Published in final edited form as:
Cognition. 1987 March ; 25(1-2): 21–52.

Acoustic-phonetic representations in word recognition*

DAVID B. PISONI and **PAUL A. LUCE**
Indiana University

Abstract

This paper reviews what is currently known about the sensory and perceptual input that is made available to the word recognition system by processes typically assumed to be related to speech sound perception. In the first section, we discuss several of the major problems that speech researchers have tried to deal with over the last thirty years. In the second section, we consider one attempt to conceptualize the speech perception process within a theoretical framework that equates processing stages with levels of linguistic analysis. This framework assumes that speech is processed through a series of analytic stages ranging from peripheral auditory processing, acoustic-phonetic and phonological analysis, to word recognition and lexical access.

Finally, in the last section, we consider several recent approaches to spoken word recognition and lexical access. We examine a number of claims surrounding the nature of the bottom-up input assumed by these models, postulated perceptual units, and the interaction of different knowledge sources in auditory word recognition. An additional goal of this paper was to establish the need to employ segmental representations in spoken word recognition.

1. Introduction

Although the problems of word recognition and the nature of lexical representations have been long-standing concerns of cognitive psychologists, these topics have not been studied extensively by investigators working in the mainstream of speech perception research. For many years, these two lines of research have remained quite distinct from each other. There are several reasons for this state of affairs. First, the bulk of research on word recognition has been concerned with investigating visual word recognition in reading, with little, if any, concern for the problems of spoken word recognition. Second, most of the interest and research effort in speech perception has been concerned with issues related to feature and phoneme perception in highly controlled environments using nonsense syllables. Such an approach is appropriate for studying “low-level” auditory and acoustic-phonetic analysis of speech, but it is not as useful in dealing with questions surrounding how words are recognized in isolation or in context or how various sources of information are used by the listener to recover the talker’s intended message.

It is now clear that many interesting and potentially quite important problems in the field of speech perception involve the interface between acoustic-phonetic processes and the processes of word recognition and lexical access. These problems deal with the nature of the acoustic cues that listeners extract from the speech signal, the processes used to integrate

*Preparation of this paper was supported, in part, by NIH research grant NS-12179, a contract with the Air Force Office of Scientific Research, Air Force Systems Command and a fellowship from the James McKeen Cattell Fund to the first author. We thank Howard Nusbaum, Beth Greene and Robert Remez for their comments and suggestions. We also thank the reviewers for many useful suggestions on an earlier draft of this paper.

Reprint requests should be sent to David B. Pisoni, Department of Psychology, Indiana University, Bloomington, IN 47405, U.S.A.

these cues, and the various types of perceptual units that are computed by the speech processing system (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Stevens & House, 1972; Studdert-Kennedy, 1974). For example, it is of considerable interest to specify precisely the kinds of representations that exist in the mental lexicon and the intermediate representations that are computed by the listener in converting the speech waveform into a symbolic representation. Are words, morphemes, phonemes or sequences of spectral templates the correct way to characterize the representations of lexical entries in spoken language understanding? Is a word accessed in the lexicon on the basis of an acoustic, phonetic or phonological code? Why are high frequency words recognized so rapidly? And, how is context used to support word recognition and facilitate access to the meaning of a word? These are a few of the questions that will need to be answered before a complete understanding of spoken word recognition will be possible.

In this paper, we consider the nature of the sensory and perceptual input that is made available to the word recognition system by processes typically assumed to be related to speech sound perception. In the first section, we summarize several of the fundamental problems that speech researchers have attempted to deal with over the last thirty-five years. We focus our discussion on the long-standing problems of invariance, linearity, and segmentation of the speech signal in order to illustrate the complex relations that exist between the speech waveform and units of linguistic description. We also consider the problems associated with identifying the basic units of perceptual analysis and the types of representations that are computed by the speech perception system. In the second section, we consider one attempt to conceptualize the speech perception process within a theoretical framework that equates levels of linguistic analysis with processing stages. Finally, in the last section, we consider several recent approaches to spoken word recognition and lexical access. Here we examine claims surrounding the nature of the bottom-up input assumed by these models, the perceptual units, and the potential interaction of different sources of information in word recognition.

2. Fundamental problems in speech perception

The fundamental problems in speech perception today are the same problems that have eluded definitive solution for more than thirty-five years (Fant, 1973; Joos, 1948). Although the intractability of these long-standing problems has led to a voluminous body of literature on the production and perception of speech, researchers are still hard-pressed to explain precisely how the human listener converts the continuously varying speech waveform into discrete linguistic units and how these units are employed to extract the linguistic message intended by the talker. Indeed, not only are we still unsure about the exact nature of the linguistic units arrived at in perceptual processing of speech, little attention has yet been paid to the problem of how the sensory and perceptual analysis of the speech waveform makes contact with representations of words in the lexicon or how these representations are used to support language understanding.

Many, if not all, of the problems in speech perception stem from the manner in which speech is produced. Phonemes are rarely, if ever, realized in the speech waveform as a linearly-ordered sequence of discrete acoustic events. This is due primarily to the fact that speakers coarticulate adjacent phonemes, so that articulation of one phoneme is affected by articulation of neighboring phonemes. It has been extremely difficult to identify the acoustic features in the speech waveform that uniquely match the perceived phonemes independently of surrounding context (see Stevens & Blumstein, 1981). The acoustic consequences of coarticulation and other sources of contextually conditioned variability result in the failure of the acoustic signal to meet two important formal conditions, invariance and linearity, which in turn give rise to the problem of segmentation.

2.1. Linearity of the speech signal

The linearity condition states that for each phoneme there must be a corresponding stretch of sound in the utterance (Chomsky & Miller, 1963). Furthermore, if phoneme X is followed by phoneme Y in the phonemic representation, the stretch of sound corresponding to phoneme X must precede the stretch of sound corresponding to phoneme Y in the physical signal. The linearity condition is clearly not met in the acoustic signal. Because of coarticulation and other contextual effects, acoustic features for adjacent phonemes are often "smeared" across phonemes in the speech waveform. Although segmentation is possible according to strictly acoustic criteria (see Fant, 1962), the number of acoustic segments is typically greater than the number of phonemes in the utterance. Moreover, no *simple* invariant mapping has been found between these purely acoustic attributes or features and perceived phonemes. This smearing, or parallel transmission of acoustic features, results in stretches of the speech waveform in which acoustic features of more than one phoneme are present (Liberman et al., 1967). Therefore, not only is there rarely a particular stretch of sound that corresponds uniquely to a given phoneme, it is also rare that the acoustic features of one phoneme always precede or follow the acoustic features of adjacent phonemes in the physical signal. For this reason, Liberman et al. (1967) have argued that speech is not a simple cipher or alphabet, but is, instead, a complex code in which "speech sounds represent a very considerable restructuring of the phonemic 'message' " (p. 4). Therefore, one of the central concerns in the field of speech perception has focused on the transformation of the continuously varying speech signal into a sequence of discrete linguistic units such as phonemes, phones, or allophones.

2.2. Acoustic-phonetic invariance

Another condition that the speech signal fails to satisfy is the principle of invariance (Chomsky & Miller, 1963). This condition states that for each phoneme X, there must be a specific set of criterial acoustic attributes or features associated with it in all contexts. These features must be present whenever X or some variant of X occurs, and they must be absent whenever some other phoneme occurs in the representation. Because of coarticulatory effects, the acoustic features of a given speech sound frequently vary as a function of the phonetic environment in which it is produced. For example, the formant transitions for syllable-initial stop consonants, which cue place of articulation (e.g., /b/ vs. /d/ vs. /g/), vary considerably depending on the following vowel (Liberman, Delattre, Cooper, & Gerstman, 1954). The formant transitions for stop consonants in syllable-initial positions, then, do not uniquely specify place of articulation across all vowels. If formant transitions are the primary cues to place of articulation for stop consonants, they must be highly context-dependent and not invariant across different phonetic contexts. In short, the problem of invariance is one of explaining perceptual constancy for speech sounds in spite of the absence of reliable acoustic correlates in the speech waveform (Stevens & Blumstein, 1981; Studdert-Kennedy, 1974).

2.3. Segmentation into higher-order units

The context-conditioned variability in the correspondence between the speech signal and phoneme also presents enormous problems for segmentation of the speech waveform into higher-order units of linguistic analysis such as syllables and words. Because of the failure to meet the linearity and invariance conditions noted above, the speech signal cannot be segmented into acoustically defined units that are independent of adjacent segments or are free from the conditioning effects of sentence-level contexts. For example, it has been difficult to determine strictly by simple physical criteria where one word ends and another begins, especially in connected speech. However, word segmentation may be possible by taking into account the systematic internal structure of words, an issue we will return to below.

