing acoustic stimuli.

In conclusion, a wide variety of processing schemes have been proposed for the temporal representation of the short-time spectra of speech sounds. For the most part, these schemes are effective for low-frequency stimuli such as vowels and voiced formant transitions, but many of them have difficulty with high-frequency sounds such as fricatives and the bursts of stop consonants. The available physiological data from ANFs do not allow us to rule out any of these schemes, so that the most valid scheme can only be identified by examining how speech sounds are processed by the central nervous system.

3.4 Rate-place and temporal representations in the cochlear nucleus

The functional organization of the central auditory system is considerably more complex than that of the auditory nerve. Whereas the auditory nerve can be considered as a two-dimensional array of fibers organized along CF and sensitivity (threshold), auditory nuclei in the brainstem contain many different types of cells interconnected by a complex pattern of projections (reviewed by Irvine, 1986). This organization is best understood for the cochlear nucleus, which contains at least six major types of cells, with some of these major cell types being further divided into sub-types These cell types are defined by a wide set of properties, including morphology, cytochemistry, intrinsic cell-membrane characteristics, regional distribution within the CN, patterns of synaptic inputs, central projections, and responses to acoustic stimuli (for reviews, see Young, 1984; Cant, 1992; Rhode and Greenberg, 1992).

Because most of these cell types cover the entire range of CFs, the CN effectively provides multiple, parallel spectro-temporal representations of the acoustic stimulus. Elucidating the functions of these parallel representations is a major task for auditory neuroscience.

Investigations of how different cell types in the cochlear nucleus respond to speech sounds (Palmer et al., 1986; Kim et al., 1986; Blackburn and Sachs, 1990; Winter and Palmer, 1990; Palmer and Winter, 1992) have focused on how the spectra of steady-state vowels are encoded. The unit type that has been the most thoroughly studied is the *primary-like* unit, whose response pattern resembles that of ANFs. Primary-like responses are recorded from bushy cells (Rhode et al., 1983; Rouiller and Ryugo, 1984), which receive giant synaptic terminals ("end bulbs") from auditory-nerve fibers. These giant endings provide multiple synapses that ensure very secure synaptic transmission, so that the response patterns of primary-like neurons closely resemble those of their ANF inputs. In particular, primary-like units show precise phase locking to pure tones similar to that of ANFs (Bourk, 1976; Rhode and Smith, 1986), with some cells even showing enhanced phase locking for low frequencies (Joris et al., 1994). Consistent with this precise phase locking, primary-like units provide a good representation of vowel formants in their fine temporal patterns of discharge (Palmer et al., 1986; Winter and Palmer, 1990; Blackburn and Sachs, 1990). On the other hand, rate-place codes for primary-like units suffer from the same dynamic range limitations as they do for low-threshold ANFs (Blackburn and Sachs, 1990). Bushy cells project to nuclei in the superior olive that play a role in the processing of binaural information (Cant, 1992). The precise phase locking of bushy cells is well adapted to this binaural processing function because interaural differences in phase are known to be an important cue for sound localization (Durlach and Colburn, 1978). Of course, the function of bushy cells in binaural circuits does not preclude their playing an additional role in speech processing.

Another CN unit type whose responses to vowels have been studied in some detail is the *chopper* unit, thus called because its response pattern shows pronounced peaks spaced at regular intervals. Chopper responses are recorded from stellate cells, which receive small synaptic terminals from many auditory-nerve fibers (Rhode et al., 1983; Rouiller and Ryugo, 1984). As such, stellate cells are more likely than bushy cells to integrate information across ANFs. Chopper units poorly phase lock to pure tones above 1 kHz (Bourk, 1976; Rhode and Smith, 1986). Consistent with this poor phase locking, the fine temporal discharge patterns of chopper units provide information about the first formant frequency, but not about higher formants whose frequencies exceed 1 kHz (Blackburn and Sachs, 1990). In contrast, chopper units provide a better rate-place representation of vowels for conversational speech levels than do low-threshold ANFs (Blackburn and Sachs, 1990). In particular, the rate-place representation for a sub-class of choppers, the "transient" choppers, is nearly invariant over a 40 dB range of stimulus levels.

Thus, chopper units (particularly transient choppers) encode the spectrum of vowels in rate-place profiles over a wide range of sound levels despite the limited dynamic range of their ANF inputs. One hypothesis for explaining the enhanced dynamic range of choppers is that they might receive an orderly convergence of inputs from ANFs having the same CF, but differing in thresholds (Winslow et al., 1987). Another (not mutually exclusive) conception is that chopper neurons receive inputs from ANFs with different CFs exerting mutually inhibitory influences (Rhode and Greenberg, 1994).

In summary, different cell types in the cochlear nucleus show distinct response patterns to vowel stimuli. Primary-like units provide a precise encoding of formant frequencies in their fine temporal patterns of discharge. Chopper units give a rate-place representation of vowel spectra over a wide range of stimulus levels. Thus, chopper and primary-like units provide complementary information about the short-time spectra of speech stimuli. The methods used for characterizing the responses of chopper and primary-like units to speech stimuli are also applicable to other types of cells in the cochlear nucleus and to cells in more central auditory nuclei. Such detailed characterizations are needed for elucidating the functions of the different cell types and neural circuits in speech processing.

4. Coding of pitch and voicing

The previous section focused on how the spectral envelope associated with the resonant properties of the vocal tract is encoded in the auditory nerve and cochlear nucleus. The fine spectral structure associated with voicing is also important in speech communication. Over short times, voiced sounds have nearly periodic waveforms. This periodicity, and the corresponding harmonicity of the spectral pattern produce a prominent pitch sensation. Variations in pitch throughout an utterance convey information about stress, grammatical structure, and speaker's attitude. Pitch may also be important for communication in the presence of competing sounds because differences in pitch help segregate voices from each other (Darwin, 1992). Pitch sensations produced by complex periodic waveforms are not unique to human speech, but are also important in music and animal vocalizations. Thus, neural mechanisms for the perception of the pitch of voice are likely to be a special instance of a general mechanism found in both humans and non-human animals for the perception of the pitch of complex acoustic stimuli. Such pitch percepts are heard even when the stimulus has no energy at the fundamental frequency (for review, see De Boer, 1976). Such missing-fundamental stimuli are of practical as well as theoretical interest because the fundamental component of speech is often lacking (as in telephone communication) or masked by low-frequency background noise in every day situations.

4.1 Temporal representation of pitch

As for the spectral envelope, the pitch of complex tones might be encoded in either the temporal or the spatial patterns of discharge of auditory neurons. The coding of voice pitch in interspike intervals of ANFs is perhaps the most directly demonstrated. For every fiber in Fig. 4A, the largest peak in the interspike interval distribution occurs at 10 msec, the pitch period of the [ae] vowel, which has a fundamental of 100 Hz. This maximum at 10 msec is even more salient in the pooled interval distribution of Fig. 4B because it is present in all fibers, while formant-related periodicities occur only in particular CF regions. Thus, for this vowel, pitch corresponds to the most frequent interspike interval in the auditory nerve. Delgutte and Cariani (1992) have shown that this result holds not only for periodic stimuli such as steady-state vowels, but also for a wide variety of inharmonic stimuli devised by psychoacousticians to test theories of pitch perception (see also Evans, 1983). These physiological results lend support to models of pitch perception based on interspike intervals (Licklider, 1951; Srulovicz and Goldstein, 1983; Moore, 1990; van Noorden, 1983; Meddis and Hewitt, 1991).

If pitch were coded in interspike intervals of ANFs, a key question is how such interval information might be processed by the central nervous system. The vast majority of cells in the cochlear nucleus show interspike intervals related to fundamental frequency of complex periodic sounds, including vowels (Kim et al., 1986; Greenberg and Rhode, 1987; Palmer and Winter, 1992; Rhode, 1995; Cariani, 1995). Particularly interesting are the responses of onset cells, which respond primarily to the onset of tones at their CF. Most onset cells phase-lock to low-frequency (< 1 kHz) tones, much as if each stimulus cycle constituted a separate onset. In response to steady-state vowels, certain onset cells show very precise phase-locking to the fundamental, basically discharging once for each stimulus cycle (Kim et al., 1986; Palmer and Winter, 1992). Thus, the temporal discharge patterns of these cells are considerably simplified compared to the responses of ANFs, which typically show multiple peaks per cycle (Fig. 3). Such simplification might aid later stages of processing in extracting pitch information. Thus, onset cells might be a component of a neural circuit involved in pitch processing, although they do not by themselves extract pitch. The hypothesis that onset cells play a role in pitch extraction needs to be tested by examining the response of these cells to inharmonic stimuli used in psychophysical experiments on pitch perception.

4.2 Rate-place representation of pitch

Alternatives to purely temporal models of pitch perception are place or "pattern recognition" models that determine pitch by identifying harmonic relationships among stimulus components (e.g. Goldstein, 1973; Terhardt, 1974). These models require an input spectral representation in which low-frequency partials are resolved, but they do not specify how such a

representation might be obtained physiologically. The simplest possibility is that this representation is provided by a rate-place code. Another possibility (not discussed here) is that a temporal-place scheme such as the ALSR might produce the required representation (Srulovicz and Goldstein, 1983; Delgutte, 1984; Miller and Sachs, 1984)

At first sight, the possibility that pitch might be derived from rate-place information appears unlikely because the patterns of average discharge rate against CF measured by Sachs and Young (1979) for vowel stimuli fail to show peaks at harmonics of the fundamental frequency, even for low stimulus levels where dynamic range limitations are not an issue (Fig. 2A). However, this negative result is likely to depend strongly on both the fundamental frequency and the species. The fundamental frequency used by Sachs and Young was 128 Hz, which is typical for a male voice. Hirahara, Cariani and Delgutte (unpublished observations) have found some evidence for rate-place cues to pitch for higher F0s (> 150 Hz), appropriate for the voices of women and children. Evidence for rate-place cues to harmonic spectral patterns is also available for nonspeech stimuli with relatively high F0s (Evans and Wilson, 1973; Smoorenburg and Linschoten, 1977). Another factor that needs to be considered is species differences in the frequency selectivity of the ear. Psychophysical data suggest that the human ear is more selective than the cat ear (Pickles, 1979, 1980). Delgutte (1995) showed that, when a rate-place model for the cat ear is modified to incorporate the human cochlear frequency map (Greenwood, 1990), the modified model does show peaks in discharge rate at the frequencies of the first 5 harmonics of F0s in the range of male voices. Thus, the lack of rate-place cues to F0 in the Sachs and Young (1979) data for the cat might not hold for the human auditory nerve. Interestingly, the F0s of cat vocalizations are near 600 Hz (Watanabe and Katsuki, 1974), in a range where rate-place cues are clearly available in the auditory nerve of this species. Thus, rate-place coding might provide a general representation for the pitch of complex tones with fundamental frequencies within the range of vocalization of each species. This hypothesis is in harmony with the view that the perception of virtual pitch and the perception of conspecific vocalizations are intimately linked (Terhardt, 1974).

In summary, information about the pitch of voice is clearly available in temporal representations, and may also be available in rate-place representations, particularly for the higher F0s. In general, temporal schemes are more plausible for the encoding of pitch than for the encoding of the formant pattern because the degradation of phase locking with increasing frequency is less likely to be a limitation in the range of fundamental frequencies of the human voice.

5. Neural correlates of speech perceptual phenomena

5.1 Context dependence in speech perception

A major issue in speech perception arises in the search for invariant acoustic correlates of phonetic categories (Liberman et al., 1967; Stevens and Blumstein, 1978; Repp, 1982): For the most part, it has not been possible to identify acoustic properties that reliably characterize a given speech sound (or phonetic feature, or syllable, or word) for all contexts. Conversely, in many cases, a given acoustic segment can be heard as different phonetic categories depending on context (Liberman et al., 1967). This context dependence is not unique to speech, but exists for the perception of objects through any sensory modality (Gibson, 1966; Marr, 1982). Nervous systems may have evolved very general mechanisms to handle such context-dependence in the stimulus.

