RESEARCH ON SPEECH PERCEPTION

Technical Report No 1.

June 15, 1976

David B. Pisoni

Principal Investigator

Department of Psychology

Indiana University

Bloomington, Indiana 47401

Supported by:

Department of Health, Education and Welfare
U.S. Public Health Service

National Institute of Mental Health
Grant No. MH-24027-03

and

National Institutes of Health
Grant No. NS-12179-01

SPEECH PERCEPTION*

David B. Pisoni

Indiana University

Bloomington, Indiana 47401

*Chapter to appear in W. K. Estes (ed.) Handbook of Learning and Cognitive Processes. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976 (In Press).

TABLE OF CONTENTS

I. INTRODUCTION

The fundamental problem in speech perception is to determine how the continuously varying acoustic stimulus produced by a speaker is converted into a sequence of discrete linguistic units by the listener so the intended message can be recovered. This problem can be broken down into a number of specific subquestions. For example, what stages of perceptual analysis intervene between presentation of the stimulus and eventual response? And what types of operations occur at each of these stages? What types of perceptual mechanisms are involved in speech perception and how do they develop during the course of language acquisition? These are just a few of the broad questions that will be considered in the present chapter.

It should be pointed out here that despite the fact that the speech signal may be of poor quality with many of the sounds slurred or distorted, the perceptual process proceeds quite smoothly. To the naive observer the perceptual process often appears to be carried out almost automatically with little conscious effort. This is not at all surprising. A good part of the speech perception process is normally unavailable for conscious inspection. Moreover, some aspects of the process are only partially dependent upon properties of the physical stimulus. As we shall see, the speech signal is so highly structured and constrained that even large distortions in the signal can be tolerated without

loss of intelligibility. This is true, in part, because the speech signal is not entirely new for the listener. As a speaker of a natural language, the listener has available a good deal of knowledge about the structure of an utterance even before it is ever produced. For present purposes we assume that the listener has two types of information at his disposal at any time. On one hand, the listener knows something about the context in which a particular utterance is produced. Knowledge of events, facts and relations about the world can be used by the listener to generate hypotheses and draw inferences from only fragmentary input. On the other hand, the listener has knowledge of his language including information about syntax, semantics and phonology which provides the means for constructing an internal representation or percept of the utterance. It is this latter kind of knowledge, especially those aspects dealing with phonetics and phonology, that will be of primary concern in this contribution.

When placed in the context of other communication systems, human language also has certain characteristics that set it apart. First, human language is symbolic and entails a dual patterning of sound and meaning (Hockett, 1958). Second, human language is grammatical thus permitting an infinite number of meanings to be expressed by the combination of a finite set of elements (Chomsky, 1957). As we shall see, both of these characteristics have placed certain

constraints on the types of signals used in speech production and the potential mechanisms available for perceptual analysis.

The present chapter will be concerned with problems and issues in speech perception rather than the details of specific experiments. Extensive literature reviews and interpretations have been carried out recently by Studdert-Kennedy (1976) and Darwin (1976) and the interested reader should consult these for further background. The major goal of this chapter is to present a point of view about the relation between certain attributes of the acoustic signal and various aspects of perceptual analysis of speech sounds. A recurrent theme is that that the attributes of sound produced by the vocal mechanism seem to be "matched", in some sense, to mechanisms involved in speech perception. Moreover, from our current knowledge of speech perception and production there is fairly strong evidence that some portions of the perceptual process may be mediated by specialized neural mechanisms that are either part of the biological endowment of the organism or develop very soon after birth.

II. LINGUISTIC STRUCTURE OF SPEECH

During the normal course of linguistic communication we are conscious of the words and sentences spoken to us but rarely the speech sounds themselves. Except under special circumstances when, for example, our attention is directed to the sound structure such as listening to a foreign accent or a child's first words, it is difficult to separate our observations from their subsequent interpretation. For the most part, the earliest stages of speech perception appear to be carried out almost automatically without conscious awareness or control by the listener.

Since most of our awareness of spoken language is based on meanings and not sounds, it is appropriate to discuss some aspects of the sound structure of language, specifically the functional categories that have been developed in linguistic science. Research on speech perception over the last twenty-five years has generally assumed that these linguistic categories formed the basic objects of perception and, as a consequence, most of the experimental work was guided by this linguistic analysis.

Although the speech signal that impinges upon the ear of the listener varies more or less continuously as a function of time, the listener, nevertheless, perceives an utterance as consisting of a sequence of discrete segments. The segments that the listener perceives are based on the

functional sound categories of the language community--the phonemes. A phoneme is usually defined as the smallest unit of speech that makes a significant difference between two linguistic forms. The phoneme, however, is an abstract concept since it does not represent a unique instance of a sound (i.e., a phonetic segment or phone) but instead refers to some derived or abstract class of sounds that functions in similar ways in a particular language.

To illustrate how the phonemic principle works, consider the differences between the words bin and pin. At the lowest level of linguistic analysis, the phonetic level, the first segment of each word is different. The difference lies both in the way the sound is articulated and in its acoustical properties. Since the phonetic differences between the [b] and [p^h] serve to distinguish different linguistic forms in English, these segments are assumed to be members of different classes of phonemes, /b/ and /p/. Thus, the phonetic differences between [b] and [p^h] are retained at a more abstract level, the phonological level, where the linguistically significant information is represented.

The situation is, however, somewhat more complicated in the case of sound segments that are phonetically different but which do not serve to contrast different linguistic forms in a specific language. For example, in English, the initial [p^h] in pin is phonetically aspirated since it is

produced with a brief puff of air when the lips are released, whereas the [p] in spin has no aspiration. At the phonetic level, the two p's are represented as distinct phonetic segments since they are produced differently and accordingly represent acoustically distinct signals. At the phonological level, however, the two sounds are considered to be allophones or variants of the same phoneme class /p/ because the feature of aspiration which distinguishes these two segments does not serve to contrast linguistic forms in English. In some languages such as Thai this feature does serve to contrast forms and in these cases the two phonetic segments [p^h] and [p] would represent functionally different phonemes.

The symbol /p/ in this example has no unique phonetic status itself; it simply stands for a class of phones having related properties. Some members of the class are in complementary distribution; phonetic segments occur in contexts in which other segments do not appear. For example, the [k] in keep is produced further forward in the vocal cavity than the [k] in coop and this difference is predictable in terms of the properties of the following vowel. Members of a class of phonetic segments also have similar articulatory and acoustic attributes. For example, the two k sounds referred to above are both voiceless velar stop consonants even though they have slightly different places of articulation.

Each utterance of a language can be represented as a sequence of discrete phonetic segments, each of which can be

further thought of as consisting of a set of distinctive features. Some idea of the feature composition of various English phonemes can be seen in Table 1. This system is based on the work of Jakobson, Fant and Halle (1952) in which

- - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - -

phonetic segments are described in terms of whether a distinctive feature is present (+) or absent (-). Both phonetic and phonological segments are assumed to consist of bundles of features such as these. The features shown in Table 1 are based on distinctions in both the articulatory and acoustic domains. A brief description of these phonetic features and some of their articulatory and acoustic correlates is given in Table 2. More recent versions of the distinctive

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

feature theory are based almost exclusively on articulatory descriptions (Chomsky and Halle, 1968).

Thus, there are two distinct levels of linguistic representation, a phonetic level and a phonological level. In modern generative phonology, a phonetic representation of an utterance consists of a distinctive feature matrix in which the columns represent phonetic segments and the rows indicate their feature specification (Chomsky and Halle, 1968). A phonological representation is more abstract than a phonetic representation. Segments that may be different at the phonetic

level are treated as functionally the same at the phonological level depending on whether the allophonic variations serve a distinctive function in the particular language. However, the distinction between phonetic and phonological levels has little importance for naive listeners who hear speech in terms of the functional categories of their language.

It should be noted here that even the description of speech at a phonetic level is not assumed to be a representation of the actual physical events of speech. A phonetic transcription is neither a physical description of the vocal tract nor a specification of the acoustic signal. At the level of phonetic segments, a great deal of abstraction and categorization has already taken place so that the speech signal is viewed as a sequence of discrete phonetic segments and features.

So far we have discussed only the phoneme which is a meaningless unit. But language also employs meaningful units. Morphemes usually have been considered the smallest meaningful units of language. A free or lexical morpheme, for example, can occur in isolation; it is a word that cannot be analyzed into smaller semantic units. On the other hand, a bound or grammatical morpheme such as the plural marker /z/ in the word dogs, can only occur in context with other morphemes. One important aspect of the dual patterning of human language is the relation between morphemes and phonemes (Hockett, 1958). All morphemes have a complex internal

structure which consists of a sequence of phonemes and, in turn, features arranged in a particular order. Differences between morphemes and hence differences in meanings are expressed by variations in the sequencing and arrangement of the constituent phonemes and their features. Interest in the phoneme as a basic unit of linguistic analysis has, therefore, been based on the assumption that morphemic relations are ultimately derived from an analysis of the phonological structure of speech.

As noted earlier, the bulk of speech perception research has assumed the psychological reality of the phoneme and the phonemic principle in linguistic descriptions. The analysis of speech as a sequence of discrete phonetic segments has guided research by providing the basis and the rationale for specifying the appropriate units of linguistic analysis and the objects of perception.

III. ACOUSTIC STRUCTURE OF SPEECH SOUNDS

Speech sounds have certain distinctive properties or attributes which provide the initial acoustic information for the earliest stages of perceptual analysis. In this section we consider some aspects of the way speech is produced by the vocal apparatus and describe several of the distinctive acoustic attributes of speech sounds.

A. Source-Filter Theory

The basic principles of sound production in the vocal tract and the acoustic filtering that is carried out there are now understood in considerable detail (Stevens and House, 1955; Fant, 1960; Flanagan, 1972). The human vocal tract may be thought of as an acoustic tube of varying cross-sectional area that extends from the glottis to the lips. The upper right-hand panel of Figure 1 shows a mid-sagittal outline of the vocal tract during the production of a neutral vowel. An additional tube, the nasal tract, can be connected to the

- - - - - - - - - - - - - -
Insert Figure 1 about here
- - - - - - - - - - - - - -

system by lowering the soft palate or velum for the production of nasal and nasal-like sounds.

The overall shape of the vocal tract can be changed rapidly by variations in the position of the lips, jaw, tongue and velum. The cross-sectional area of the vocal tract at its point of maximum constriction can be varied from

complete closure, as in the production of a stop consonant to about 20 cm^2, as in the production of an open vowel. When the velum is lowered, the nasal tract is also excited and produces changes in the spectral properties of the radiated sound output.

Sound is generated in the vocal tract by either forcing air through the glottis (i.e., space between the vocal folds) to produce a quasi-periodic sound source or by creating a noisy turbulence in the vicinity of a constriction in the vocal tract. Both sound sources can be used to excite the vocal tract above the larynx. For some sounds like non-nasalized vowels there is a direct acoustic transmission path between glottis and lips whereas for other sounds such as nasals there are significant side branches in the transmission path.

The vocal system acts as a time-varying filter with resonant properties that influence the sound waves generated in the tract. The sound pressure radiated from the lips may be thought of simply as the product of the source spectrum $S_{(f)}$, the vocal tract transfer function $T_{(f)}$ and the radiation characteristic of the vocal tract $R_{(f)}$. The spectrum of the radiated sound pressure $P_{(f)}$ is given by the following equation:

$$P_{(f)} = S_{(f)} \times T_{(f)} \times R_{(f)}$$

Each of these components is shown separately in the left-hand column of Figure 1. The spectrum envelope ($p_{(f)}$) of the radiated sound pressure which displays the relative distribution of energy at different frequencies is shown in the bottom right-hand corner of the figure. Thus, the speech production mechanism consists of two relatively independent components: (1) mechanisms which contribute primarily to the generation of sound energy and (2) mechanisms which function to modify the sound energy.

For a periodic sound such as a vowel, the sound source consists of a line spectrum with components at multiples of the fundamental frequency ($|S_{(f)}|$ in Figure 1). The amplitude of these components decreases by about 12 dB per octave at high frequencies. When the vocal tract is excited by this source, it acts as a filter to reinforce some frequencies and suppress others. The vocal tract transfer function shown in the middle panel on the left of Figure 1 can be characterized by a number of natural frequencies or formants which change as the shape of the vocal tract changes from one articulatory position to another. As a consequence, the radiated sound output reflects the resonant frequencies that are favored by the system. These formant frequencies appear as peaks in the spectrum. Finally, the radiation characteristic reflects the relation between acoustic volume velocity at the mouth opening and sound

pressure at a distance from the lips. This effect occurs primarily at low frequencies: the slope of the spectrum envelope of the radiated sound pressure drops by about 6 dB per octave as shown in the lower right-hand panel of the figure. The result is to reduce the intensity differences between low and high frequency harmonics originally displayed in $|S_{(f)}|$. In summary, the vocal tract acts as a linear time-varying filter that imposes its transmission properties on the frequency spectra of the sound sources generated in the vocal tract. As the vocal tract changes shape during the production of different sounds, the properties of the transfer function change and accordingly the sound output changes.

B. Attributes of Speech Sounds

One method of describing speech sounds is in terms of the degree of vocal tract constriction employed for their production. This can be ordered along a continuum. At one extreme are the vowels and vowel-like sounds which are produced with a relatively unconstricted vocal tract. Liquids, glides and fricatives have intermediate constrictions, whereas stop consonants represent the other extreme with complete closure of the vocal tract at some point of articulation. Figure 2 shows midsagittal outlines of the vocal tract shapes for the vowels [i], [a] and [u], curves showing the cross-sectional area of the vocal tract as a

function of distance from the glottis (i.e., area functions) and their respective spectrum envelopes. When the tongue

- - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - -

body is high and fronted as in [i], the oral cavity is relatively constricted, whereas when the tongue is low as in [a] and [u], the oral cavity is relatively large. On the other hand, the pharyngeal cavity is relatively large for [i] and [u], but constricted for [a]. The cross-sectional area functions of the vocal tract reflect these differences in position of the tongue body and degree of constriction in the vocal tract as illustrated in the middle panel of the figure. The effect of these differences in vocal tract shape on the spectra of the sound output is shown by the differences in the spectrum envelopes on the far right of the figure.

The frequency of the first formant (F1) is low for [i] and [u] reflecting, in part, the relatively large pharyngeal cavity compared to [a] where F1 is high. On the other hand, the frequency of F2 is high for [i] due to the narrow oral cavity and low for [u] as a result of lip rounding which increases the overall length of the oral cavity. Although there is a simple relation between formant frequency and cavity affiliation for the vowels [i], [a] and [u], it is not possible to associate a particular formant frequency with a specific cavity resonance. For other vowels, both

oral and pharyngeal cavities influence the resonance frequencies of the formants (see Fant, 1960, for further discussion).

The production of speech may be thought of as consisting of a sequence of maneuvers from one idealized articulatory position or target to another. During connected speech the articulatory apparatus frequently moves so rapidly from one position to another that the target positions often are not fully reached. Moreover, instructions for new target configurations often begin to be implemented before a previous target has reached its idealized value. Thus, there is "undershoot" from the ideal articulatory configurations.

As noted above, vowels are produced with a relatively unconstricted vocal tract that generates sounds with well-defined formant structures. From acoustic analysis and synthesis experiments, it has been found that the relative positions of the lowest two or three formant frequencies are sufficient to distinguish different vowels in both production and perception (see Peterson & Barney, 1952; Delattre, Liberman, Cooper and Gerstman, 1952; Stevens & House, 1955).

In the production of consonants, however, the vocal tract is often highly constricted or even occluded at some point alont its length. Consider, for example, the production of a fricative sound in which there is a turbulent noise source generated at the point of a constriction. The spectrum of this noise source is continuous since energy is

distributed over all frequencies rather than being restricted to only harmonics of the fundamental. The spectrum of the radiated sound output for a fricative sound is the product of the source function and the vocal-tract transfer function as was the case with the vowels. However, the sound source for fricatives is located above the glottis: the sound output is influenced not only by structures above the source but also to a lesser extent by structures below the sound source in addition to specific properties of the constriction itself. Differences in the production of fricatives are reflected in the relative frequency of noise, its bandwidth and overall intensity. These attributes have been shown to be important perceptual cues for different fricative sounds (Harris, 1958; Strevens, 1960; Delattre, Liberman & Cooper, 1964; Heinz & Stevens, 1961).

The production of stop consonants, liquids and glides is characterized by total or virtual closure of the tract, followed by opening (release). During closure no sound is produced and pressure builds up behind the closure. At release, of an initial stop, for example, there is an abrupt change in the vocal tract which results in a rapid spectral change over a very brief period of time as the articulators move toward the position appropriate for the next sound. These rapid transitions affect the rate of frequency change and position of the formant frequencies systematically in terms of the place and type of closure (see Stevens & House,

1956; Liberman, Delattre, Gerstman and Cooper, 1956). Differences in place of articulation among the stop and nasal consonants are cued primarily in terms of the direction and extent of the second and third formant transitions. The liquids, /r/ and /l/, usually have a brief steady-state period which is followed by relatively slow formant transitions into the following vowel. The segment /r/ differs from /l/ in terms of the changes in the third formant transition. The glides /w/ and /j/ are distinguished from the stops in terms of the duration and rate of change of the transitions of the first two formants. These transitions are longer and usually slower for the glides.

The voicing distinction in consonants has received a great deal of attention in the literature. For consonants in final position, voicing can be cued by the duration of the preceding vowel (Denes, 1955; Raphael, 1972). The duration of articulatory closure has been shown to cue voicing differences between stops in intervocalic position (Lisker, 1957). Voice onset time (VOT), a complex articulatory timing dimension, has been shown to characterize the voicing distinctions between stop consonants in initial position [b,d,g] vs. [p,t,k]. Acoustically, VOT involves simultaneous changes in the relative onset of voiced excitation, the amplitude of the F1 transition and the presence of aspiration in the higher formants during this time period.

While knowledge of the acoustic cues for all segmental phonemes is far from complete, sufficient information is available for reasonably good speech synthesis by rule. With a set of general phonological rules as well as specific phonetic-acoustic rules, it possible to take as input some discrete phonemic representation and provide as output a continuous speech signal (Mattingly, 1968; Klatt, 1975).

The acoustical structure of speech can be illustrated by examining a sound spectrogram of an utterance such as the one shown in Figure 3. As customary, time is represented on the abscissa and frequency on the ordinate. The relative

- - - - - - - - - - - - - -

Insert Figure 3 about here

- - - - - - - - - - - - - -

concentration of energy at each frequency is shown by the degree of darkness on the trace. The horizontal bars represent concentrations of energy that occur at the natural resonant frequencies of the vocal tract and are called formants. These can most easily be seen in the lower frequency regions during the production of vowel and vowel-like sounds. The closely spaced repetitive striations which occur during vowels and other voiced sounds reflect the presence of individual pulses of air passing through the glottis (i.e., glottal pulses) whereas the more randomly structured portions of the spectrogram reflect the presence of turbulance or noise.

Inspection of this figure will reveal that there are some discrete sound segments present in a sound spectrogram but these acoustic segments do not always correspond to the linguistic segments which result from perceptual analysis or to larger units such as morphemes. Thus, from this form of analysis, it became quite clear to many investigators that, at the acoustic level, phonetic segments are not represented discretely in time like beads on a string or bricks in a wall. Instead, phonetic segments are merged together with each other so that there is a complex mapping of acoustic attribute to phonetic segment.

One particularly engaging description of the nature of speech is given by Charles Hockett in his well-known easter egg passage:

> Imagine a row of easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point, the belt carries the row of eggs between the two rollers of a wringer, which quite effectively smash them and rub them more or less into each other. The flow of eggs before the wringer represents the series of impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread out albumine, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer. (Hockett, 1955, p. 210)

Although this is an amusing analogy to the acoustical structure of speech, our understanding of the relation between phoneme and sound is not as impoverished as it might suggest.

For example, some idea of the relation between acoustic cues and phonetic segments can be seen in Figure 4 which shows stylized acoustic patterns[1] for nine consonant-vowel syllables all followed by the vowel [a]. If converted to sound on a pattern playback device (Cooper, Liberman & Borst, 1951) these patterns produce reasonably intelligible approximations of the intended syllables. As shown in this figure, the acoustic cues that distinguish place, manner and voicing among these consonants show a systematic relation when the vowel is the same in each syllable. In these syllables, place of articulation is cued by differences in the direction and extent of the F2 transition. Stops are distinguished from nasals by the presence of a low frequency nasal murmur. The voicing differences between [b,d,g] and [p,t,k] are distinguished by the relative onset of F1 to F2.

- - - - - - - - - - - - - -
Insert Figure 4 about here
- - - - - - - - - - - - - -

The situation is, however, more complicated when the vowel context of the syllable changes. As shown in Figure 5, the acoustic cues for a particular consonant, for example, the /d/ segment in the middle panel, vary and are co-articulated[2] with the following vowel and, as a consequence,

- - - - - - - - - - - - - -
Insert Figure 5 about here
- - - - - - - - - - - - - -

it has been extremely difficult to find a simple invariant attribute that corresponds uniquely to the same stop consonant in all vowel environments (Cooper, Delattre, Liberman, Borst & Gerstman, 1952). The results of early synthesis experiments with two-formant patterns such as these, therefore, failed to reveal acoustic invariants for the various places of articulation in stop consonants independently of context.

It cannot be concluded, however, from these results that there are no invariant acoustic attributes for stop consonants in natural speech. Rather, a more reasonable conclusion, given our current knowledge, is that the invariant attributes are probably not to be found in terms of relatively simple acoustic properties as displayed in a sound spectrogram. Recent work employing somewhat more complex and natural-like synthetic stimuli has, in fact, suggested that there may be rather general properties of the acoustic signal that uniquely specify a particular place of articulation in stop consonants independently of vowel context (Stevens & Blumstein, 1976). The extent to which invariant, context independent acoustic cues to phonemes can be established has been a topic of some controversy in the speech perception literature and is an issue that will be considered again in the next section.

C. Quantal Aspects of Speech Production

What regions of the articulatory space are used to form the phonetic segments and features employed in natural languages? Although the vocal apparatus can theoretically assume a relatively large number of articulatory positions, only a small number of preferred or "natural" regions are actually used in the phonological systems of natural languages. Stevens (1972) has proposed that speech sounds produced in these so-called natural regions have certain quantal properties or attributes that appear to be good candidates for the inventory of phonetic features in a given language. According to Stevens, all phonetic features that occur in languages probably have their basis in acoustic attributes that have such quantal properties: perturbations in articulation at one of these natural regions produce only small changes in the acoustic output. These acoustic attributes are assumed to be well-matched to the auditory system (see the categorical perception section below).

The basic line of reasoning behind Stevens' Quantal Theory of Speech Production can be illustrated by reference to Figure 6. Assume that one could manipulate some articulatory paramater continuously along a particular dimension.

- - - - - - - - - - - - - -

Insert Figure 6 about here

- - - - - - - - - - - - - -

And further assume that one could obtain a measure of some specific acoustic parameter of the speech signal that would be controlled by changes in this articulatory parameter. As the articulatory parameter is varied continuously, one might expect to find continuous changes in the output of this acoustic parameter. However, what seems to happen is shown schematically in Figure 6. There are places where very small changes in the articulatory parameter, as in region II, produce large variations in the acoustic parameter whereas in other places, as region III, large changes in the articulatory parameter produce only small changes in the acoustic output. Relations such as these between vocal tract shape and sound output have been studied in some detail by Stevens (1972, 1973) and others with the use of computer simulation models of the vocal tract (see also Lindblom and Sundberg, 1969; Liljencrants and Lindblom, 1972). Small changes have been made systematically in an articulatory parameter and the resultant effects on the properties of sound output calculated quite precisely.

To see how the simulation was accomplished it will be helpful to briefly examine the resonance characteristics of the vocal tract. Figure 7 shows an approximation of the vocal tract shape for the vowel [a] in terms of a two tube resonator (Fant, 1960). The spectrum envelope produced by

- - - - - - - - - - - - - -

Insert Figure 7 about here

- - - - - - - - - - - - - -

this configuration is shown below. The back or pharyngeal cavity is constricted when compared with the front or oral cavity. Using a vocal tract simulation, Stevens found that over a wide range, variations in the lengths of the cavities (i.e., d_1 and d_2) and the cross-sectional areas (i.e., A_1 and A_2) did not affect the formant frequencies of the vowel to any large extent. The results of this simulation are shown in Figure 8 where the calculated formant frequencies are plotted as a function of the length of d_1. Notice that

- - - - - - - - - - - - - -
Insert Figure 8 about here
- - - - - - - - - - - - - -

there is an area in the middle of this figure where variations in d_1 produce only small changes in the formant frequencies of F1 and F2. On the other hand, there are regions at the extremes where a small variation in d_1 produces a much larger shift in the formant frequencies, especially the values of the second formant.