2.4. Units of perceptual analysis

In addition to the problems of linearity, invariance, and segmentation, there is one other troublesome problem that arises from the coarticulation of speech. This problem involves the relationship between units of perceptual analysis and the units assumed from linguistic analysis. It has been suggested that limitations on channel capacity in the auditory system require that raw sensory information must be recoded into some abstract representation that can be used for subsequent analysis (Liberman et al., 1967). However, what constitutes these abstract units of analysis has been a topic of long-standing debate in the field of speech research. Many investigators have argued for the primacy of the phonetic feature, the phoneme, and the word in the perceptual processing of speech. Other researchers have even proposed units as large as the clause or sentence (Bever, Lackner, & Kirk, 1969; Miller, 1962). In our view, much of the debate over the choice of a basic perceptual unit in language processing is somewhat misguided, for as the level of linguistic processing changes, so do the units of perceptual analysis. The question of whether there is *one* basic or primary unit is to a large extent the wrong question to ask, in our view, because there are, in fact, many units used by the speech processing mechanism. If anything, it is the interaction among the various units that presents a challenge to the perceptual theorist, not the identification or delineation of the one basic unit of perceptual analysis. For some purposes, abstract units such as phonemes are sufficient to capture important distinctions and generalizations within and across constraint domains (Allen, 1985). For other purposes, more parametric representations of the speech waveform may be more appropriate.

A crucial question remains, however, concerning what units are *obligatory* or necessary in the perceptual processing of speech and what the nature of these units may be. Although no one unit may be primary, it is still necessary to specify what units are employed at all in speech perception, word recognition, and lexical access. This problem has arisen primarily in attempts to specify the initial acoustic-phonetic representation of speech. Because of the highly encoded nature of phonemes in the speech waveform resulting from coarticulation (Fischer-Jorgensen, 1954; Liberman et al., 1967), a number of researchers have abandoned the notion that a segmental representation is actually perceived by the listener during speech processing. Alternative accounts, to name a few, have proposed syllables (Cole & Scott, 1974a,b; Massaro & Oden, 1980; Studdert-Kennedy, 1974, 1980), context-sensitive allophones (Wickelgrcn, 1969, 1976), and context-sensitive spectra (Klatt, 1980) as the minimal units of encoding the speech signal. These approaches have generally attempted to circumvent the problem of specifying how a phonemic segmental representation is constructed from the speech waveform in which the conditions of linearity and invariance are not met (see, however, Studdert-Kennedy, 1974). Although one or more of these approaches to the problem of the unit of analysis may be correct in some form, we believe that there is still considerable evidence that can be marshalled to support the claim that at initial stages of speech processing, some type of segmental representation is derived (see below). The question remains, however, concerning what the initial unit of speech encoding is and how it is computed by the processing system. For purposes of the present discussion, it is sufficient simply to note here that this issue has not been resolved satisfactorily even among researchers in the field. Nevertheless, research has continued despite the ambiguity and disagreements over the basic processing units in speech perception. It is our feeling that units like phonemes which are defined within linguistic theory are probably not good candidates for processing units in the real-time analysis of speech. However, units like phones, allophones or context-sensitive diphones may be more appropriate to capture important generalizations about speech and to serve as perceptual units during the earliest stages of speech perception.

3. Perceptual processing of speech

Speech perception has commonly been viewed as a process encompassing various stages of analysis in the transformation of the speech signal to the intended message (Studdert-Kennedy, 1974, 1976). This componential analysis of speech perception has proven very useful in establishing a conceptual framework from which to approach the study of spoken language understanding. Although the exact nature of each of the postulated stages and the interactions among the stages are still tentative, they are nevertheless theoretically justifiable on linguistic grounds. Studdert-Kennedy (1974) was the first to advocate this approach. He proposed four conceptual stages of analysis: (1) auditory, (2) phonetic, (3) phonological, and (4) lexical, syntactic, and semantic. In our discussion of the stages of perceptual processing of speech, we have added a fifth stage of processing—peripheral auditory analysis—to emphasize several recent approaches to speech perception that focus on the earliest transformations of the speech signal by the peripheral auditory system. Conceptually, this stage of processing actually constitutes a subcomponent of the stage of auditory analysis proposed by Studdert-Kennedy.

3.1. Peripheral auditory analysis

Over the last three or four years, a great deal of new research has been reported in the literature on how the peripheral auditory system encodes speech signals (see Carlson & Granstrom, 1982). Research on the peripheral processing of speech signals comes from two different directions. First, a number of important physiological studies using animals have been carried out to describe, in fairly precise terms, how speech signals are coded in the peripheral auditory system (Delgutte, 1980, 1982). These studies have examined auditory-nerve activity in response to simple speech signals such as steady-state vowels and stop consonants in consonant-vowel syllables. The goal of this work has been to identify reliable and salient properties in the discharge patterns of auditory-nerve fibers that correspond, in some direct way, to the important acoustic properties or attributes of speech sounds (Sachs & Young, 1979).

Pursuing a second approach to the peripheral processing of speech, several researchers have begun to develop psychophysically-based models of speech processing (Klatt, 1982). These models explicitly incorporate well-known psychoacoustic data in their descriptions of the filtering that is carried out by the peripheral auditory system (Searle, Jacobson, & Rayment, 1979; Zwicker, Terhardt, & Paulus, 1979). The goal of this line of research is to develop representations of the speech signal that take into account known psychophysical facts about hearing such as critical bands, upward spread of masking, and the growth of loudness (Klatt, 1982).

Searle et al. (1979) have addressed questions related to the appropriate bandwidth of the filters used by human listeners to process speech stimuli. Reviewing evidence from psychophysical and physiological studies, Searle et al. propose that the human peripheral auditory system analyzes auditory stimuli with approximately a $\frac{1}{3}$-octave frequency resolution. The choice of $\frac{1}{3}$-octave bandwidths is motivated not only by the psychophysical and physiological data, but also by the properties of human speech. Because bandwidth is proportional to frequency, $\frac{1}{3}$-octave bandwidths allow spectral resolution of low frequencies as well as temporal resolution at high frequencies. Spectral resolution of low frequencies enables separation of the first and second formants while temporal resolution of high frequencies provides accurate timing information for rapid onsets of bursts. Reasoning from known filtering properties of the human peripheral auditory system, Searle et al. were able to construct a phoneme recognizer with quite high levels of accuracy at discriminating initial stop consonants in consonant-vowel syllables, thereby demonstrating the degree to which

speech recognition may be improved once psychologically reasonable transformations of the speech signal are incorporated in peripheral representations of speech.

The recent interest and extensive research efforts in developing new and presumably more appropriate and valid representations of speech signals derives, in part, from the assumption that a more detailed examination of these auditory representations should, in principle, provide researchers with a great deal more relevant information about the distinctive perceptual dimensions that underlie speech sounds (Stevens, 1980). It has been further assumed that information contained in these so-called neuroacoustic and psychoacoustic representations will contribute in important ways to finally resolving the acoustic-phonetic invariance problem in speech (Goldhor, 1983). Although new and important findings will no doubt come from continued research on how speech signals are processed in the auditory periphery, one should not be misled into believing that these new research efforts on the processing of speech by the auditory nerve will provide all the needed solutions in the field of speech processing. On the contrary, a great deal more research is still needed on questions concerning the central auditory mechanisms used in pattern recognition and higher level sources of information in speech perception (Klatt, 1982).

3.2. Central auditory analysis

Following the initial transformation of the speech signal by the peripheral auditory system, acoustic information about spectral structure, fundamental frequency, changes in source function, overall intensity, and duration of the signal, as well as amplitude onsets and offsets is extracted and coded by the auditory system (Stevens, 1980). These spectral and temporal patterns of the speech signal are assumed to be preserved in sensory memory for a brief period of time during which acoustic feature analysis is carried out (see Pisoni & Sawusch, 1975). The results of auditory analysis provide “speech cues”; that is, auditory-based representations of the speech signal that are subsequently used for phonetic classification.

A great deal of research over the last thirty-five years has been devoted to the description of acoustic cues to phonetic segments. (Reviews may be found in Darwin, 1976; Pisoni, 1978; Studdert-Kennedy, 1974, 1980.) Typically, many acoustic cues map onto a single phonetic feature. For example, Lisker (1978) has listed sixteen possible cues to voicing of intervocalic stop consonants. In general, however, a few basic cues can be listed that serve to signal place, manner, and voicing of consonants and frontness-backness and height of vowels. For example, for stop consonants (/b/, /d/, /g/, /p/, /t/, /k/), place of articulation may be signalled by the direction and extent of formant transitions, by the gross spectral shape of the release burst at onset, by the frequency of the spectral maximum at the burst, and by the bandwidth of the burst (see Cole & Scott, 1974a,b; Delattre, Liberman, & Cooper, 1955; Dorman, Studdert-Kennedy, & Raphael, 1977; Liberman, Delattre, & Cooper, 1952; Liberman, Delattre, Cooper, & Gerstman, 1954; Stevens & Blumstein, 1978).

Voicing of initial stops may be signalled by voice-onset time, frequency of the first formant transition, and amplitude of the burst (see Abramson & Lisker, 1965; Lisker & Abramson, 1964; Stevens & Klatt, 1974; Summerfield & Haggard, 1974). Among the many cues signalling voicing of post-vocalic stops are closure duration (in post-stressed syllable-medial position), duration of the preceding vowel, extent of formant transitions, and voicing into closure (Denes, 1955; Fitch, 1981; House, 1961; Lisker, 1957, 1978; Port, 1977, 1979; Raphael, 1972; Raphael & Dorman, 1980). For any given phonetic contrast, then, it is clear that multiple acoustic events are involved in signalling the contrast, and it is at the stage of auditory analysis that such cues are extracted.