Neural mechanisms such as adaptation, facilitation, and long-lasting inhibition may underlie certain context-dependencies in speech perception. Figure 6 shows spectrograms for a set of stimuli designed to investigate the context-dependent encoding of speech (Delgutte and Kiang, 1984c). These six stimuli share a common part which, by itself, sounds like [da]. They differ in that this common part is preceded by different contexts, yielding stimuli sounding like [na], [ša], [sa], [ada] and [šta] as well as the basic [da]. Figure 7 shows the response patterns of an auditory-nerve fiber for five of these stimuli. For [da], the response pattern shows a clear peak resulting from adaptation at the beginning of the formant transitions. The amplitude of this peak is reduced for the stimuli in which the context elicits a pronounced response (particularly [ša] and [ada]). Thus, the contrast between these stimuli is more salient in the neural response, where differences are present during both the context and the transitions, than in the acoustic waveform, where differences occur only during the context. In general, the greater and longer-lasting the response to the context, the smaller the response to the formant transitions (Delgutte, 1980; Delgutte and Kiang, 1984c), consistent with the properties of auditory-nerve adaptation (Smith, 1979).

Central auditory neurons exhibit more complex forms of context-dependencies than auditory-nerve adaptation. An example is shown in Fig. 8 for an inferior-colliculus neuron in response the stimuli of Fig. 6. The response pattern for [da] shows a pronounced peak at the onset of the formant transition, followed by a pause and then relatively sustained activity during the steady-state [a]. The peak at the onset of the transitions is both decreased in amplitude and altered in shape for all the other stimuli. These findings differ from those of Fig. 7 in that, whereas for the ANF the peak amplitude is always inversely related to the amount of activity in response to the preceding context, this simple relationship does not hold for the IC neuron. In particular, the peak in discharge rate is almost entirely eliminated for [na] despite the weak response during [n]. This strong effect of a context which, by itself produces virtually no spike

discharges suggests that a form of long-lasting inhibition may be involved. This interpretation is consistent with evidence for long-lasting inhibition in IC neurons in response to pairs of click stimuli presented in succession (Carney and Yin, 1989). Overall, the temporal interactions for this neuron are considerably more complex than those resulting from simple adaptation.

Evidence for complex temporal interactions is not limited to the IC, but can also be found for more peripheral stages of processing such as the CN. Rupert et al. (1977) and Caspary et al. (1977) studied the responses of single units in the dorsal cochlear nucleus to brief excerpts from sustained vowels presented in rapid succession. They found that the response to a particular vowel could be suppressed when preceded by another vowel that, by itself, produced no response. Even though these artificial stimuli show abrupt spectral changes that never occur in natural speech, these results point to the existence of inhibitory interactions clearly distinct from adaptation, and lasting for durations comparable to those of speech sounds. This conclusion is supported by studies of "forward masking" in CN neurons using pairs of tone stimuli presented in succession (Boettcher et al., 1990; Palombi et al., 1994; Shore, 1995). These studies show that for certain unit types, particularly chopper, onset, and build-up neurons, forward masking does not obey the functional relationships established by Smith (1979) for adaptation in ANFs, suggesting the existence of additional inhibitory or facilitatory mechanisms.

The significance of these context-sensitive interactions in neural responses is that they might underlie certain trading relations in speech perception (Delgutte, 1982). Trading relations arise when multiple acoustic cues contribute to a phonetic distinction, such that a change in one cue can be compensated for by an opposite change in another cue without changing the phonetic identity of the stimulus (Repp, 1982). For example, two cues for the [aša]-[acha] distinction are the rise time of the frication noise and the duration of the silent interval preceding the noise (Dorman et al., 1979). Using a model of peripheral auditory processing incorporating adaptation, Delgutte (1982) showed that both decreasing the rise time and increasing the duration of the silent interval result in a more prominent adaptation peak at the onset of the frication noise in the response patterns of model ANFs. Thus, in this case, the model neural response showed greater invariance over different realizations of the phonetic category than did the acoustic stimulus.

Adaptation in ANFs is admittedly too simple a neural mechanism to explain more than a handful of trading relations in speech perception. However, neurons in the CN and IC possess more complex forms of temporal interactions that might account for more general trading relations and context dependencies. Some of these neurons may be sensitive to combinations of stimulus components that are widely separated in both time and frequency, thereby providing the neural machinery necessary for the perception of highly context-dependent stimuli such as speech.

5.2 Categorical perception

Human listeners are often better at resolving small differences between speech stimuli that lie near the perceptual boundary between two phonetic categories than they are at resolving stimuli that are far from the boundary (Liberman et al., 1967; Abramson and Lisker, 1970). One of the best-studied examples of such natural perceptual boundaries is for stop consonants differing along a VOT continuum (e.g. [da] vs. [ta]). In many languages, stimuli with VOTs shorter than 20-40 msec are heard as voiced (e.g. [da]), while stimuli with longer VOTs are heard are unvoiced (e.g. [ta]) (Abramson and Lisker, 1970). Pairs of stimuli having VOTs near the 20-40 msec boundary are more easily discriminable than stimuli having either shorter or longer VOTs (Abramson and Lisker, 1970). This enhanced psychophysical acuity for VOTs near the perceptual boundary is not unique to humans, but is also found for chinchillas (Kuhl and Miller, 1978; Kuhl, 1981), monkeys (Kuhl and Padden, 1982), and birds (Dooling et al., 1989; Kluender, 1991), suggesting that it may reflect a very general property of the vertebrate auditory system. Such natural perceptual boundaries may be the basis for certain forms of categorical perception, and may have guided the evolution of the repertoire of sounds used for speech communication (Kuhl, 1981).

A neural correlate of the perceptual boundary along the VOT continuum was identified by Sinex and his colleagues in the discharge patterns of chinchilla ANFs (Sinex and McDonald, 1988, 1989; Sinex et al., 1991). Figure 9 shows representative results. Each panel shows the population response patterns of a sample of low-CF fibers for a pair of stimuli differing in VOT along a [da]-[ta] continuum. The two stimuli in the middle panel span the category boundary between [da] and [ta], while both stimuli in the top panel are normally identified as [da], and both stimuli in the bottom panel as [ta]. The population response to stimuli with VOTs of 30-40 msec shows a rapid rise/peak in discharge rate that is stable across all low-CF fibers. This rise is either less rapid or less stable in response to stimuli that have either lower or higher VOTs. As a result of these differences, the population response patterns are more clearly distinct for the pair of stimuli whose elements belong to two different phonetic categories than for pairs whose elements belong to the same category. Thus, there exists a correlate of the enhanced psychophysical acuity near the 30-40 msec VOT boundary in the response pattern of the *population* of ANFs.

A neural correlate of categorical perception along the VOT continuum was also identified by Steinschneider et al. (1994, 1995) in the primary auditory cortex of the macaque. We have seen in Section 2.2 a class of "double onset" units in the auditory cortex respond to stop-vowels syllables with two bursts of activity: a first burst at consonantal release, and a second one at voicing onset (Steinschneider et al., 1994). In response to stimuli differing in VOT along a [da]-

[ta] continuum, a subset of this class of multi-units only shows the second burst of activity if the VOT exceeds 30-40 msec (Steinschneider et al., 1995). In effect, the second burst of activity is suppressed, perhaps through a form of long-lasting inhibition, when it occurs less than 30-40 msec after the first burst. Regardless of the mechanism involved, responses to stimuli identified as [da] show a single burst of activity, while responses to [ta] stimuli show two bursts.

The auditory-nerve correlate of categorical perception found by Sinex et al. (1991) and cortical correlate of Steinschneider et al. (1995) differ fundamentally in that the former is a property of the population of ANFs, while the latter is found for single recording sites. Nevertheless, the cortical correlate can be interpreted as a transformation of the auditory-nerve correlate. Such a transformation might be accomplished by coincidence-detector neurons that would discharge only when they receive nearly simultaneous inputs from ANFs spanning a wide range of CFs. The cortical correlate of Steinschneider et al. (1995) may be appealing because its behavior is consistent with that of a feature detector. However, only a fraction of cortical units show the categorical effect, so that, in principle, stimuli within a phonetic category could be easily discriminated using VOT information from the majority of units that do not show a categorical effect. This difficulty does not arise for the auditory-nerve correlate of Sinex et al. (1991) because it inherently depends on the entire population of fibers. Overall, the results of Sinex et al. (1991) and Steinschneider et al. (1995) offer promise that relatively simple neural mechanisms might underlie certain forms of categorical perception, and provide further support for the view that properties of the auditory system may have influenced the selection of the repertoire of phonetic features (Kuhl, 1981).

Since this chapter was written, a new study of the encoding of VOT in the primary auditory cortex has appeared (Eggermont, 1995). As Steinschneider et al. (1995) found for the unanesthetized macaque, Eggermont found that single units in the auditory cortex of the anesthetized cat respond with "double onset" patterns to long-VOT stimuli, and with a single onset for short VOTs. However, for the neural population as a whole, there was no tendency for the cross-over between the two response patterns to occur at VOTs near the 30-msec phonetic boundary, in contrast to the Steinschneider et al. report. This negative result illustrates the need for caution when seeking correlates of perceptual phenomena in single "feature detector" neurons, as opposed to populations of neurons.

6. Conclusion

Taken together, studies of the neural processing of speech show that a rich array of cues to phonetic distinctions is available in the discharge patterns of ANFs and CN cells. However, much less is known about which of these cues are actually utilized by more central stages in the auditory nervous system, and if they are used, how this information is processed. For example,

while fine time patterns of discharge of ANFs contain highly-redundant information about the formant pattern of vowels, there is little agreement as to which schemes might be used for processing this information in the central nervous system.

Despite this limitation in our knowledge, two major conclusions emerge from this survey of neural processing of speech. First, many features of the neural responses to speech stimuli can be understood in terms of responses to simpler stimuli that share some acoustic characteristics with speech. For example, ANF responses to pure tones and two-tone stimuli help to understand responses to vowels. Responses of brainstem auditory neurons to AM stimuli help to elucidate responses to acoustic transients in speech. Thus, there is no evidence that speech is treated as a "special" stimulus in the auditory nerve, the brainstem or the midbrain. This observation suggests that a productive approach to understanding the neural processing of speech is to combine speech stimuli with nonspeech analogs in order to identify general neural mechanisms.

The second major conclusion is that the acoustic characteristics of speech that are phonetically the most important are prominently and robustly encoded in neural responses. For example, a majority of ANFs phase lock to one or more of the formant frequencies of vowels, providing a very robust representation of the formant pattern. Phonetically-important changes in amplitude and spectrum are encoded by prominent peaks in discharge rate in the response patterns of ANFs, and may produce even more prominent responses for phasic cells in the central nervous system. For stimuli differing along a VOT continuum, pairs of stimuli that are psychophysically the most discriminable produce the most distinct neural responses. These observations support the view that the auditory system shows predispositions for the particular set of acoustic features used for phonetic contrasts.