Stevens (1972, 1973) has employed these simulation techniques to study other vowels as well as several types of consonants. One general conclusion to be drawn from Stevens' work is that there are certain places in the vocal tract which show quantal relations between articulation and the attributes of the sound output. Perturbations in articulation apparently have only small effects on the acoustic signal and therefore would presumably affect perception only minimally.

One of the proposals that Stevens has made about these findings is that the sound segments used in languages are selected from a small range of features or distinctive attributes. A careful examination of these attributes reveals that they have a natural basis in terms of certain vocal tract configurations. These configurations are chosen from precisely those regions where variability in articulation can be tolerated without the sound output being appreciably affected.

Using the same types of simulation procedures, Liljencrants and Lindblom (1972) have studied the vowel systems of a number of languages in order to predict their phonetic structure. Since human vowels represent only a small subset of the possible combinations of formant frequencies, it was of interest to see how the vowels were distributed in acoustic space. Liljencrants and Lindblom found that the vowel systems of quite diverse languages can be described in terms of a simple principle of "maximal perceptual contrast" as defined by the linear distance in mel units between the points representative of two vowels[3]. The whole vowel space for a particular language seems to be organized in terms of sound contrasts that are highly discriminable. This is true of languages having as few as three vowels to as many as twelve. These two sets of findings are important because they indicate that the sound systems of natural languages are matched in some sense to attributes of both perception and production.

It should be noted here that not all of the potential distinctive features at the phonetic level are used distinctively at a phonological level in any particular language. Similarly, no natural language has as many phonemes in its phonological system as there are logically possible combinations of utilized distinctive features. The sound systems of language appear to have evolved the way they are for several reasons. First, the distinctions are easily pronouncable for talkers. Second, these articulatory distinctions generate acoustic attributes that are highly distinctive resulting in sounds that can be identified and discriminated under very poor listening conditions. In the next section we consider the results of experiments dealing with the perception of speech sounds.

IV. PERCEPTION OF SPEECH SOUNDS

The process of speech production imposes certain well-defined constraints on the resulting acoustic waveform. As noted, these constraints are derived from the anatomy and physiology of the vocal mechanism and the associated resonant properties. Although a good deal of work remains to be done in understanding and modeling the speech perception process, there is a number of well established empirical findings which define the problems in the field of speech perception and set it apart from other related areas such as auditory psychophysics and general auditory perception. It is possible to describe the research in speech perception in terms of two general lines of investigation: (1) studies aimed at establishing the acoustic cues to the perception of speech sound segments and (2) studies aimed primarily at demonstrating the effects of manipulating syntactic and semantic variables on speech perception. The remainder of this chapter will be concerned primarily with experiments that fall into the first class of studies. Four major areas of research dealing with the perception of phonetic segments will be discussed: (1) early research on the acoustic cues for stop consonants, (2) experiments on the identification and discrimination of speech sounds, (3) research on developmental aspects of speech perception in young infants, and, finally, (4) recent work on the role of feature detectors in

speech perception. The material described in these sections is quite selective in order to focus on some of the major problems and theoretical issues that have been studied in speech perception over the last few years.

The study of speech perception differs in several ways from the study of general auditory perception. First, the signals typically used to study the functioning of the auditory system are simple, discrete and usually well defined mathematically. Moreover, they typically vary along only a single dimension. In contrast, speech sounds involve complex spectral relations that vary as a function of time; changes that occur in a single parameter often affect the perception of other attributes of the stimulus (Lane, 1962). Secondly, most of the research in auditory psychophysics over the last two decades has been concerned with the discriminative capacities of the transducer and the peripheral auditory mechanism. In the perception of speech, the relevant mechanisms are, for the most part, centrally located. Moreover, experiments in auditory psychophysics have commonly focused on experimental tasks involving discrimination rather than absolute identification. This is rarely the situation of listeners when they perceive and understand speech. In fact, the listener must almost always attempt to identify, on an absolute basis, a particular stretch of speech. Thus, it is generally believed that a good deal of what we have learned from traditional auditory

psychophysics is only marginally relevant to the study of the complex cognitive processes that are involved in speech perception.

In addition to differences in the signal, there are also marked differences in the way speech and non-speech sounds are processed by listeners. For the most part, when people are presented with speech signals they respond to them as linguistic entities rather than as auditory events in the environment. Speech signals are categorized and labeled almost immediately with reference to the listener's linguistic background. Moreover, as we shall see, a listener's ability to discriminate certain speech sounds is often a function of the extent to which the particular acoustic distinction under study plays a functional role in the listener's linguistic system.

A. Invariant Acoustic Cues for Stop Consonants

One of the most firmly established findings in the speech perception literature is that the acoustic correlates of a number of consonantal features are highly dependent on context. This is especially true for the stop consonants (b, d, g; p, t, k) which show the greatest amount of contextual variability. As a result of co-articulation effects, one sound segment often carries information about two or more successive phonemes in an utterance. And, conversely, a single phoneme often exerts an influence on

several successive sound segments in the acoustic signal. As noted earlier, listeners perceive speech as consisting of a sequence of discrete segments arrayed in time although the physical signal often varies continuously. The earliest experimental studies in speech perception were aimed at uncovering the relation or mapping between attributes of the acoustic signal and phonetic units derived from perceptual analysis. The outcome of these initial studies indicated that there were few discrete isolatable and invariant sound segments in the physical signal that correspond uniquely to perceived phonemes. The lack of correspondence between attributes of the acoustic signal and units of linguistic analysis has been, and still currently is one of the most important and controversial issues in speech perception. Because of the prominence of this issue, it is appropriate to review first some of the early findings and then briefly discuss some of the more recent attempts that have been made to resolve this issue.

The initial work on the acoustic cues for phonemes involved two related procedures. First, spectrographic analyses were carried out on minimal pairs to identify the potentially important acoustic attributes that distinguished these utterances. Then synthesis experiments were conducted to verify the significance of these acoustic cues in perception. In the first study to use these combined methods,

Liberman, Delattre and Cooper (1952) examined the relation between the frequency of a noise burst and the perception of the voiceless consonants p, t and k. An examination of sound spectrograms of real speech showed that the voiceless stops could potentially be distinguished by the frequency of a brief burst of noise, the acoustic counterpart of the articulatory explosion at the release of stop closure. Liberman *et al.* (1952) systematically varied the frequency of a synthetic noise burst before a number of different two formant vowels and observed its effect on perception. The actual stimulus patterns employed in this experiment are shown in Figure 9. An example of one typical stimulus pattern is shown in panel C of this figure. The stimuli were presented one at a time to subjects who were told to identify

- - - - - - - - - - - - - -
Insert Figure 9 about here
- - - - - - - - - - - - - -

each stimulus as either p, t, or k. The results of this experiment are shown in Figure 10. Subjects' identification

- - - - - - - - - - - - - -
Insert Figure 10 about here
- - - - - - - - - - - - - -

of a particular stop consonant varied not only according to the frequency of the noise burst but also in terms of the relation of the burst to the vowel with which it was paired. As shown in this figure, high frequency bursts above 3.2 KHz were heard as /t/ in all vowel environments whereas bursts

at other frequencies were heard as either /p/ or /k/ depending on the frequency of the burst in relation to the formant frequencies of the following vowel. Thus, placing the identical burst before two different vowels changes the way in which the burst is perceived by listeners.

The results of this initial study with synthetic speech were replicated with real speech stimuli in a study by Schatz (1954). She found that release bursts spliced from [ki], [ka] and [ku] were perceived as different stops depending upon the following vowel context. For example, the release burst excised from [ki] is perceived as /t/ when spliced before [a] but is perceived as /p/ before [u]. The findings of these two experiments therefore indicate that identification of the consonants does not depend exclusively on the absolute frequency of the burst but rather depends on the attributes of the burst in relation to the vowel which follows.

In another study, Cooper *et al.* (1952) studied the role of formant transitions in the perception of stop consonants. Variations in the first formant transition were found to provide cues to voicing and manner whereas variations in the second formant transition were found to provide cues to place of articulation among the stops /b,d,g/ and /p,t,k/. Cooper *et al.* (1952) also studied a range of second formant transitions with the same vowels that were used in the previous burst experiment. Subjects were required to identify

the stimuli as b, d or g. The results indicated that, while most Ss heard rising transitions as /b/ in almost all vowel contexts, perception of the falling transitions as either /d/ or /g/ varied as a function of the following vowel. Thus, as in the previous burst experiments, the perception of the formant transitions as a particular phoneme depended on the following vowel.

Three important conclusions have been drawn from the results of these early perceptual experiments. First, with regard to stop consonants, no invariant acoustic cues could be identified which corresponded uniquely to the same phoneme in all environments. Second, because burst and transition cues depend, to a large extent, on properties of the following vowel, the minimal acoustic unit seemed more likely to be about the size of a consonant-vowel syllable than an isolated phoneme. Indeed, as Cooper *et al.* (1952) remark, "one may not always be able to find the phoneme in the speech wave, because it may not exist there in free form". Finally, within the context of these experiments which have employed relatively simple synthetic stimuli, it appeared that the acoustic information for a particular phoneme was encoded into the sound stream in a complex way because no one-to-one correspondence could be found between perceived phoneme and acoustic segment.

Although these early findings failed to uncover the invariant acoustic attributes for phonemes (e.g., stop consonants) and emphasized a complex relation between

acoustic attribute and phonetic unit, numerous investigators have continued, nevertheless, to search for a description of the acoustic signal which would reveal invariant attributes for phonemes (Fant, 1960, 1962, 1973; Stevens, 1967, 1973). The experimental literature on this topic during the 1950's and 1960's is extensive and cannot be reviewed here. To illustrate the nature of the problem, however, we consider briefly two different approaches to the invariance issue.

For a number of years, primarily because of their interest in distinctive feature theory and its emphasis on relational invariants, Stevens (1967, 1971, 1973) and Fant (1960, 1973) have sought to find acoustic invariants for distinctive features. Their argument is that there are well-defined acoustic correlates for the principal places of articulation for consonants, particularly the stops, and that previous research with synthetic speech has tended to obscure some of the important cues found in natural speech. According to Fant and Stevens, these invariant acoustic attributes are contained in the rapid changes in spectral energy that occur during the first 10-30 msec after the release. Both investigators emphasize that burst and formant transitions, which previously were assumed to be independent cues, should be regarded as a single integrated stimulus event or cue. Thus, burst and transitions constitute an overall spectral change at onset.

Based primarily on acoustic analyses of CV syllables, Fant (1969) has claimed invariant spectral relations for labials, post-dentals and velars. For labials (i.e., [b] and [p]), spectral energy is weak and spread with a major concentration at low frequencies; for post-dentals ([d] and [t]), spectral energy is stronger and spread although the major concentration occurs at high frequencies; for velars ([g] and [k]), the spectral energy is compact and concentrated at mid-frequencies.

Focusing on the rapid spectrum changes that accompany the release of a stop consonant, Stevens (1967, 1973) has also argued for invariant spectral patterns for place of articulation. According to Stevens, labials can be characterized by low frequency onsets followed by upward or rising changes in spectral energy whereas post-dentals have high frequency onsets followed by downward or falling changes in spectral energy. Velars have spectral energy concentrated narrowly in the mid-frequency range at onset followed by spreading of energy to frequencies above and below this region. Some examples of these spectral relations are shown schematically in Figure 11. Stevens (1975) believes that these rapid spectrum changes at onset "identify features of place of

- - - - - - - - - - - - - -

Insert Figure 11 about here

- - - - - - - - - - - - - -

articulation without reference to acoustic events remote from this point in time" and that "these cues are absolute

properties of the speech signal and are context-independent (p. 319)".

More recently, however, Stevens and Blumstein (1976) have modified this earlier account somewhat and now claim that the invariant properties for stops in initial position involve simply the location and diffuseness of spectral energy at stimulus onset. This position is now closer to Fant's views summarized above. The main direction of both Stevens' and Fant's position has been to focus on somewhat more complex integrated acoustic attributes of consonants rather than simple isolated cues. By following this strategy it is assumed that an integrated acoustic pattern will show invariance when each of its components fails to do so when considered separately in isolation.

The work of Stevens and Fant is important because it represents the only serious attempt to specify the invariants for stops directly in the acoustic signal. It is clear, however, despite Stevens' remark to the contrary, that the types of invariant attributes being proposed here are relational invariants not absolute invariants that occur independently of context. The patterns that Stevens and Fant describe are derived from examining frequency relations over some short period of time and are, therefore, not absolutely invariant in all contexts (see below).

Although working independently of Fant and Stevens, Cole and Scott, (1974a, b) have also argued for context-

independent or invariant cues for stop consonants. Their position is, however, quite different from that of Fant and Stevens. Cole and Scott claim that invariant cues to place of articulation are present in the initial portion of the consonant energy in the release burst. This claim is based on the results of a tape splicing experiment in which Cole and Scott transposed the initial consonant energy from the six stops between the vowels /i/ and /u/. Their results showed that listeners could identify the original consonants accurately in all cases except when the consonant energy from /ki/ and /gi/ was transposed to /u/.

At first glance these results as well as the claims made by Cole and Scott appear to be in sharp conflict with the findings of the earlier experiments carried out by Liberman *et al.* (1952) with synthetic stimuli and Schatz (1954) with natural speech. The discrepancy, however, can be accounted for quite easily and the claims dismissed by a careful examination of the stimuli obtained after the tape splicing procedure. In these experiments Cole and Scott transposed not only the release bursts, as in the earlier studies, but also the aspirated formant transitions which are, in fact, co-articulated with the following vowel. Thus, it is quite likely that sufficient information was present in these stimuli for subjects to identify the consonant from its original context even before it was transposed to another vowel. Indeed, Schatz (1954) remarked on precisely this

point in her tape-splicing experiment some twenty years earlier stating that "these 'voiceless formants' in the aspiration are prominent enough so that the vowel is easily identifiable even when the entire voiced portion of the syllable has been removed and the voiceless part is heard along (p. 52)". More recently, Winitz, Scheib and Reeds (1972) obtained comparable results when listeners were asked to identify only the burst and aspiaration portions of /p,t,k/ spoken before the vowels /i,a,u/. The results of Cole and Scott, therefore, are not at all in conflict with the previous work by Liberman *et al.* and Schatz. The results and claims have little bearing on the issue of invariance since the burst and aspirated formant transitions vary as a function of the following vowel of the syllable and consequently are context-dependent cues.

The arguments advanced in favor of invariant context-independent attributes for stop consonants have recently been reexamined in great detail by Dorman, Studdert-Kennedy and Raphael (1976). In contrast to the views of Fant and Stevens, these investigators argue, based on new perceptual data, that burst and transition cues are highly dependent on context. Dorman *et al.* removed these cues from CVC syllables containing /b,d,g/ spoken before nine different vowels and then recombined them with the same vowels spoken in VC syllables in order to determine whether bursts and transitions were sufficient cues to the three places of articula-

tion. The results indicated that the importance of release bursts and transitions, as cues to place, varied substantially with the particular consonant, the vowel and the speaker and that no single cue was sufficient for the recognition of a particular place feature in all contexts. In another experiment, Dorman *et al.* transposed the release bursts that were removed from each CVC syllable across all nine VC syllables for a given place of articulation. Although the release bursts were, to a large degree, invariant in their perceptual effect before most vowels, they were sufficient cues to place of articulation in only a small number of cases. Dorman *et al.* observed in these experiments that the effects of burst and transition were reciprocally related. In some contexts where one cue (e.g., the burst) was sufficient for a particular place distinction, the other cue (e.g., the transitions) was not and vice versa. For example, the burst was found to be a strong cue for /b/ before rounded vowels (i.e., /u/). In this context, however, the formant transitions are brief and served only as weak cues to place. On the other hand, formants transitions are strong cues for /b/ before middle, unrounded vowels such as /a/ whereas the burst was only a weak cue. According to Dorman *et al.*, the burst will be an effective cue to place only if its frequency is close to the main formant of the following vowel. If burst frequency differs substantially from the main formant, the transitions will then become strong cues to place. Thus,

as Dorman *et al.* point out, burst and transition serve basically similar functions by providing information about the consonantal release and spectral changes into the following vowel.

Based on these more recent findings, Dorman *et al.* conclude that the sufficiency of each cue varies as a function of context. Thus, any account of the perception of place of articulation for stop consonants will, therefore, have to be relative rather than absolute and will consequently require reference to the following vowel.

It is worth pointing out here that the invariance problem in speech perception has, by no means, been resolved yet. The work summarized in this section has been limited to only the stop consonants in initial position in stressed CV syllables. Some idea of the magnitude of the invariance problem in perception can be obtained by considering the contextual effects for stop consonants that appear in other phonetic environments such as medial and final position of syllables as well as consonant clusters. Moreover, the problem becomes enormous when we add to it the contextual variability found for other classes of speech sounds such as fricatives, liquids, nasals and vowels as well as the inherent variability associated with phonetic context, speaking rate and individual talker differences.

B. The Speech Mode and Categorical Perception

The earliest experiments in speech perception showed that listeners respond to speech sounds quite differently from other auditory signals. Liberman and his colleagues at Haskins Laboratories found that listeners perceived synthetic speech stimuli varying between [b], [d] and [g] as members of distinct categories (Liberman, Harris, Hoffman and Griffith, 1957). When these same listeners were required to discriminate pairs of these sounds, they could discriminate stimuli drawn from different phonetic categories but could not discriminate stimuli drawn from the same phonetic category. The obtained discrimination functions showed marked discontinuities at places along the stimulus continuum that were correlated with changes in identification.

The ideal case of this form of perception, "categorical perception", is illustrated in Figure 12. In an experiment such as this, two or more phonetic segments are selected to

- - - - - - - - - - - - - -

Insert Figure 12 about here

- - - - - - - - - - - - - -

represent end points and a continuum of synthetic stimuli is generated. Subjects are required to carry out two tasks: identification and descrimination. In the identification task, stimuli are selected from the continuum and presented one at a time in random order for labeling into categories defined by the experimenter. In the discrimination task, pairs of stimuli are selected from the continuum and presented to listeners for some discriminative response.

The basic finding of the categorical perception experiments is that listeners can discriminate between two speech sounds which have been identified as different phonemes much better than two stimuli which have been identified as the same phoneme even though the acoustic differences are comparable. At the time, the categorical perception results were considered by Liberman and others to be quite unusual when compared with the results typically obtained in most psychophysical experiments with non-speech stimuli. In general, stimuli that lie along a single continuum are perceived continuously resulting in discrimination functions that are monotonic with the physical scale. As is well known, there are capacity limitations on information transmission in terms of absolute identification (Miller, 1956). Listeners can discriminate many more acoustic stimuli than they can identify in absolute terms (Pollack, 1952, 1953). However, in the case of categorical perception the situation is quite different. The listener's differential discrimination appears to be no better than his absolute identification. In the extreme case of categorical perception a listener's discrimination performance can be predicted from his identification function under the strong assumption that the listener can discriminate between two stimuli only to the extent that these stimuli are identified as different on an absolute basis (Liberman *et al.*, 1957).

These initial findings with stop consonants led to a similar experiment with vowels that varied in acoustically equal steps through the range /I/, /ɛ/ and /ae/. Fry, Abramson, Eimas and Liberman (1962) reported that these stimuli were perceived continuously, much like non-speech stimuli. The discrimination functions did not yield discontinuities along the stimulus continuum which were related to changes in identification but were relatively flat across the whole continuum. Moreover, it was observed that vowels were, in general, more discriminable than stop consonants indicating that listeners could perceive many more intraphonemic differences.

The differences in perception between stop consonants and steady-state vowels have been assumed to reflect two basic modes of perception, a categorical mode and a continuous mode. Categorical perception reflects a mode of perception in which each acoustic pattern is <u>always</u> and <u>only</u> perceived as a token of a particular phonetic type (Studdert-Kennedy, 1974). Listeners can discriminate between two different acoustic patterns if the stimuli have been categorized into different phonetic categories, but they cannot discriminate two different acoustic patterns that have been categorized into the same phonetic category. Information about the acoustic properties of these stimuli appears to be unavailable for purely auditory judgements as a consequence of phonetic classification. What remains available to the decision

process is a more abstract and permanent code based on the listener's interpretation of the stimulus event (see Pisoni, 1971; Pisoni and Tash, 1974).

Although the stimulus generalization as reflected in categorical perception might seem to be a more primitive form of stimulus control, it may provide, on the other hand, a more efficient mode of response for absolute and rapid decisions concerning the presence or absence or particular attributes such as those required in the processing of connected speech. Indeed, the great interest expressed in categorical perception of speech presumably derives from the assumption that listeners do make categorical decisions in listening to continuous speech.

Continuous perception, on the other hand, may be thought of as reflecting an auditory mode of perception where discrimination is independent of category assignment. Although listeners can assign acoustically different stimuli to the same category, they may still discriminate between tokens selected from the same category. Thus, an auditory, non-phonetic basis for discrimination is available to the listener.

For a number of years the categorical perception results were assumed to be unique to speech perception and primarily a consequence of phonetic categorization. Indeed, the differences in perception between consonants and vowels has led liberman (1970) to argue strongly for a specialized mode

of perception, a "speech mode", to characterize the way these stimuli are perceived. Other findings have suggested that a specialized perceptual mechanism--a "special speech decoder" may exist for processing speech sounds (Studdert-Kennedy and Shankweiler, 1970).

The differences in perception between consonants and vowels were originally interpreted as supporting a Motor Theory of Speech Perception (Liberman, Cooper, Harris and MacNeilage, 1963). According to the strong motor theory, reference to articulation is a mediating stage in the perceptual process between the incoming acoustic signal and its recognition. As discussed earlier, stop consonants are produced in a discontinuous way by a constriction at a particular place in the vocal tract, whereas vowels are produced by continuous changes in the overall shape of the vocal tract. The strong version of the motor theory assumed that although the appropriate acoustic cues for consonants could be described by an acoustic continuum, these stimuli were perceived discontinuously (i.e., categorically) because the articulations underlying the production of these sounds is essentially discontinuous. In contrast, vowels were perceived continuously because of the continuous changes in the articulators from one position to another. Although the original accounts of the motor theory (Liberman, 1957; Liberman *et al.*, 1963) assumed that articulatory movements and their sensory effects mediate perception, more recent versions of the theory (Liberman, Cooper, Shankweiler and

Studdert-Kennedy, 1967; Studdert-Kennedy, 1974) have placed the reference at the level of the neuromotor commands that activate the articulators.

The major reason for proposing a motor theory was to attempt to resolve the lack of invariance between acoustic attribute and perceived phoneme. According to Liberman (1957) there is a simpler relation between articulation and perception than between acoustic attribute and perception[4]. At that time it was claimed that the articulatory movements and motor commands for a particular phoneme showed less contextual variability than the resultant acoustic manifestation of the phoneme. However, this argument has not found support in the subsequent electromyographic research on speech production (see MacNeilage, 1970; Harris, 1974). If anything, these findings show the ubiquitous nature of variability at all stages of speech production.

The interpretation that categorical perception reflects a specialized mode of perception unique to speech as well as the claims associated with the motor theory have come under strong criticism from a number of directions in the last few years. Several investigators have argued that the differences in perception between consonants and vowels reflect differences in the psychophysical properties of the acoustic cues which distinguish these two classes of speech sounds (Lane, 1965; Pisoni, 1971; Studdert-Kennedy, 1974). For the stop consonants there is a relatively complex rela-

tion between phoneme and its representation as sound; the essential acoustic cues are contained in the rapidly changing spectrum at onset (i.e., release burst and formant transitions) which is weak, relatively brief in duration (30-50 msec.) and transient in nature. On the other hand, the cues to the vowels involve changes in the steady-state frequencies of the first three formants which have a relatively long duration and more uniform spectral properties as well as greater intensity. As support for this, Fujisaki and Kawashima (1969, 1970) and Pisoni (1971, 1975) have shown that the differences in perception between consonants and vowels are due, in part, to the duration of the acoustic cues. Vowels of very short duration (i.e., 40-50 msec.) are perceived more categorically than identical stimuli having longer durations.

Other findings have shown that categorical perception is also due, in part, to encoding processes in short-term memory that result from the particular type of discrimination task used in these experiments (Pisoni, 1971, 1973, 1975). The ABX procedure has been used in almost all of the speech perception experiments demonstrating categorical perception. In this task the subject is presented with three sounds successively, ABA or ABB. A and B are always acoustically different and the subject has to indicate whether the third sound is identical to the first or second sound. This is basically a recognition memory paradigm. In order to solve

the discrimination task, the subject is forced to encode the individual stimuli in temporal succession and then base his decision on the encoded representations that have been maintained in short-term memory rather than to respond to the magnitudes of difference between stimuli within an ABX triad. In a number of experiments Pisoni (1971, 1973) has shown that differences between categorical and continuous modes of perception are crucially dependent on the memory requirements of the particular discrimination procedure and the level of encoding required to solve the task (Pisoni, 1971, 1973).