3.3. Acoustic-phonetic analysis

The acoustic-phonetic level, the first level at which linguistic processing is accomplished, is assumed to be the next stage of perceptual analysis. Here the speech cues from the previous level of analysis are mapped onto distinctive phonetic features. Phonetic features may be thought of as abstract perceptual and memory codes that stand for combinations of both specific acoustic attributes on the one hand, and their articulatory antecedents on the other hand. In the phonetic and phonological literature, it has been convenient to describe these features in terms of articulatory descriptions and labels primarily because this notation captures linguistically relevant distinctions at the phonetic and phonological levels. One description of speech at this level consists of a phonetic matrix in which the columns represent discrete phonetic segments and the rows indicate the phonetic feature composition of each segment (Chomsky & Halle, 1968).

The acoustic-phonetic level of analysis has received a great deal of attention in connection with the hypothesis that specialized phonetic feature detectors may be operative at this stage of speech processing (Abbs & Sussman, 1971; Eimas & Corbit, 1973; see Diehl, 1981, for a review). The notion of feature detectors in speech perception was originally proposed by Eimas and Corbit (1973) on the basis of two sources of evidence. The first line of evidence came from research on infant speech perception that demonstrated that 1- and 4-month-old infants discriminate certain speech and non-speech stimuli in much the same way that adults do (Eimas, Siqueland, Jusczyk, & Vigorito, 1971). Infants discriminated speech contrasts in a categorical-like manner, such that between-category stimuli (e.g., /bae/ and /dae/) were discriminated better than within-category stimuli (e.g., two different tokens of /bae/). Infant discrimination of comparable non-speech stimuli, however, was not demonstrably superior for between-category stimuli than within-category stimuli (Eimas, 1974). Because of the striking similarity of the infants' and adults' performance on both the speech and nonspeech stimuli, Eimas (1974) proposed that infants are equipped at birth with feature detectors specialized for processing speech stimuli. (See Aslin, Pisoni, & Jusczyk, 1983, for further discussions of these and related issues in infant speech perception.)

A second line of evidence for the feature detector hypothesis comes from studies on the selective adaptation of speech, the first of which was conducted by Eimas and Corbit (1973). For reviews see Cooper (1975) and Eimas and Miller (1978). Eimas and Corbit (1973) interpreted their demonstration of selective adaptation of speech as evidence for the operation of specialized phonetic feature detectors. They reasoned that the repeated presentation of an endpoint stimulus fatigued detectors tuned to the phonetic features of the stimulus. Fatigue of the detector sensitive to the voicing feature causes a shift in the identification function toward either the voiced or voiceless end of the continuum, depending on the adaptor used. Eimas and Corbit concluded that the results from the infant data as well as the demonstration of selective adaptation for speech supported the notion of specialized feature detectors at the level of phonetic analysis.

The Eimas et al. (1971) and Eimas and Corbit (1973) studies inspired a large number of studies on both infant perception and selective adaptation, a review of which is well beyond the scope of the present discussion (see the references cited above for reviews). Suffice it to note here that the notion of specialized phonetic feature detectors has been abandoned. Studies on both infant perception and selective adaptation have since shown that the previously demonstrated effects lie not at the level of phonetic analysis, but rather at an earlier stage or stages of auditory feature analysis (Eimas & Miller, 1978; Ganong, 1978; Sawusch, 1977a,b; see also Remez, 1979). In an elegant study by Sawusch and Jusczyk (1981), the locus of selective adaptation effects at the auditory level was clearly identified. Sawusch and Jusczyk found that adaptation followed the spectral characteristics of the

adaptor and not the perceived phonetic identity, thus clearly placing the effects of selective adaptation at the level of auditory analysis.

3.4. Phonological analysis

At the level of phonological analysis, the phonetic features and segments from the previous level are converted into phonological segments. The phonological component provides information about the sound structure of a given language that is imposed on the phonetic matrix to derive a phonological matrix (Chomsky & Halle, 1968). Thus, the phonological rules that are applied to the phonetic input at this level determine the extent to which the phonetic segments function as distinctive elements in the language and the extent to which these attributes may be predicted from either language-specific rules or language universal principles. Thus, predictable and redundant phonetic details can be accounted for systematically at this level. Allophonic variations present at the phonetic level are also eliminated and only phonologically distinctive information is coded for subsequent processing.

Historically, the output of the phonological component was believed by linguists to be a linearly-ordered sequence of phonemes in which syllables played no role in phonological organization (Chomsky & Halle, 1968). Recently, however, some phonologists (Clements & Keyser, 1983; Halle & Vergnaud, 1980; Kahn, 1976; Selkirk, 1980) have postulated a hierarchical representation of the internal structure of the syllable in which the major constituents are an *onset*—an optional initial consonant or consonant cluster— and a *rime*—the rest of the syllable excluding inflectional endings. Some preliminary behavioral evidence (Treiman, 1983) indicates that such constituents may be psychologically real. It may be, then, that both phonemic and syllabic tiers of organization are computed at this stage of analysis (see Halle, 1985, for a similar proposal).

Two aspects of the phonological level in this processing scheme are worthy of mention here. The first concerns the suggestion, already alluded to above, that a segmental representation is computed by listeners in the on-line processing of speech. As we mentioned earlier, several researchers have abandoned the notion of a distinct phonemic level of representation in speech perception, primarily due to the difficulties encountered in identifying linearly-ordered, invariant acoustic segments in the waveform that correspond to phonemes (e.g., Klatt, 1979, 1980; Massaro & Oden, 1980; Wicklegren, 1969, 1976). If sufficient evidence for a phonetic level of representation cannot be rallied, it would be superfluous to postulate a phonological level of analysis in any conceptual framework for speech perception. However, we believe that a number of compelling arguments can be made to demonstrate the need for an abstract segmental representation at some level of the speech perception process (see also Studdert-Kennedy, 1976, 1980). Because the assumption of segmental representations has played such an important role in linguistics and especially in speech perception and word recognition over the last thirty-five years, below we present a brief defense of the existence of segmental representations in speech processing.

3.5. Higher-order analysis of speech

Beyond the level of phonological analysis, several additional levels of “higher-order” analysis are carried out on the recoded speech signal. First, we assume that word recognition and lexical access accept as input some segmental representation of the speech signal. This representation could consist of phones, allophones or context-sensitive phoneme-like units. In word recognition, patterns from lower levels of analysis are then matched to representations of words residing in long-term memory. Lexical access takes place when the meanings of words are contacted in long-term semantic memory (see other papers in this issue).

Second, a word's functional, semantic, and syntactic roles are also derived from some segmental representation of the speech signal in order to parse and interpret the utterance. Prosodic information is interpreted as well in order to organize the utterance in short-term memory, to identify syntactic boundaries, to predict upcoming stress patterns, and to evaluate certain pragmatic aspects of the conversational situation. In short, a great deal of analysis subsequent to the phonological level is necessary to recover the speaker's intended message.

The precise roles of these higher levels of analysis in guiding the earlier levels of processing is a topic of considerable interest among researchers. Some of the questions currently under examination concern the degree to which higher levels of processing interact with the initial acoustic-phonetic and phonological analyses, the role of higher level sources of information in predicting upcoming speech input, and the degree to which other sources of information can compensate for misperceptions and impoverished acoustic-phonetic information in the signal. Although many of these issues were formerly believed to be beyond the immediate concern of researchers in speech perception, a growing number of theorists are realizing the need to specify the contributions of higher levels of analysis in order to understand more fully the speech perception process. Perhaps more important, there is a real need to establish a point of contact in language comprehension between the early sensory-based acoustic-phonetic input and the language processing system itself. The primary locus of this interface appears to lie at the level of processing corresponding to word recognition and lexical access (see Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978).

4. Segmental representations in speech perception

For a number of years there has been a continuing debate concerning the role of segmental representations in speech perception and spoken word recognition. Several theorists have totally abandoned an intermediate segmental level of representation in favor of direct access models. In these models, words are recognized without an intermediate analysis of their "internal structure" into units like allophones, phonemes, diphones or demisyllables. Proponents of this view have argued that their recognition models do not require the postulation or use of these intermediate representations and that human listeners do not actually employ these units in the real-time analysis of spoken language. In this section, we argue against this position and summarize evidence from several different areas supporting the existence of these processing units in speech perception and spoken word recognition. While some theorists have attempted to ignore or even to deny the existence of these units, we suggest that they are, in fact, *tacitly* assumed by all contemporary models of word recognition. Without this assumption, it would not be possible to recover the internal structure of words and access their meanings. Based on several sources of evidence, we argue that the output from the speech perception system consists of some form of segmental representation. Furthermore, it is this representation that forms the input to processes involved in word recognition and lexical access.

The first general line of evidence we offer in support of segmental representations in speech perception comes from linguistics. One of the fundamental assumptions of linguistic analysis is that the continuously varying speech waveform can be represented as a sequence of discrete units such as features, phones, allophones, phonemes, and morphemes. This assumption is central to our current conceptions of language as a system of rules that governs the sound patterns and sequences used to encode meanings (Chomsky & Halle, 1968). The very existence of phonological phenomena such as alternation, systematic regularity, and diachronic and synchronic sound changes require, ipso facto, that some type of segmental level be postulated in order to capture significant linguistic generalizations. In describing the sound structure of a given language, then, a level of segmental representation

is required in order to account for the idiosyncratic and predictable regularities in the sound pattern of that language (see Kenstowicz & Kisseberth, 1979). Whether these segmental units are actually used by human listeners in the real-time analysis of spoken language is, however, another matter.