For the most part, physiological studies of the auditory processing of speech have not helped to address fundamental issues in speech perception such as variability in acoustic correlates of phonetic categories or the nature of internal representations. It might be hoped that, since the auditory system can easily distinguish phonetic elements, invariant correlates of phonetic categories might be more easily identified from neural responses than from the acoustic signal. In certain respects, this task may in fact be more difficult for neural responses. For example, stimulus level generally has a strong effect on neural responses to acoustic stimuli, while speech perception remains remarkably stable over a very wide range of levels. On the other hand, it is clear that the responses of auditory neurons at the level of the brainstem and above are much more than smeared reflections of the spectrogram, and that there exists a wide variety of neural mechanisms adapted for processing highly context-dependent stimuli such as speech. These mechanisms include adaptation, long-lasting inhibition, coincidence detection, and lateral inhibition. Some neural correlates of classic speech perceptual phenomena such as

categorical perception and trading relations are beginning to emerge. Further studies of the neural processing of speech, particularly those dealing with the central auditory system, are likely to contribute much more to our understanding of fundamental issues in speech perception.

Practical benefits of physiological studies may come even earlier than their theoretical contributions to our understanding of speech perception. Because single-unit techniques provide detailed, signal-oriented descriptions of neural responses, they may be more likely to inspire novel signal processing algorithms than traditional psychophysical models. Neurophysiological studies together with computational and psychophysical approaches may help in developing hearing aids and auditory implants that would provide better speech reception in adverse environments. They may also aid in the design of artificial systems for the coding, transmission and recognition of speech that perform more like the auditory system.

Acknowledgments

I thank P.A. Cariani and T. Hirahara for assistance in collecting physiological data, and B.E. Norris for expert figure preparation. M.C. Brown, P.A. Cariani, B.M. Hammond, J.R. Iversen, S. Kalluri, R.Y. Litovsky and M.F. McKinney made valuable comments on the manuscript. Preparation of this chapter was supported by NIH Grants DC00119 and DC02258.

References

Abramson, A.S., and Lisker, L. (1970). Discriminability along the voicing continuum: cross-language tests. *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 569-573). Prague: Academia.

Ainsworth, W. (ed) (1992). *Advances in Speech, Hearing and Language Processing, Vol. 3*, UK: JAI Press.

Blackburn, C.C., and Sachs, M.B. (1990). The representation of the steady-state vowel /ε/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology* 63, 1191-1212.

Blumstein, S.E., and Stevens, K.N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* 67, 648-662.

Boettcher, F.A., Salvi, R.J., and Saunders, S.S. (1990). Recovery from short-term adaptation in single neurons in the cochlear nucleus. *Hearing Research* 48, 125-144.

Bourk, T.R. (1976). Electrical responses of neural units in the anteroventral cochlear nucleus of the cat. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Bregman, A.S. (1990). *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

Brugge, J.F., and Reale, R.A. (1985). Auditory cortex. In A. Peters and E.G. Jones (eds), *Cerebral Cortex, Vol. 4* (pp. 229-271). New-York: Plenum.

Cant, N.B. (1992). The cochlear nucleus: Neuronal types and their synaptic organization. In D.B. Webster, A.N. Popper and R.R. Fay (eds), *The Mammalian Auditory Pathway: Neuroanatomy* (pp. 66-116*)*. New-York: Springer-Verlag.

Cariani, P.A. (1995). Physiological correlates of periodicity pitch in the cochlear nucleus. *Abstracts of the Association for Research in Otolaryngology* 18, 128.

Cariani, P.A., and Delgutte, B. (1993). Response of auditory-nerve fibers to concurrent vowels with same and different fundamental frequencies. *Abstracts of the Association for Research in Otolaryngology* 16, 373.

Carlson, R., and Granström, B. (eds) (1982). *The Representation of Speech in the Peripheral Auditory System.* Amsterdam: Elsevier.

Carlson, R., Fant, C.G.M., and Granström, B. (1975). Two-formant models, pitch and vowel perception. In C.G.M. Fant and M.A.A. Tatham (eds), *Auditory Analysis and Perception of Speech* (pp. 55-82. London: Academic.

Carney, L.H. (1990). Sensitivities of cells in the anteroventral cochlear nucleus of cat to spatiotemporal discharge pattern discharge patterns across primary afferents. *Journal of Neurophysiology* 64, 437-456.

Carney, L.H. (1994). Spatiotemporal encoding of sound level: Models for normal encoding and recruitment of loudness. *Hearing Research* 76, 31-44.

Carney, L.H., and Geisler, C.D. (1986). A temporal analysis of auditory-nerve fiber responses to spoken stop consonant vowel syllables. *Journal of the Acoustical Society of America* 79, 1896-1914.

Carney, L.H., and Yin, T.C.T. (1989). Responses of low-frequency cells in the inferior colliculus to interaural time differences of clicks: Excitatory and inhibitory components. *Journal of Neurophysiology* 62, 144-161.

Caspary, D.M, Rupert, A.L., and Moushegian, G. (1977). Neuronal coding of vowel sounds in the cochlear nuclei. *Experimental Neurology* 54, 414-431.

Chistovich, L.A., Lublinskaya, V.V., Malinnikova, T.G., Ogorodnikova, E.A., Stoljarova, E.I., and Zhukov, S.J.S. (1982). Temporal processing of peripheral auditory patterns of speech. In R. Carlson and B. Granström (eds), *The Representation of Speech in the Peripheral Auditory System* (pp. 165-180). Amsterdam: Elsevier.

Cooke, M., Beet, S., and Crawford, M. (eds) (1993). *Visual Representations of Speech Signals.* New York: Wiley.

Darwin, C.J. (1992). Listening to two things at once. In M.E.H. Schouten (ed) *The Auditory Processing of Speech* (pp. 133-147). Berlin: Mouton-De Gruyter.

De Boer, E. (1976). On the "residue" and auditory pitch perception. In W.D. Keidel and W.D. Neff (eds), *Handbook of Sensory Physiology, V/3* (pp. 479-583). Berlin: Springer-Verlag.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955). Acoustic loci and transitional cues for stop consonants. *Journal of the Acoustical Society of America* 27, 769-773.

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America* 68, 843-857.

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In R. Carlson and B. Granström (eds) *The Representation of Speech in the Peripheral Auditory System* (pp. 131-150). Amsterdam: Elsevier.

Delgutte, B. (1984). Speech coding in the auditory nerve II: Processing schemes for vowel-like sounds. *Journal of the Acoustical Society of America* 75, 879-886.

Delgutte, B. (1986). Analysis of French stop consonants with a model of the peripheral auditory system. In: J.S. Perkell and D.H. Klatt (eds) *Invariance and Variability of Speech Processes* (pp. 163-177). Hillsdale, NJ: Erlbaum.

Delgutte, B. (1987). Peripheral auditory processing of speech information: Implications from a physiological study of intensity discrimination. In M.E.H. Schouten (ed), *The Psychophysics of Speech Perception* (pp. 333-353) Dordrecht: Nijhof.

Delgutte, B. (1995). Physiological models for basic auditory percepts. In H. Hawkins and T. McMullen (eds) *Auditory Computation.* New-York: Springer-Verlag (in press).

Delgutte, B., and Cariani, P.A. (1992). Coding of the pitch of harmonic and inharmonic complex tones in the interspike intervals of auditory-nerve fibers. In M.E.H. Schouten (ed) *The Auditory Processing of Speech* (pp. 37-45). Berlin: Mouton-De Gruyter.

Delgutte, B., and Kiang, N.Y.S. (1984a). Speech coding in the auditory nerve I: Vowel-like sounds. *Journal of the Acoustical Society of America* 75, 866-878.

Delgutte, B., and Kiang, N.Y.S. (1984b). Speech coding in the auditory nerve III: Voiceless fricative consonants. *Journal of the Acoustical Society of America* 75, 887-896.

Delgutte, B., and Kiang, N.Y.S. (1984c). Speech coding in the auditory nerve IV: Sounds with consonant-like dynamic characteristics. *Journal of the Acoustical Society of America* 75, 897-907.

Delgutte, B., and Kiang, N.Y.S. (1984d). Speech coding in the auditory nerve V: Vowels in background noise. *Journal of the Acoustical Society of America* 75, 908-918.

Deng, L., and Geisler, C.D. (1987). Response of auditory-nerve fibers to nasal consonant-vowel syllables. *Journal of the Acoustical Society of America* 82, 1977-1988.

Deng, L., Geisler, C.G., and Greenberg, S. (1988). A composite model of the auditory periphery for the processing of speech. *Journal of Phonetics* 16, 109-123.

Dewson, J.H. III (1964). Speech sound discrimination by cats. *Science* 144, 555-556.

Dooling, R.J., Okanoya, K., and Brown, S.D. (1989). Speech perception by budgerigars *(melopsitaccus undulatus)*: the voiced-voiceless distinction. *Perception and Psychophysics* 46, 65-71.

Dorman, M.F., Raphael, L.J., and Liberman, A.M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America* 65, 1518-1532.

Durlach, N.I. and Colburn, H.S. (1978). Binaural phenomena. In E.C. Carterette and M.P. Friedman (eds) *Handbook of Perception* (pp. 365-466). New York: Academic.

Ehret, G. (1992). Preadaptations in the auditory system of mammals for phonetic recognition. In M.E.H. Schouten (ed) *The Auditory Processing of Speech* (pp. 99-112). Berlin: Mouton-De Gruyter.

Evans, E.F. (1981). The dynamic range problem: Place and time coding at the level of the cochlear nerve and nucleus. In J. Syka and L. Aitkin (eds), *Neuronal Mechanisms of Hearing* (pp. 69-95). New York: Plenum.

Evans, E.F. (1983). Pitch and cochlear nerve fibre temporal discharge patterns. In R. Klinke and R. Hartmann (eds), *Hearing - Physiological bases and psychophysics* (pp. 140-146). Berlin: Springer-Verlag.

Evans, E.F., and Wilson, J.P. (1973). The frequency selectivity of the cochlea. In A.R. Møller (ed), *Basic Mechanisms in Hearing* (pp. 519-554). London: Academic.

Fant, C.G.M. (1973). *Speech Sounds and Features.* Cambridge, MA: MIT Press.

Flanagan, J.L. (1972). *Speech Analysis, Synthesis and Perception.* New-York: Springer-Verlag.

Frisina, R.D., Smith, R.L., and Chamberlain, S.C. (1990). Encoding of amplitude modulation in the gerbil cochlear nucleus: I. A hierarchy of enhancement. *Hearing Research* 44, 99-122.

Geisler, C.D., and Gamble T. (1989). Responses of "high-spontaneous" auditory-nerve fibers to consonant-vowel syllables in noise. *Journal of the Acoustical Society of America* 85, 1639-1652

Ghitza, O. (1988). Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. *Journal of Phonetics* 16, 109-124.

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems.* New-York: Houghton-Mifflin.

Goldstein, J.L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America* 54, 1496-1516.

Goldstein, J.L., and Srulovicz, P. (1977). Auditory-nerve spike intervals as an adequate basis for aural spectrum analysis. In E.F. Evans and J.P. Wilson (eds), *Psychophysics and Physiology of Hearing* (pp. 337-345). London: Academic.

Greenberg, S. (1988). The ear as a speech analyzer. *Journal of Phonetics* 16, 139-149.

Greenberg, S., and Rhode, W.S. (1987). Periodicity coding in cochlear nerve and ventral cochlear nucleus. In W.A. Yost and C.S. Watson (eds), *Auditory Processing of Complex Sounds* (pp. 225-23). Hillsdale, NJ: Erlbaum.

Greenwood, D.D. (1990). A cochlear frequency-position function for several species - 29 years later. *Journal of the Acoustical Society of America* 87, 2592-2605.

Guinan, J.J., and Gifford, M.L. (1988). Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory nerve fibers. III. Tuning curves and thresholds at CF. *Hearing Research* 37, 29-46.

Harris, D.M., and Dallos, P. (1979). Forward masking of auditory-nerve fiber responses. *Journal of Neurophysiology* 42, 1083-1107.