Several recent experiments employing non-speech stimuli have also suggested that categorical perception may not be peculiar to speech sounds or a specialized speech mode as once supposed, but may be a more general property of cognitive processes that involve categorization and coding of the stimulus input (see Bruner, 1957). For example, Cutting and Rosner (1974) demonstrated categorical perception effects for non-speech musical sounds varying in rise-time that could be labeled as a "pluck" or a "bow". Miller, Wier, Pastore, Kelly and Dooling (1976) have shown comparable categorical perception effects for non-speech stimuli varying in the onset of a noise preceding a buzz. In another study, Pisoni (1976) has reported similar results for stimuli differing in the relative onset time of two component tones. In each

case, these non-speech experiments showed that discrimination was better for pairs of stimuli selected from different categories than pairs of stimuli selected from the same category. Moreover, discrimination of stimuli selected from within a category was very nearly close to chance performance as predicted by the categorical perception model.

The results obtained with non-speech stimuli have provided some insight into the underlying basis of categorical perception for speech stimuli. These non-speech experiments have succeeded in demonstrating categorical perception when previous attempts have failed primarily for three reasons. First, the investigators employed relatively complex acoustic stimuli in which only a single component was varied relative to the remainder of the stimulus complex. In most of the previous non-speech experiments only simple stimuli were used. Second, while these complex stimuli may be characterized as varying in linear steps along some nominally physical continuum, on both psychophysical and perceptual grounds, the stimulus continuum that is generated results in several distinctive perceptual attributes or qualities that are present for some stimuli but not others. These perceptual attributes, in turn, define quantal regions along the stimulus continuum that are separated by natural psychophysical boundaries; within these regions sensitivity is low whereas between these regions it is high. Finally, because the stimulus continuum can be partitioned into several perceptually

distinctive classes, subjects can easily employ a set of labels or descriptive categories to encode these signals in short-term memory. These codes can then be assigned to stimuli presented in the subsequent ABX discrimination task.

Categorical perception of both speech and complex non-speech signals, therefore, can be explained, in part, by the presence of well-defined psychophysical boundaries which separate stimuli into distinctive perceptual categories and by the use of verbal labels which can be used to encode these attributes in short-term memory. Thus, this account of categorical perception involves two distinct components, a sensory component and a memory or labeling component. Previous explanations have stressed only the labeling component of categorical perception (Liberman *et al.*, 1967). If categorical perception were based only on labels and encoding processes in short-term memory as Fujisaki and Kawashima (1969, 1970) and Massaro (1976) have argued as well as others, we would expect to find comparable categorical-like discrimination functions for vowels and other steady-state signals that can be labeled easily. However, the available evidence indicates that while labeling and memory effects may account for some aspects of vowel discrimination, particularly the results obtained with very short vowels, it cannot account for all of the relevant findings. For example, comparable categorical-like discrimination functions have been obtained for stop consonants and non-speech signals

without specific labeling instructions to subjects. While it is possible to argue that subjects did use labels for the speech stimuli in this task, it seems unlikely that they could have used them with the non-speech signals which also showed marked discontinuities in the shape of the discrimination functions. Thus, these discriminations findings provide support for the presence of a real sensory effect in terms of a well-defined psychophysical boundary and, accordingly, a perceptual notch in the stimulus continuum at the category boundary.

But what is the basis for the distinctive perceptual attributes of speech sounds? One approach can be found in the work of Stevens (1972) on the Quantal Theory discussed earlier. According to Stevens, phonetic features are grounded in a close match between articulatory and auditory capacities. The acoustic attributes common to a phonetic category are determined, in part, by articulatory constraints on speech production and, in part, by the distinctiveness of the resulting acoustic signals in perception. Thus, the strongest evidence for a perceptual match between speech perception and production according to Stevens, is the categorical perception findings. The acoustic correlates of certain phonetic features that show quantal properties in production are also precisely those features that show categorical-like discrimination in perceptual experiments.

It should be pointed out, however, that the category boundaries for phonetic features are not inherently fixed perceptual thresholds since the boundaries and the resulting shifts in sensitivity in this region of the continuum are also a function of linguistic experience. Indeed, as Popper (1972) has suggested, "people who speak different languages may tune their auditory systems differently (p. 218)". Cross-language research has shown, in fact, that the categorizations imposed on synthetic stimuli are based on the acoustic attributes of the stimuli and the linguistic experience of the listener. To take one example, Abramson and Lisker (1965) generated a continuum of synthetic stimuli varying in voice onset time between /da/ and /ta/ and presented them for labeling to listeners of three different language backgrounds. The labeling functions for English, Spanish and Thai subjects are shown in Figure 13. As shown, these

- - - - - - - - - - - - - -
Insert Figure 13 about here
- - - - - - - - - - - - - -

listeners categorized the same stimuli in quite different ways depending on the phonological structure of their language. The phoneme boundaries are not only placed somewhat differently along the continuum in each case, but the Thai subjects show an additional category. This result was expected since in Thai a phonological distinction is made between the voiceless aspirated stop [t^h] and the voiceless unaspirated stop [t]. This phonetic difference is not

realized phonologically in either English or Spanish and consequently fails to play a role in the listener's identification and discrimination. The phonological systems of different languages, therefore, make use of the acoustic and phonetic distinctions that exist between different speech sounds in somewhat different ways.

In summary, several implications can be drawn from the categorical perception research. First, the perception of speech sounds appears to have certain quantal properties much like those observed earlier in the production of speech: listeners treat acoustically different sounds as functionally the same. Second, the categorical perception results can be thought of as representing a phonetic mode of perception in which the listener responds to speech signals in terms of the auditory features deployed in his own linguistic system. The extent to which these perceptual processes are innately determined or modified by the environment has been a topic of great interest in speech perception, as will be seen below.

C. Speech Perception in Infants

Much of what we currently know about the development of language and the acquisition of phonology is based on data obtained in studies of speech production (see McNeill, 1970; Menyuk, 1971; Jakobson, 1968). One conclusion that has been drawn by a number of investigators is that the developmental process proceeds from the general to the specific and

gradually involves a greater and greater differentiation of language skills. Within the last five years, a number of pioneering studies by Eimas and others has demonstrated that infants as young as one month of age are capable of making fine discriminations among a number of the distinctive attributes of speech sounds. These results obviously call into question the validity of the differentiation assumption. While increasing differentiation may very well be true of the development of productive language skills, it may not be true of the perception of speech. The recent evidence from studies of infant speech perception points to a loss of discriminative abilities over time for certain speech sounds if specific experience with these distinctions fails to take place in the local environment.

The procedure used in these speech perception experiments involved a discrimination paradigm in which the infant was first familiarized with a particular stimulus and then shifted to another stimulus. If the infant showed an increased response rate after the shift, it was assumed that the infant could discriminate the difference between the two stimuli. The actual procedure was a modification of the conjugate reinforcement methodology developed by Siqueland and DeLucia (1969). Each criterion response is reinforced by the presentation of a synthetic speech sound. In the infant speech perception studies carried out by Eimas, a nonnutritive sucking response was employed (Eimas,

Siqueland, Jusczyk and Vigorito, 1971; Eimas, 1974, 1975). During the course of the experiment, the infant becomes aware of the contingency between the sucking response and presentation of a stimulus pattern. After reaching some asymptote, response rate declines presumably because the stimulus loses some of its original novel or reinforcing properties. After a period of satiation, a second stimulus is presented repeatedly to the infant under the same contingent arrangement. After the shift, if the infants show a response pattern reliably different from a control group in which no stimulus change was introduced, the results are taken as evidence that the infants can discriminate the difference between the two stimuli.

In the first experiment using this procedure, Eimas *et al.* (1971) studied the voicing feature which distinguishes [b] from [p^h]. The stimuli were synthetically produced CV syllables and varied in acoustically equal steps of voice-onset time between [ba] and [p^ha]. The results revealed two important findings. First, infants could discriminate between two speech sounds selected from different phonetic categories. Second, infants could <u>not</u> reliably discriminate between two acoustically different stimuli selected from within the same phonetic category. The latter finding is particularly relevant since it permitted Eimas and his collaborators to argue that their infants perceived the voicing distinction in a more nearly categorical manner and,

therefore, in a linguistic mode comparable to that found in the adult studies.

Other findings have shown that infants can also discriminate between synthetic stimuli varying in place of articulation which is represented acoustically in terms of differences in the second and third formant transitions (Moffitt, 1971; Morse, 1972). In another study, Eimas (1974) has also reported that infants can discriminate between stimuli varying in place of articulation between [b], [d], and [g] and that they do this in a categorical-like manner too. The discrimination data obtained in this study are completely consistent with the three major distinctions found universally for place of articulation in stop consonants.

Two cross-language studies of infant speech perception have also been carried out recently on the voicing distinction using similar synthetic stimuli and methodology. In one study, Lasky, Syrdal-Lasky and Klein (1975) found that infants from Spanish-speaking environments could discriminate three categories along the voicing continuum. One boundary occurred between +20 and +60 msec and another between -20 and -60 msec. The first boundary is consistent with the discrimination findings of Lisker and Abramson (1970) for English-speaking adults and the previous results of Eimas *et al.* with infants and suggests a possible innate or sensory basis to the distinction. However, the presence of a second boundary in the discrimination data was of particular

interest. Spanish-speaking adults distinguish between only two categories of voicing and, based on the adult discrimination data of Lisker and Abramson (1970), their phoneme boundary does not correspond to the VOT values of either of the two boundaries found with these infants. The implication of these findings is that infants are capable of perceiving three major voicing distinctions in the absence of any specific experience with these particular voicing contrasts in the environment.

In another related study, Streeter (1976) found that Kikuyu infants are capable of discriminating voicing differences between labial stops which are not used phonologically by the adults in their language environment. In Kikuyu, a Bantu language spoken in Kenya, there is only one labial stop with a VOT value in the range of -60 msec (i.e., a pre-voiced stop). However, the Kikuyu infants could discriminate differences between three voicing categories corresponding roughly to the same ones found in the Lasky *et al.* study. Thus, the discrimination of these voicing contrasts can also be made in the absence of relevant linguistic experience with the specific sound contrasts. The results of the cross-language experiments suggest that infants may be predisposed, in some sense, to deal with these acoustic attributes with only a very limited exposure to the specific sounds and well before any experience in producing these distinctions.

The developmental course of speech perception, however, may be somewhat different from other forms of perceptual development in which it is assumed that environmental experience serves primarily to sharpen the discriminative capacities of an organism (Gibson and Gibson, 1955; Gibson, 1969). Since the child is capable of making relevant discriminations between the important distinctive acoustic attributes of speech at a very early age, the effects of linguistic experience may be restricted primarily to learning that particular distinctions are not functional within a child's language environment. Thus, the course of development may not involve learning to make finer and finer discriminations among stimulus attributes but may be more analogous to the effects of acquired similarity or equivalence (Gibson and Gibson, 1955; Liberman, Harris, Kinney and Lane, 1961). As Eimas (1976) has suggested "the course of development of phonetic competence is one characterized by a loss of abilities over time if specific experience is not forthcoming". Like the adult, if the phonetic distinctions are not used phonologically in the language, sensitivity to the relevant acoustic attributes is lowered and the child will fail to respond differentially to them. One proposal to account for the infant results is considered in the next section in terms of a detector system with feature detectors each sensitive to a restricted range of acoustic information. These detectors are assumed to be available innately to the

infant for processing the relevant acoustic attributes of speech although they can be modified substantially by specific linguistic experience in the environment.

The speech perception research on infants has also generated a great deal of interest in the types of acoustic attributes that are used for discrimination. As a result, a number of investigators have begun to study other organisms whose peripheral auditory system is similar to humans. Since these organisms have no spoken language and, therefore, lack a phonological system, it has been assumed that their discriminative behavior to speech stimuli would be determined exclusively by the acoustic and psychophysical attributes of the speech signals. For example, in one study, Burdick and Miller (1975) found that with appropriate training four chinchillas could respond differentially to the vowels /a/ and /i/ selected from an ensemble of tokens produced by different talkers at different pitch levels. These results are not surprising since they demonstrate that common spectral relations characterize similar vowels. As noted earlier, these relations are closely associated with the relative values of the steady-state formant frequencies.

In another study, Kuhl and Miller (1975) reported that chinchillas could learn to respond differentially to the consonants /d/ and /t/ in syllables produced by four talkers in three vowel contexts. Furthermore, the training experience with these stimuli generalized to synthetic speech sounds

varying in voice-onset time. Identification functions were obtained for the chinchilla that were quite similar to data obtained with humans for the identical stimulus continuum; the stimuli were partitioned into two discrete categories with a sharp crossover point at the boundary. Thus, these results indicate that there are quantal regions in perception which have well-defined psychophysical properties. In the case of voice-onset time, several changes in the acoustic attributes of these stimuli occur at precisely the region where the boundary between voiced and voiceless stops occur in a number of languages (see Stevens and Klatt, 1974; Lisker, 1975).

The experiments with chinchillas are no doubt of some interest in demonstrating that complex spectral and temporal relations are present in the acoustic waveform and that non-human organisms can be trained to respond differentially to them. But what inferences can be drawn from these results with regard to categorical perception? Unfortunately, Kuhl and Miller did not obtain discrimination data from their chinchillas so we can only guess at how well they might have done on this task. We would expect, however, that chinchillas could probably discriminate between pairs of stimuli selected from within a phonetic category since, presumably they do not code these sounds phonetically. Accordingly, discrimination should be based on the coding of only lower-level acoustic attributes. The results of Morse and Snowdon (1975)

and Sinnott (1974), although obtained with monkeys, are consistent with this prediction. Using somewhat different experimental procedures, these investigators found that monkeys could discriminate between pairs of speech sounds selected from within a phonetic category and they could do this better than would be predicted by the labeling hypothesis derived from the categorical perception model. Thus, while the chinchilla and monkey can be trained to respond differentially to speech sounds as acoustic signals, they may not code these signals as phonetic events as humans do.

D. Property Detectors in Speech Perception

To explain the results of the categorical-like discrimination found in infants, Eimas (1974) proposed an approach to speech perception based on the idea of feature detectors finely tuned to restricted ranges of acoustic information in the speech signal. This particular idea did not originate with Eimas since a number of other investigators has remarked on the possibility of some sort of feature detecting mechanism in speech perception (see, for example, Whitfield, 1965; Liberman *et al.*, 1967; Abbs and Sussman, 1971; Lieberman, 1970; Stevens, 1972). However, it was left to Eimas and Corbit (1973) to introduce an experimental paradigm, selective adaptation, to speech perception that could reveal the workings of these hypothesized detectors in some detail (see Cooper, 1975, for an extensive review). In selective adaptation, repetitive presentation of a stimulus

alters the perception of a set of test stimuli. For example, in the initial study, Eimas and Corbit (1973) investigated the voicing feature and showed that adaptation with the syllable [ba] caused the locus of the phonetic category boundary between [ba] and [p^ha] to shift towards the [ba] end of the continuum. Stimuli near the boundary which were identified as [ba] when the listener was in an unadapted state were subsequently labeled as [p^ha] after adaptation with [ba]. Similar findings were obtained when [p^ha] was used as an adaptor; the locus of the phonetic boundary shifted toward the [p^ha] end of the stimulus continuum. Eimas and Corbit also showed that these results were not specific to the syllables or phonetic segments in the test series but were due rather to the presence of a specific attribute or feature in the consonants. This conclusion was based on cross-series adaptation in which the voiceless bilabial stop [p^ha] produced approximately equivalent effects on the identification functions for a series of alveolar stop consonants (i.e., [d] and [t^h]) as it did for the series of bilabial stops (i.e., [b] and [p^h]). In both cases, the locus of the phonetic boundary shifted toward the voiceless end of the continuum. These results are shown in Figure 14 for one of the subjects in the Eimas and Corbit study.

- - - - - - - - - - - - - - -

Insert Figure 14 about here

- - - - - - - - - - - - - - -

In another experiment Eimas and Corbit showed that the peak in the discrimination function also shifts after adaptation suggesting that the shifts in the labeling function are not simply due to a response bias introduced by changing the stimulus probabilities. In a different study, Eimas, Cooper and Corbit (1973) showed that the adaptation effects are centrally located since presentation of the adaptor and test stimuli to different ears still produced large and reliable shifts in the locus of the phonetic boundary.

Eimas and Corbit (1973) interpreted the selective adaptation findings as support for the hypothesis that the perception of voicing involves two distinct types of feature detectors organized as opponent pairs, a voiced detector (+V) and a voiceless detector (-V). Each detector is assumed to be selectively tuned to a range of partially overlapping voice onset-time values. When a stimulus containing a particular VOT value is presented repetitively, it fatigues the detector most sensitive to that range of the feature and, accordingly, its sensitivity is reduced. After adaptation, the opponent or unadapted detector provides a greater output to the decision process in identification than the adapted detector and, accordingly, produces a shift in the locus of the phonetic category boundary.

At the time, Eimas and his collaborators argued that the selective adaptation results provided convincing support for the existence of detectors specialized for processing

phonetic features rather than the acoustic attributes that form the basis for these phonetic distinctions. However, their conclusions have been shown to be premature since more recent work has demonstrated that the adaptation effects are related more to spectral similarity between test series and adaptor than to phonetic identity (see Cooper and Blumstein, 1974; Tartter and Eimas, 1975; Bailey, 1975; Pisoni and Tash, 1975). One of the questions that is currently under extensive investigation is whether the selective adaptation results are due to fatigue of mechanisms that process the acoustic attributes of speech stimuli, their more abstract phonetic features or both. The specific details of the arguments are intricate and need not concern us here. The important point of the adaptation work is that the perceptual system responds to certain acoustic attributes of speech stimuli and these attributes turn out to underlie precisely the relevant distinctions made between phonetic segments.

One of the intriguing questions that the infant speech perception work has raised is the extent to which environmental input determines the development and sensitivity of these hypothesized feature detectors. There is an extensive literature on the role of early experience in the development of the visual system which indicates that early environmental experience can modify the selectivity of cortical cells in kittens (e.g., Hirsch and Spinelli, 1970; Blakemore and

Cooper, 1970). The analogy to this developmental work has already been drawn by Eimas (1976), who argues that the lack of experience with specific phonetic contrasts in the local environment during language acquisition has the effect of modifying the appropriate detectors by reducing their sensitivity. Some detectors originally designed to process certain phonetic distinctions may be captured or subsumed by other detectors after exposure to specific acoustic stimuli from the linguistic environment. The nonspecific detectors might, therefore, assume the specificity for only those attributes present in the stimuli to which they are exposed. The poor within category discrimination of speech sounds found with adults and infants in the categorical perception experiments may not only be due to phonetic coding of these signals but may also be a consequence of the modification of the needed discrimination mechanism (i.e., a feature detector). This could then account for the troughs found in the discrimination functions. Thus, considering the infant research, it may well be the case that the general program of development of speech perception is genetically determined with experience in the environment playing only a role in the tuning and alignment of the system. Moreover, the presence of linguistic universals, particularly in terms of the relatively small number of phonetic features found across many different languages, lends some support to this contention and implies that at

least some of the structural mechanisms underlying speech perception are part of the biological endowment of the organism.

In summary, the work on selective adaptation provides strong evidence for the existence of some type of feature detecting system in perception. These findings are important since they provide a way of determining precisely to what types of acoustic attributes the system responds and eventually may help to determine how this information is used by the perceptual system. However, the work on selective adaptation and the resulting feature detector models of speech perception as well as the infant perceptual research are still in the very earliest stages of development. Numerous questions still remain to be resolved such as specifying the locus of adaptation effects and determining whether there are separate classes of detectors that respond to auditory and phonetic feature information as well as detailing the role of early experience in speech perception. Work on these problems is currently being carried out in a number of laboratories and a good deal of progress can be anticipated over the next few years.

V. BASIC ISSUES IN SPEECH PERCEPTION

The previous section summarized several of the important empirical findings in speech perception, particularly work dealing with segmental perception. In this section, some of the major issues that have emerged from this work will be reviewed briefly. Emphasis will be placed on the relevance of these issues to theoretical accounts of speech perception which will be considered in the final section of the chapter.

The basic issues in speech perception are, in principle, no different from the problems encountered in other areas of perception. In general, these deal with the problems of perceptual constancy to diverse stimulation and of perceptual contrast to identical stimulation and how the physical environment is represented internally (Bruner, 1957).

A. Linearity, Invariance and Segmentation

One of the most important problems in speech perception is that the speech signal fails to meet the conditions of linearity and invariance (Chomsky and Miller, 1963). As a consequence, the recognition problem becomes quite a complicated task for humans as well as machines. The linearity condition assumes that for each phoneme there must be a particular stretch of sound in the utterance and if phoneme X is to the left of phoneme Y in the phonemic representation, the stretch of sound associated with X must precede the stretch of sound associated with Y in the

physical signal. The invariance condition assumes 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.

As noted in earlier sections, it has been extremely difficult to establish acoustic segments or features which match the perceived phonemes independently of context. As a result of co-articulation in speech production there is a great deal of contextual variability. Often a single acoustic segment contains information about several neighboring linguistic segments, and, conversely, the same linguistic segment is often represented acoustically in quite different ways depending on the surrounding phonetic context, rate of speaking and talker.

The context-conditioned variability between acoustic signal and phoneme resulting from co-articulation also presents enormous problems for segmentation of speech. Because of the failure to meet the linearity and invariance conditions, it has been difficult to segment speech into acoustically defined units that are independent of adjacent segments. Although some 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 and, moreover, no simple invariant mapping

has been found between these acoustic attributes and perceived phonemes.

The lack of invariance and segmentation problems suggest something peculiar about speech as an acoustic stimulus. Certainly, relations between segments of the acoustic signal and units of linguistic analysis are complex and this, in turn, places certain constraints on the classes of perceptual theories that might be proposed for speech perception. For example, filter or template-matching theories are generally believed to be poor candidates for speech perception primarily because linguistic segments cannot be defined exclusively by attributes of the acoustic signal. As Chomsky and Miller (1963) have remarked, if both the invariance and linearity conditions were met, the task of building machines capable of recognizing various phonemes in human speech would be greatly simplified. It would surely be a simple enough matter to arrange the appropriate filters in a network in order to construct a recognition device. Although numerous attempts have been made along these lines in the past, the results have been generally unsuccessful because of the inherent variability of the physical signal. It is therefore not at all surprising that passive theories of recognition involving template matching and filtering are held in poor regard as potential models of human speech perception.

To deal with the contextual variability of phonemes and the complex relation between acoustic signal and phonetic

segment, a number of investigators have proposed models based on specialized feature detectors. It was originally assumed that these complex phonetic feature detectors processed quite diverse acoustic inputs equivalently and thus provided one way to dodge the invariance and segmentation problems. As noted earlier, however, the selective adaptation results indicate that if there are feature detectors they appear to be quite sensitive to context. Moreover, most of the evidence adduced in support of specific phonetic feature detectors is weak and subject to alternative acoustically based interpretations (Pisoni and Tash, 1975). Other investigators have preferred "active" theories which employ higher-level linguistic information in the earliest stages of perceptual analysis. A brief account of this approach will be considered in the next section which deals with models of speech perception.

B. Articulation and the Internal Representation of Speech

There is good agreement among investigators that speech is represented internally as a sequence of discrete segments and features although there is somewhat less agreement as to the exact description of these features. Arguments have been proposed for feature systems based on distinctions in the acoustic domain and in the articulatory domain as well as systems which combine both types of distinctions (Jakobson, Fant and Halle, 1952; Chomsky and Halle, 1968;

Wickelgren, 1969). The suspicion that articulation may play a role in speech perception goes back to at least the 1600's (Stevens & House, 1972). In this section several reasons for proposing articulatory-motor involvement in speech perception will be considered briefly. Because this has been a controversial topic in the field of speech perception it is appropriate to review these claims at this point.

The first concern with articulation in speech perception is basically historical. Early acoustic phonetic work and much of the experimental research on speech perception that followed was guided by the traditional articulatory descriptions developed by phoneticians (Bell, 1867). The acoustic cues underlying the perception of different phonetic segments could be described in terms of the articulatory gestures and dimensions that distinguished these sounds in production (Delattre, 1951). Thus, speech sounds seemed to be naturally arranged in terms of a few simple and relatively independent dimensions which turned out to provide the same distinctions in perception that the articulatory dimensions provided in production and suggested that the two processes might be linked in some yet unknown way.

Ladefoged (1972) has also pointed out the need for some absolute frame of reference in the phonetic description of languages. Articulatory categories such as labial, alveolar or velar refer to observable phenomena which can be compared across different languages and can be used to express

commonalities among attributes used in these languages. Thus, articulatory descriptions could be used to express the similarities and differences among phonetic entities at a common level. Acoustic descriptions of speech were clearly too complex to capture the necessary linguistic distinctions between sounds.

