Chapter 7

PHONETIC CODING AND SERIAL ORDER*

WAYNE A. WICKELGREN

I. INTRODUCTION

This chapter will discuss the theoretical units of encoding that are useful in accounting for speech recognition and articulation phenomena. In doing this, it is necessary to give some attention to the processes (mechanisms) by which speech recognition and articulation are accomplished.

Although the ultimate goal of speech is to communicate meanings from one person to another, the primary concern of the present chapter is largely with the phonetic (structural) means by which this communication of ideas is accomplished. Thus, although there will be some discussion of the inter-

* The work presented in this chapter was supported by Grant MH17958 from The National Institute of Mental Health.

action of syntax and semantics with the phonetic aspects of speech communication, the emphasis will be on the nature of phonetic representation and its role in speech recognition and articulation.

The assumption that speech and language can be subdivided into phonetics and phonology, on the one hand, and syntax and semantics, on the other hand, is an assumption frequently made in linguistics and speech communication. This assumption has considerable empirical motivation, some of which will be made clear in the present chapter. What I call the phonetics of speech and language consists of that part of speech communication starting with a set of concept representatives (formatives, words) at some high level of the speaker's nervous system, the various stages of decoding of this set of concepts into a sequence of phonetic segments, and specification of the articulatory features of each segment, then continuing to an acoustic representation of the information, thereafter to an auditory representation in the nervous system of the hearer, and finally terminating in a set of concept representatives in the hearer.

I assume that, when a speaker plans an utterance, he selects concepts and organizes them into a syntactic structure in a manner that is part of the syntactic and semantic domain of speech and language. Similarly, when a hearer understands a spoken message, I assume that at the terminal stage he has a set of concepts organized into a syntactic and semantic structure, hopefully in a manner that in some way corresponds to what the speaker intended to communicate. This aspect of speech recognition is in the syntactic and semantic domain and outside the scope of the present chapter. The phonetic domain, which is the scope of the present chapter, includes everything in speech articulation and recognition between the syntactic–semantic domain of the speaker and that of the hearer. In particular, it will be assumed that a string of concept representatives is transformed into a string consisting of a much larger number of phonetic segments (phonemes, allophones, or syllables). These segments are, in turn, analyzable as a set of simultaneously present features. The decoding of concept representatives into segments and sets of features constitutes the necessary information for controlling the articulatory musculature to produce speech. Having produced speech, one has a physical acoustic representation of the message. This physical acoustic representation of the message is transformed into a set of simultaneously present auditory features in the hearer. These features change over time and can be organized at some stage in the speech recognition process into a set of segments in the auditory recognition process. If a sufficient number of auditory segments composing a word are recognized in an utterance, then the word will be recognized. In the present chapter I will not be concerned with anything beyond the recognition of individual words. Thus, in particular, I will not be concerned with how we recognize phrase structure at the most superficial syntactic level, to say nothing of how we recognize what Chomsky refers to as deep or underlying structure in an utterance.

The present chapter will attempt to account for the principal facts of speech recognition and articulation. No attempt will be made to account for phonological phenomena. Precise definition of phonology in linguistics varies from linguist to linguist. Chomsky and Halle's (1968) definition of phonology is in some ways similar to the definition of the phonetic domain of language just offered. However, the types of phenomena for which we are attempting to account are almost completely different. The following are examples of the types of regularity for which phonological theories attempt to account: (*a*) the accent patterns of different words, especially different words that are composed of some of the same basic concepts (stems, morphemes), such as *permít* and *pérmit, tormént* and *tórment;* (*b*) regularities in the combination of stems and affixes, such as in forming the plurals of nouns and the past tenses of verbs; (*c*) the phonetic relationships between various words that share a common stem (morpheme), such as *divine* and *divinity, communicate* and *communication;* and (*d*) stress patterns over phrases and sentences. Finally, Chomsky and Halle's (1968) phonology also attempts to characterize lexical elements by the minimum number of phonetic features, with the complete set of phonetic features necessary to specify the articulatory segments being supplied by the phonological rules. All of these matters are outside the scope of the present discussion.

II. LEVELS OF LINGUISTIC CODING

The primary units of linguistic coding of concern in this chapter are the concept level, the segment level, the articulatory feature level, and the auditory feature level. In addition, phrases, clauses, sentences, and the possibility of abstract feature representatives will be briefly discussed.