The second general line of evidence in support of segmental representations in speech perception is more psychological in nature. Psychological evidence for the hypothesis of a segmental level of representation in speech perception comes from a number of diverse sources. One source of evidence comes from observations of speakers of languages with no orthography who are attempting to develop writing systems. In his well-known article "The psychological reality of phonemes," Sapir (1963) cites several examples of cases in which the orthographic choices of an illiterate speaker revealed a conscious awareness of the phonological structure of his language. More recently, Read (1971) has described a number of examples of children who have invented their own orthographies spontaneously. The children's initial encounters with print show a systematic awareness of the segmental structure of language, thereby demonstrating an ability to analyze spoken language into representations of discrete segments such as phones, allophones, or phonemes. Indeed, it has been recently suggested (Liberman, Shankweiler, Fischer, & Carter, 1974; Rozin & Gleitman, 1977; Treiman, 1980) that young children's ability to learn to read an alphabetic writing system like English orthography is highly dependent on the development of phonemic analysis skills, that is, skills that permit the child to consciously analyze speech into segmental units.

The existence of language games based on insertion of a sound sequence at specifiable points in a word, the movement of a sound or sound sequence from one point to another in a word, or the deletion of a sound or sound sequence all provide additional support for the existence of segmental representations of the internal structure of words (see Treiman, 1983, 1985). The existence of rhymes and the metrical structure of poetry also entail the awareness, in one way or another, that words have an internal structure and organization to them and that this structure can be represented as a linear sequence of discrete units distributed in time.

An examination of errors in speech production provides additional evidence that words are represented in the lexicon in terms of some sort of segmental representation. The high frequency of single segment speech errors such as substitutions and exchanges provide evidence of the phonological structure of the language (Fromkin, 1973, 1980; Garrett, 1976, 1980; Shattuck-Hufnagel & Klatt, 1979; Stemberger, 1982). It has been difficult, if not impossible, to explain these kinds of errors without assuming some kind of segmental representation in the organization of the lexicon used for speech production.

Over the years there have been many perceptual findings that can be interpreted as support for an analysis of speech into segmental representations. Perhaps the most compelling data have come from numerous experiments involving an analysis of errors and confusions in short-term memory and of the errors produced in listening to words and nonsense syllables presented in noise (Conrad, 1964; Klatt, 1968; Miller & Nicely, 1955; Wang & Bilger, 1973; Wickelgren, 1965, 1966). While some of these findings were originally interpreted as support for various types of feature systems, they also provide strong evidence for the claim that the listener carries out an analysis of the internal structure of the stimulus input into dimensions used for encoding and storage in memory. While these findings can be interpreted as support for segmental analysis of spoken input, they have also been subject to alternative interpretations because of the specific tasks involved. As we noted earlier, the size of the perceptual unit changes as the level of analysis shifts according to the experimental task and instructions to subjects. If perceptual and short-term memory data

were the only findings that could be cited in support of segmental representations, one might be less inclined to accept a segmental level of representation in speech perception and spoken word recognition. However, there are other converging sources of evidence from perceptual studies that provide additional support for this view.

For example, there have been numerous reports describing the phoneme restoration effect (Samuel, 1981a,b; Warren, 1970), a phenomenon demonstrating the on-line synthesis of the segmental properties of fluent speech by the listener. Numerous studies have also been carried out using the phoneme monitoring task in which subjects are required to detect the presence of a specified target phoneme while listening to sentences or short utterances (see Foss, Harwood, & Blank, 1980). Although some earlier findings (Foss & Swinney, 1973; Morton & Long, 1976) suggested that listeners first recognize the word and then carry out an analysis of the segments within the word, other more recent findings (Foss & Blank, 1980) indicate that subjects can detect phonemes in nonwords that are not present in the lexicon (see also Foss & Gernsbacher, 1983). Thus, subjects can detect phonemes based on two sources of knowledge, information from the sensory input and information developed from their knowledge of the phonological structure of the language (Dell & Newman, 1980).

A large body of data has also been collected on the detection of mispronunciations in fluent speech (see, for example, Cole, 1973; Cole & Jakimik, 1978, 1980). While these findings have been interpreted as support for the primacy of word recognition in speech perception (Cole, 1973), these results can just as easily be used to support the claim that listeners can gain access to the internal structure of words in terms of their segmental representations, and that they can do this while listening to continuous speech.

Finally, in terms of perceptual data, there is a small body of data on misperceptions of fluent speech (Bond & Games, 1980; Bond & Robey, 1983). The errors collected in these studies suggest that a very large portion of the misperceptions involve segments rather than whole words.

Having considered several sources of evidence for positing a phonological level of analysis, we now briefly consider what types of analyses are performed at this level. We have argued that at the level of phonological analysis, a segmental representation is derived based on the acoustic-phonetic features computed at the previous level of analysis. That is, the phonetic segment is converted into an abstract, systematic segmental representation (Chomsky & Halle, 1968). The process of converting information at the phonetic level into systematic segmental representations is complex, in that these representations are abstract entities that may be realized in any number of ways at the phonetic level. For example in fluent speech, the word “and” may be produced as [ænd], [æen], or [ən] (Oshika, Zue, Weeks, Neu, & Aurbach, 1975).

At the phonological level of analysis, the listener applies his knowledge of the phonological rules of the language in mapping phonetic representations onto more abstract segmental representations. Much of the variability at the phonetic level is inherently rule-governed, so that the listener may greatly simplify the task of deriving abstract representations from the acoustic-phonetic waveform by employing his knowledge of phonology and morphology. Oshika et al. (1975) have proposed a general class of phonological rules that attempt to describe this systematic pronunciation variation in order to illustrate the role that knowledge of phonological rules may play in speech processing. Among the rules proposed by Oshika et al. are (1) vowel reduction, (2) alveolar flapping, (3) palatalization, (4) homorganic stop deletion, and (5) geminate reduction. Each of these rules describes general phenomena found in continuous speech. For example, the vowel reduction rule describes the tendency for unstressed vowels to be realized as [ə]. Thus, although “and” may frequently be realized

as [ən] at the phonetic level of analysis, the listener may "recover" the underlying representation by employing his knowledge of the phonological rule governing vowel reduction. Likewise, a listener's knowledge of the circumstances under which obstruent alveolars palatalize may allow him to recover the representation "did you" from its phonetic realization as [dɪ ʃ yu]. In this way, much of the task of translating acoustic-phonetic information into some type of segmental representation is simplified by knowledge of a few general rules that account for much of the variability at the phonetic level (see Church, 1987, this issue).

5. Spoken word recognition and lexical access

Among the most important recent trends in the field of speech perception has been the increased interest in theories of auditory word recognition and lexical access. Although much basic work is still being conducted on fundamental problems in acoustic-phonetics, many researchers have begun to expand their domain of inquiry to include the processes by which spoken words are recognized and meanings are retrieved from long-term memory. In our view, speech perception is not synonymous with phoneme perception, although much of the early work emphasized this orientation. There can be little doubt in anyone's mind that speech perception is an extremely complex process involving many levels of processing from phonetic perception to semantic interpretation. To isolate one level of processing for investigation, while ignoring the possible contributions of and interaction with other levels, is, in our view, somewhat myopic, and may lead to grossly incorrect theories. What we learn about word recognition, for example, may inform our theories of phonetic perception, and vice versa. Of course, analysis of the speech perception process is made much easier by the division of our domain of inquiry into isolatable subcomponents. However, investigating one subcomponent (e.g., phonetic perception) to the exclusion of others (e.g., word recognition) would appear to limit our insights into the process as a whole, as well as lead us to postulate theories at one level that are clearly untenable or unparsimonious given what we know about processing at other levels of analysis.

A good deal of the work carried out over the last thirty-five years in speech perception has been concerned with the "primary recognition problem"; that is, how the *form* of a spoken utterance is recognized or identified from an analysis of the acoustic waveform (Fry, 1956). Conscious identification and awareness of all of the segments in a word is probably not necessary or even obligatory for word recognition to take place, although it is certainly possible under special circumstances when a listener's attention is directed specifically to the sound structure of an utterance. Under normal listening conditions, the human listener may not have to *identify* all of the phonetic input to recognize the words in an utterance. Context and other constraints can serve to narrow down the available choices so that only a small portion of the acoustic waveform need be identified for word recognition to take place successfully.

Although the theories of word recognition and lexical access that we discuss below are too vague to render any significant insights into the nature of speech sound perception at this time, they are indicative of a growing trend to consider speech perception in a broader framework of spoken language processing. In addition, these theories represent what might be called a "new" interest among some speech researchers, namely, the way in which acoustic-phonetic information is used to contact lexical items in long-term memory. In this final section, we briefly review five current approaches to word recognition and lexical access. Each of these accounts was proposed to deal with somewhat different empirical issues in word recognition and lexical access, but each addresses a current topic of some interest in word recognition, namely, the extent to which higher-level knowledge sources come to bear on the perception of spoken words. Another issue addressed, in part at least, by

each of these theories is the means by which lexical items are activated in memory. Here we are interested in specifying the nature of the bottom-up input and the processing units assumed by each theory. More specifically, we are interested in the degree to which these models assume the existence of segmental representations either tacitly or explicitly in order to solve the primary recognition problem.