Finally, another motivation for emphasizing the articulatory domain is the view that the speech production mechanism, which reflects the articulatory capabilities of human talkers, has been the contributing factor in the development of speech sound systems. For example, Peterson and Shoup (1966) note that "there is considerable reason to believe that the phonological aspects of speech are primarily organized in terms of the possibilities and constraints of the motor . . mechanism with which speech is produced (p. 7)". Indeed, speech sounds constitute a highly distinctive class of acoustic signals in the environment. They are produced by a sound source that has well defined acoustic constraints for the listener (Fant, 1960; Stevens, 1972). In most normal situations, the listener is also a speaker and presumably has access to knowledge of the correlation between vocal tract shape and resulting acoustic output. Thus, although the human auditory system is capable of processing a relatively large number of acoustic signals selected from a wide range of frequencies, only a very small number of distinctive acoustic attributes are actually employed in

the phonological systems of human languages. The acoustic properties of these particular attributes are highly constrained by limitations of the production mechanism and not the perceptual mechanism, at least not the peripheral auditory system.

C. Units of Perceptual Analysis

Another important issue in speech perception is the choice of a minimal unit of perceptual analysis. Because of limitations on channel capacity, especially in the auditory system, raw sensory information must be categorized and recoded into some more permanent form that can be used for subsequent analysis. But what is the basic or "natural" coding unit for speech perception? Many investigators have argued for the primacy of the feature, phoneme, syllable or word as their candidate for the basic perceptual unit. Other investigators, motivated chiefly by early work in generative linguistic theory, have proposed much larger units for perceptual analysis such as clauses or sentences (Miller, 1962; Bever, Lackner & Kirk, 1969). The current debate over the choice of a perceptual unit can be resolved if a strict distinction were made concerning the level of linguistic analysis under consideration. The size of the processing unit in speech perception apparently varies from feature to segment to clause as the level of linguistic processing changes and thus the question of whether there is one basic or primary unit may be inappropriate.

However, it is worth considering briefly an issue that has received some attention in recent years, namely, the arguments in favor of the syllable as the basic unit of analysis. A number of investigators have proposed that the syllable is a more basic perceptual or linguistic unit than the phoneme (Savin and Bever, 1970; Massaro, 1972). The claim is that phonemes are more abstract entities than syllables because some phonemes cannot exist independently as articulatory and acoustic units whereas syllables presumably can. In support of their claim, Savin and Bever (1970) presented reaction times obtained in a target monitoring task which showed that subjects respond faster to syllable targets than phoneme targets. Similar experiments also have been carried out by Foss and Swinney (1973) and McNeill and Lindig (1973) who argue that the reaction time results are due to the differential speed with which various sized units (i.e., phonemes, syllables, words, sentences) can become available to consciousness for a decision and not whether one unit is perceived earlier and, therefore, is more basic than another unit.

Taking a different tact, Massaro (1972) argued that syllables are the basic units of speech perception because phonemes are not represented discretely in the acoustic signal and accordingly cannot serve as perceptual units. He also has claimed that by assuming the syllable as the earliest unit of perceptual analysis the problems connected

mechanism requires minimal acoustic information distributed over at least a syllable-sized unit in order to analyze the constituent phonetic segments. This assumption says nothing about the size of the earliest perceptual unit but only that the minimal information for a phoneme requires the analysis of acoustic information spread over several adjacent segments.

D. Prosody in Speech Perception

Most of the research in speech perception, as well as the theoretical emphasis, has been concerned with segmental analysis of phonemes. A seriously neglected topic has been the prosodic or suprasegmental attributes of speech which involve differences in pitch, intensity and duration of segments. At the present time there is a wide gap between the research conducted on isolated segments and features and prosodic factors in speech perception (see Cohen and Nooteboom, 1975). How prosodic factors might be used in perception has not been considered in detail, although it is clear that this information serves as a possible link or interface between phonetic segments and features on the one hand and grammatical processes at higher levels on the other (see Huggins, 1972a, for a review).

There is evidence that differences in fundamental frequency provide important cues to segmentation of speech into constituents that are suitable candidates for syntactic analysis. Lea (1973) has found from acoustic analysis of

connected speech that a decrease in fundamental frequency (F_o) usually occurred at the end of each major syntactic constituent in a sentence, and an increase in F_o occurred near the beginning of the following constituent. Thus, prosodic cues may be used to cue syntactic structures. Lindblom and Svensson (1973) and Svensson (1974) have carried out investigations on the role of prosody in identifying various syntactic structures in the absence of segmental cues. Their findings indicate that a good deal of information about the surface syntactic structure of an utterance may be conveyed by prosodic features and that unambiguous identification can take place even with an incomplete specification of the acoustic properties of the phonetic segments.

Other evidence from studies on the acoustic analysis of speech indicates that the durations of phonetic segments vary in stressed and unstressed syllables as well as in various syntactic environments. For example, Oller (1973) has found substantial lengthening effects for both consonants and vowels in a number of environments, including utterances with various intonation patterns, word-final and phrase-final positions as well as utterance-final position. Klatt (1974) has shown that the duration of the segment [s] is longer in prestressed position and shorter before unstressed vowels and in word-final position. In another study, Klatt (1975) also found that vowel lengthening effects occur at

the end of major syntactic units such as the boundary between a noun phrase and a verb phrase.

It is clear that durational effects such as these argue against uncovering invariant acoustic attributes for phonemes that are identical in all phonetic environments. More importantly, however, these systematic lengthening effects may provide additional cues to the higher-order syntactic structure of the sentence. Much research remains to be done on this particular problem. It would be of interest to know, for example, the extent to which syntactic and semantic variables influence the durations of phonetic segments and the precision with which listeners can and do use this information in the recognition process (see Huggins, 1972b; Klatt and Cooper, 1975; Cooper, 1976).

E. Higher-Level Contributions to Speech Perception

A well-documented finding in speech perception is that words presented in sentential context are more intelligible than the same words presented in isolation (Miller, Heise, and Lichten, 1951; Pollack and Pickett, 1964). The usual interpretation of these findings is that syntax and semantics serve only to narrow down the number of possible response alternatives available to the listener (Miller, 1962). It is assumed that phonetic segments and features are recognized more or less directly from a set of physical properties in the sound wave and that processing proceeds serially on the basis of

independent decisions about individual speech segments at each successive level (Licklider, 1952; Halle and Stevens, 1962).

But what role does syntax and semantics serve in speech perception? Chomsky (1964) has argued that it is not possible to describe a language adequately by starting with only a description of the sound system without reference to the function of these sounds as linguistic entities. That is, more information than a phonetic sequence is necessary to establish the identity of a phoneme. This information presumably involves the contribution of syntactic and semantic variables to the recognition process.

Several recent studies have examined this problem experimentally. For example, Shockey and Reddy (1974) determined how well phonetically trained listeners could recover the correct phonemes in the absence of higher-order information such as syntax and semantics. To accomplish this, listeners heard recorded utterances of sentences taken from unfamiliar languages and were required to provide a phonetic transcription. The only basis for segmentation and recognition of these phonemes was therefore the information contained in the physical signal, since higher-level information was eliminated. The results of Shockey and Reddy's experiment are of interest. As it happened, the transcription task was quite difficult; only about 56 percent of the segments in the original utterances could be identified correctly. With an accuracy

in this range, it can be concluded that higher-levels of linguistic information must provide a good deal of additional information even to the earliest stages of perceptual analysis.[5]

In a similar experiment, Klatt and Stevens (1973) attempted to recognize a set of unknown sentences by visual examination of sound spectrograms along with a machine-aided lexical look-up program that was implemented on a computer. Although the syllable structure of an utterance could be identified reasonably well from a spectrographic representation, only 33 percent of the segments could be transcribed correctly whereas only another 40 percent of the segments could be partially transcribed. Klatt and Stevens (1973) emphasized that the problem of recognizing segments in sentences from spectrograms is more difficult than in isolated words because in sentences word boundaries are not as clearly marked and co-articulatory effects occur between adjacent words. Moreover, the duration of a segment is shorter in sentences than in isolated words and there is more vowel reduction. All these observations lead Klatt and Stevens to conclude that it is doubtful whether accurate recognition from spectrograms can be carried out in the absence of higher-level constraints. They suggest that these constraints can be used to verify decisions about segments and predict missing or distorted information based on the previous generation of some hypothesis or search set.

In another study, Marslen-Wilson (1975) found more direct evidence for the use of higher-levels of processing in speech perception. Subjects shadowed sentences in which specific changes were made in the syntactic, semantic or phonetic structure. Analysis of the restoration of disrupted words to their original form was shown to be dependent on the semantic and syntactic context variables. According to Marslen-Wilson (1975), the listener analyzes the incoming information at all levels of analysis so that decisions at any level can affect processing at other levels.

The results of these studies as well as the findings on prosody indicate that the perception of connected speech does not rely exclusively on the analysis and recognition of segmental acoustic features. When speech perception is viewed only from the vantage point of the acoustic signal and the phoneme, the task of finding invariant attributes becomes the primary focus. However, if prosodic information and higher-order variables are included in the perceptual process the scope of the potential models widens appreciably.

F. Speech Perception as a Specialized Process

For a number of years, Liberman and others have argued that speech perception is a specialized process requiring the postulation of specialized neural mechanisms and processes for perceptual analysis. But what is the evidence for a specialized mode of perception unique to speech? Some of the original support for this view came from the early

categorical perception experiments with synthetic speech sounds. As noted in previous sections, this form of perception was thought to be unique to only certain types of speech sounds, namely stop consonants. Recent findings, however, have shown comparable categorical effects for complex non-speech signals suggesting that the original interpretation of categorical perception was probably wrong and that the findings are a special case of a more general phenomenon involving the coding of complex acoustic signals.

Other evidence adduced in support of a specialized neural mechanism has come from dichotic listening experiments in which laterality effects have been demonstrated for competing acoustic signals (Studdert-Kennedy and Shankweiler, 1970). These results have been interpreted as evidence for hemispheric specialization in the perception of speech signals and processing of linguistic information. It has been known for over 100 years that the left hemisphere of most right-handed adults is the language-dominant hemisphere, specialized for linguistic analysis. The dichotic listening results provide behavioral support for this asymmetry even at the phonetic feature level and have been used as strong evidence for the assumption of a specialized neural mechanism for processing speech. Unfortunately, little is currently known about the specific types of processes or operations that the left hemisphere performs in perceiving speech and language other than that it is clearly different from the processing carried out in the right hemisphere.

The recent findings on infant speech perception have also been interpreted as supporting the view that some aspects of perceptual analysis of speech are innately determined and therefore may be due to some specialized neural process. The experiments demonstrating categorical-like discrimination with infants are, however, subject to precisely the same criticisms as the categorical perception experiments with adults. The discrimination findings may not be due exclusively to phonetic coding and linguistic processing as Eimas has argued, but may be the result of the specific psychophysical properties and memory representations of the specific speech signals employed. The cross-language experiments with infants as well as the labeling results obtained with the chinchilla provide additional support for this conclusion. Young infants may be predisposed in some sense to respond to speech signals in a phonetically relevant manner but alternative explanations based only on auditory processing capabilities have not been adequately ruled out yet. Moreover, our understanding of the exact contribution of the linguistic environment to the course of phonetic development, as well as the developmental process itself, is still quite meager at the present time. Thus, the infant speech perception findings while important in their own right are somewhat ambiguous with regard to providing convincing support for the speciality issue.

Finally, additional support for the speciality of speech has come from rational and logical considerations dealing with attempts to communicate language by sounds other than speech. It has been argued by Liberman and his colleagues that speech sounds are uniquely efficient vehicles for transmitting phonemic information in language primarily because speech represents a complex code rather than a simple cipher or alphabet. The rate at which phonemic information can be perceived in speech is known to be well above the temporal resolving power of the ear if listeners had to process only isolated acoustic segments which stood in one-to-one relation with phonemes (Liberman *et al.*, 1967). Liberman argues, however, that speech sounds are a code since they represent a substantial restructuring of phonemic information in the acoustic waveform.

It is reasonable to conclude that while there is some evidence for a specialized neural mechanism in speech perception, the exact nature of its operation and course of its development remain somewhat elusive problems at the present time. The blanket assertions of speciality have been shown to be generally inadequate as explanatory principles as more and more becomes known about the psychophysical and perceptual aspects of speech signals and the physiological mechanisms involved in speech production. Future work should be directed towards determining in precisely what ways speech perception conforms to general perceptual processes,

regardless of modality, and in what ways speech perception will clearly require the postulation of additional specialized explanatory theories and mechanisms for perceptual analysis.

VI. LEVELS OF PROCESSING AND THEORIES IN SPEECH PERCEPTION

Current theories of speech perception are quite general and vague and, for the most part, not terribly well-developed, at least by the standards in other areas of experimental psychology. Indeed, it is not unreasonable to characterize these theories as only preliminary attempts at specifying what a possible model of human speech perception might entail. A few quotes should make this point clear:

> Since this symposium is concerned with models, we should say at the outset that we do not have a model in the strict sense, though we are in search of one. (Liberman, Cooper, Harris, MacNeilage, and Studdert-Kennedy, 1967, p. 68)

> Any attempt to propose a model for the perception of speech is deemed to become highly speculative in character and the present contribution is no exception. (Fant, 1967, p. 111)

> Since we are still far from an understanding of the neurophysiological processes involved, any model, that can be proposed must be a functional model, and one can only speculate on the relation between components of the functional model and the neural events at the periphery of the auditory system and in the central nervous system. (Stevens and House, 1972, p. 47)

> We have no models specified in enough detail for serious test. (Studdert-Kennedy, 1976, p. 000)

In this section we summarize first the most widely accepted views of the levels of processing in speech perception as they have developed over the past few years, and then discuss several prominent theoretical views that have

influenced current thinking about models of the perceptual process.

A. Speech Perception as a Process

It has become common in recent years to think of speech perception as a process involving a series of levels between the initial acoustic waveform that impinges upon the ear and its final conceptual representation in the mind of the listener (Studdert-Kennedy, 1974, 1976). Most investigators assume that these levels are hierarchically organized although the exact relation to a definite sequence of processing stages is still a matter of some controversy. One view of these levels of processing is shown in Figure 15.

- - - - - - - - - - - - - - -
Insert Figure 15 about here
- - - - - - - - - - - - - - -

At the very lowest level of analysis in this figure is the acoustic structure of an utterance which may be thought of as the time-varying acoustic waveform. At the highest level is the conceptual representation of the utterance as a linguistic object. Such a representation is constructed from the listener's knowledge of events, relations and facts about the world that are present in long-term memory as well as his specific linguistic knowledge. These two levels, corresponding to sound and meaning, respectively, are linked by a number of intermediate levels of analysis which serve to transform and recode the initial acoustic input into

successively more abstract and often more elaborated representations.

B. Levels of Processing in Speech Perception

The view of speech perception suggested by Figure 15 assumed that speech is processed through a series of independent stages corresponding to the levels identified in the Figure. Although the specific levels have been derived from rational principles and linguistic considerations, the arrangement of the processing stages and their specific function is still a topic of current investigation (see Pisoni and Sawusch, 1975). As such, we focus primarily on some of the levels of analysis that have been identified and are generally agreed upon and postpone a discussion of processing stages until the end of this section.

1. Auditory Level - The auditory level is assumed to be the first and earliest stage in the perceptual analysis of speech. At this level the acoustic waveform is transformed or recoded into some neural representation in the nervous system. Acoustic information about spectral structure, fundamental frequency, overall intensity and duration of the signal as well as amplitude onsets and offsets is extracted and coded by the auditory system. These properties are assumed to be preserved in sensory memory for a brief period of time until subsequent operations can be carried out to recode this information into a more permanent form in short-term memory.

2. Phonetic Level - The phonetic level is assumed to be the next stage of analysis. Here features and segments necessary for phonetic classification are abstracted or derived from the auditory representations of the acoustic signal at the previous stage and placed in short-term memory. As discussed earlier, there is a many-to-many mapping of the auditory attributes derived from the previous level and the phonetic features identified at this level. Phonetic features may be thought of simply 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. However, it has been convenient to describe these features in terms of articulatory descriptions and labels primarily since this notation captures linguistically relevant distinctions at a phonetic and phonological level. The 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 and Halle, 1968). Some segmentation is assumed to already have taken place such that segments and some juncture boundaries are indicated in the matrix representation.

3. Phonological Level - At this level the phonetic segments from the previous level are converted into phonological segments or phonemes. The phonological component provides information about the sound structure of a given language which is imposed on the phonetic matrix to derive

a phonological matrix. 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. And it is also at this level that allophonic variations present at the phonetic level are eliminated and only phonologically distinctive information is coded for subsequent processing.

4. Higher Levels of Analysis - There are also several additional levels of analysis involving lexical search, syntactic analysis and semantic interpretation of the original input. These higher levels serve to generate the structure into which the phonological segments are placed as well as specifying the grammatical organization of the input. Since the primary concern of this chapter has been with the lower levels of speech perception, we will not consider these higher levels of analysis any further in this discussion. However, the reader should not be led to believe that their contribution to the lower levels of perceptual analysis is unimportant. As noted earlier, the information generated at these levels may be used to constrain phonetic and phonological interpretations as well as to guide lexical search and subsequent word verification processes.

C. Models of Speech Perception

Historically, models of speech perception were derived from the taxonomic principles of segmentation and classification. It was generally assumed that the level of linguistic structure corresponding roughly to the phoneme could be discovered entirely by the use of these cateloguing procedures applied automatically to a corpus of data. Moreover, according to this view, a separate level of phonological structure could be described independently of higher-level grammatical and syntactic considerations. The process of phoneme recognition was assumed to take place passively by means of successive filtering of the acoustic signal or by template matching of the acoustic input against stored representations in memory.

More recent models, developed primarily under the impetus of generative linguistic theory, have stressed a much more active or dynamic view of the recognition process. Models proposed along these lines assume phoneme recognition is carried out by matching some representation of the input signal against an internally generated representation. The presence of active feedback loops in the system permits the recognition of a wide variety of diverse signals since the parameters can be adjusted continually over time. In this final section we consider briefly a number of the prominent theories in speech perception and discuss several aspects of their approach to the major problems raised earlier in the chapter.

1. Motor Theory of Speech Perception - The basic assumption of the motor theory is that "speech is perceived by processes that are also involved in its production" (Liberman *et al.*, 1967). This view of speech perception was motivated by the observation that the listener is also a speaker and it would be more economical to assume that the speaker-hearer uses only one common process for language processing rather than two separate processes. As mentioned earlier, one of the central problems in speech perception at the phonemic level and the major reason for postulating a motor theory is the issue of the lack of invariance between the acoustic signal and its phonemic representation. Advocates of the motor theory argue that one possible way of resolving the invariance problem is to assume that the same perceptual response to widely differing acoustical signal arises because the intended pattern is produced by the same articulation or underlying motor commands to the articulators. Similarly, different perceptual responses to fairly similar acoustic signals arise from different underlying articulations or motor commands.

Although the motor theory has occupied an overwhelmingly dominant place in contemporary accounts of speech perception, the link between empirical data and theory has not been very strong. Indeed, much of the early support for the motor theory was based on recurrent demonstrations of the same experimental outcome using near identical techniques and a

very restricted set of stimuli. As discussed earlier, the primary data used to support the early versions of the motor theory came from perceptual experiments that found differences between synthetic stop consonants and steady-state vowels. If we set aside the consonant and vowel differences which may be due to other factors, there seems to be very little direct empirical support for some active mediation of articulatory knowledge or information during perceptual processing. Most of the current arguments for motor theory rest on parsimony, logic and faith rather than a firm empirical foundation. The most serious problem for the motor theory rests on the failure to specify the level of perceptual analysis where articulatory knowledge is employed in recognition.

2. Analysis-by-Synthesis - The analysis-by-synthesis model proposed by Stevens is much more explicit than the motor theory (Stevens, 1960; Stevens and Halle, 1967; Stevens and House, 1972). The basic assumption of the model is similar to the motor theory: there exist close ties between the processes of speech production and perception, and there are components and operations common to both. According to Stevens, the perceptual process begins initially with peripheral processing of the speech signal to yield a description in terms of auditory patterns. In cases where phonetic features are not strongly context-dependent, the auditory pattern will provide a relatively direct mapping of those features during preliminary analysis. The output of

preliminary analysis is a rough matrix of phonetic segments and features which is then transferred to a control system. Recognition of some features is thus assumed to take place by relatively direct operations on the acoustic information output from peripheral analysis. However, when there are no invariant attributes to identify a phonetic feature, additional processing is required. In analysis-by-synthesis an hypothesis concerning the representation of the utterance in terms of phonetic segments and features (i.e., an abstract distinctive feature matrix) is constructed. This representation then forms the input to a set of generative rules which produce candidate patterns which are subsequently compared against the original patterns. Results of this match are sent to a control component which transfers the phonetic description to higher stages of linguistic analysis. Analysis-by synthesis is simply a more carefully specified version of the motor theory except that the comparison process takes place at the neuroacoustic level rather than the neuromotor level. Analysis-by-synthesis like the motor theory is quite abstract and little direct experimental evidence has been found to support this view.

3. Fant's Auditory Theory - Fant's theory of speech perception is not terribly well-developed. He objects strongly to the "active" motor-type theories on the grounds that the evidence used to support these theories is not conclusive (Fant, 1967). Fant claims that all the arguments

brought forth in support of the motor theory would fit just as well into sensory-based theories in which the decoding process proceeds without postulating the active mediation of speech motor centers. The basic idea in Fant's approach is that the motor and sensory functions become more and more involved as one proceeds from the peripheral to central stages of analysis. He assumes that the final destination is the concept of a "message" which comprises brain centers common to both perception and production. However, according to Fant, there are separate sensory (auditory) and motor (articulatory) branches, although he leaves the possibility of interaction between these two blocks open. Auditory input is first processed by the ear and subject to some primary auditory analysis. These incoming auditory signals are then submitted to some kind of direct encoding into distinctive auditory features (Fant, 1962). Auditory features are then combined together in some unspecified way to form phonemes, syllables, morphemes and words. Although much of Fant's concern has been on continued acoustical investigations of distinctive features of phonemes, the model is much too gross to be tested in any serious way. Moreover, the problems of invariance and segmentation which are central issues in speech perception still remain to be resolved by the model.

4. Stage Theories - In line with a number of other investigators, Bondarko *et al.* (1970) have proposed a model containing a series of hierarchically organized stages of

perceptual analysis. The following stages are proposed: (a) auditory analysis, (b) phonetic analysis, (c) morphological analysis, (d) syntactic analysis and (e) semantic analysis. Only the first two stages in their model will be considered here. The first stage involves auditory analysis of the signal and provides a description of the stimulus in terms of "auditory features". Some form of spectral analysis is assumed to take place at this stage although the presence of auditory feature detectors which respond differentially to specific acoustic attributes is also considered a possibility. Thus, some low-level auditory feature analysis may take place relatively close to the periphery. The second stage in the model involves phonetic analysis. The output of this stage is abstract and may be either an acoustic or articulatory representation of the speech signal. The representation of the information at this stage is thought to be based on phonetic segments or distinctive features. Provision is also made for the use of various types of normalization routines so that variability from different talkers can be reduced for subsequent processes of recognition. The model deals with the invariance problem by assuming that multiple decisions can be made on the basis of outputs from a number of auditory feature detectors. Thus, the model is designed to process multiple cues simultaneously and in parallel rather than to focus on only isolated auditory features (i.e., acoustic cues).

5. A Novel Theory of Speech Perception - Chomsky and Miller (1963) and Chomsky and Halle (1968) have proposed a "novel theory of speech perception" that also merits some attention. According to these authors, speech perception is an "active process" in which acoustic information is utilized to form hypotheses about the structure of sentences. These hypotheses which are based, in part, on expectations and knowledge of the language, are then subsequently used to generate syntactic structures which can be compared against the original input. In this way, the listener uses knowledge of phonological principles to determine the phonetic properties of the sentence. As Chomsky and Halle have put it, a description of the perceptual process might be as follows:

> The hearer makes use of certain cues and certain expectations to determine the syntactic structure and semantic content of an utterance. Given a hypothesis as to its syntactic structure--in particular its surface structure--he uses the phonological principles that he controls to determine a phonetic shape. The hypothesis will then be accepted if it is not too radically at variance with the acoustic material...Given acceptance of such a hypothesis, what the hearer 'hears' is what is internally generated by the rules. That is, he will 'hear' the phonetic shape determined by the postulated syntactic structure and the internalized rules. (Chomsky and Halle, 1968, p. 24)

This view of the perceptual process is, of course, a version of the analysis by synthesis theory summarized earlier although the comparison process takes place at the phonological and syntactic levels rather than the neuro-

acoustic level. While this theory is very abstract, it emphasizes an important point that has often been overlooked in speech perception work. A listener's final interpretation of a speech signal depends upon a number of variables including the listener's linguistic knowledge as well as many extragrammatical factors that determine what the listener expects to hear in a given situation.

One preliminary account of speech perception based on this approach is illustrated schematically in Figure 16.