Concepts are units of meaning, and the concept is usually represented in language by a single word or very short phrase. Thus, *dog* is a concept, *little black dog* is a concept, *sit* is a concept, *slowly* is a concept, etc. Crudely speaking, the concept level is the word level, though many concepts can be referred to by more than one word or phrase and the same word often stands for a variety of different concepts (sometimes similar to one another, sometimes not). Consider the word *bank* meaning a financial institution or as in a river bank.

The spoken representative of a concept (word) is an ordered set of articulatory gestures that translate themselves into a temporally varying acoustic output. It is usually assumed that the articulatory and auditory representations of a word can be considered to be an ordered set of phones, allophones, phonemes, or syllables, though at peripheral articulatory and acoustic levels, it is often rather difficult to specify exactly where one segment ends and the next segment begins. However, it seems likely that, at some level of the nervous system, words are "spelled" in terms of smaller structural units in either ordered or unordered sets. These smaller structural units can be referred to in a very general way as *segments.*

An articulatory segment must produce a set of instructions to the articulatory musculature to assume a particular position appropriate, say, for production of the consonant phoneme /p/. Thus, a segment representative must activate a set of simultaneously present articulatory feature representatives. For example, the phoneme /p/ is unvoiced (vocal cords not vibrating), produced by closure of the lips (frontal place of articulation), not nasal (velum is closed so that little or no sound is transmitted through the nose), and the manner of articulation is that of the stop consonant (vocal tract completely closed at some point in the articulation). All of these articulatory features must be present for an individual to produce the segment that we identify as /p/. Of course, we cannot hear stop consonants such as /p/ in isolation, unaccompanied by a preceding or succeeding vowel, but that is largely irrelevant to the discussion of articulatory features.

However we analyze the temporally distributed acoustic information associated with a word, it is clear that we will have to analyze that information in terms of a number of auditory features characteristic of each segment of the word. Such features include the principal frequency components present at any point in time and the changes in these frequency components that occur over short periods of time. The detectors for these auditory features are considered to comprise the auditory feature level of the verbal modality.

III. CONCEPTS AND WORDS

A. Neural Representation of Attributes and Concepts

No-one knows at the present time what the best definition of the term "concept" will turn out to be for analysis of the human mind. However, I will give what I consider to be the best definition available at present.

Work in sensory neurophysiology has made it quite clear that the nervous system encodes peripheral sensory input and motor output in terms of certain innate attributes. There are single neurons that respond vigorously when a stimulus with the appropriate properties (features, attributes) is presented. For example, there are known to be neurons in the auditory system that respond selectively to particular frequencies, neurons encoding frequency changes (e.g., Whitfield & Evans, 1965), and neurons that encode the direction of a sound source by responding to phase and intensity differences in the signals arriving at the left and right ears (e.g., Hall, 1965). On the motor side, a single motor neuron controls the activity of a single motor unit in a muscle. At higher levels of the motor system, there is evidence to indicate that certain more complex patterns of movement may be encoded in single neurons as well, though there will surely be redundant coding across many neurons, just as there is within the sensory nervous systems. Thus, it is reasonable to assume that on both the sensory and motor side there are feature representatives: units in the nervous system that represent particular features (attributes) of the sensory input and motor output. There is evidence that many of these attribute representatives are innately specified (e.g., Hubel & Wiesel, 1963).

Even if much of the peripheral representation of attributes is innate in human beings, it is clear that there cannot be representation of all human concepts by innately specified single neurons. The reason for this is that there are too many different types of concepts that human beings appear able to possess, perhaps an infinite number. Setting reasonable upper limits on the rate at which human beings might be expected to learn new concepts, say one new concept every second there are still more than enough neurons in the human brain (there are around 10^{10}) to have one neuron for every concept, provided that the specification of a neuron to stand for a concept is learned rather than innate. There is no experimental evidence to support the hypothesis that all concepts are represented by single neurons in the manner in which a single neuron can represent a straight line at a certain angle at a certain position in the visual field. However, this is a relatively clear, precise way to think of the representation of concepts in the nervous system, and I know of no other equivalently clear way to think of this representation. Thus, it seems reasonable to me to adopt this as a working hypothesis; but one should note that there is no direct experimental support for this hypothesis and many theorists consider the hypothesis outrageous.