Throughout the following discussion, we make a distinction between word recognition and lexical access. When speaking of *word recognition,* we refer explicitly to those processes responsible for generating a pattern from the acoustic-phonetic information in the speech waveform and matching this pattern to patterns previously stored in memory (i.e., for words) or to patterns generated by rule (i.e., for pseudowords). Word recognition, in our view, is synonymous with the term *form perception* as discussed by Bradley and Forster (1987, this issue) and *phonetic perception* as discussed by Liberman et al. (1967) and Studdert-Kennedy (1974, 1980). When speaking of *lexical access,* we refer explicitly to those processes that are responsible for contacting the appropriate lexical information in memory once a pattern match has been accomplished. Lexical access, then, is that process by which information about words stored in the mental lexicon is retrieved. More detailed discussions of several of these models can be found in other contributions to this issue.

It should be clear from the distinction we have drawn between the processes of word recognition and lexical access that most contemporary models of word recognition that claim to be concerned with lexical access are actually models of word recognition. Little, if any, work has been devoted to describing the structure and organization of the mental lexicon (see however Johnson-Laird, 1975; Miller & Johnson-Laird, 1976). Moreover, it should also be obvious that assumptions about word recognition and the input to the lexicon are probably not independent from assumptions made about the structure and organization of the lexicon itself (see Bradley & Forster, 1987, this issue; Luce, 1986). Indeed, it may be difficult or impossible to separate the processes of word recognition and lexical access from their products. To take one example, segmentation of the speech waveform into discrete linguistic units such as phonemes or words has always been a troublesome problem to deal with because of the continuous nature of the speech signal. However, segmentation may very well be a natural by-product of the recognition process itself (Reddy, 1976). As we learn more about the sources of variability in speech, it is becoming clear that the variability in the speech waveform is extremely systematic and potentially quite useful to the recognition process (Church, 1983; Elman & McClelland, 1984). Indeed, the recent findings of Church have demonstrated how knowledge of allophonic variation can aid in phonological parsing and lexical retrieval and therefore reduce the search process in locating the correct lexical entry (see Church, 1987, this issue).

6. Models of word recognition

6.1. Logogen theory

In Morton's (1969, 1979, 1982) logogen theory, passive sensing devices called "logogens" represent each word in the mental lexicon. Each logogen contains all of the information about a given word, such as its meaning, its possible syntactic functions, and its phonetic and orthographic structure. A logogen monitors for relevant sensory and/or contextual information and, once such information is encountered, the activation level of the logogen is raised. Upon sufficient activation, a logogen crosses threshold, at which time the information about the word that the logogen represents is made available to the response system.

One important feature of the logogen theory is that logogens monitor all possible sources of information, including higher-level semantic and syntactic information as well as lower

level sensory information. Thus, information from any level can combine to push a logogen over its threshold. In this way, logogen theory is a highly *interactive* model of word recognition. For example, a word of high frequency, which has a starting threshold lower than words of lower frequency, may require very little sensory input if syntactic and semantic sources of information strongly favor the word. Likewise, a word of low frequency with few associated higher-level expectations may require considerable sensory input for the activation level to reach threshold. Thus, it may not really matter what sort of information activates a logogen, so long as the threshold is exceeded.

According to logogen theory, word recognition is accomplished when the activation threshold of a logogen is reached. As we have seen, logogen theory portrays the word recognition process as highly interactive. Lexical access, in our terminology, is achieved when the information contained within the logogen is made available to the response system. Thus, lexical access is a fairly automatic process once the word has been recognized. It is of interest to note that not only are interactive knowledge sources at play at the level of word recognition, but word frequency is handled at this stage as well. Words of higher frequency have lower activation thresholds than those of lower frequency (see, however, Luce, 1986).

The specific details of logogen theory have changed somewhat over the years, although the basic mechanisms have remained unchanged. For example, Morton (1982) has recently broken the logogen system into separate visual and auditory subsystems. Nevertheless, the fundamental notion of a passive threshold device that monitors information from a variety of sources has remained. As it stands, logogen theory, like many of the theories we will discuss, is extremely vague. At best, the theory helps to conceptualize how an interactive system may work and how word frequency and contextual effects in word recognition may be accounted for. However, the theory says very little, if anything, about precisely how acoustic-phonetic and higher-level sources of information are integrated, the time-course of word recognition, the nature of the perceptual units, or the role of the lexicon in word recognition.

6.2. Cohort theory

Marslen-Wilson's (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978) cohort theory posits two stages in the word recognition process, one autonomous and one interactive. In the first, autonomous stage of word recognition, acoustic-phonetic information at the beginning of an input word activates all words in memory that share this word-initial information. For example, if the word "slave" is presented to the system, all words beginning with /s/ are activated, such as "sight," "save," "sling," and so on. The words activated on the basis of word-initial information comprise the "cohort." Activation of the cohort is an autonomous process in the sense that *only* acoustic-phonetic information can serve to specify the members of a cohort. At this stage of the model, then, word recognition is a completely data-driven or bottom-up process.

The key to this approach is the notion of a set of "word initial cohorts" or recognition candidates which are defined by the acoustic-phonetic commonality of the initial sound sequences of words. A particular word is "recognized" at that point—the "critical recognition point"—where the word is *uniquely* distinguished from any other word in the language beginning with the same *initial sound sequence.* The theory accounts for the facilitatory effects of context in word recognition by assuming, as in Morton's logogen model, that context influences the recognition of a particular word. However, unlike the logogen model, context is used to deactivate candidate words and therefore reduce the size of a word initial cohort set that is active at any time. The interaction of context with the sensory input is assumed to occur at the level of word recognition (see Marslen-Wilson &

Welsh, 1978). Processing at early sensory levels is assumed to occur automatically and is not influenced by other higher-order sources of information. In cohort theory, the set of word initial cohorts that is activated is defined, in principle, by the internal segmental structure of the linear arrangement of speech sounds. To say, as Marslen-Wilson has done, that his theory makes no claim about the structure of the input to the word recognition system in terms of segmental representations is simply to deny the existence of a prima facie assumption that is central to the organization of his word initial cohort set. The theory, as currently formulated, would never work if the internal structure of words could not be described formally as a sequence of segment-like units.

Once a cohort structure is activated, all possible sources of information may come to bear on selection of the appropriate word from the cohort. Thus, further acoustic-phonetic information may eliminate “sight” and “save” from the cohort, leaving only words that begin with /sl/, such as “sling” and “slave.” Note that word recognition based on acoustic-phonetic information is assumed to operate in a strictly left-to-right fashion. At this stage of word recognition, however, higher-level sources of information may also come into play to eliminate candidates from the set of hypothesized word cohorts. Thus, if “sling” is inconsistent with the presently available semantic or syntactic information, it will be eliminated from the cohort. At this second stage of word recognition, the theory is highly interactive. Upon isolation of a single word in the cohort, word recognition is accomplished.

Marslen-Wilson’s cohort theory has attracted a considerable amount of attention in the last few years, presumably because of its relatively precise description of the word recognition process, its novel claim that all words in the mental lexicon sharing initial acoustic-phonetic information with the input word are activated in the initial stage of the word recognition process, and because of the priority it affords to the beginnings of words, a popular notion in the literature (see also Cole & Jakimik, 1980).

The theory is not without its shortcomings, however. For example, the original version of cohort theory incorporates no mechanism by which word frequency can be accounted for (see, however, Marslen-Wilson, 1987, this issue). Do high frequency words have higher activation levels in the cohort structure or are high frequency words simply more likely to be selected as candidates for a cohort than low frequency words? This last possibility seems unlikely, for the system would then be hard pressed to account for recognition of low frequency words that may be excluded a priori from the cohort structure. Perhaps associating various activation levels with word candidates would be more appropriate, but the theory as it stands has no means of accounting for differential activation levels.

Another problem with cohort theory is error recovery. For example, if “foundation” is perceived as “thoundation,” due to a mispronunciation or misperception, the word-initial cohort will not contain the word candidate “foundation.” Marslen-Wilson allows for some *residual* activation of acoustically similar word candidates in the cohort structure so a second pass through the cohort structure may be possible to attempt a best match, but as it currently stands the theory does not specify how such off-line error recovery may be accomplished.

6.3. Forster’s autonomous search model

In contrast to Morton’s logogen theory and Marslen-Wilson’s cohort theory, Forster’s (1976, 1979) theory of word recognition and lexical access is autonomous in the strictest sense. Whereas Morton and Marslen-Wilson allow parallel processing of information at some stage, in Forster’s theory, linguistic processing is completely serial. Forster’s theory posits three separate linguistic processors: a lexical processor, a syntactic processor, and a message processor. In addition, the latest version of Forster’s theory incorporates a third, non-linguistic processor, the General Processing System (GPS).

In the first stage of Forster's model, information from peripheral perceptual systems is submitted to the lexical processor. The lexical processor then attempts to locate an entry in three peripheral access files: an orthographic file (for visual input), a phonetic file (for auditory input), and syntactic-semantic file (for both visual and auditory input). Search of the peripheral access files is assumed to proceed by frequency, with higher frequency words being searched prior to lower frequency words. Once an entry is located in the peripheral access files, a pointer is retrieved by which the entry is located in the master lexicon. Thus, word recognition is accomplished at the level of the peripheral access files. Once an entry is located in the peripheral files, lexical access is accomplished by locating the entry in the master lexicon.