- - - - - - - - - - - - - - - -
Insert Figure 16 about here
- - - - - - - - - - - - - - - -

Processing stages corresponding to syntactic, lexical and semantic levels are assumed to be arranged more or less in parallel. Auditory input first enters the system and is processed by the peripheral auditory mechanism. Sensory information is then passed on to several intermediate stages shown as the "recognition device" in this figure. Various operations associated with the feature detection and recognition process are assumed to be carried out here in a relatively automatic fashion. The output of the recognition device is a rough and tentative classification of the features and some of the segments in the input. Some features can be recognized by fairly straightforward operations on the input signals whereas other features may be analyzed in only a very gross way pending additional information about the phonological and syntactic structure of

the sentence. The output of the recognition device is then passed on to the phonological component where phonological decoding rules are employed in conjunction with syntactic and semantic information to derive a lexical representation. At this stage, information about segments, stress assignment, duration of each segment and fundamental frequency is used to arrive at a tentative representation of the lexical items in the input sentence. In this model, the word recognition process is thought to allow considerable ambiguity pending information from higher levels.[6]

The earliest stages of speech perception are assumed to occur relatively automatically and are not under the conscious control of the listener. There is a good deal of uncertainty about the exact composition of the features and segments of the input, the placement of word boundaries and the specific syntactic structure of the utterance. Although very little information is generally available in the waveform about the exact phonological or syntactic structure of the input, some initial acoustic information is necessary to formulate hypotheses for generating a syntactic structure. This initial structure provides the necessary information for specifying the segments and features of the phonological representation. One consequence of this view is that feature recognition and segmentation may result from similar processing operations. Similarly, word boundary placements may be determined simply as a by-product of a listener's solution to the word recognition problem.

The approach to speech perception proposed here is rather speculative and in need of empirical support. Nevertheless, it is consistent with the emphasis of other active approaches to perception in which the structure is imposed on the input by the listener during the perceptual process (Neisser, 1967). Thus, the listener does not identify speech feature-by-feature or phoneme-by-phoneme in some strictly serial order from bottom to top; rather, a gross and tentative analysis is performed initially with the final decision delayed until information from higher levels is available.

One common property of the diverse theoretical approaches reviewed in this section is that they all are quite general and vague and not easily tested empirically, at least in any obvious way. Because speech perception is a complex process involving several interrelated components, it has been difficult to formalize all relevant aspects of the process into a coherent model. The goal of future research in speech perception over the next few years is, however, quite clear; a global theory with a specific model should be developed that is on the one hand rich enough to account for the relevant phenomena in speech perception, yet, on the other, precise enough to be subject to empirical test and modeling.

ACKNOWLEDGEMENTS

Preparation of this chapter was supported, in part, by NIMH Grant MH-24027 and NIH Grant NS-12179 to Indiana University and, in part, by NIH Grants NS-07040 and NS-04332 to the Research Laboratory of Electronics, Massachusetts Institute of Technology, where the final manuscript was completed. I thank Professor Kenneth Stevens and Dr. Dennis Klatt for their advice and kind hospitality at M.I.T., Professor Michael Studdert-Kennedy and Professor Peter Eimas for their comments and criticisms on an earlier draft of the chapter and Professor W.K. Estes for his patience and thoughtful suggestions during the time this chapter was being written.

BIBLIOGRAPHY

Abbs, J.H. & Sussman, H.M. Neurophysiological detectors and speech perception: A discussion of theoretical implications. Journal of Speech and Hearing Research, 1971, 14, 23-36.

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

Bailey, P.J. Perceptual adaptation in speech: Some properties of detectors for acoustical cues to phonetic distinctions. Unpublished doctoral dissertation, University of Cambridge, 1975.

Bell, A.M. Visible speech and the science of universal alphabetics. London: Simpkin, Marshall & Co., 1867.

Bever, T.G., Lackner, J. & Kirk, R. The underlying structures of sentences are the primary units of immediate speech processing. Perception & Psychophysics, 1969, 5, 225-231.

Blakemore, C. and Cooper, G.F. Development of the brain depends on the visual environment. Nature, 1970, 228, 477-478.

Bondarko, L.V. *et al.*. A Model of Speech Perception in Humans. Working Papers in Linguistics No. 6. Computer & Information Science Research Center, Ohio State Univ., Columbus, Ohio. Technical Report 70-12, 1970.

Bruner, J.S. Neural mechanisms in perception. Psychological Review, 1957, 64, 340-358.

Brudick, C.K. and Miller, J.D. Speech perception by the chinchilla: Discrimination of sustained /a/ and /i/. Journal of the Acoustical Society of America, 1975, 58, 415-427.

Cairns, H.S. and Kamerman, J. Lexical information processing during sentence comprehension. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 170-179.

Chomsky, N. Syntactic Structures. The Hague, Netherlands: Mouton, 1957.

Chomsky, N. Current issues in linguistic theory. In: Fodor, J.A. and Katz, J.J. (eds.) The Structure of Language. Englewood Cliffs, New Jersey: Prentice-Hall, 1964, pp. 50-118.

Chomsky, N. and Halle, M. The Sound Pattern of English. New York: Harper and Row, 1968.

Chomsky, N. and Miller, G.A. Introduction to the formal analysis of natural languages. In: Luce, R.D., Bush, R. and Galanter, E. (eds.) Handbook of Mathematical Psychology. Vol. 2. New York: John Wiley & Sons, 1963, pp. 269-321.

Cohen, A. and Nooteboom, S. (eds.) Structure and Process in Speech Perception. Heidelberg: Springer-Verlag, 1975.

Cole, R.A. and Scott, B. The phantom in the phoneme: Invariant cues for stop consonants. Perception and Psychophysics, 1974, 15, 101-107. (a)

Cole, R.A. and Scott, B. Toward a theory of speech perception. Psychological Review, 1974, 4, 348-374. (b)

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M. and Gerstman, L.J. Come experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 1952, 24, 597-606.

Cooper, F.S., Liberman, A.M. and Borst, J.M. The inter-conversion of audible and visible patterns as a basis for research in the perception of speech. Proceedings of the National Academy of Sciences, 1951, 37, 318-325.

Cooper, W.E. Selective adaptation to speech. In: F. Restle, R.M. Shiffrin, N.J. Castellan, H. Lindman and D.B. Pisoni (eds.) Cognitive Theory: Vol. 1. Hillsdale, N.J.: Erlbaum Associates, 1975, pp. 23-54.

Cooper, W.E. Syntactic control of timing in speech production. Unpublished doctoral thesis, Massachusetts Institute of Technology, 1976.

Cooper, W.E. and Blumstein, S.E. A "labial" feature analyzer in speech perception. Perception and Psychophysics, 1974, 15, 591-600.

Cutting, J.E. and Rosner, B.S. Categories and boundaries in speech and music. Perception and Psychophysics, 1974, 16, 564-570.

Daniloff, R. and Moll, K. Coarticulation of lip rounding. Journal of Speech and Hearing Research, 1968, 11, 707-721.

Darwin, C.J. The perception of speech. In: Carterette, E.C. and Friedman, M.P. (eds.) Handbook of Perception. New York: Academic Press, 1976, pp. 000-000.

Delattre, P. The physiological interpretation of sound spectrograms. Publication of the Modern Language Association, 1951, 66, 864-875.

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

Delattre, P.C., Liberman, A.M. and Cooper, F.S. Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. Studia Linguistica, 1964, 104-121.

Delattre, P.C., Liberman, A.M., Cooper, F.S. and Gerstman, L.J. Observations on one- and two-formant vowels synthesized from spectrographic patterns. Word, 1952, 8, 195-210.

Denes, P. Effect of duration on percetpion of voicing. Journal of the Acoustical Society of America, 1955, 27, 761-764.

Dorman, M., Studdert-Kennedy, M. and Raphael, L. The invariance problem in initial voiced stop consonants: Release bursts and formant transitions as functionally equivalent context-dependent cues. Haskins Laboratories Status Report on Speech Research, 1976, SR-45, 000-000.

Eimas, P.D. Auditory and linguistic processing of cues for place of articulation by infants. Perception and Psychophysics, 1974, 16, 513-521.

Eimas, P.D. Speech perception in early infancy. In: Cohen, L.B. and Salapatek, P. (eds.) Infant Perception. New York: Academic Press, 1975. (a)

Eimas, P.D. Auditory and phonetic coding of the cues for speech: Discrimination of the r-l distinction by young infants. Perception and Psychophysics, 1975, 18, 341-347. (b)

Eimas, P.D. Developmental aspects of speech perception. In Held, R., Leibowitz, H. and Teuber, H.L. (eds.) Handbook of Sensory Physiology: Perception. New York: Springer-Verlag, 1976.

Eimas, P.D., Cooper, W.E. and Corbit, J.D. Some properties of linguistic feature detectors. Perception and and Psychophysics, 1973, 13, 2, 247-252.

Eimas, P.D. and Corbit, J.D. Selective adaptation of linguistic feature detectors. Cognitive Psychology, 1973, 4, 99-109.

Eimas, P.D., Siquiland, E.R., Jusczyk, P. and Vigorito, J. Speech perception in infants. Science, 1971, 171, 303-

Fant, G. Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

Fant, G. Descriptive analysis of the acoustic aspects of speech. Logos, 1962, 5, 3-17.

Fant, G. Auditory patterns of speech. In: Wathen-Dunn, W. (ed.) Models for the Perception of Speech and Visual Form. Cambridge, Massachusetts: M.I.T. Press, 1967.

Fant, G. Stops in CV-syllables. Speech Transmission Laboratory Quarterly Progress and Status Report No. 4, 1969, 1-25.

Fant, G. Speech Sounds and Features. Cambridge, Massachusetts: M.I.T. Press, 1973.

Flanagan, J.L. Speech Analysis, Synthesis and Perception (2nd edition). New York: Academic Press, 1972.

Foss, D.J. and Swinney, D.A. On the physiological reality of the phoneme: Perception, identification and consciousness. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 246-257.

Fry, D.B., Abramson, A.S., Eimas, P.D. and Liberman, A.M. The identification and discrimination of synthetic vowels. Language and Speech, 1962, 5, 4, 171-189.

Fujisaki, H. and Kawashima, T. On the modes and mechanisms of speech perception. Annual Report of the Engineering Research Institute, Vol. 28, Faculty of Engineering, University of Tokyo, 1969, pp. 67-73.

Fujisaki, H. and Kawashima, T. Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute, Vol. 29, Faculty of Engineering, University of Tokyo, 1970, pp. 207-214.

Gibson, E.J. Principles of Perceptual Learning and Development. New York: Appleton Century Crofts, 1969.

Gibson, J.J. and Gibson, E.J. Perceptual learning: Differentiation or enrichment? Psychological Review, 1955, 62, 32-41.

Halle, M. and Stevens, K.N. Speech recognition: A model and a program for research. IRE Transactions of the Professional Group on Information Theory, 1962, IT-8, 155-159.

Harris, K.S. Cues for the discrimination of American English fricatives in spoken syllables. Language and Speech, 1958, 1, 1-7.

Harris, K.S. Physiological aspects of articulatory behavior. In: Sebeok, T.A. (ed.) Current Trends in Linguistics: Vol. 12, The Hague: Mouton, 1974.

Heinz, J.M. and Stevens, K.N. On the properties of voiceless fricative consonants. Journal of the Acoustical Society of America, 1961, 33, 589-596.

Hirsch, H.V.B. and Spinelli, D.N. Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. Science, 1970, 168, 869-871.

Hockett, C.F. Manual of Phonology. Indiana University Publications in Anthropology and Linguistics, No. 11, Bloomington, Indiana 1955.

Hockett, C.F. A Course in Modern Linguistics. New York: The MacMillan Company, 1958.

Huggins, A.W.F. On the perception of temporal phenomena in speech. Journal of the Acoustical Society of America, 1972, 51, 1279-1290. (a)

Huggins, A.W.F. Just noticeable differences for segment duration in natural speech. Journal of the Acoustical Society of America, 1972, 51, 1270-1278. (b)

Jakobson, R. Child Language, Aphasia and Phonological Universals. The Hague: Mouton, 1968.

Jakobson, R., Fant, G. and Halle, M. Preliminaries to Speech Analysis. Technical Report No. 13, Acoustics Laboratory, Massachusetts Institute of Technology, May, 1952.

Klatt, D.H. The duration of [s] in English words. Journal of Speech and Hearing Research, 1974, 17, 51-63.

Klatt, D.H. Structure of a phonological rule component for a synthesis-by-rule program. Paper presented at the Acoustical Society of America, San Francisco, November, 1975. (a)

Klatt, D.H. Vowel lengthening is syntactically determined in a connected discourse. Journal of Phonetics, 1975, 3, 129-140. (b)

Klatt, D.H. and Cooper, W.E. Perception of segment duration in sentence contexts. In: Cohen, A. and Nooteboom, S. (eds.) Structure and Process in Speech Perception. Heidelberg: Springer-Verlag, 1975, pp. 000-000.

Klatt, D.H. and Stevens, K.N. On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment. IEEE Transactions on Audio and Electroacoustics, 1973, AU-21, 210-217.

Kuhl, P.K. and Miller, J.D. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science, 1975, 190, 69-72.

Ladefoged, P. Phonetic prerequisites for a distinctive feature theory. In: Valdman, A. (ed.) Papers in Linguistics and Phonetics to the Memory of Pierre Delattre. The Hague: Mouton, 1972, 273-285.

Lane, H.L. Psychophysical parameters of vowel perception. Psychological Monographs, 1962, 76 (44, whole No. 563).

Lane, H.L. The motor theory of speech perception: A critical review. Psychological Review, 1965, 72, 275-309.

Lasky, R.E., Syrdal-Lasky, A. and Klein, R.E. VOT discrimination by four to six and a half month old infants from Spanish environments. Journal of Experimental Child Psychology, 1975, 20, 215-225.

Lea, W.A. An approach to syntactic recognition without phonemics. IEEE Transactions on Audio and Electroacoustics, 1973, AU-21, No. 3, 249-258.

Liberman, A.M. Some results of research on speech perception. Journal of the Acoustical Society of America, 1957, 29, 117-123.

Liberman, A.M. Some characteristics of perception in the speech mode. In: Hamburg, D.A. (ed.) Perception and Its Disorders, Proceedings of A.R.N.M.D. Baltimore: Williams and Wilkins Co., 1970, pp. 238-254. (a)

Liberman, A.M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 4, 301-323. (b)

Liberman, A.M., Cooper, F.S., Harris, K.S. and MacNeilage, P.F. A motor theory of speech perception. In: Fant, G. (ed.) Proceedings of the Speech Communication Seminar. Stockholm. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, 1963.

Liberman, A.M., Cooper, F.S., Harris, K.S., MacNeilage, P.F. and Studdert-Kennedy, M. Some observations on a model for speech perception. In: Wathen-Dunn, W. (ed.) Models for the Perception of Speech and Visual Form. Cambridge: M.I.T. Press, 1967.

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

Liberman, A.M., Delattre, P.C., and Cooper, F.S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. American Journal of Psychology, 1952, 65, 497-516.

Liberman, A.M., Delattre, P.C., Gerstman, L.J. and Cooper, F.S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology, 1956, 52, 127-137.

Liberman, A.M., Harris, K.S., Hoffman, H.S. and Griffith, B.C. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 1957, 54, 358-368.

Liberman, A.M., Harris, K.S., Kinney, J.A. and Lane, H.L. The discrimination of relative onset time of the components of certain speech and non-speech patterns. Journal of Experimental Psychology, 1961, 61, 379-388.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P.C., and Cooper, F.S. Minimal rules for synthesizing speech. Journal of the Acoustical Society of America, 1959, 31, 1490-1499.

Licklider, J.C.R. On the process of speech perception. Journal of the Acoustical Society of America, 1952, 24, 590-594.

Lieberman, P. Towards a unified phonetic theory. Linguistic Inquiry, 1970, 1, 3, 307-322.

Lieberman, P., Crelin, E.S. and Klatt, D.H. Phonetic ability and related anatomy of the newborn and adult human. Neanderthal man, and the chimpanzee. American Anthropologist, 1972, 74, 4, 287-307.

Liljencrants, J. and Lindblom, B. Numerical simulation of vowel quality systems: The role of perceptual contrast. Language, 1972, 48, 839-862.

Lindblom, B.E.F. Spectrographic study of vowel reduction. Journal of the Acoustical Society of America, 1963, 35, 1773-1781.

Lindblom, B. and Sundberg, J. A quantitative model of vowel production on the distinctive features of Swedish vowels. Speech Transmission Laboratory, Quarterly Progress and Status Report, No. 1, 1969, 14-30.

Lindblom, B.E.F. and Svensson, S.-G. Interaction between segmental and nonsegmental factors in speech recognition. IEEE Transactions on Audio and Electroacoustics, 1973, AU-21, 536-545.

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

Lisker, L. Is it VOT or a first-formant transition detector? Journal of the Acoustical Society of America, 1975, 57, 6, 1547-1551.

Lisker, L. and Abramson, A.S. The voicing dimension: Some experiments in comparative phonetics. In: Proceedings of the Sixth International Congress of Phonetic Sciences, Prague. Prague: Academia, 1970, pp. 563-567.

MacNeilage, P.F. Motor control of serial ordering of speech. Psychological Review, 1970, 77, 182-196.

Malmberg, B. The phonetic basis for syllable division. Studia Linguistica, 1955, 9, 80-87.

Marslen-Wilson, W.D. Sentence perception as an interactive parallel process. Science, 1975, 189, 226-228.

Massaro, D.W. Perceptual images, processing time, and perceptual units in auditory perception. Psychological Review, 1972, 79, 124-145.

Massaro, D.W. Auditory information processing and short-term memory. In: Estes, W.K. (ed.) Handbook of Learning and Cognitive Processes. Hillsdale, New Jersey: Erlbaum Associates, 1976, pp. 000-000.

Mattingly, I.G. Synthesis by rule of general American English. Supplement to Status Report on Speech Research, Haskins Laboratories, New York, 1968.

McNeill, D. The Acquisition of Language. New York: Harper and Row, 1970.

McNeill, D. and Lindig, L. The perceptual reality of phonemes, syllables, words and sentences. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 419-430.

Menyuk, P. The Acquisition and Development of Language. Englewood Cliffs: Prentice-Hall, 1971.

Miller, G.A. The magic number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Miller, G.A. Decision units in the perception of speech. IRE Transactions on Information Theory, 1962, IT-8, 81-83.

Miller, G.A., Heise, G.A. and Lichten, W. The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology, 1951, 41, 329-335.

Miller, J.D., Wier, C.C., Pastore, R., Kelly, W.J. and Dooling, R.J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 00, 000-000.

Moffitt, A.R. Consonant cue perception by twenty to twenty-four week old infants. Child Development, 1971, 42, 717-731.

Morse, P.A. The discrimination of speech and nonspeech stimuli in early infancy. Journal of Experimental Child Psychology, 1972, 14, 477-492.

Morse, P.A. and Snowdon, C.T. An investigation of categorical speech discrimination by rhesus monkeys. Perception and Psychophysics, 1975, 17, 9-16.

Neisser, U. Cognitive Psychology. New York: Appleton-Century, 1967.

Ohman, S.E.G. Coarticulation in VCV utterances: Spectrographic measurements. Journal of the Acoustical Society of America, 1966, 39, 151-168.

Oller, D.K. The effect of position in utterance on speech segment duration in English. Journal of the Acoustical Society of America, 1973, 54, 1235-1247.

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

Peterson, G.E. and Shoup, J.E. A physiological theory of phonetics. Journal of Speech and Hearing Research, 1966, 9, 1, 5-67.

Pisoni, D.B. On the nature of categorical perception of speech sounds. Supplement to Status Report on Speech Research, SR-27, Haskins Laboratories, New Haven, November, 1971.

Pisoni, D.B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception and Psychophysics, 1973, 13, 253-260.

Pisoni, D.B. Auditory short-term memory and vowel perception. Memory and Cognition, 1975, 3, 7-18.

Pisoni, D.B. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. Progress Report No. 118, Research Laboratory of Electronics, M.I.T., June, 1976, 000-000.

Pisoni, D.B. and Sawusch, J.R. Some stages of processing in speech perception. In: Cohen, A. and Nooteboom, S. (eds.) Structure and Process in Speech Perception. Heidelberg: Springer-Verlag, 1975, pp. 16-34.

Pisoni, D.B. and Tash, J.B. Reaction times to comparisons within and across phonetic categories. Perception and Psychophysics, 1974, 15, 285-290.

Pisoni, D.B. and Tash, J.B. Auditory property detectors and processing place features in stop consonants. Perception and Psychophysics, 1975, 18, 401-408.

Pollack, I. The information in elementary auditory displays. Journal of the Acoustical Society of America, 1952, 24, 745-749.

Pollack, I. The information in elementary auditory displays II. Journal of the Acoustical Society of America, 1953, 25, 765-769.

Pollack, I. and Pickett, J.M. The intelligibility of excerpts from conversation. Language and Speech, 1964, 6, 165-171.

Popper, R.D. Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. Journal of Psycholinguistic Research, 1972, 1, 205-219.

Raphael, L.J. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. Journal of the Acoustical Society of America, 1972, 51, 1296-1303.

Savin, H.B. and Bever, T.G. The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 1970, 9, 295-302.

Schatz, C. The role of context in the perception of stops. Language, 1954, 30, 47-56.

Shockey, L. and Reddy, R. Quantitative analysis of speech perception: Results from transcription of connected speech from unfamiliar languages. Paper presented at the Speech Communication Seminar, Stockholm, Sweden, August 1-3, 1974.

Sinnott, J.M. A comparison of speech sound discrimination in humans and monkeys. Unpublished doctoral dissertation, University of Michigan, 1974.

Siqueland, E.R. and DeLucia, C.A. Visual reinforcement of non-nutritive sucking in human infants. Science, 1969, 165, 1144-1146.

Stevens, K.N. Toward a model for speech recognition. Journal of The Acoustical Society of America, 1960, 32, 1, 47-55.

Stevens, K.N. Acoustic correlates of certain consonantal features. Paper presented at Conference on Speech Communication and Processing, M.I.T., Cambridge, Ma., November 6-8, 1967.

Stevens, K.N. Airflow and turbulence noise for fricative and stop consonants: Static considerations. Journal of the Acoustical Society of America, 1971, 50, 1180-1192.

Stevens, K.N. The quantal nature of speech: Evidence from articulatory-acoustic data. In: David, Jr., E.E. and Denes, P.B. (eds.) Human Communication: A Unified View. New York: McGraw-Hill, 1972.

Stevens, K.N. Further theoretical and experimental bases for quantal places of articulation for consonants. Quarterly Progress Report no. 108, Research Laboratory of Electronics, M.I.T., 1973, 247-252. (a)

Stevens, K.N. The potential role of property detectors in the perception of consonants. Paper presented at the Symposium on Auditory Analysis and Perception of Speech. Leningrad, USSR, August, 1973. (b)

Stevens, K.N. The potential role of property detectors in the perception of consonants. In: Fant, G. and Tatham, M.A.A. (eds.) Auditory Analysis and Perception of Speech. New York: Academic Press, 1975, pp. 303-330.

Stevens, K.N. and Blumstein, S.E. Context-independent properties for place of articulation in stop consonants. Paper presented at the 91st meeting of the Acoustical Society of America, Washington, D.C., April, 1976.

Stevens, K.N. and Halle, M. Remarks on analysis by synthesis and distinctive features. In: Wathen-Dunn, W. (ed.) Models for the Perception of Speech and Visual Form. Cambridge: M.I.T. Press, 1967.

Stevens, K.N. and House, A.S. Development of a quantitative description of vowel articulation. Journal of the Acoustical Society of America, 1955, 27, 484-493.

Stevens, K.N. and House, A.S. Studies of formant transitions using a vocal tract analog. Journal of the Acoustical Society of America, 1956, 28, 578-585.

Stevens, K.N. and House, A.S. Speech perception. In: Tobias, J. (ed.) Foundations of Modern Auditory Theory: Volume II. New York: Academic Press, 1972, pp. 1-62.

Stevens, K.N. and Klatt, D.H. Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 1974, 55, 653-659.

Stevens, S.S. and Volkmann, J. The relation of pitch to frequency: A revised scale. American Journal of Psychology, 1940, 53, 329-353.

Streeter, L.A. Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. Nature, 1976, 259, 39-41.

Strevens, P. Spectra of fricative noise. Language and Speech, 1960, 32-49.

Studdert-Kennedy, M. The perception of speech. In: Sebeok, T.A. (ed.) Current Trends in Linguistics, Vol. XII, The Hague, Mouton, 1974.

Studdert-Kennedy, M. Speech perception. In: Lass, N.J. (ed.) Contemporary Issues in Experimental Phonetics, New York: Academic Press, 1976.

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S. and Cooper, F.S. Motor theory of speech perception: A reply to Lane's critical review. Psychological Review, 1970, 77, 234-249.

Studdert-Kennedy, M. and Shankweiler, D. Hemispheric specialization for speech perception. Journal of the Acoustical Society of America, 1970, 48, 2, 570-594.

Svensson, S-G. Prosody and grammar in speech perception. Monographs from the Institute of Linguistics University of Stockholm, Institute of Linguistics, Stockholm, Sweden, 1974.

Tartter, V.C. and Eimas, P.D. The role of auditory feature detectors in the perception of speech. Perception and Psychophysics, 1975, 18, 293-298.