Hashimoto, Y., Katayama, Y., Murata, K, and Tanigushi, I. (1975). Pitch synchronous response of cat cochlear nerve fibers to speech sounds. *Japanese Journal of Physiology* 25, 633-644.

Helmholtz, H.L.F. von (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Braunschweig: Vieweg und Sohn.

Houtgast, T., and Steeneken, H.J.M. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica* 28, 66-73.

Irvine, D.R.F. (1986). *The Auditory Brainstem.* Berlin: Springer-Verlag.

Johnson, D.H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *Journal of the Acoustical Society of America* 68, 1115-1122.

Joris, P.X., Carney, L.H., Smith, P.H., Yin, T.C.T. (1994). Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Response to tones at the characteristic frequency. *Journal of Neurophysiology* 71, 1022-1051.

Kawase, T., Delgutte, B., and Liberman, M.C. (1993). Anti masking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones. *Journal of Neurophysiology* 70, 2533-2549.

Kiang, N.Y.S. (1975). Stimulus representation in the discharge patterns of auditory neurons. In D.B. Tower (ed), *The Nervous System, Vol. 3: Human Communication and its Disorders* (pp. 81-96). New-York: Raven.

Kiang, N.Y.S., and Moxon, E.C. (1974). Tails of tuning curves of auditory-nerve fibers. *Journal of the Acoustical Society of America* 55, 620-630.

Kiang, N.Y.S., Watanabe, T., Thomas, E.C., Clark, L.F. (1965). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve.* Research Monograph #35, MIT Press, Cambridge, MA.

Kim, D.O., Rhode, W.S., and Greenberg, S.R. (1986). Responses of cochlear nucleus neurons to speech signals: Neural encoding of pitch, intensity and other parameters. In B.C.J. Moore and R.D. Patterson (eds), *Auditory Frequency Selectivity* (pp. 281-288). New-York: Plenum.

Kluender, K.R. (1991). Effects of first formant onset on voicing judgments result from processes not specific to humans. *Journal of the Acoustical Society of America* 90, 83-96.

Kluender, K.R. (1994). Speech perception as a tractable problem in cognitive neuroscience. In *Handbook of Psycholinguistics* (pp. 173-217). London: Academic.

Kluender, K.R., Diehl, R.L., and Killeen, P.R. (1987). Japanese quails can learn phonetic categories. *Science* 237, 1195-1197.

Kuhl, P.L. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America* 70, 340-349.

Kuhl, P.L., and Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America* 63, 905-917.

Kuhl, P.L., and Padden, D.M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception and Psychophysics* 32, 542-550.

Langner, G. (1992). A review: Periodicity coding in the auditory system. *Hearing Research* 60, 115-142.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* 74, 431-461.

Liberman, A.M., and Mattingly, I.G. (1989). A specialization for speech perception. *Science* 243, 489-494.

Liberman, M.C. (1978). Auditory-nerve response from cats raised in a low-noise environment. *Journal of the Acoustical Society of America* 63, 442-455.

Liberman, M.C. (1982a). Single-neuron labeling in the cat auditory nerve. *Science* 216, 1239-1241.

Liberman, M.C. (1982b). The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency. *Journal of the Acoustical Society of America* 72, 1441-1449.

Liberman, M.C. (1991). Central projections of auditory-nerve fibers of differing spontaneous rates. I. Antero-ventral cochlear nucleus. *Journal of Comparative Neurology* 313, 240-258.

Licklider, J.C.R. (1951). The duplex theory of pitch perception. *Experientia* 7, 128-137.

Lisker, L., and Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustic measurements. *Word* 20, 484-422.

Loeb, G.E., White, M.W., and Merzenich, M.M. (1983). Spatial crosscorrelation: A proposed mechanism for acoustic pitch perception. *Biological Cybernetics* 47, 149-163.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Meddis, R., and Hewitt, M.J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification. *Journal of the Acoustical Society of America* 89, 2866-2882.

Meddis, R., and Hewitt, M.J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America* 91, 233-245.

Miller, J.L., and Jusczyk, P.W. (1990). Seeking the neurobiological bases of speech perception. Cognition, 33, 111-137.

Miller, M.I., and Sachs, M.B. (1983). Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America* 74, 502-517.

Miller, M.I., and Sachs, M.B. (1984). Representation of voiced pitch in the discharge patterns of auditory-nerve fibers. *Hearing Research* 14, 257-279.

Moore, B.C.J. (1990). *Introduction to the Psychology of Hearing.* London: Academic.

Moore, B.C.J., Glasberg, B.R., Plack, C.J., and Biswas, A.K. (1988). The shape of the ear's temporal window. *Journal of the Acoustical Society of America* 83, 1102-1116.

Moore, J.K. (1987). The human auditory brainstem: A comparative view. *Hearing Research* 29, 1-32.

Moore, T.J., and Cashin, J.L. (1976). Response of cochlear-nucleus neurons to synthetic speech. *Journal of the Acoustical Society of America* 59, 1443-1449.

Palmer, A.R. (1990). The representation of spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers. *Journal of the Acoustical Society of America* 88, 1412-1426.

Palmer, A..R., Rees, A., and Caird, D. (1990). Interaural delay sensitivity to tones and broadband signals in the guinea-pig inferior colliculus. *Hearing Research* 50, 71-86.

Palmer, A.R., and Winter, I.M. (1992). Cochlear nerve and cochlear nucleus responses to the fundamental frequency of voiced speech sounds and harmonic complex tones. In Y. Cazals, L. Demany, and K. Horner (eds) *Auditory Physiology and Perception* (pp. 231-240). Oxford: Pergamon.

Palmer, A.R., Winter, I.M., and Darwin, C.J. (1986). The representation of steady-state vowel sounds in the temporal discharge patterns of the guinea-pig cochlear nerve and primarylike cochlear nucleus neurons. *Journal of the Acoustical Society of America* 79, 100-113.

Palmer, A.R., Winter, I.M., Jiang, G, and James, N. (1994). Across-frequency integration by neurones in the ventral cochlear nucleus. In: G.A. Manley, G.M. Klump, C. Köppl, H. Fastl, H. Oeckinghous (eds) *Advances in Hearing Research.* Singapore: World Scientific (in press).

Palombi, P.S., Backoff, P.M., and Caspary, D.M. (1994). Paired tone facilitation in dorsal cochlear nucleus neurons: A short-term potentiation model testable in vivo. *Hearing Research* 75, 175-183.

Peterson, G.E., and Barney, H.L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America* 24, 175-184.

Pickles, J.O. (1979). Psychophysical frequency resolution in the cat as determined by simultaneous masking, and its relation to auditory-nerve resolution. *Journal of the Acoustical Society of America* 66, 1725-1732.

Pickles, J.O. (1980). Psychophysical frequency resolution in the cat studied with forward masking. In G. van den Brink and F.A. Bilsen (eds), *Psychophysical, Physiological, and Behavioral Studies in Hearing* (pp. 118-125). Delft: Delft U.P.

Reale, R.A., and Geisler, C.D. (1980). Auditory-nerve fiber encoding of two-tone approximations to steady-state vowels. *Journal of the Acoustical Society of America 67,* 891-902.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.C. (1981). Speech perception without traditional speech cues. *Science* 212, 947-950.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin* 92, 129-132.

Rhode, W.S. (1995). Interspike intervals as a correlate of periodicity pitch in cat cochlear nucleus. *Journal of the Acoustical Society of America* 97, 2413-2429.

Rhode, W.S., and Greenberg, S. (1992). Physiology of the cochlear nuclei. In A.N. Popper and R.R. Fay (Eds), *The Mammalian Auditory Pathway: Neurophysiology* (pp. 94-152*).* New-York: Springer-Verlag.

Rhode, W.S., and Greenberg, S. (1994). Lateral suppression and inhibition in the cochlear nucleus of the cat. *Journal of Neurophysiology* 71, 493-514.

Rhode, W.S., Oertel, D., and Smith, P.H. (1983). Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *Journal of Comparative Neurology* 213, 448-463.

Rhode, W.S., and Smith, P.H. (1986). Encoding time and intensity in the ventral cochlear nucleus of the cat. *Journal of Neurophysiology* 56, 262-286.

Rose, J.E., Brugge, J.F., Anderson, D.J., and Hind, J.E. (1967). Phase-locked responses to low-frequency tones in single auditory-nerve fibers of the squirrel monkey. *Journal of Neurophysiology* 30, 769-793.

Rothman, J.S., Young, E.D., and Manis, P.B. (1993). Convergence of auditory-nerve fibers onto bushy cells in the ventral cochlear nucleus: Implications of a computational model. *Journal of Neurophysiology* 70, 2562-2583.

Rouiller, D.K., and Ryugo, D. (1984). Intracellular marking of physiologically characterized cells in the ventral cochlear nucleus. *Journal of Comparative Neurology* 225, 167-186.

Rupert, A.L., Caspary, D.M., and Moushegian, G. (1977). Response characteristics of cochlear nucleus neurons to vowel sounds. *Annals of Otology* 86, 37-48.

Sachs, M.B. (1984). Speech encoding in the auditory nerve. In C. Berlin (ed), *Hearing Science* (pp. 263-308). San Diego: College Hill.

Sachs, M.B., and Abbas, P.J. (1974). Rate versus level functions for auditory-nerve fiber in cats: tone burst stimuli. *Journal of the Acoustical Society of America* 56, 1835-1847.

Sachs, M.B., Voigt, H.F., and Young, E.D. (1983). Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology* 50, 27-45.

Sachs, M.B., Winslow, R.L., and Blackburn, C.C. (1988). Representation of speech in the auditory periphery. In G.M. Edelman, W.E. Gall, and W.M. Cowan (eds), *Auditory Function* (pp. 747-774*)*. New York: Wiley.

Sachs, M.B., and Young, E.D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America* 66, 470-479.

Sachs, M.B., and Young, E.D. (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *Journal of the Acoustical Society of America* 68, 858-875.

Schalk, T., and Sachs, M.B. (1980). Nonlinearities in auditory-nerve fiber response to band limited noise. *Journal of the Acoustical Society of America* 67, 903-913.

Schouten, M.E.H. (ed) (1987). *The Psychophysics of Speech Perception*. Dordrecht: Nijhof.

Schouten, M.E.H. (ed) (1992). *The Auditory Processing of Speech.* Berlin: Mouton-De Gruyter.

Schwartz, J.L., Beautemps, D., Arrouas, Y., and Escudier, P. (1992). Auditory analysis of speech gestures. In M.E.H. Schouten (ed) *The Auditory Processing of Speech* (pp. 239-252). Berlin: Mouton-De Gruyter.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics* 16, 55-76.

Shamma, S. (1985). Speech processing in the auditory system. II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America* 78, 1622-1632.

Shamma, S. (1988). The acoustic features of speech sounds in a model of auditory processing: Vowels and voiceless fricatives. *Journal of Phonetics* 16, 77-91.

Shore, S.E. (1995) Recovery of forward-masked responses in ventral cochlear nucleus neurons. *Hearing Research* 82, 31-43.

Siebert, W.M. (1970) Frequency discrimination in the auditory system: Place or periodicity mechanism? *Proceedings of the IEEE* 58, 723-730.

Sinex, D.G. (1993). Auditory nerve fiber representation of cues to voicing in syllable-final stop consonants. *Journal of the Acoustical Society of America* 94, 1351-1362.

Sinex, D.G., and Geisler, C.D. (1983). Responses of auditory-nerve fibers to consonant-vowel syllables. *Journal of the Acoustical Society of America* 73, 602-615.

Sinex, D.G., and McDonald, L. (1988). Average discharge rate representation of voice onset time in the chinchilla auditory nerve. *Journal of the Acoustical Society of America* 83, 1817-1927.