Upon location of an item in the master lexicon, information regarding the location of that item in the master list is passed on to the syntactic processor, which attempts to build a syntactic structure. From the syntactic processor information is passed to the message processor which attempts to build a conceptual structure for the intended message. Each of the three processors—the lexical processor, the syntactic processor, and the message processor—can pass information to the GPS. However, the GPS cannot influence processing in any of the three dedicated linguistic processors. The GPS serves to incorporate general conceptual knowledge with the output of the information from the linguistic processors in making a decision (or response).

Forster's theory is autonomous in two senses. First, the lexical processor is independent of the syntactic and message processors, and the syntactic processor is independent of the message processor. Second, the entire linguistic system is independent of the general cognitive system. This strictly serial and autonomous characterization of language processing means that word recognition and lexical access are in no way influenced by higher-level knowledge sources and are exclusively bottom-up or data-driven processes. Forster's model is attractive because of its relative specificity and the apparently testable claims it makes regarding the autonomy of its processors. Forster's model also attempts to describe word recognition and lexical access in the context of sentence processing. In addition, the model incorporates a specific explanation of the word frequency effect, namely that entries in the peripheral access files are organized according to frequency and that search proceeds from high to low frequency entries.

6.4. Elman and McClelland's interactive-activation theory

Elman and McClelland's (1984, 1986) model is based on a system of simple processing units called "nodes." Nodes may stand for features, phonemes, or words. However, nodes at each level are alike in that each has an activation level representing the degree to which the input is consistent with the unit the node stands for. In addition, each node has a resting level and a threshold. In the presence of confirmatory evidence, the activation level of a node rises toward its threshold; in the absence of such evidence, activation decays toward the resting level of the node (see also McClelland & Rumelhart, 1981).

Nodes within this system are highly interconnected and when a given node reaches threshold, it may then influence other nodes to which it is connected. Connections *between* nodes are of two types: excitatory and inhibitory. Thus, a node that has reached threshold may raise the activation of some of the nodes to which it is connected while lowering the activation of other nodes. Connections between levels are exclusively excitatory and are bidirectional. Thus, phoneme nodes may excite word nodes and word nodes may in turn excite phoneme nodes. For example, the phoneme nodes corresponding to /l/ and /e/ may excite the word node "lake," and the word node "lake" may then excite the phoneme nodes /l/, /e/ and /k/. Connections *within* levels are inhibitory and bidirectional. Thus, activation of

the phoneme node /l/ may inhibit activation of the phoneme node /b/, lowering the probability that the word node “bake” will raise its activation level.

The Elman and McClelland model illustrates how a highly interactive system may be conceptualized (see also McClelland & Elman, 1986). In addition, it incorporates notions of excitation *and* inhibition. By so doing, it directly incorporates a mechanism that reduces the possibility that nodes inconsistent with the evidence will be activated while at the same time allowing for positive evidence at one level to influence activation of nodes at another. Although Elman and McClelland’s model is very interactive, it is not without constraints. Namely, connections between levels are only excitatory and within levels only inhibitory.

Elman and McClelland’s model explicitly assumes a segmental representation for speech. The entire organization of the network is based on the existence of different processing units at each level corresponding to acoustic-phonetic features or cues, segmental phonemes and finally words. Because of the architecture of the system, words have a much more complex structure than other elements. Thus, word nodes not only reflect activation of the word as a whole but also activation of each of the constituent phonemes of the word and their component features.

There are two interesting features of this model that are worth noting here. First, although coarticulation effects and contextual variability have been considered by theorists as “noise” that is imposed on an ideal discrete phonetic transcription of speech by the speech production apparatus, Elman and McClelland’s model treats this variability, what they call “lawful variability,” as a source of useful information and provides a “graceful” way to account for the effects of context in speech perception (Elman & McClelland, 1984). Second, there is no explicit segmentation of the input speech waveform at any time during processing in their model. The segmentation into phones or allophones simply falls out naturally as a result of the labeling process itself. Thus, the problem of dealing with segmentation directly is avoided by permitting the activation of all feature and phoneme nodes and simply observing the consequences at the word level.

6.5. Klatt’s LAFS model

Whereas Elman and McClelland’s model allows for interaction between and within levels of nodes, Klatt’s *L*exical *A*ccess *F*rom *S*pectra (LAFS) model assumes direct, noninteractive access of lexical entries based on context-sensitive spectral sections (Klatt, 1980). Klatt’s model assumes that adult listeners have a dictionary of all lawful diphone sequences in long-term memory. Associated with each diphone sequence is a prototypical spectral representation. Klatt proposes spectral representations of diphone sequences to overcome the contextual variability of individual segments. To a certain extent, then, Klatt tries to overcome the problem of the lack of acoustic-phonetic invariance in speech by precompiling coarticulatory effects directly into the representations residing in memory.

In Klatt’s LAFS model, the listener computes spectral representations of an input word and compares these representations to the prototypes in memory. Word recognition is accomplished when a best match is found between the input spectra and the diphone representations. In this portion of the model, word recognition is accomplished directly on the basis of spectral representations of the sensory input. There is a means by which phonetic transcriptions can be obtained intermediate to lexical access (i.e., via the SCRIBER module), but in most circumstances access is direct, with no intermediate levels of computation corresponding to segments or phonemes.

One important aspect of Klatt’s LAFS model is that it explicitly avoids any need to compute a distinct level of representation corresponding to discrete phonemic segments. Instead,

LAFS uses a precompiled, acoustically-based lexicon of all possible words in a network of diphone power spectra. These spectral templates are assumed to be context-sensitive units much like “Wick-elphones” because they are assumed to represent the acoustic correlates of phones in different phonetic environments (Wickelgren, 1969). Diphones in the LAFS system accomplish this by encoding the spectral characteristics of the segments themselves and the transitions from the middle of one segment to the middle of the next segment.

Klatt argues that diphone concatenation is sufficient to capture much of the context-dependent variability observed for phonetic segments in spoken words. Word recognition in this model is accomplished by computing a power spectrum of the input speech signal every 10 ms and then comparing this input spectrum to spectral templates stored in a precompiled network. The basic idea of LAFS, adapted from the Harpy system, is to find the path through the network that best represents the observed input spectra (Klatt, 1977). This single path is then assumed to represent the optimal phonetic transcription of the input signal.

Elman and McClelland’s and Klatt’s models fall on either end of a continuum of theories of word recognition and lexical access. Elman and McClelland’s theory represents the class of theories that emphasize *interactive* systems in which many different levels of information play a role in word recognition and lexical access. In this sense, their model is closest to those of Morton and Marslen-Wilson, although Marslen-Wilson’s cohort theory does incorporate an initial autonomous stage of processing. Klatt’s model, on the other hand, represents the class of models in which lexical access is accomplished almost entirely on the basis of bottom-up acoustic-phonetic information. In this sense, Klatt’s model resembles Forster’s approach. However, Forster’s model does posit intermediate levels of analysis in the word recognition process, unlike Klatt’s LAFS, which assumes *direct* mapping of power spectra onto words in a precompiled network. One of the central questions to be addressed with regard to current theories of word recognition involves the extent to which word recognition involves interactive knowledge sources and the manner in which these processes interact with processes involved in speech sound perception (see other contributions to this issue).

7. Summary and conclusions

In this paper, we have attempted to review briefly what is currently known about the nature of the input to the word recognition system provided by mechanisms employed in speech sound perception. After considering several of the basic issues in speech perception such as linearity, invariance and segmentation, we described several stages of perceptual analysis within a conceptual framework. This framework assumed that speech is processed through a series of analytic stages ranging from peripheral auditory processing, acoustic-phonetic and phonological analysis to word recognition and lexical access. Finally, we examined several contemporary approaches to word recognition in order to make explicit some of the major assumptions regarding the nature of the input to the word recognition process. An additional goal of the paper was to establish the need for segmental phonemic representations in spoken word recognition. This is the point in spoken language processing that serves to interface the initial sensory information in the speech waveform with the representation of words in the lexicon. An examination of current models revealed the extent to which segmental representations are assumed either explicitly or tacitly in mediating word recognition and lexical access.

References

Abbs JH, Sussman HM. Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. Journal of Speech and Hearing Research. 1971; 14:23–36. [PubMed: 4994487]

Abramson, AS.; Lisker, L. Voice onset time in stop consonants: Acoustic analysis and synthesis. Proceedings of the 5th International Congress of Acoustics; Liege. 1965.

Allen J. A perspective on man-machine communication by speech. Proceedings of the IEEE. 1985; 74(11):1541–1550.

Aslin, RN.; Pisoni, DB.; Jusczyk, PW. Auditory development and speech perception in infancy. In: Mussen, P., editor. Carmichael's manual of child psychology. 4. Vol. 2 . 1983. Haith, MM.; Campos, JJ., editors. Infancy and the biology of development. Vol. 2. New York: Wiley;

Bever TG, Lackner J, Kirk R. The underlying structures of sentences are the primary units of immediate speech processing. Perception & Psychophysics. 1969; 5:225–231.

Bond, ZS.; Garnes, S. Misperceptions of fluent speech. In: Cole, RA., editor. Perception and production of fluent speech. Hillsdale, N.J: Erlbaum; 1980.

Bond, ZS.; Robey, RR. The phonetic structure of errors in the perception of fluent speech. In: Lass, NJ., editor. Speech and language: Advances in basic research and practice. Vol. 9. New York: Academic Press; 1983.