Treon, M.A. Fricative and plosive perception-identification as a function of phonetic context in CVCVC utterances. Language and Speech, 1970, 13, 54-64.

Whitfield, I.C. "Edges" in auditory information processing. In: Proceedings of the XXIII International Congress of Physiological Sciences. Tokyo, September, 1965, pp. 245-247.

Wickelgren, W.A. Auditory or articulatory coding in verbal short-term memory. Psychological Review, 1969, 76, 232-235.

Winitz, H., Scheib, M.E. and Reeds, J.H. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. Journal of the Acoustical Society of America, 1972, 51, 1309-1317.

FOOTNOTES

[1]It should be noted here that these are acoustic patterns that listeners interpret as speech sounds. An examination of natural speech utterances will reveal a great deal of additional information which is not shown in these displays and which we can presume is employed by listeners in perceiving speech.

[2]There are two types of co-articulation effects in speech, forward and backward. Each results from distinctly different phenomena. Backward effects are due entirely to the non-instantaneous response of the production system due to physical inertia or muscle delay etc. On the other hand, forward co-articulation effects are assumed to be due to some higher-level look-ahead mechanism or anticipation process.

[3]A mel is a psychophysical unit reflecting equal sense distances of pitch. It is linear at low frequencies and approximately logarithmic at frequencies above 1000 Hz. The scaling of formant frequencies from linear units to mels is based on the assumption that this transformation reflects the response of the auditory system to frequency differences. The mel scale was constructed by Stevens and Volkmann (1940) on the basis of subjective pitch evaluations by naive listeners.

[4]With regard to the relation between articulation and acoustics, Liberman (1957) raised the following question: "When articulation and soundwave go their separate ways, which way does the perception go?" and in response replied that "the perception always goes with articulation" (p. 121). Fant (1960) argues that Liberman is wrong since "articulation and sound waves never go separate ways" (p. 218).

[5]It should be pointed out here that exotic languages were deliberately used in this study. A better test might have been to require the phoneticians to transcribe non-sense syllable sentences which have appropriate intonation. In this case, they would surely have done very much better. Phonetic inventory size as well as familiarity were factors that were not controlled in this experiment.

[6]We should note here that some investigators have argued that the word recognition process can take place in the absence of, or at least prior to, syntactic analysis (see Cairns and Kamerman, 1975). Given the comments made by Chomsky (1964) on the role of syntactic descriptions in phonological analysis, it is hard to see the rationale behind the claims for a strictly autonomous level of lexical analysis.

Table 1.

Binary Distinctive Feature Representations

for Some English Phonemes

FEATURES	VOWELS							CONSONANTS																					
	u	o	ɔ	a	i	e	æ	p	b	m	f	v	k	g	t	d	θ	ð	n	s	z	č	ǰ	š	ž	h		
1. Vocalic-Nonvocalic	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		
2. Consonantal-Nonconsonantal	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−		
3. Compact-Diffuse	−	−	+	+	−	−	+	−	−	−	−	−	+	+	−	−	−	−	−	−	−	+	+	+	+				
4. Tense-Lax								+	−		+	−	+	−	+	−	+	−		+	−	+	−	+	−	+			
5. Voiced-Voiceless	+	+	+	+	+	+	+	−	+	+	−	+	−	+	−	+	−	+	+	−	+	−	+	−	+	−			
6. Nasal-Oral	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−		
7. Continuant-Interrupted	+	+	+	+	+	+	+	−	−	−	+	+	−	−	−	−	+	+	−	+	+	−	−	+	+				
8. Strident-Mellow	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	−	+	+	+	+	+	+				
9. Checked-Unchecked																													
10. Grave-Acute	+	+	+	+	−	−	−	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−				
11. Flat-Plain	+	+	+	−	−	−	−																						
12. Sharp-Plain																													

Table 2

Some Articulatory and Acoustic Correlates of the Distinctive Features
(Adapted from Jakobson, Fant and Halle, 1952)

DISTINCTIVE FEATURE	ARTICULATORY CORRELATES	ACOUSTIC CORRELATES
1. Vocalic-Nonvocalic	Periodic excitation and open vocal tract vs. narrowed or constricted vocal tract.	Presence vs. absence of a well-defined formant structure.
2. Consonantal-Nonconsonantal	Presence of a constriction or occlusion in the midline of the oral cavity vs. lesser degrees of narrowing in the central path of the oral cavity.	Overall lower energy and presence of a rapid spectrum change at the release of the consonantal configuration vs. higher energy and absence of rapid spectrum change.
3. Compact-Diffuse	Compact sounds have a higher ratio of volume of front cavity to back cavity than diffuse sounds.	Higher vs. lower concentration of energy in the central region of the spectrum as well as an increase vs. decrease of total energy.
4. Tense-Lax	Tense sounds are articulated with greater distinctiveness and oral pressure than lax sounds. For some vowels there is a greater deviation of the vocal tract from its neutral position.	Tense sounds are longer in duration and have greater energy than the corresponding lax sounds.
5. Voiced-Voiceless	Voiced sounds are produced by vibrating the vocal cords whereas voiceless sounds are produced without such vibration.	Presence vs. absence of periodic low-frequency excitation. The spectrum of voiced sounds contains harmonic components of the fundamental.

Table 2 (continued)

DISTINCTIVE FEATURE	ARTICULATORY CORRELATES	ACOUSTIC CORRELATES
6. Nasal-Oral	Nasal sounds are produced with lowered velum whereas oral sounds are produced with a raised velum which shuts off the nasal cavity from the rest of the vocal tract.	Nasal sounds have wider spectral peaks and the presence of additional nasal resonances.
7. Continuant-Interrupted	Continuants are produced with a vocal tract having no complete closure; interrupted sounds have a complete closure at some point between glottis and lips.	Continuants have smooth envelopes at onset whereas interrupted sounds have an abrupt onset of energy.
8. Strident-Mellow	This feature is restricted to consonantal sounds. Strident sounds are generated by directing an airstream at an obstacle in the vocal tract thus producing greater turbulence of the airstream.	Strident sounds are characterized by turbulent noise well above the intensity of the vowel in the high frequency range vs. mellow sounds which have much lower intensity.
9. Checked-Unchecked	The airstream is checked by the compression or closure of the glottis.	Checked sounds have an abrupt decay or sharper termination than unchecked sounds.
10. Grave-Acute	Grave sounds are articulated with constriction at the periphery of oral cavity (i.e. labials, velars) vs. acute sounds which are produced in the central region of the oral cavity.	Grave sounds have predominance of energy in lower half of the spectrum whereas acute sounds have predominant energy in upper half.
11. Flat-Plain	Flattening is generated by reduction of the lip orifice by rounding with an increase in the length of the lip constriction and therefore an increase in the overall length of the vocal tract.	Flat sounds manifest themselves by a downward shift of the frequencies of the formants.

Table 2 (continued)

DISTINCTIVE FEATURE	ARTICULATORY CORRELATES	ACOUSTIC CORRELATES
12. Sharp-Plain	For sharp sounds, the oral cavity is reduced in size by raising a part of the tongue against the palate. Also known as "palatalization", this feature is made simultaneously with the main articulation of a consonant.	Sharp sounds as compared with plain ones are characterized by a slight rise in the frequency of the second formant and to some extent the higher formants.

FIGURE CAPTIONS

Figure 1. Contribution of source spectrum, vocal tract transfer function and radiation characteristic to the spectrum envelope of the radiated sound pressure. (Courtesy of Dr. Dennis Klatt)

Figure 2. Illustrations of the outline shape of mid-saggital sections of the vocal tract, cross-sectional area functions and the corresponding vocal tract transfer functions for the vowels [i], [a] and [u]. (From Lieberman, Crelin and Klatt, 1972 with permission of the authors and publisher).

Figure 3. Sound spectrogram of the utterance: "Handbook of Learning and Cognitive Processes".

Figure 4. Schematized sound spectrograms for nine consonant-vowel syllables illustrating the acoustic cues for place, manner and voicing. (From Liberman, Ingeman, Lisker, Delattre and Cooper, 1959 with permission of the authors and publisher.)

Figure 5. Schematized sound spectrograms showing the formant transitions that are appropriate for the voiced stop consonants [b], [d] and [g] before various vowels. (From Delattre, Liberman and Cooper, 1955 with permission of the authors and publisher.)

Figure 6. Hypothetical quantal relations between a parameter that describes some aspect of articulation and the resulting acoustic parameter of speech. (From Stevens, 1972, with permission of the author and publisher.)

Figure 7. (a) A Two-tube resonator approximating the vocal tract configuration for the vowel [a]. A_1 and A_2 represent the cross-sectional areas of the pharyngeal and oral cavities, respectively. (b) An approximation of the spectrum envelope for the vowel [a] produced by the configuration shown above. The peaks in the function represent the center frequencies of the formants. (From Stevens, 1972, with permission of the author and publisher.)

Figure 8. Results of a vocal tract simulation showing the relation between frequencies of the first and second formants (F1 and F2) when the length of the back tube, d_1, is varied between 5 and 15 cm. (From Stevens, 1972, with permission of the author and publisher.)

Figure 9. Schematized spectrographic patterns showing the frequency positions of synthetic release bursts (A), the vowel formants (B) and an example of one of the resulting test syllables (C). (From Liberman, Delattre and Cooper, 1951, with permission of the authors and publisher.)

Figure 10. Results showing the distribution of /p/, /t/ or /k/ responses as a function of the burst position and formant frequencies of the vowel. (From Liberman, Delattre and Cooper, 1951, with permission of the authors and publisher.)

Figure 11. Hypothetical invariant patterns for labial, post-dental and velar stop consonants in initial position. (Adapted from Stevens, 1975.)

Figure 12. Idealized form of categorical perception showing the identification function (left ordinate) and the discrimination function (right ordinate). (From Studdert-Kennedy, Liberman, Harris and Cooper, 1970, with permission of the authors and publisher.)

Figure 13. Labeling functions for a set of stimuli varying in voice onset time presented to speakers of English, Spanish and Thai. (From Abramson and Lisker, 1965, with permission of the authors and publisher.)

Figure 14. Identification functions obtained with and without adaptation for a single subject. The functions for the [b,p] series are shown on the left whereas the ones for the [d, t] series are on the right. The lower two functions in each panel show the results of cross-series adaptation when the adaptor

was not selected from the test series. In each panel, the solid lines show the unadapted identification functions, the dotted and dashed lines show the functions obtained after adaptation. (From Eimas and Corbit, 1973, with permission of the authors and publisher.)

Figure 15. Hierarchial organization of levels of processing in speech perception. (Adapted from Liberman, 1970.)

Figure 16. Tentative organization of some processing stages in a functional model of speech perception.

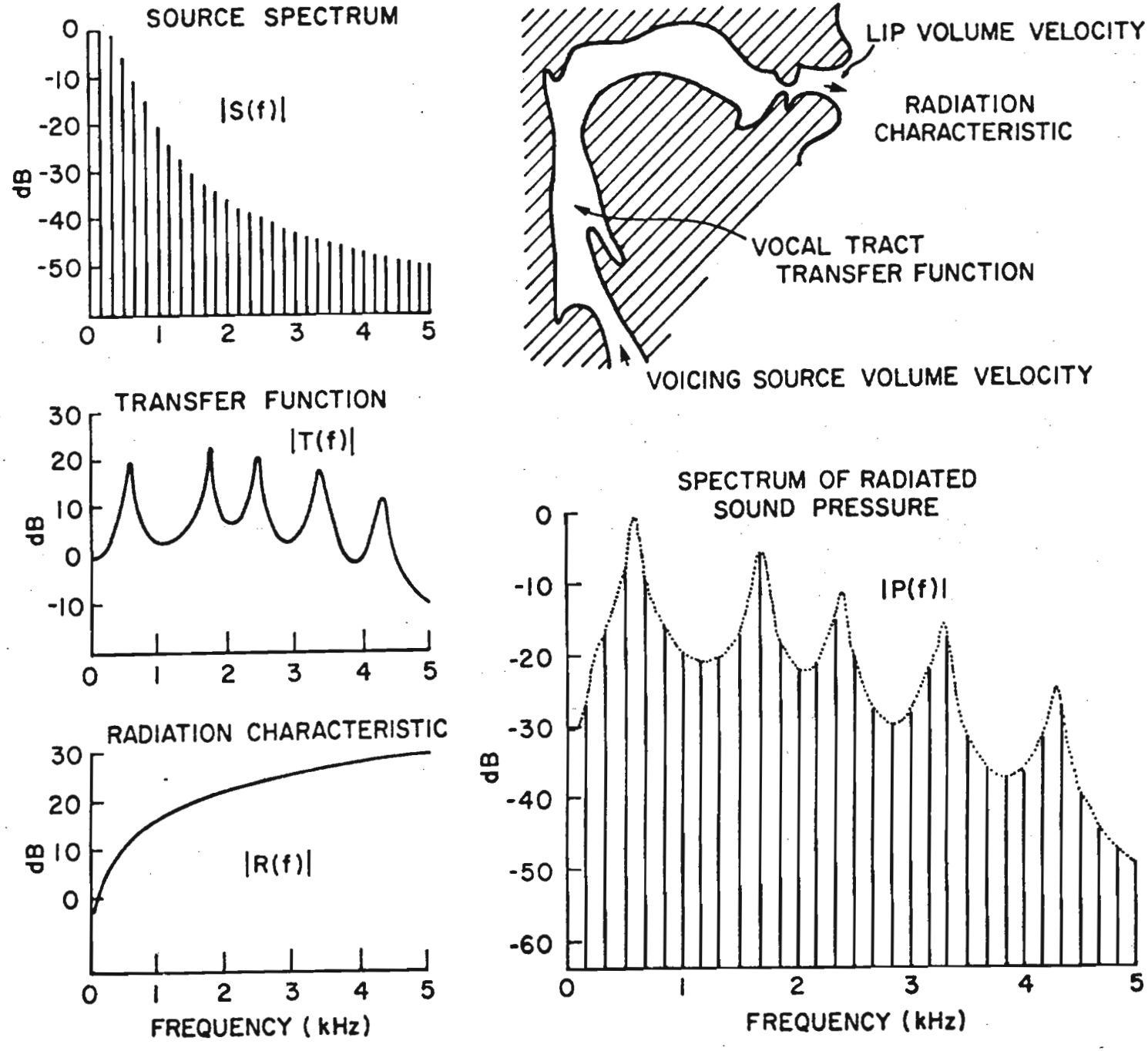
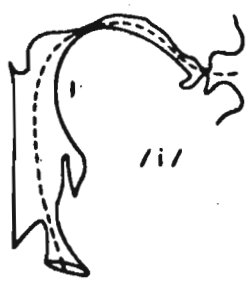
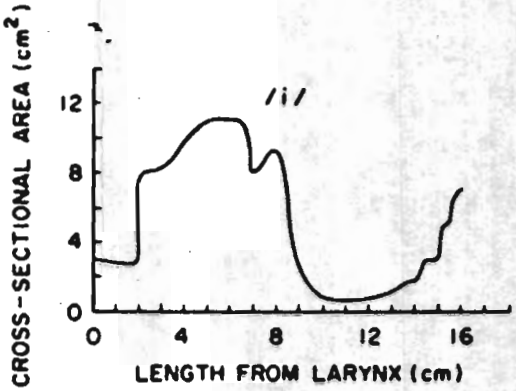
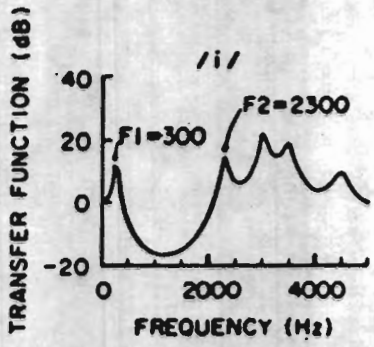
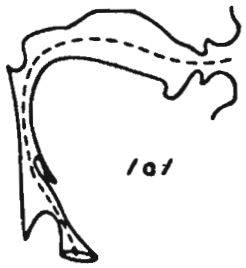

SOURCE SPECTRUM
|S(f)|
dB
0
-10
-20
-30
-40
-50
0 1 2 3 4 5
LIP VOLUME VELOCITY
RADIATION
CHARACTERISTIC
VOCAL TRACT
TRANSFER FUNCTION
VOICING SOURCE VOLUME VELOCITY
TRANSFER FUNCTION
|T(f)|
dB
30
20
10
0
-10
0 1 2 3 4 5
SPECTRUM OF RADIATED
SOUND PRESSURE
|P(f)|
dB
0
-10
-20
-30
-40
-50
-60
0 1 2 3 4 5
FREQUENCY (kHz)
RADIATION CHARACTERISTIC
|R(f)|
dB
30
20
10
0
0 1 2 3 4 5
FREQUENCY (kHz)

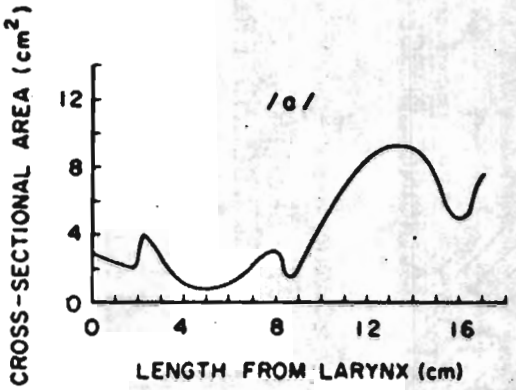
FIGURE 1.

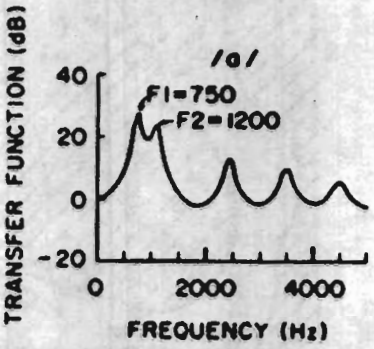


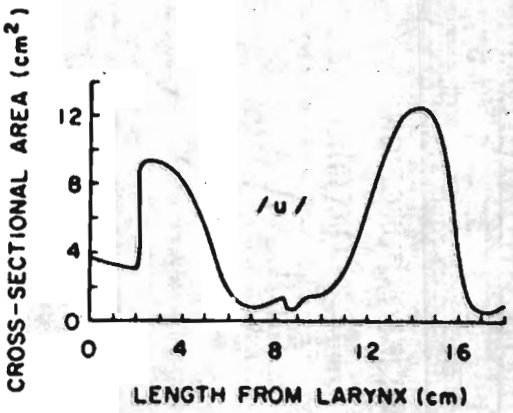


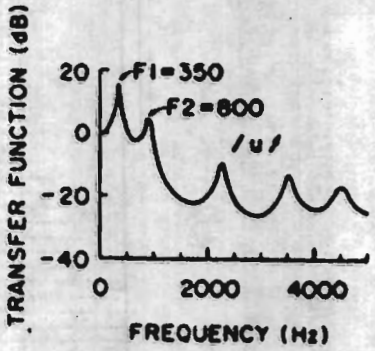












MIDSAGGITAL SECTION OF THE VOCAL TRACT

CROSS-SECTIONAL AREA FUNCTION OF THE VOCAL TRACT

MAGNITUDE OF THE VOCAL TRACT TRANSFER FUNCTION

FIGURE 2.

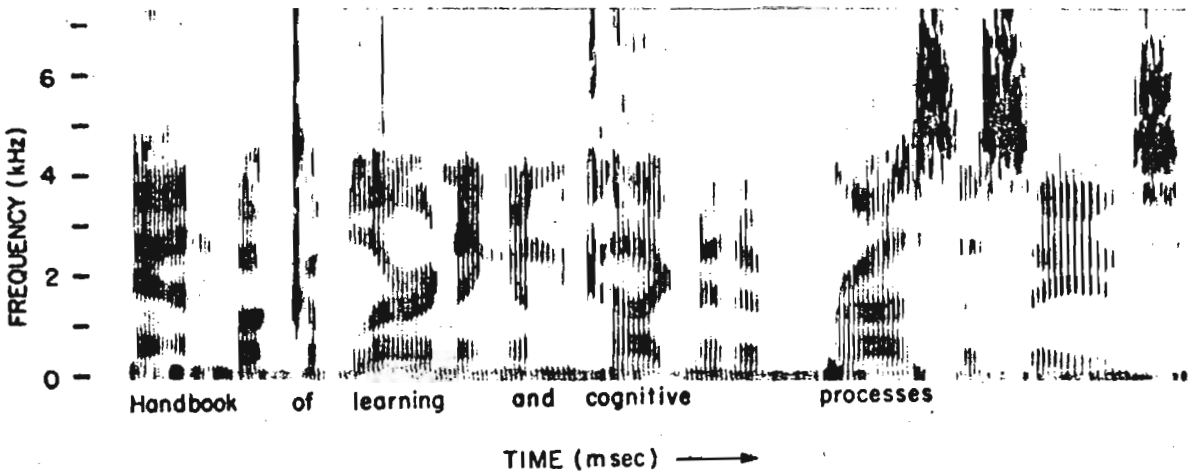

FREQUENCY (kHz)
6
4
2
0
Handbook of learning and cognitive processes
TIME (msec)

FIGURE 3.

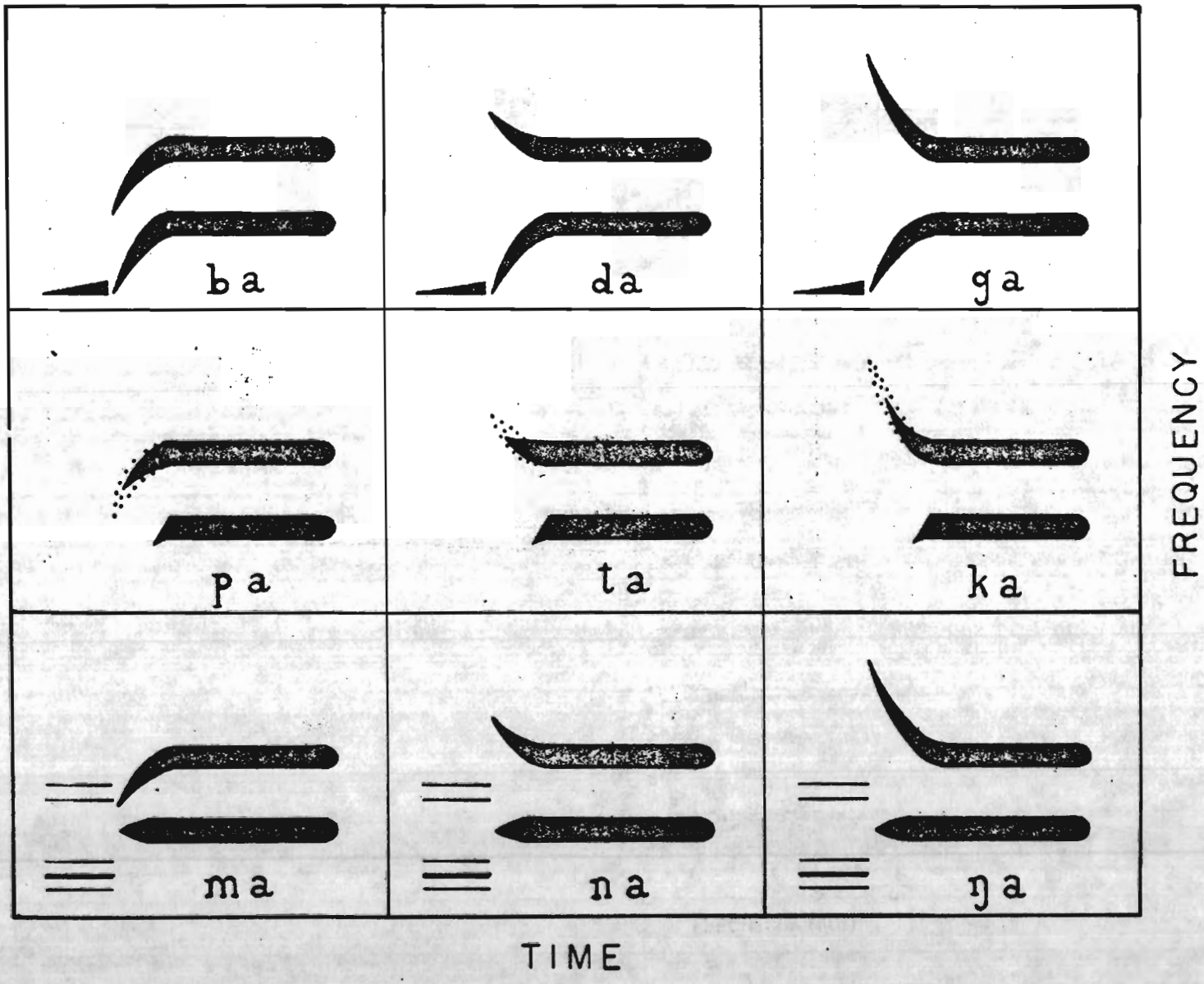


FIGURE 4.