Sinex, D.G., and McDonald, L. (1989). Synchronized discharge rate representation of voice onset time in the chinchilla auditory nerve. *Journal of the Acoustical Society of America* 85, 1995-2004.

Sinex, D.G., McDonald, L., and Mott, J.B. (1991). Neural correlates of nonmonotonic temporal acuity for voice onset time. *Journal of the Acoustical Society of America* 90, 2441-2449.

Sinnott, J.M., Beecher, M.D., Moody, D.B., and Stebbins, W.C. (1976). Speech sound discrimination by monkeys and humans. *Journal of the Acoustical Society of America* 60, 687-695.

Slaney, M., and Lyon, R.F. (1993). On the importance of time - A temporal representation of sound. In M. Cooke, S. Beet, M. Crawford (eds), *Visual Representations of Speech Signals* (pp. 95-116). New York: Wiley.

Smith, R.L. (1979). Adaptation, saturation and physiological masking in single auditory-nerve fibers. *Journal of the Acoustical Society of America* 65, 166-178.

Smoorenburg, G.F. (1987). Discussion of the physiological correlates of speech perception. In M.E.H. Schouten (ed), *The Psychophysics of Speech Perception* (pp. 393-399) Dordrecht: Nijhof.

Smoorenburg, G.F., and Linschoten, D.H. (1977). A neurophysiological study on auditory frequency analysis of complex tones. In E.F. Evans and J.P. Wilson (eds), *Psychophysics and Physiology of Hearing* (pp. 175-184). London: Academic.

Srulovicz. P., and Goldstein, J.L. (1983). A central spectrum model: A synthesis of auditory nerve timing and place cues in monaural communication of frequency spectrum. *Journal of the Acoustical Society of America* 73, 1266-1276.

Steinschneider, Schroeder, C.E., M., Arezzo, J.C., and Vaughan, H.G. (1994). Speech-evoked activity in primary auditory cortex: effects of voice onset time. *Electroencephalography and Clinical Neurophysiology* 92, 30-43.

Steinschneider, Schroeder, C.E., M., Arezzo, J.C., and Vaughan, H.G. (1995). Physiologic correlates of voice onset time boundary in primary auditory cortex (A1) of the awake monkey: Temporal response patterns. *Brain and Language* 48, 326-340.

Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America* 68, 836-842.

Stevens, K.N., and Blumstein, S.E. (1978). Invariant cues for place of articulation of stop consonants. *Journal of the Acoustical Society of America* 64, 1358-1368.

Suga, N. (1964). Recovery cycles and responses to frequency-modulated tone pulses in auditory neurones of echolocating bats. *Journal of Physiology (London)* 175, 50-80.

Suga, N. (1992). Philosophy and stimulus design for neuroethology of complex-sound processing. *Philosophical Transactions of the Royal Society of London* B 336, 423-428.

Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustical Society of America* 55, 1061-1069.

Van Noorden, L. (1982). Two channel pitch perception. In M. Clynes (ed), *Music, Mind, and Brain,* New-York: Plenum.

Voigt, H.F., Sachs, M.B., and Young, E.D. (1982). Representation of whispered vowels in discharge patterns of auditory-nerve fibers. *Hearing Research* 8, 49-58.

Warr, W.B. (1992). Organization of olivocochlear efferent systems in mammals. In D.B. Webster, A.N. Popper and R.R. Fay (Eds), *The Mammalian Auditory Pathway: Neuroanatomy* (pp. 410-448*)*. New-York: Springer-Verlag.

Warren, E.H. and Liberman, M.C. (1989). Effects of contralateral sound on auditory-nerve responses. I. Contributions of cochlear efferents. *Hearing Research* 37, 89-104.

Watanabe, T., and Katsuki, Y. (1974). Response patterns of single auditory neurons of the cat to species-specific vocalization. *Japanese Journal of Physiology* 24, 135-155.

Watanabe, T., and Sakai, H. (1973). Response of the collicular auditory neurons to human speech. I. Response to monosyllable /ta/. *Proceedings of the Japanese Academy* 49, 291-296.

Watanabe, T., and Sakai, H. (1978). Responses of the cat's collicular auditory neuron to human speech. *Journal of the Acoustical Society of America* 64, 333-337.

Wever, E.G., and Bray, C.W. (1930). Auditory nerve impulses. *Science* 71, 215-217.

Whitfield, I.C., and Evans, E.F. (1965). Responses of auditory cortical neurones to stimuli of changing frequency. *Journal of Neurophysiology* 28, 655-672.

Wiederhold, M.L. and Kiang, N.Y.S. (1970). Effects of electrical stimulation of the crossed olivocochlear bundle on single auditory nerve fibers in cat. *Journal of the Acoustical Society of America* 48, 950-965.

Winslow, R.L., Barta, P.E., and Sachs, M.B. (1987). Rate coding in the auditory nerve. In W.A. Yost and C.S. Watson (eds), *Auditory Processing of Complex Sounds* (pp. 212-224). Hillsdale, NJ: Erlbaum.

Winslow, R.L. and Sachs, M.B. (1987). Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. *Journal of Neurophysiology* 57, 1002-1021.

Winter, I.M., and Palmer, A.R. (1990). Temporal responses of primarylike anteroventral cochlear nucleus units to the steady-state vowel /i/. *Journal of the Acoustical Society of America* 88, 1437-1441.

Winter, I.M., Robertson, D., Yates, G.K. (1990). Diversity of characteristic frequency rate-intensity functions in guinea pig auditory nerve fibers. *Hearing Research* 45, 191-202.

Wu, Z.L., Schwartz, J.L., and Escudier, P. (1992). Physiologically-based modules and detection of articulatory-based acoustic events. In W. Ainsworth (ed), *Advances in Speech, Hearing and Language Processing, Vol. 3*, UK: JAI Press.

Young, E.D. (1984). Response characteristics of neurons of the cochlear nuclei. In C. Berlin (ed), *Hearing Science* (pp. 423-460). San Diego: College Hill.

Young, E.D., and Sachs, M.B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America* 66, 1381-1403.

Figure Captions

Fig. 1. Neurogram and spectrogram for a speech utterance produced by a female speaker. A. Neurogram display of the activity of the cat auditory nerve in response to the utterance. Each trace represents the average post-stimulus-time histogram for 2-7 auditory-nerve fibers whose CFs are located in a 1/2 octave band centered at the vertical ordinate. All histograms were computed with a bin width of 1 msec, and have been normalized to the same maximum in order to emphasize temporal patterns. The stimulus level was such that the most intense vowels were at 50 dB SPL. B. Broadband spectrogram of the utterance. Filled arrows point to rapid increases in amplitude in the low frequencies (and their neural correlates on top), while open arrows point to rapid increases in amplitude in the high frequencies. The ovals show the second-formant movement in "green" and its neural correlate.

Fig. 2 (Modified from Sachs and Young, 1979). A-C. Normalized average discharge rate against CF for a large sample of auditory-nerve fibers from the same cat in response to a synthetic [ε] vowel. Each symbol shows the average rate for one fiber having a spontaneous discharge rate (SR) greater than 1/sec. The line is a moving-window average of the data points. Average discharge rates are normalized so that 0 corresponds to SR, and 1 to the maximum rate for a pure tone at the CF. Each panel shows data for one stimulus level. D. Power spectrum of the [ε] stimulus, which had a fundamental frequency of 128 Hz and a duration of 400 msec. The formant frequencies are 512, 1792, and 2432 Hz. E and F: Same as B and C respectively for fibers with SRs smaller than 1/sec.

Fig. 3. A. Neurogram display of the auditory-nerve activity in response to a synthetic [ae] vowel presented at 60 dB SPL. Each trace shows a smoothed period histogram for one auditory-nerve fiber whose CF was approximately equal to the vertical ordinate. The histogram bin width is 50 μsec, and its base period is 20 msec, corresponding to two pitch periods of the vowel stimulus. Brackets indicate CF regions in which ANFs phase-lock primarily to the first or second formant frequency. B. Waveform of two pitch periods of the [ae] stimulus, which had a 100-Hz fundamental. The power spectrum is shown in Fig. 5A.

Fig. 4. A. Interval-place representation of the response of the auditory nerve to the same [ae] stimulus as in Fig. 3. Each trace shows an all-order interspike interval (also called autocorrelation histogram) for one auditory-nerve fiber. Fibers are arranged vertically by CF as in Fig. 3. The histogram bin width is 50 μsec. Arrows indicate interspike intervals corresponding to the reciprocal of the first and second frequencies. B. Pooled interspike interval distribution obtained by summing all-order interspike interval histograms for 57 auditory-nerve fibers whose CFs ranged from 200 to 15,000 Hz.

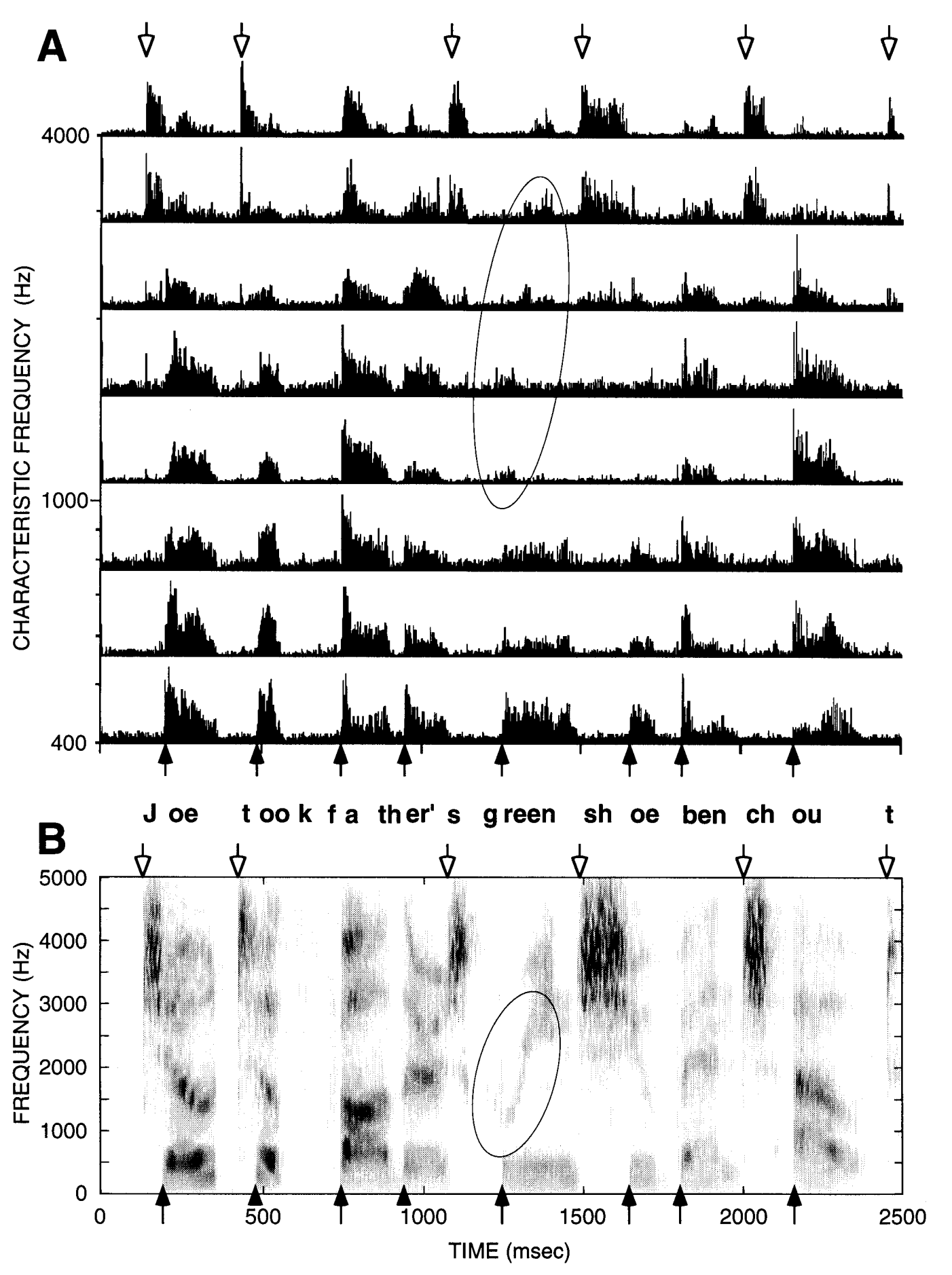
Fig. 5. A. Power spectrum of the [ae] vowel stimulus of Fig. 3 and 4. The frequencies of the first three formants are 750, 1450, and 2450 Hz. B. Average localized synchronized rate (ALSR) computed from period histograms such as those of Fig. 3 for 57 ANFs. The ALSR at frequency *F* was evaluated by averaging the *F*-components of the period histograms for all ANFs whose CFs lie in a 0.4-octave band centered at *F* (Young and Sachs, 1979). C. Power spectrum of the pooled interspike interval distribution shown in the bottom of Fig. 4.