Bradley DC, Forster KI. A reader's view of listening. Cognition. 1987; 25 this issue.

Carlson, R.; Granstrom, B., editors. The representation of speech in the peripheral auditory system. New York: Elsevier Biomedical Press; 1982.

Chomsky, N.; Halle, M. The sound pattern of English. New York: Harper and Row; 1968.

Chomsky, N.; Miller, GA. Introduction to formal analysis of natural languages. In: Luce, RD.; Bush, R.; Galanter, E., editors. Handbook of mathematical psychology. Vol. 2. New York: Wiley; 1963.

Church, KW. Phrase-structure parsing: A method for taking advantage of allophonic constraints. Bloomington, Ind: Indiana University Linguistics Club; 1983.

Church KW. Phonological parsing and lexical retrieval. Cognition. 1987; 25 this issue.

Clements, GN.; Keyser, SJ. CV phonology: A generative theory of the syllable. Cambridge, Mass: MIT Press; 1983.

Cole RA. Listening for mispronunciations: a measure of what we hear during speech. Perception & Psychophysics. 1973; 13:153–156.

Cole, RA.; Jakimik, J. Understanding speech: How words are heard. In: Underwood, G., editor. Strategies of information processing. New York: Academic Press; 1978.

Cole, RA.; Jakimik, J. A model of speech perception. In: Cole, RA., editor. Perception and production of fluent speech. Hillsdale, N.J: Erlbaum; 1980.

Cole RA, Scott B. The phantom in the phoneme: Invariant cues for stop consonants. Perception & Psychophysics. 1974a; 15:101–107.

Cole RA, Scott B. Toward a theory of speech perception. Psychological Review. 1974b; 81:348–374. [PubMed: 4607301]

Conrad R. Acoustic confusions in immediate memory. British Journal of Psychology. 1964; 55:75–84.

Cooper, WE. Selective adaptation to speech. In: Restle, F.; Shiffrin, RM.; Castellan, NJ.; Lindman, HR.; Pisoni, DB., editors. Cognitive theory. Vol. 1. Hillsdale, N.J: Erlbaum; 1975.

Darwin, DJ. The perception of speech. In: Carterette, EC.; Friedman, MP., editors. Handbook of perception. New York: Academic Press; 1976.

Delattre PC, Liberman AM, Cooper FS. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America. 1955; 27:769–773.

Delgutte B. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America. 1980; 68:843–857. [PubMed: 7419820]

Delgutte, B. Some correlates of phonetic distinctions at the level of the auditory nerve. In: Carlson, R.; Granstrom, B., editors. The representation of speech in the peripheral auditory system. New York: Elsevier Biomedical Press; 1982.

Dell GS, Newman JE. Detecting phonemes in fluent speech. Journal of Verbal Learning and Verbal Behavior. 1980; 19:608–623.

Denes P. Effect of duration on the perception of voicing. Journal of the Acoustical Society of America. 1955; 27:761–764.

Diehl RL. Feature detectors for speech: A critical reappraisal. Psychological Bulletin. 1981; 89:1–18. [PubMed: 7195048]

Dorman M, Studdert-Kennedy M, Raphael L. Stop consonant recognition: Release bursts and formant transitions as functionally equivalent context-dependent cues. Perception & Psychophysics. 1977; 22:109–122.

Eimas PD. Auditory and linguistic processing of cues for place of articulation by infants. Perception & Psychophysics. 1974; 16:513–521.

Eimas PD, Corbit JD. Selective adaptation of linguistic feature detectors. Cognitive Psychology. 1973; 4:99–109.

Eimas, PD.; Miller, JL. Effects of selective adaptation on the perception of speech and visual patterns: Evidence for feature detectors. In: Pick, HL.; Walk, RD., editors. Perception and experience. New York: Plenum; 1978.

Eimas PD, Siqueland ER, Jusczyk P, Vigorito J. Speech perception in infants. Science. 1971; 171:303–306. [PubMed: 5538846]

Elman, JL.; McClelland, JL. Exploiting lawful variability in the speech waveform. In: Perkell, JS.; Klatt, DH., editors. lnvariance and variability in speech processes. Hillsdale, N.J: Erlbaum; 1986.

Elman, JL.; McClelland, JL. Speech perception as a cognitive process: The interactive activation model. In: Lass, NJ., editor. Speech and language: Advances in basic research and practice. Vol. 10. New York: Academic Press; 1984. p. 337-374.

Fant G. Descriptive analysis of the acoustic aspects of speech. Logos. 1962; 5:3–17. [PubMed: 13891546]

Fant, G. Speech sounds and features. Cambridge, Mass: MIT Press; 1973.

Fischer-Jorgensen E. Acoustic analysis of stop consonants. Miscellanea Phonetica. 1954; 2:42–59.

Fitch, HL. Haskins Laboratories Status Report on Speech Research. Vol. SR-65 . New Haven: Haskins Laboratories; 1981. Distinguishing temporal information for speaking rate from temporal information for intervocalic stop consonant voicing; p. 1-32.

Forster, KI. Accessing the mental lexicon. In: Wales, RJ.; Walker, E., editors. New approaches to language mechanisms. Amsterdam: North-Holland; 1976.

Forster, KI. Levels of processing and the structure of the language processor. In: Cooper, WE.; Walker, ECT., editors. Sentence processing: Psycholinguistic studies presented to Merrill Garrett. Hillsdale, N.J: Erlbaum; 1979.

Foss DJ, Blank MA. Identifying the speech codes. Cognitive Psychology. 1980; 12:1–31. [PubMed: 7351123]

Foss DJ, Gernsbacher MA. Cracking the dual code: Towards a unitary model of phoneme identification. Journal of Verbal Learning and Verbal Behavior. 1983; 22:609–632.

Foss, DJ.; Harwood, DA.; Blank, MA. Deciphering decoding decisions: Data and devices. In: Cole, RA., editor. Perception and production of fluent speech. Hillsdale, N.J: Erlbaum; 1980.

Foss DJ, Swinney DA. On the psychological reality of the phoneme: Perception, identification, and consciousness. Journal of Verbal Learning and Verbal Behavior. 1973; 12:246–257.

Fromkin, V. Speech errors as linguistic evidence. The Hague: Mouton; 1973.

Fromkin, V. Errors in linguistic performance. New York: Academic Press; 1980.

Fry, DW. Perception and recognition. In: Halle, M.; Lunt, HG.; McClean, H.; van Schoonefeld, CH., editors. For Roman Jakobson. The Hague: Mouton; 1956.

Ganong WF. The selective effects of burst-cued stops. Perception & Psychophysics. 1978; 24:71–83. [PubMed: 693243]

Garrett, MF. Syntactic processes in sentence production. In: Wales, RJ.; Walker, E., editors. New approaches to language mechanisms. Amsterdam: North-Holland; 1976.

Garrett, MF. Levels of processing in sentence production. In: Butterworth, B., editor. Language production. Vol. 1. New York: Academic Press; 1980.

Goldhor R. A speech signal processing system based on a peripheral auditory model. Proceedings of IEEE. 1983; 1CASSP-83:1368–1371.

Halle, M. Speculations about the representation of words in memory. In: Fromkin, V., editor. Linguistic phonetics: Papers presented to Peter Ladefoged. New York: Academic Press; 1985.

Halle M, Vergnaud JR. Three dimensional phonology. Journal of Linguistic Research. 1980; 1:83–105.

House AS. On vowel duration. Journal of the Acoustical Society of America. 1961; 33:1174–1178.

Johnson-Laird, PN. Meaning and the mental lexicon. In: Kennedy, A.; Wilkes, A., editors. Studies in long-term memory. London: Wiley; 1975. p. 123-142.

Joos MA. Acoustic phonetics. Language. 1948; 24(Suppl):1–136.

Kahn, D. Syllable-based generalizations in English phonology. Bloomington, Ind: Indiana University Linguistics Club; 1976.

Kenstowicz, M.; Kisseberth, C. Generative phonology. New York: Academic Press; 1979.

Klatt DH. Structure of confusions in short-term memory between English consonants. Journal of the Acoustical Society of America. 1968; 44:401–407. [PubMed: 5665520]

Klatt DH. Review of the ARPA speech understanding project. Journal of the Acoustical Society of America. 1977; 62:1345–1366.

Klatt DH. Speech perception: A model of acoustic-phonetic analysis and lexical access. Journal of Phonetics. 1979; 7:279–312.

Klatt, DH. Speech perception: A model of acoustic-phonetic analysis and lexical access. In: Cole, RA., editor. Perception and production of fluent speech. Hillsdale, N.J: Erlbaum; 1980.

Klatt, DH. Speech processing strategies based on auditory models. In: Carlson, R.; Granstrom, B., editors. The representation of speech in the peripheral auditory system. New York: Elsevier Biomedical Press; 1982.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 74:431–461. [PubMed: 4170865]

Liberman AM, Delattre PC, Cooper FS. The role of selected stimulus variables in the perception of the unvoiced stop consonants. American Journal of Psychology. 1952; 52:127–137.

Liberman AM, Delattre PC, Cooper FS, Gerstman LH. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs. 1954; 68:1–13.

Liberman IY, Shankweiler D, Fischer FW, Carter B. Explicit syllable and phoneme segmentation in the young child. Journal of Experimental Child Psychology. 1974; 18:201–212.