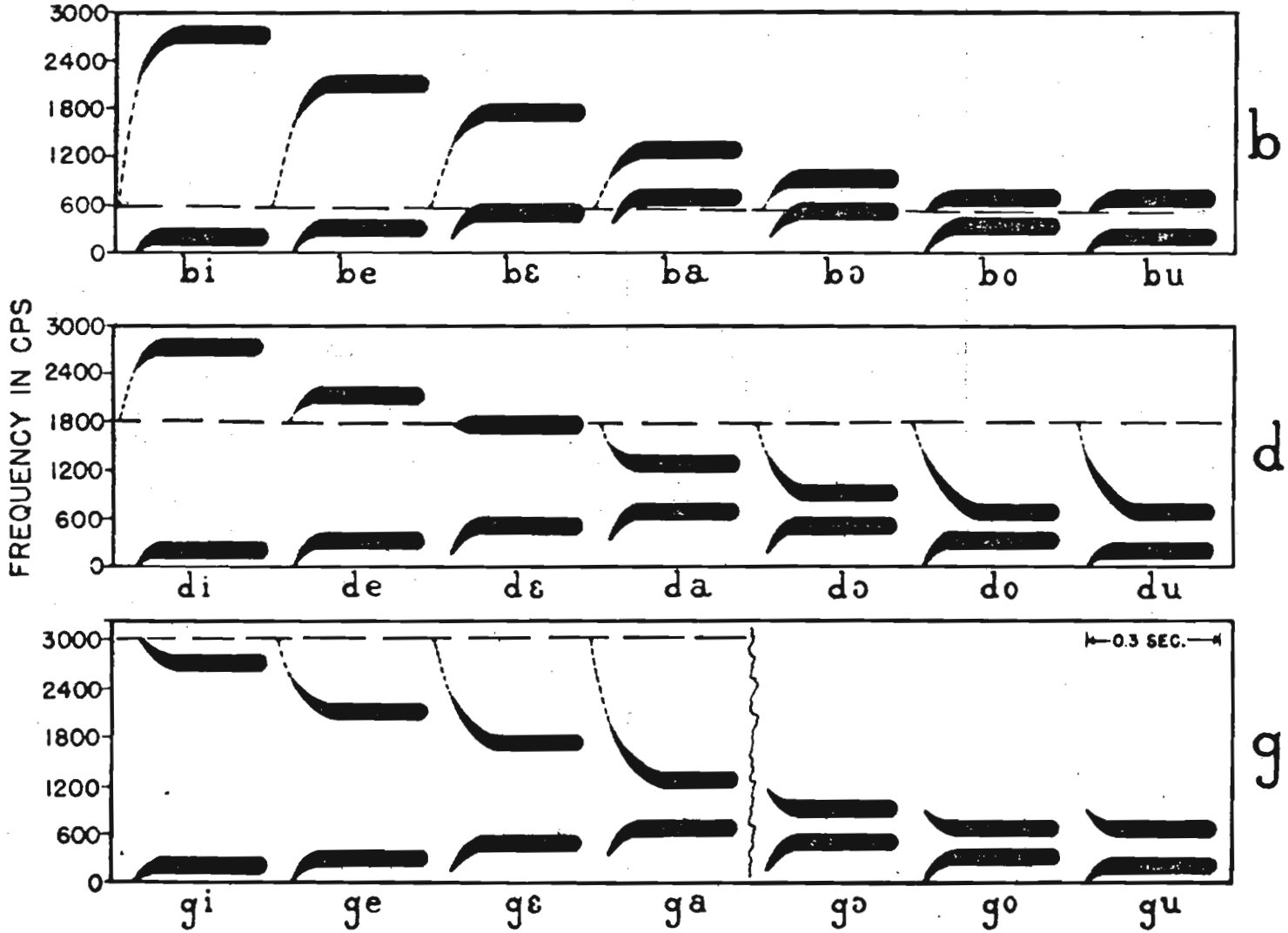

FREQUENCY IN CPS
3000
2400
1800
1200
600
0
bi
be
bɛ
ba
bɔ
bo
bu
b
di
de
dɛ
da
dɔ
do
du
d
gi
ge
gɛ
ga
gɔ
go
gu
g
0.3 SEC.

FIGURE 5.

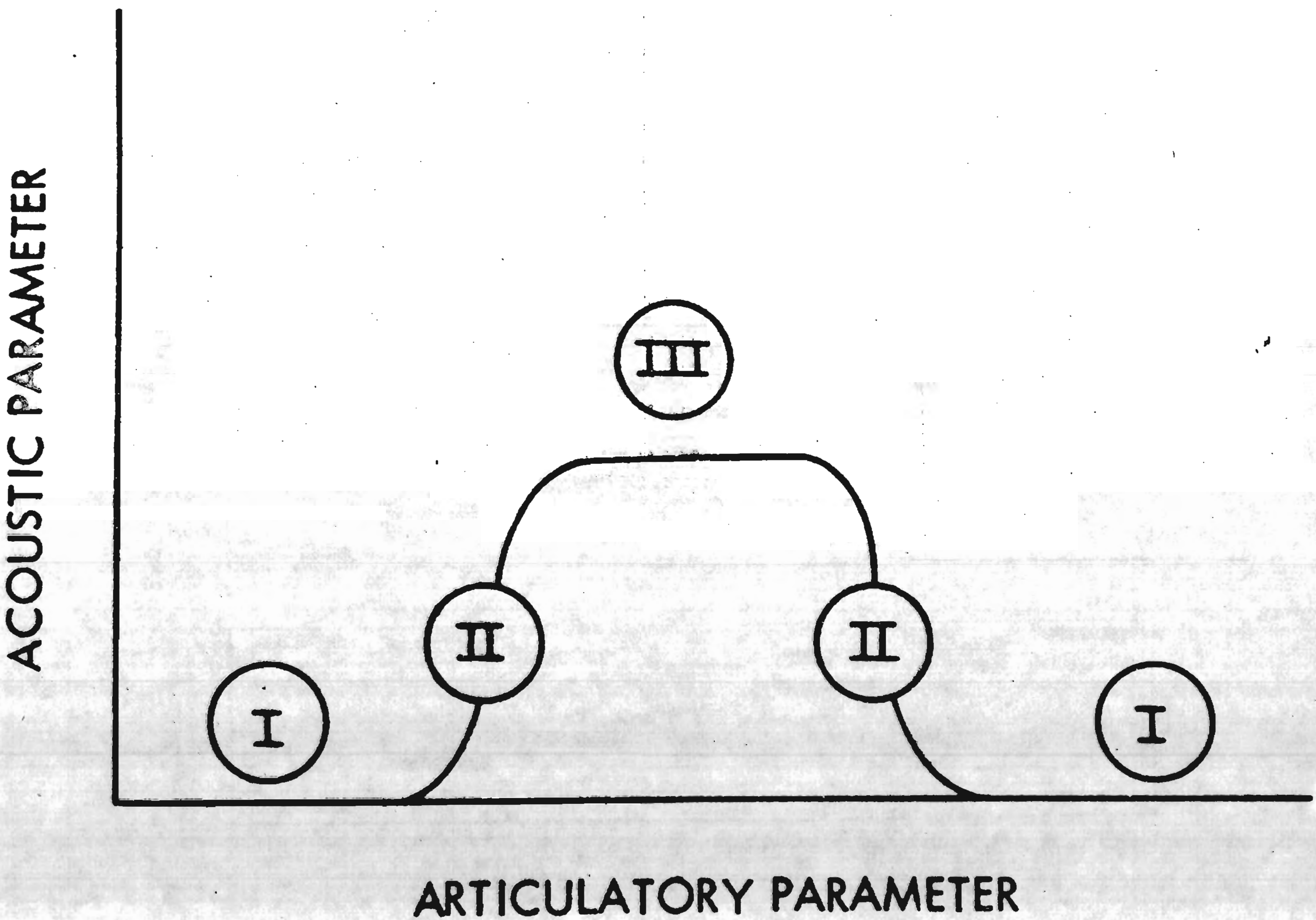
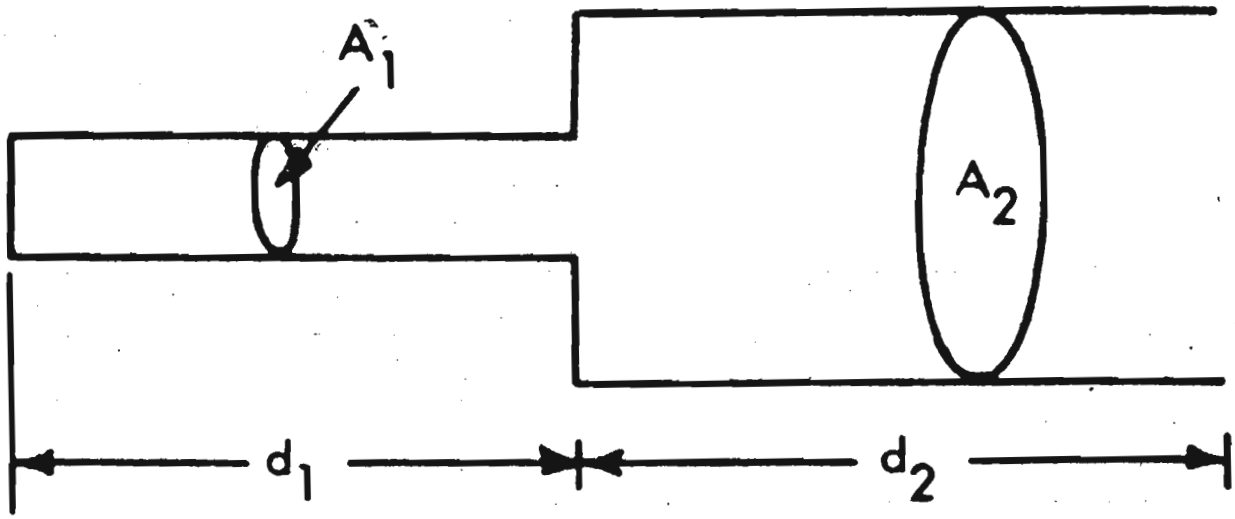
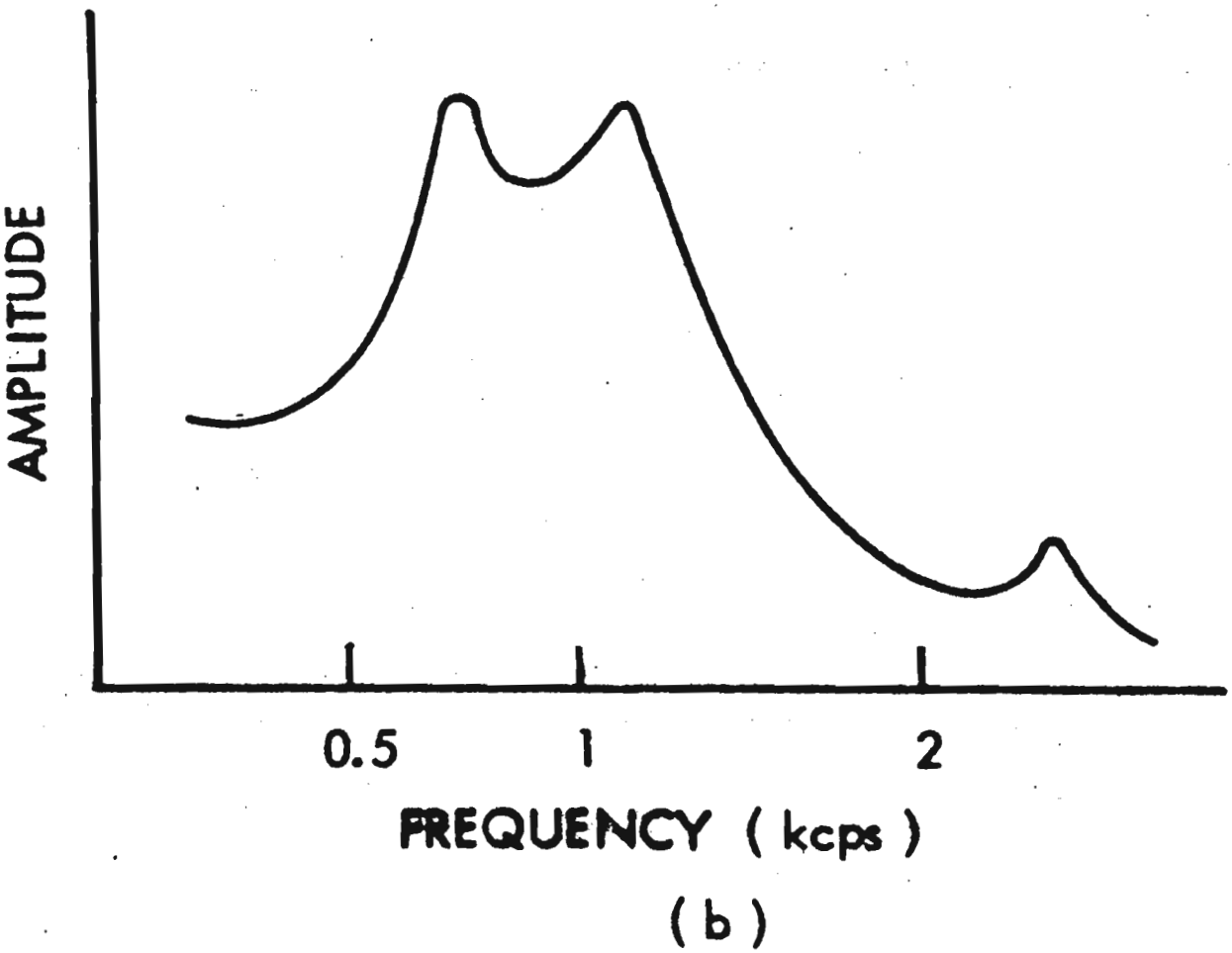


FIGURE 6.



(a)



(b)

FIGURE 7.

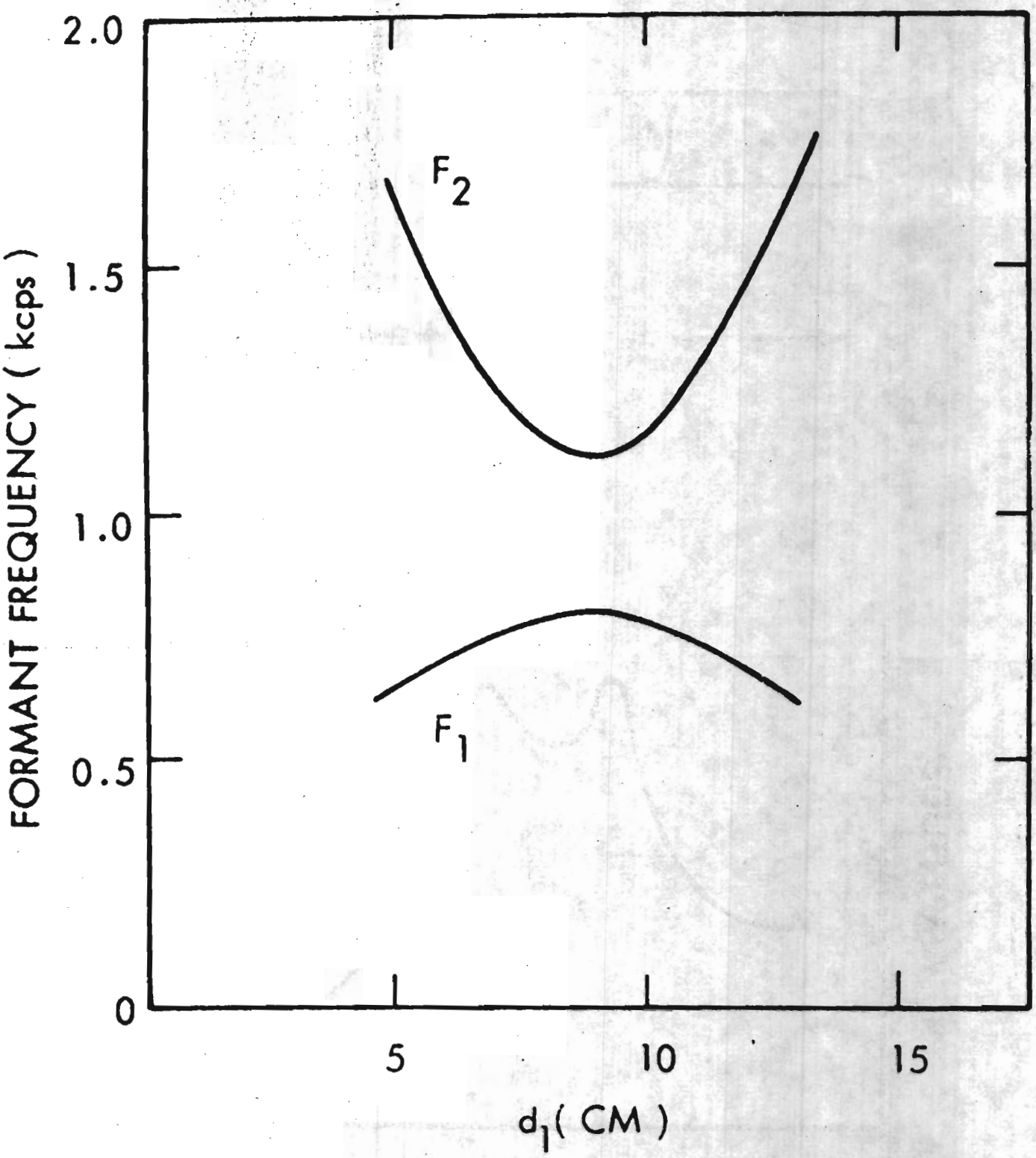


FIGURE 8.

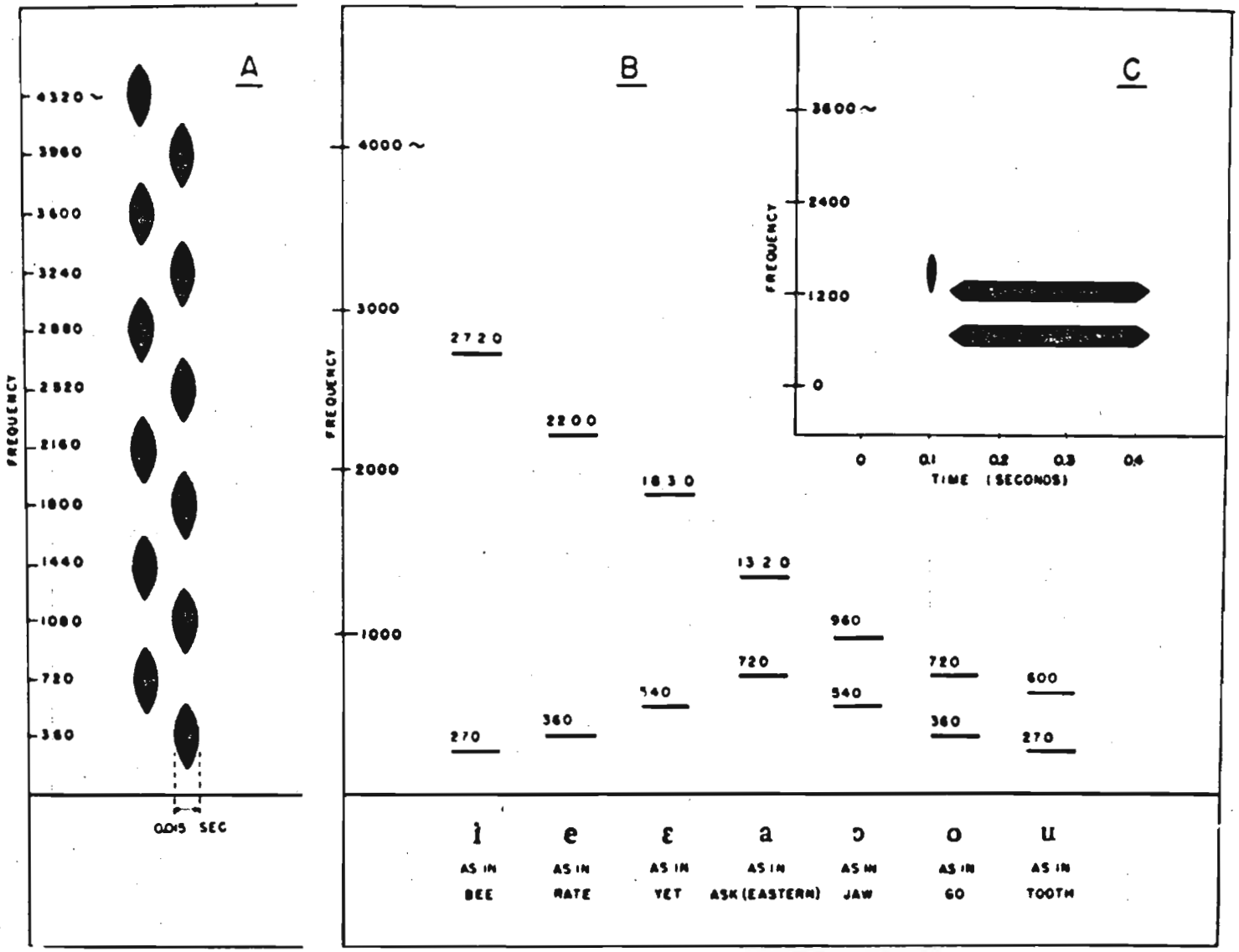

A
4320 ~
3960
3600
3240
2880
2520
2160
1800
1440
1080
720
360
FREQUENCY
0.015 SEC
B
4000 ~
3000
2000
1000
FREQUENCY
2720
2200
1830
1320
960
720
540
720
600
540
360
360
270
270
i AS IN BEE
e AS IN RATE
ɛ AS IN YET
a AS IN ASK (EASTERN)
ɔ AS IN JAW
o AS IN GO
u AS IN TOOTH
C
3600 ~
2400
1200
0
FREQUENCY
0
0.1
0.2
0.3
0.4
TIME (SECONDS)

FIGURE 9.

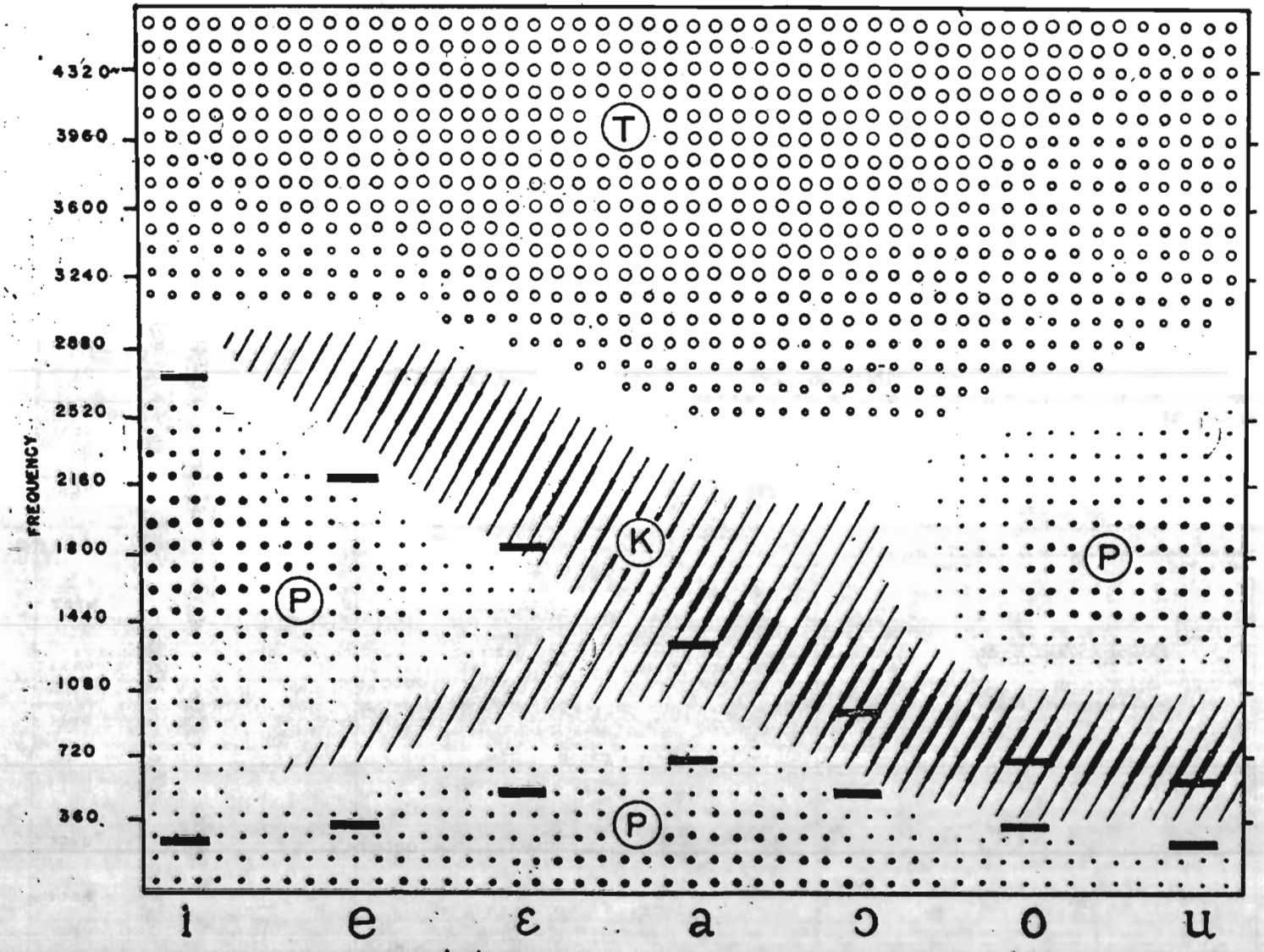

FREQUENCY
4320
3960
3600
3240
2880
2520
2160
1800
1440
1080
720
360
T
K
P
P
P
i
e
ɛ
a
ɔ
o
u

FIGURE 10.