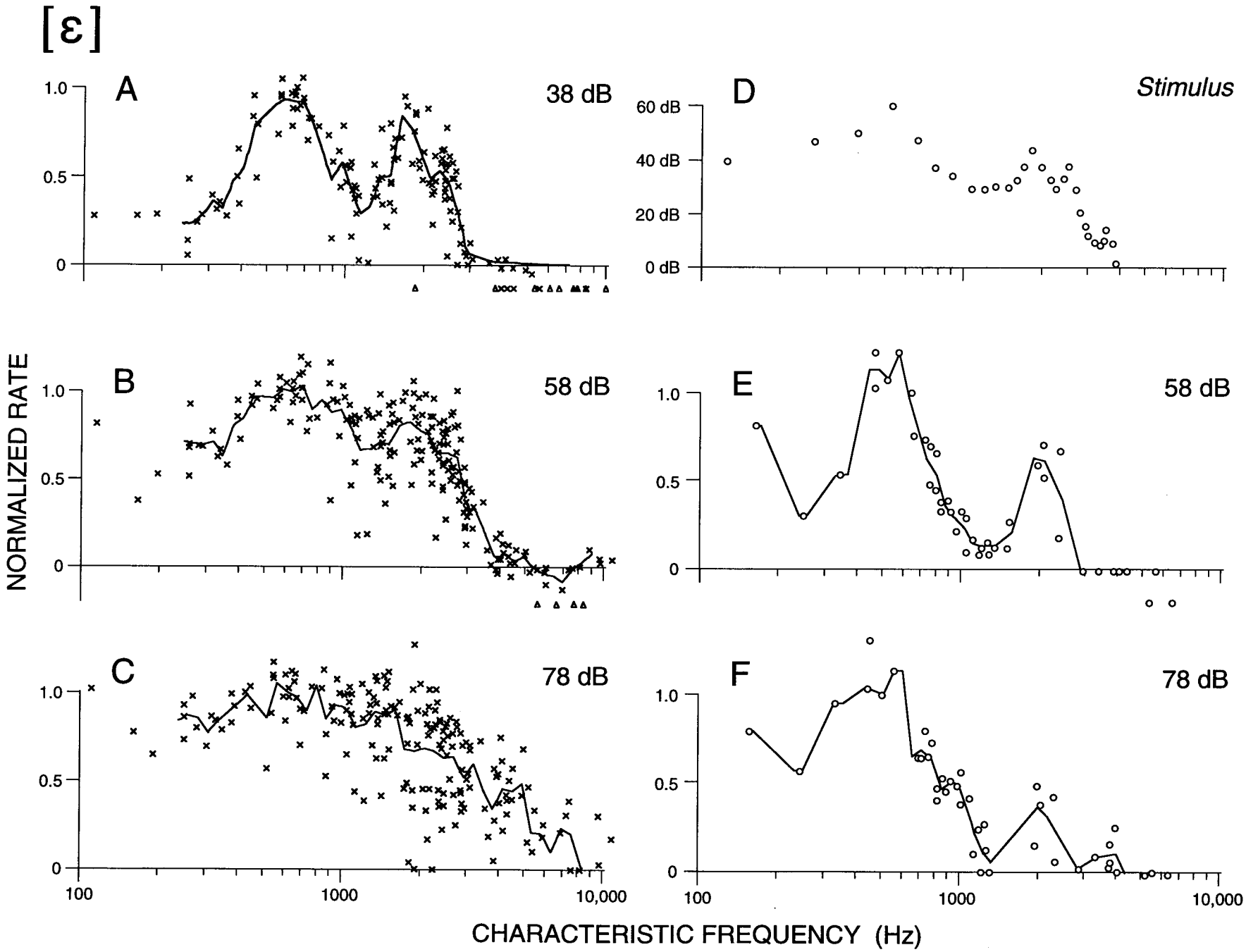
Fig. 6. Broadband spectrograms of 6 synthetic speech stimuli designed for studying the context dependence of the neural response to speech. The spectrograms are aligned so that the common [da] segment occur at the same time.

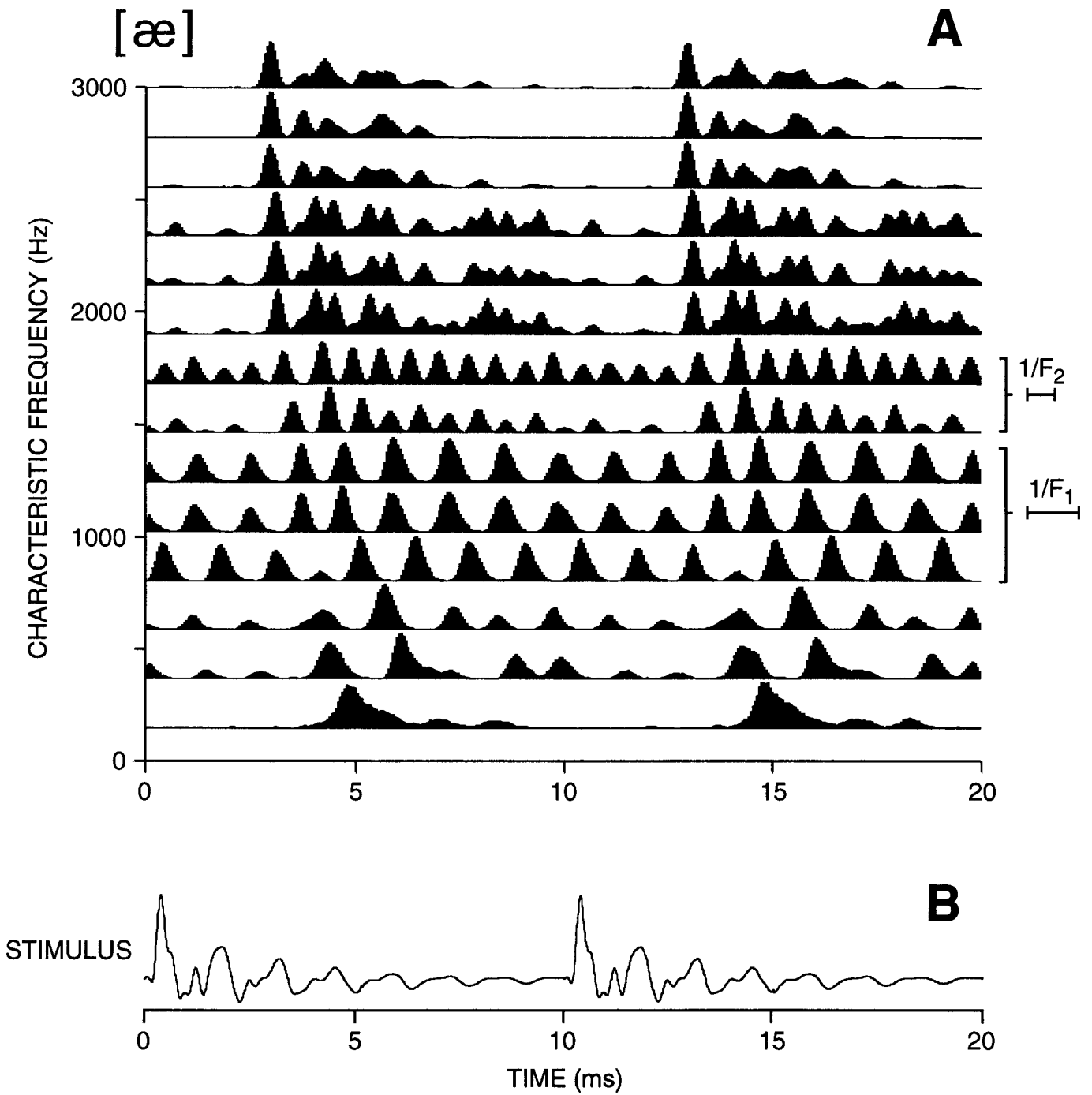
Fig. 7. Response patterns of a high-SR auditory-nerve fiber for 5 of the 6 stimuli of Fig. 6 presented at 60 dB SPL. The shaded area indicates the interval of formant transitions. The PST histograms have a bin width of 1 msec.

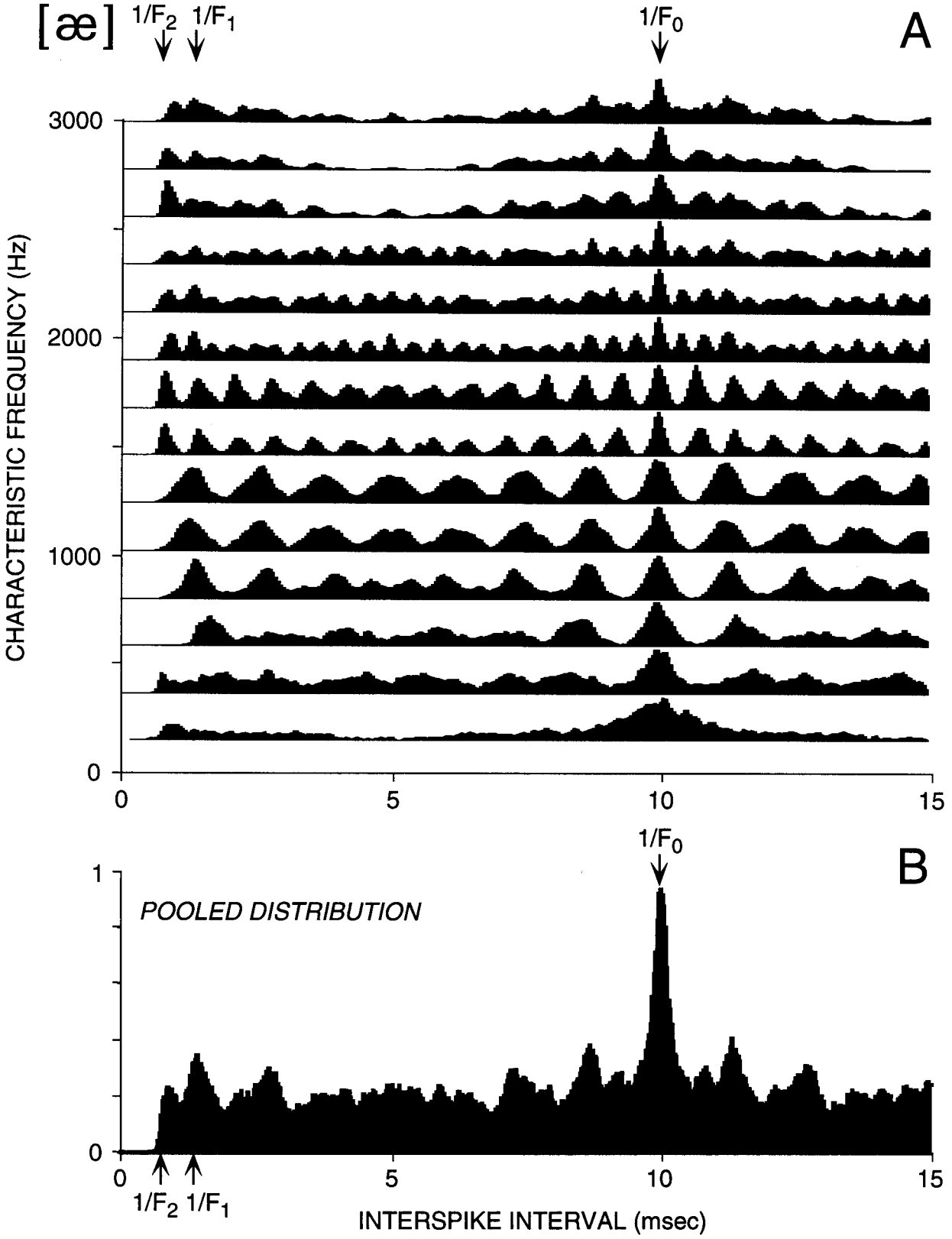
Fig. 8. Response patterns of a binaurally-excited (EE) inferior-colliculus unit for the stimuli of Fig. 6 presented diotically at 60 dB SPL. The PST histograms have a bin width of 1 msec.