Lisker L. Closure duration and the intervocalic voiced-voiceless distinction in English. Language. 1957; 33:42–49.

Lisker, L. Haskins Laboratories Status Report on Speech Research. Vol. SR-65 . New Haven: Haskins Laboratories; 1978. Rapid vs rabid: A catalogue of acoustic features that may cue the distinction; p. 127-132.

Lisker L, Abramson AS. A cross language study of voicing in initial stops: Acoustical measurements. Word. 1964; 20:384–422.

Luce, PA. Research on Speech Perception, Technical Report No 6. Bloomington, Ind: Department of Psychology, Speech Research Laboratory; 1986. Neighborhoods of words in the mental lexicon.

Marslen-Wilson WD. Parallel processing in spoken word recognition. Cognition. 1987; 25 this issue.

Marslen-Wilson WD, Tyler LK. The temporal structure of spoken language understanding. Cognition. 1980; 8:1–71. [PubMed: 7363578]

Marslen-Wilson WD, Welsh A. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology. 1978; 10:29–63.

Massaro, DW.; Oden, GC. Speech perception: A framework for research and theory. In: Lass, NJ., editor. Speech and language: Advances in basic research and practice. Vol. 3. New York: Academic Press; 1980. p. 129-165.

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18:1–86. [PubMed: 3753912]

McClelland JL, Rumelhart DE. An interactive-activation model of context effects in letter perception, Part I: An account of basic findings. Psychological Review. 1981; 88:375–407.

Miller GA. Decision units in the perception of speech. IRE Transactions on Information Theory. 1962; IT-8:81–83.

Miller, GA.; Johnson-Laird, PN. Language and perception. Cambridge, Mass: Harvard University Press; 1976.

Miller GA, Nicely PE. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America. 1955; 27:338–352.

Miller MI, Sachs MB. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America. 1983; 74:502–517. [PubMed: 6619427]

Morton J. Interaction of information in word recognition. Psychological Review. 1969; 76:165–178.

Morton, J. Word recognition. In: Morton, J.; Marshall, JD., editors. Psycholinguistics 2: Structures and processes. Cambridge, Mass: MIT Press; 1979. p. 107-156.

Morton, J. Disintegrating the lexicon: An information processing approach. In: Mehler, J.; Walker, E.; Garrett, M., editors. On mental representation. Hillsdale, N.J: Erlbaum; 1982.

Morton J, Long J. Effect of word transitional probability on phoneme identification. Journal of Verbal Learning and Verbal Behavior. 1976; 15:43–52.

Oshika BT, Zue VW, Weeks RV, Neu H, Aurbach J. The role of phonological rules in speech understanding research. IEEE Transactions on Acoustics Speech, and Signal Processing, ASSP-. 1975; 23:104–112.

Pisoni, DB. Speech perception. In: Estes, WK., editor. Handbook of learning and cognitive processes. Vol. 6. Hillsdale, N.J: Erlbaum; 1978.

Pisoni DB. In defense of segmental representations in speech processing. Journal of the Acoustical Society of America. 1983; 69:S32.

Pisoni DB. Speech perception: Some new directions in research and theory. Journal of the Acoustical Society of America. 1985; 78:381–388. [PubMed: 4031245]

Pisoni DB, Nusbaum HC, Luce PA, Slowiaczek LM. Speech perception, word recognition and the structure of the lexicon. Speech Communication. 1985; 4:75–95.

Pisoni, DB.; Sawusch, JR. Some stages of processing in speech perception. In: Cohen, A.; Nooteboom, S., editors. Structure and process in speech perception. Heidelberg: Springer-Verlag; 1975. p. 16-34.

Port, RF. The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words. Bloomington, Ind: Indiana University Linguistics Club; 1977.

Port RF. Influence of tempo on stop closure duration as a cue for voicing and place. Journal of Phonetics. 1979; 7:45–56.

Raphael LJ. Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. Journal of the Acoustical Society of America. 1972; 51:1296–1303. [PubMed: 5032946]

Raphael LJ, Dorman MF. Silence as a cue to the perception of syllable-initial and syllable-final stop consonants. Journal of Phonetics. 1980; 8:269–275.

Read C. Preschool children's knowledge of English phonology. Harvard Educational Review. 1971; 41:1–34.

Reddy DR. Speech recognition by machine: A review. Proceedings of the IEEE. 1976; 64:501–523.

Remez RE. Adaptation of the category boundary between speech and nonspeech: A case against feature detectors. Cognitive Psychology. 1979; 11:38–57. [PubMed: 761448]

Rozin, P.; Gleitman, LR. The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In: Reber, AS.; Scarborough, DL., editors. Toward a psychology of reading. Hillsdale, N.J: Erlbaum; 1977.

Sachs MB, Young ED. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. Journal of the Acoustical Society of America. 1979; 66:470–479. [PubMed: 512208]

Samuel AG. Phonemic restoration: Insights from a new methodology. Journal of Experimental Psychology: General. 1981a; 110:474–494. [PubMed: 6459403]

Samuel AG. The role of bottom-up confirmation in the phonemic restoration illusion. Journal of Experimental Psychology: Human Perception and Performance. 1981b; 7:1124–1131. [PubMed: 6457110]

Sapir, E. The psychological reality of phonemes. In: Mandelbaum, D., editor. Selected writings of Edward Sapir. Berkeley: University of California Press; 1963.

Sawusch JR. Peripheral and central processes in selective adaptation of place of articulation in stop consonants. Journal of the Acoustical Society of America. 1977a; 62:738–750. [PubMed: 903514]

Sawusch JR. Processing place information in stop consonants. Perception & Psychophysics. 1977b; 22:417–426.

Sawusch JR, Jusczyk PW. Adaptation and contrast in the perception of voicing. Journal of Experimental Psychology: Human Perception and Performance. 1981; 7:408–421. [PubMed: 6453933]

Searle CL, Jacobson JF, Rayment SG. Stop consonant discrimination based on human audition. Journal of the Acoustical Society of America. 1979; 65:799–809. [PubMed: 447910]

Selkirk E. The role of prosodic categories in English word stress. Linguistic Inquiry. 1980; 11:563–603.

Shattuck-Hufnagel S, Klatt DH. The limited use of distinctive features and markedness in speech production: Evidence from speech error data. Journal of Verbal Learning and Verbal Behavior. 1979; 18:41–45.

Stemberger, JP. Unpublished doctoral dissertation. University of California; San Diego: 1982. The lexicon in a model of language production.

Stevens KN. Acoustic correlates of some phonetic categories. Journal of the Acoustical Society of America. 1980; 68:836–842. [PubMed: 7419819]

Stevens KN, Blumstein SE. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America. 1978; 64:1358–1368. [PubMed: 744836]

Stevens, KN.; Blumstein, SE. The search for invariant acoustic correlates of phonetic features. In: Eimas, PD.; Miller, JL., editors. Perspectives on the study of speech. Hillsdale, N.J: Erlbaum; 1981.

Stevens KN, Klatt DH. Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America. 1974; 55:653–659. [PubMed: 4819867]

Stevens, KN.; House, AS. Speech perception. In: Tobias, J., editor. Foundations of modern auditory theory. Vol. II. New York: Academic Press; 1972.

Studdert-Kennedy, M. The perception of speech. In: Sebeok, TA., editor. Current trends in linguistics. The Hague: Mouton; 1974.

Studdert-Kennedy, M. Speech perception. In: Lass, NJ., editor. Contemporary issues in experimental phonetics. New York: Academic Press; 1976.

Studdert-Kennedy M. Speech perception. Language and Speech. 1980; 23:45–66. [PubMed: 6999263]

Summerfield Q, Haggard MP. Perceptual processing of multiple cues and contexts: Effects of following vowel upon stop consonant voicing. Journal of Phonetics. 1974; 2:279–295.

Treiman, R. Unpublished doctoral dissertation. University of Pennsylvania; 1980. The phonemic analysis ability of preschool children.

Treiman R. The structure of spoken syllables: Evidence from novel word games. Cognition. 1983; 15:49–74. [PubMed: 6686514]

Treiman R. Onsets and rimes as units of spoken syllables: Evidence from children. Journal of Experimental Child Psychology. 1985; 39:161–181. [PubMed: 3989458]

Wang MD, Bilger RC. Consonant confusions in noise: A study of perceptual features. Journal of the Acoustical Society of America. 1973; 54:1248–1266. [PubMed: 4765809]

Warren RM. Perceptual restoration of missing speech sounds. Science. 1970; 176:392–393. [PubMed: 5409744]

Wickelgren WA. Acoustic similarity and retroactive interference in short-term memory. Journal of Verbal Learning and Verbal Behavior. 1965; 4:53–61.

Wickelgren WA. Phonemic similarity and interference in short-term memory for single letters. Journal of Experimental Psychology. 1966; 71:396–404. [PubMed: 5908822]

Wickelgren WA. Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychological Review. 1969; 76:1–15.

Wickelgren, WA. Phonetic coding and serial order. In: Carterette, EC.; Friedman, MP., editors. Handbook of perception. Vol. 7. New York: Academic Press; 1976.

Young ED, Sachs MB. 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. 1979; 66:1381–1403. [PubMed: 500976]

Zwicker E, Terhardt E, Paulus E. Automatic speech recognition using psychoacoustic models. Journal of the Acoustical Society of America. 1979; 65:487–498. [PubMed: 489818]