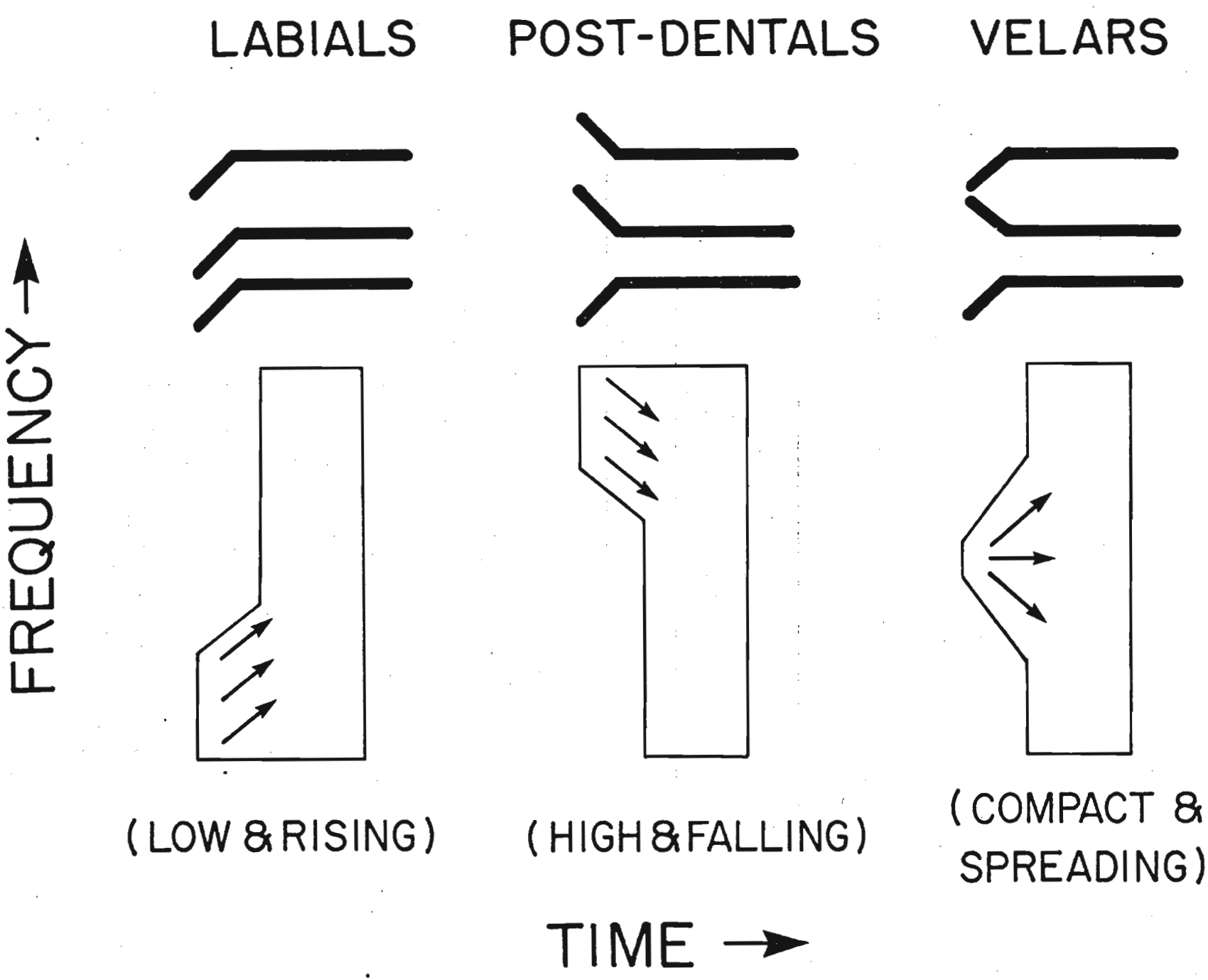

LABIALS
POST-DENTALS
VELARS
FREQUENCY →
(LOW & RISING)
(HIGH & FALLING)
(COMPACT & SPREADING)
TIME →

FIGURE 11.

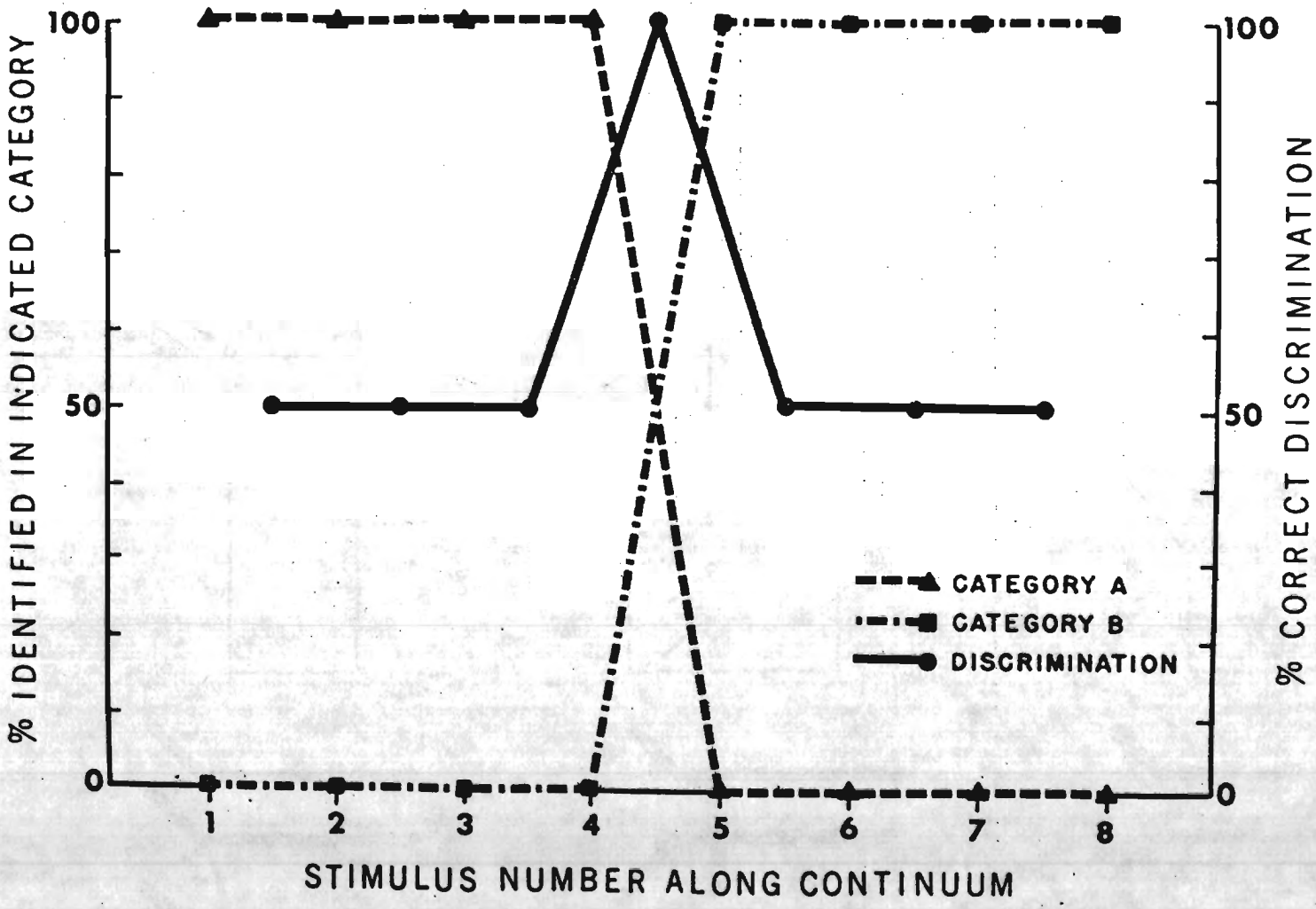

100
50
0
% IDENTIFIED IN INDICATED CATEGORY
100
50
0
% CORRECT DISCRIMINATION
CATEGORY A
CATEGORY B
DISCRIMINATION
1
2
3
4
5
6
7
8
STIMULUS NUMBER ALONG CONTINUUM

FIGURE 12.

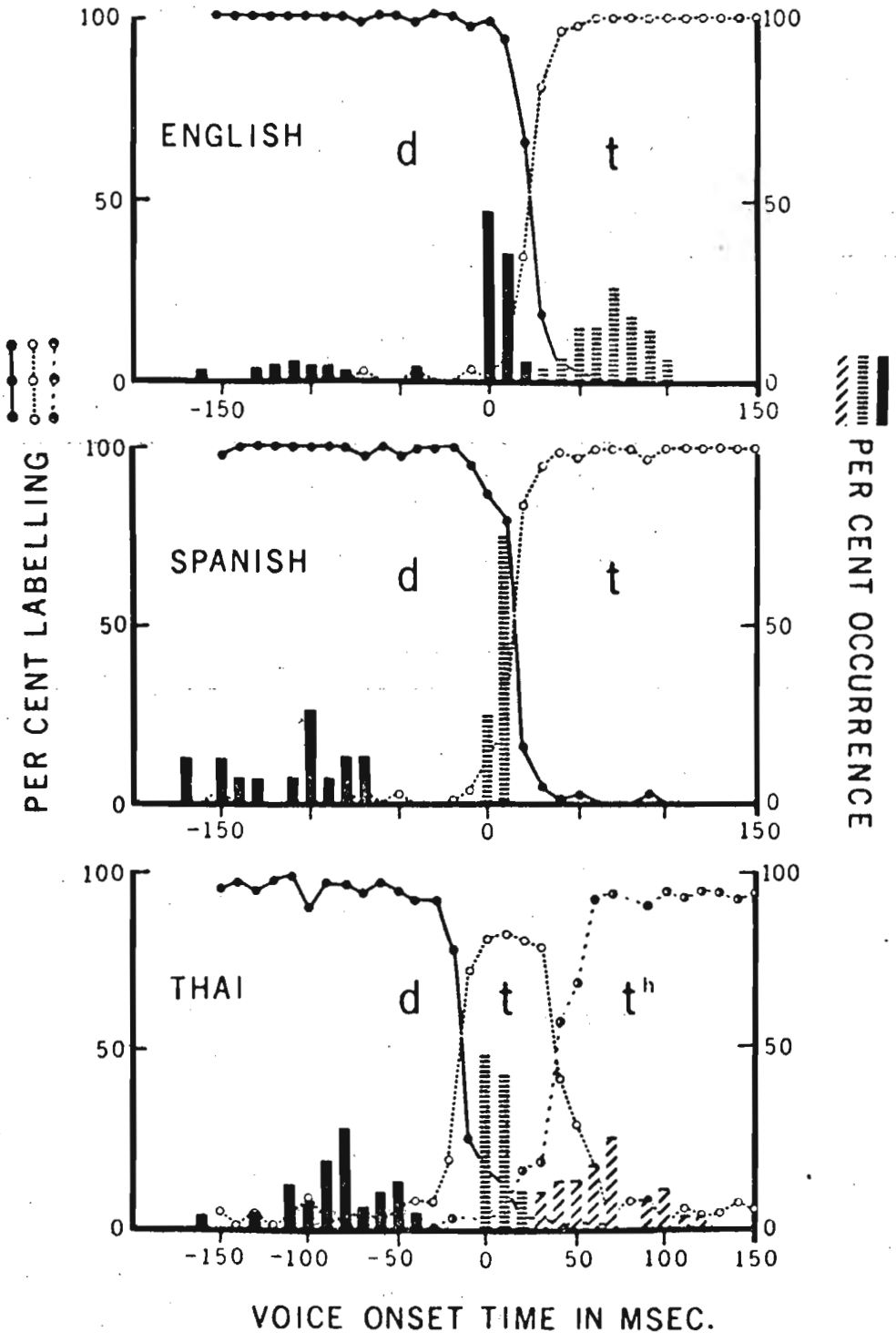

ENGLISH
d
t
SPANISH
d
t
THAI
d
t
t^h
100
50
0
-150
-100
-50
0
50
100
150
PER CENT LABELLING
PER CENT OCCURRENCE
VOICE ONSET TIME IN MSEC.

FIGURE 13.

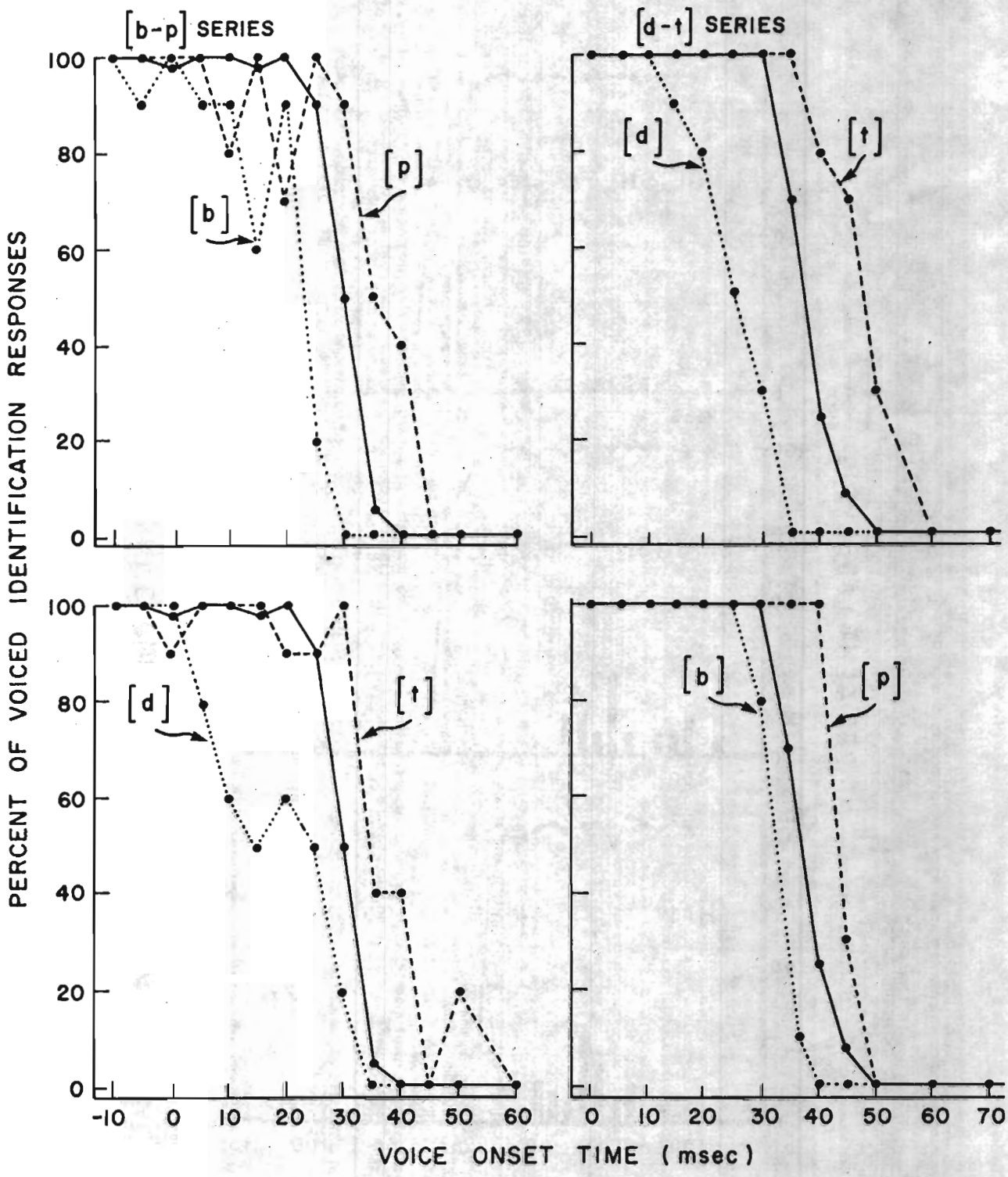

[b-p] SERIES
[d-t] SERIES
[b]
[p]
[d]
[t]
[d]
[t]
[b]
[p]
PERCENT OF VOICED IDENTIFICATION RESPONSES
100
80
60
40
20
0
-10
0
10
20
30
40
50
60
70
VOICE ONSET TIME (msec)

FIGURE 14.

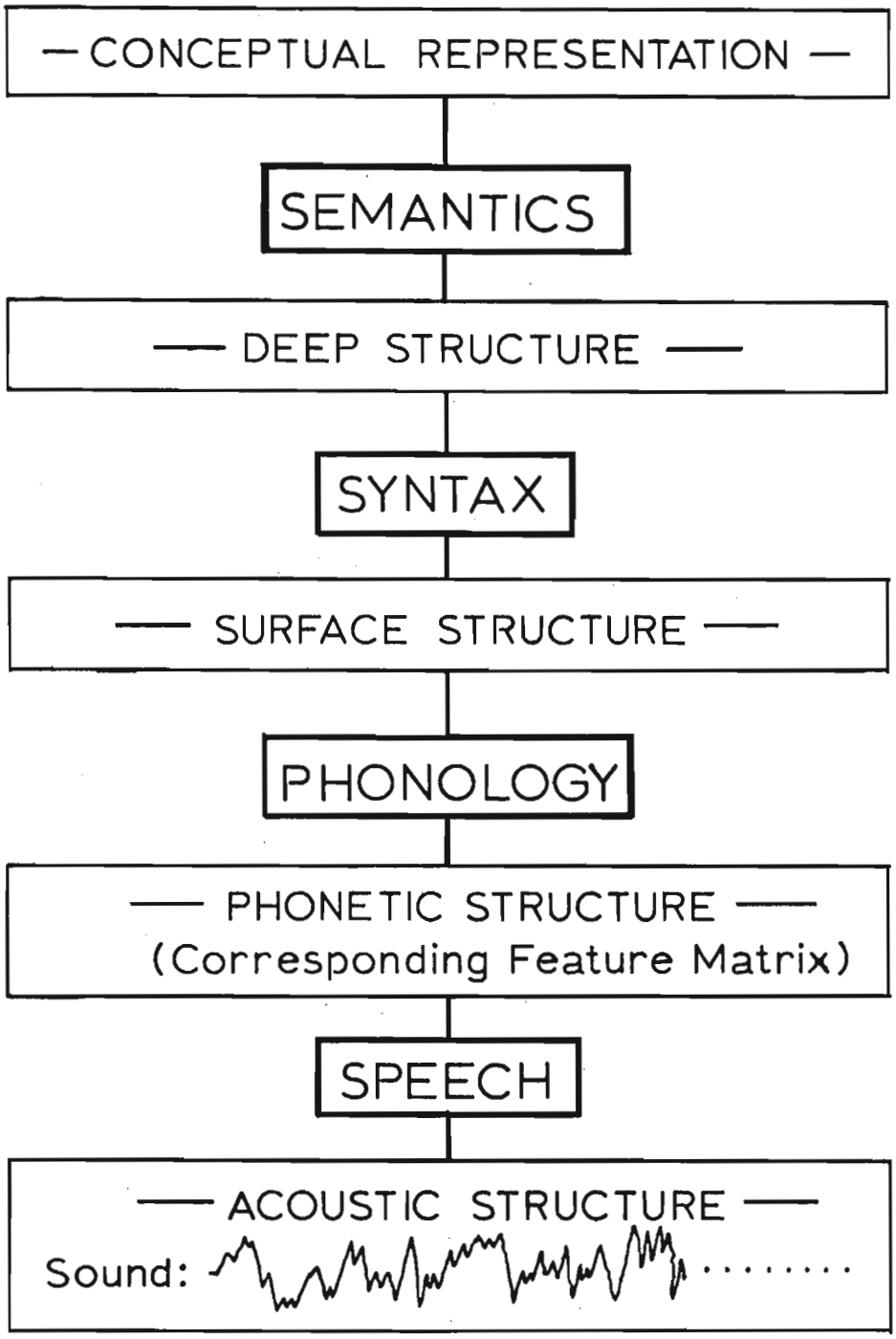

— CONCEPTUAL REPRESENTATION —
SEMANTICS
— DEEP STRUCTURE —
SYNTAX
— SURFACE STRUCTURE —
PHONOLOGY
— PHONETIC STRUCTURE —
(Corresponding Feature Matrix)
SPEECH
— ACOUSTIC STRUCTURE —
Sound:

FIGURE 15.

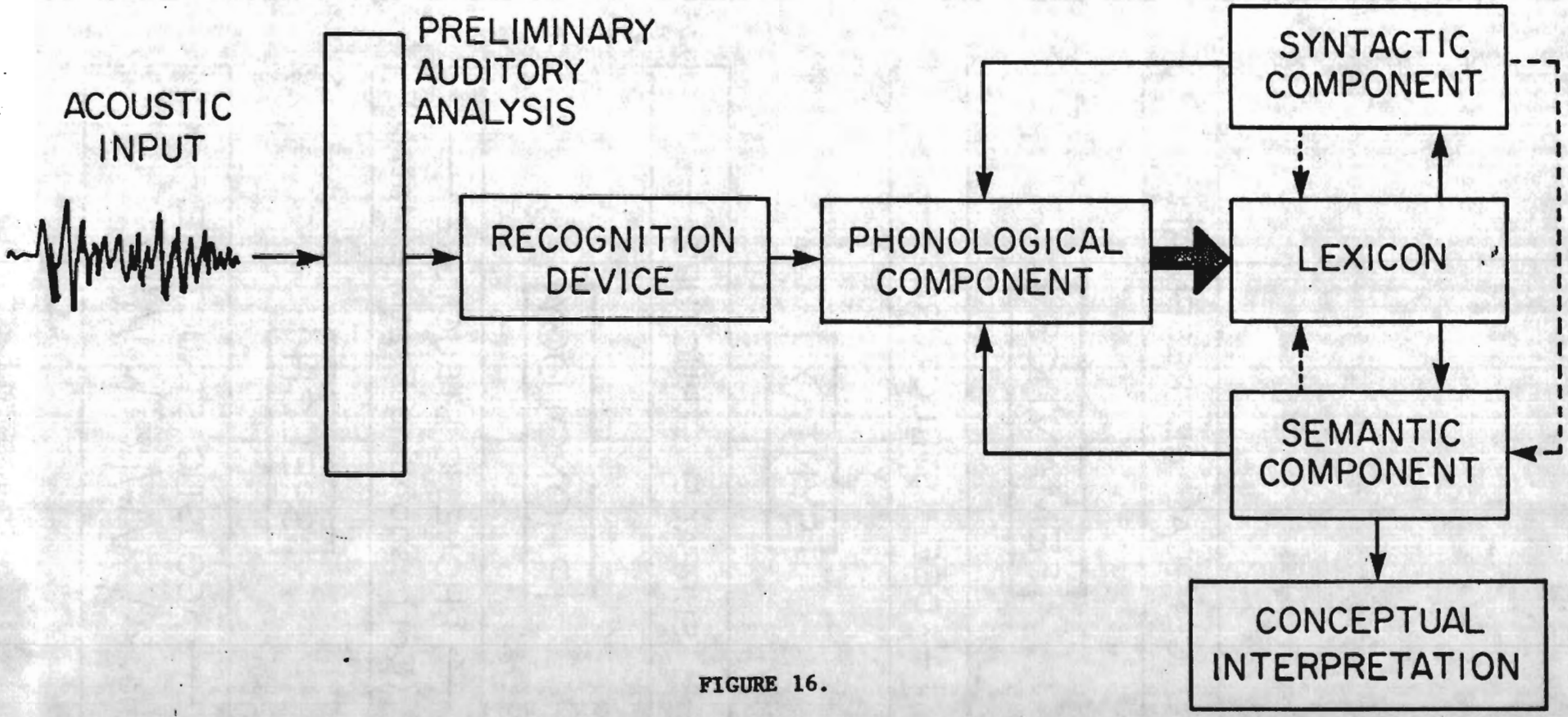

ACOUSTIC
INPUT
PRELIMINARY
AUDITORY
ANALYSIS
RECOGNITION
DEVICE
PHONOLOGICAL
COMPONENT
SYNTACTIC
COMPONENT
LEXICON
SEMANTIC
COMPONENT
CONCEPTUAL
INTERPRETATION

FIGURE 16.