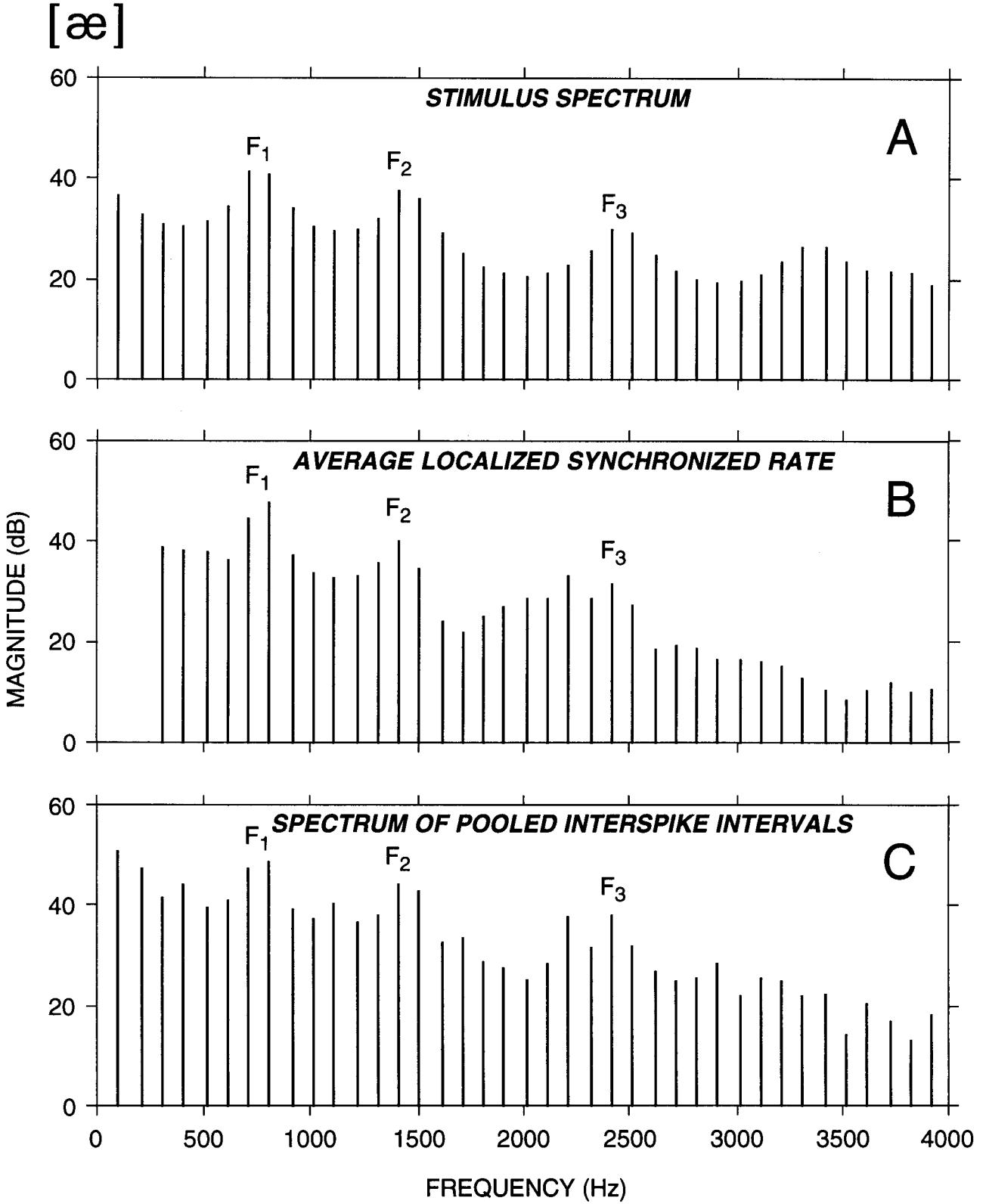
Fig. 9 (From Sinex et al., 1991). Comparison of the population ANF responses elicited by pairs of synthetic syllables differing in VOT by 10 msec. Each crossed area encloses the mean ± 1 standard deviation of the PST histograms of 11 low-CF (< 1 kHz) ANFs from one chinchilla. A. Population response patterns for two stimuli clearly identified at [da]. B. Population response patterns for two stimuli located near the phonetic boundary between [da] and [ta]. C. Population response patterns for two stimuli identified at [ta].

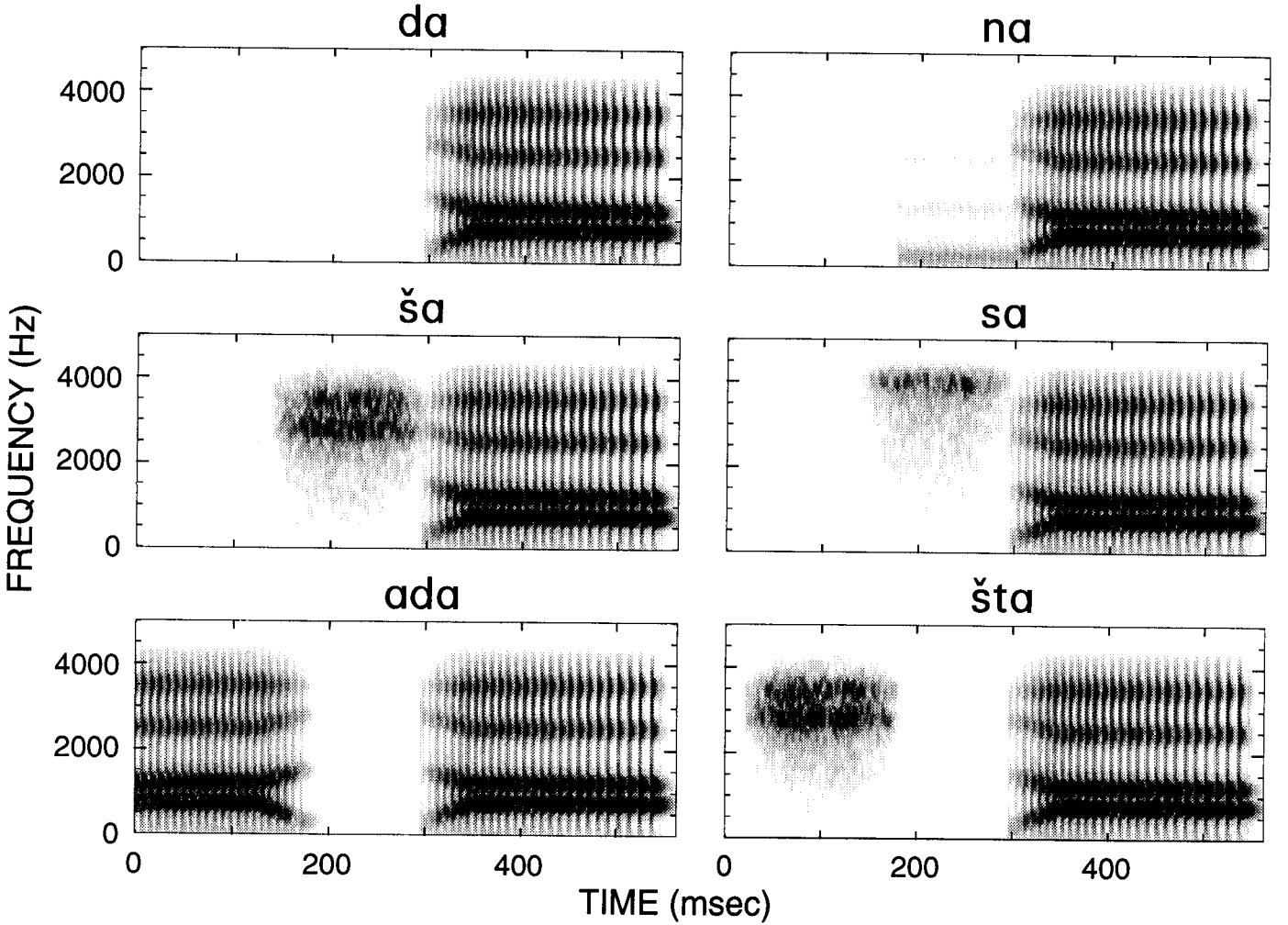

A
4000
1000
400
CHARACTERISTIC FREQUENCY (Hz)
J oe t oo k f a th er' s g reen sh oe ben ch ou t
B
5000
4000
3000
2000
1000
0
FREQUENCY (Hz)
0
500
1000
1500
2000
2500
TIME (msec)

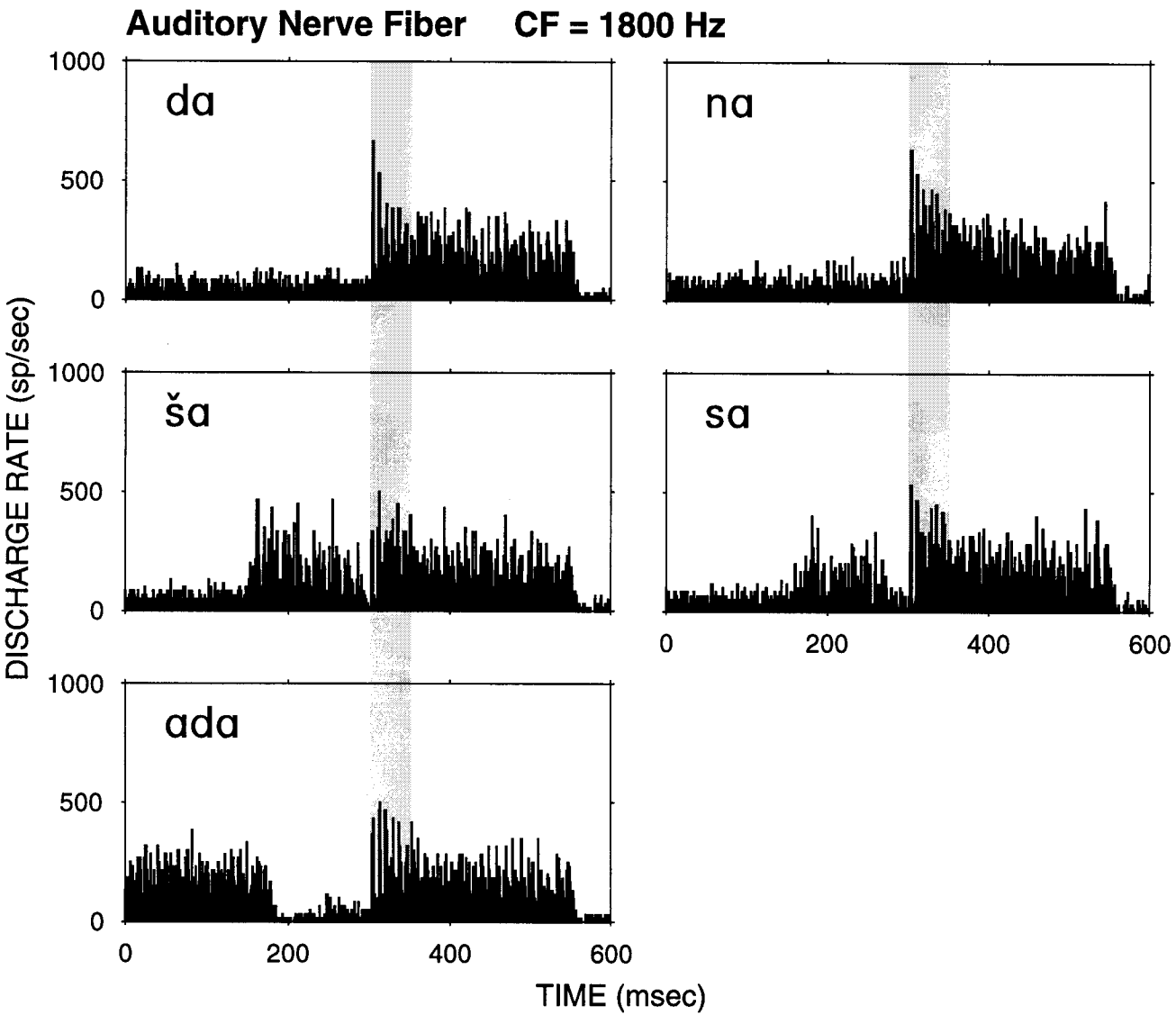

[ε]
A
38 dB
B
58 dB
C
78 dB
D
Stimulus
60 dB
40 dB
20 dB
0 dB
E
58 dB
F
78 dB
1.0
0.5
0
NORMALIZED RATE
100
1000
10,000
CHARACTERISTIC FREQUENCY (Hz)

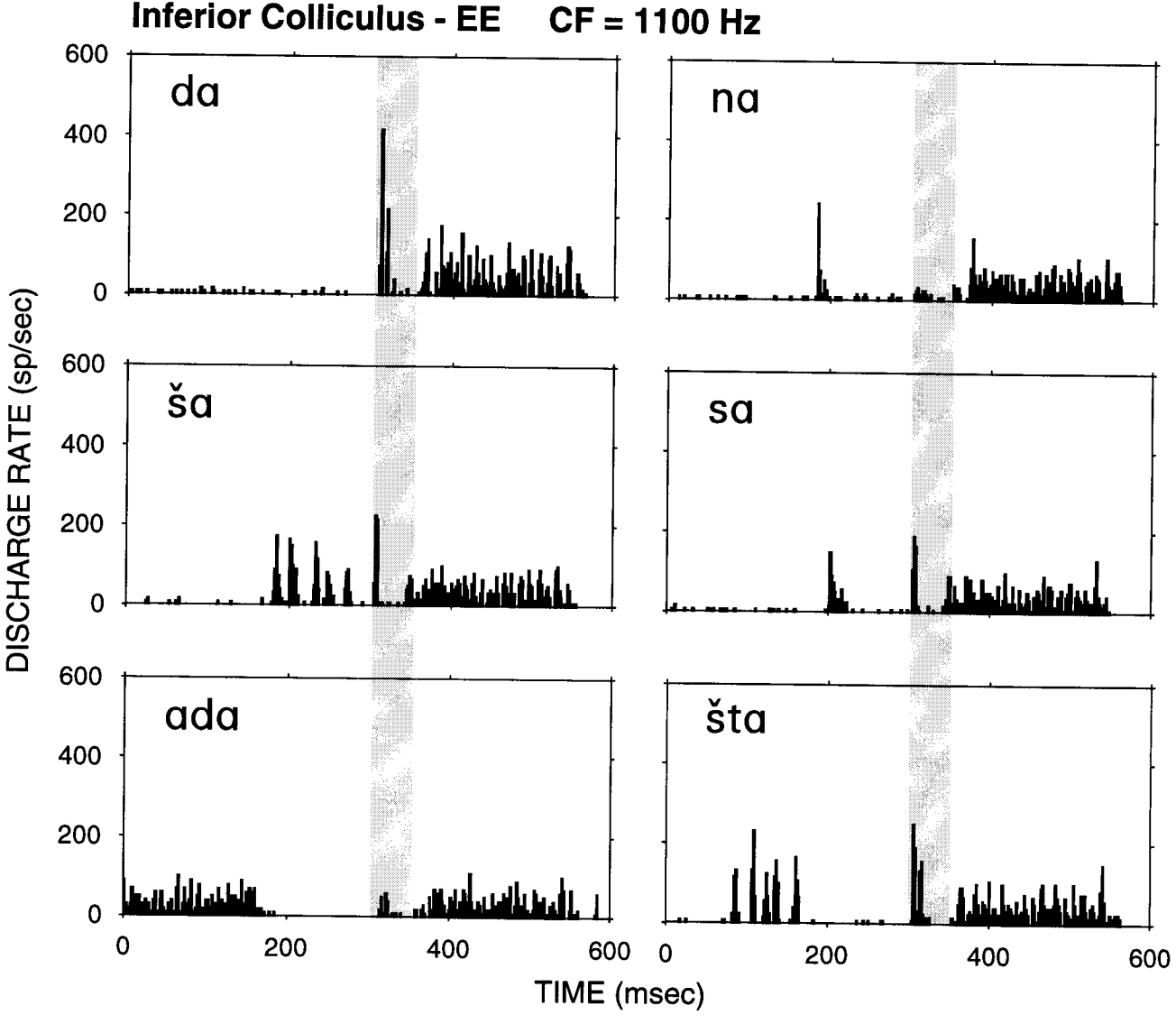

[æ]
A
3000
2000
1000
0
CHARACTERISTIC FREQUENCY (Hz)
1/F2
1/F1
0
5
10
15
20
B
STIMULUS
0
5
10
15
20
TIME (ms)

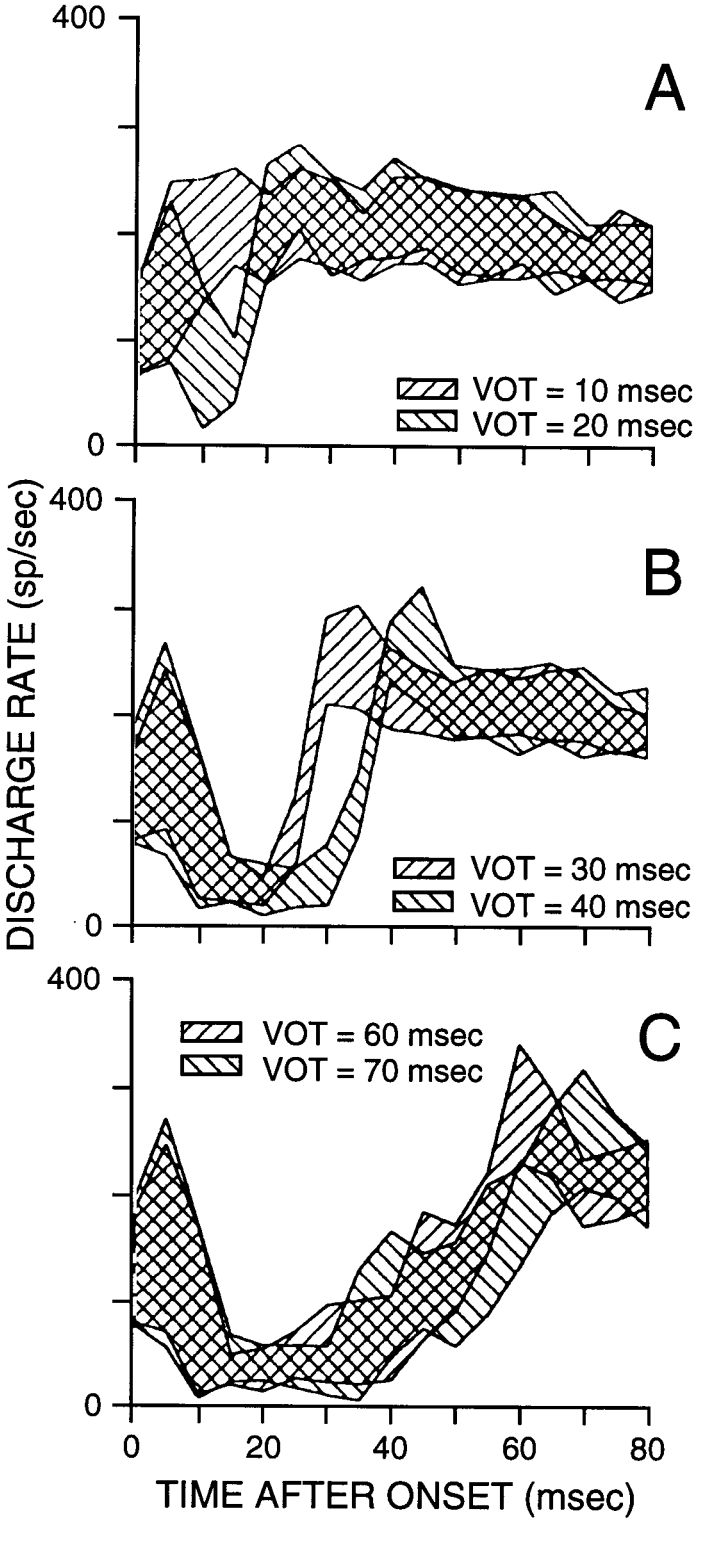

[æ]
1/F2
1/F1
1/F0
A
3000
2000
1000
0
CHARACTERISTIC FREQUENCY (Hz)
0
5
10
15
1/F0
B
1
POOLED DISTRIBUTION
0
0
1/F2
1/F1
5
10
15
INTERSPIKE INTERVAL (msec)


[æ]
STIMULUS SPECTRUM
A
AVERAGE LOCALIZED SYNCHRONIZED RATE
B
SPECTRUM OF POOLED INTERSPIKE INTERVALS
C
F1
F2
F3
MAGNITUDE (dB)
FREQUENCY (Hz)
0
20
40
60
0
500
1000
1500
2000
2500
3000
3500
4000


da
na
ša
sa
ada
šta
4000
2000
0
FREQUENCY (Hz)
0
200
400
TIME (msec)


Auditory Nerve Fiber CF = 1800 Hz
da
na
ša
sa
ada
1000
500
0
0
200
400
600
DISCHARGE RATE (sp/sec)
TIME (msec)


Inferior Colliculus - EE CF = 1100 Hz
da
na
ša
sa
ada
šta
600
400
200
0
DISCHARGE RATE (sp/sec)
0
200
400
600
TIME (msec)


400
A
VOT = 10 msec
VOT = 20 msec
0
400
B
VOT = 30 msec
VOT = 40 msec
0
DISCHARGE RATE (sp/sec)
400
C
VOT = 60 msec
VOT = 70 msec
0
0
20
40
60
80
TIME AFTER ONSET (msec)