B. Psychological Definition of Concepts

Irrespective of the neural representation of concepts, we still need to formulate some hypothesis concerning the logical relation between concepts and attributes. It will not do to say that concepts are simple conjunctions of attributes. It may be possible for a precise thinker to give some definition of a concept such as the concept of the species dog in terms of a conjunction of attributes possessed by all dogs. However, this is not the same as saying that whenever we perceive or think of the concept *dog,* it is because that conjunction of attribute representatives has been activated. Rather, we can think of the concept *dog* when we see a particular dog in a particular perspective with a conjunction of attributes that are quite different from those that would be present seeing the same dog from a different perspective, seeing a different dog, seeing a different species of dog, seeing the word *dog,* or hearing the word *dog.*

To explain how we are able to recognize examples of concepts, it is sufficient to consider that a concept is a disjunction of a large number of conjunctions of attributes. That is, there are large number of sets of attributes that are individually sufficient to activate the concept *dog* in our minds. It should be noted that defining a concept to be a disjunction of conjunctions of attributes is not necessarily the ideal way to define concepts. However, it appears to me at present to be a sufficient way of defining concepts, and I know of no other way that is sufficient to handle the problems of concept recognition. This description of concepts and the arguments for it are described in greater detail in Wickelgren (1969b).

It is worth pointing out that the representation of concepts proposed by Katz and Fodor (1963) is essentially a theory that characterizes a concept by a simple conjunction of attributes, which we have noted is not adequate to account for the perceptual recognition of concepts. What role if any such "dictionary" or "scientific" definitions of concepts may play in human thought or human perception and cognition is another question, but the dictionary definition is clearly not sufficient to define concepts as people actually use them.

Furthermore, Katz and Fodor (1963) postulate word representatives, which encode all the different possible meanings of a concept, rather than concept representatives, which encode only a single meaning.

Another important issue in discussing the representation of concepts concerns whether or not there is any unique representative for each concept separate from the set of representatives of any of its component attributes. As we have seen, no single conjunction of attributes is sufficient to define concepts as people use them, but perhaps a large number of sets of attributes would be sufficient to stand for the concept, without any need to specify a single unique representative for any of these sets of attributes. The issue can be stated in terms of whether there is a single neuron somewhere that stands for the one or more conjunctions of attributes that define the concept or whether the representation of the concept is simply by the set of representatives for each of its attributes. This problem is discussed at some length in Wickelgren (1969b), where it is concluded that there would be severe associative interference problems if the only representation of a concept was by a set of attribute representatives with a no single concept representative. Attribute representation of concepts simply creates too much confusion in an associative memory. Thus, it appears to be necessary to assume that individuals can "chunk," in the sense of Miller (1956), a set of attributes to form a new representative for each chunk. Then perhaps humans associate several different chunk representatives into a single concept representative to stand for the disjunction of these chunks. According to this theory, the process of concept learning results from two component processes: (*a*) chunking of the attributes within any given conjunction and (*b*) associating the different chunks together to form their disjunction, which is the concept.

C. Evidence for Concept Representatives

There is abundant evidence for the existence of concept representatives that stand for meanings in addition to representation of verbal concepts only by their segmental components (phonemes, allophones, syllables, etc.). In long-term recognition memory for single words, false-recognition errors are consistently higher for words that are either semantically or phonetically similar to previously presented words than they are for control (neutral) words (Ainsfeld & Knapp, 1968; Grossman & Eagle, 1970; Kimble, 1968; Klein, 1970; Underwood, 1965).

Although I know of no formal experimental evidence to support this, personal experience suggests that associations learned to a concept signaled by a particular set of cues will transfer to the same concept signaled by a different set of cues. Since the cues may have nothing in common from a perceptual or structural point of view this can only be mediated by some type of concept representative that uses semantic (meaningful) encoding.

Additional evidence for the existence of concept representatives comes from studies in conceptual recognition of words. Miller, Heise, and Lichten (1951), Rubenstein and Pollack (1963), and Stowe, Harris, and Hampton (1963) found that content words (nouns, verbs, adjectives, adverbs) are more easily recognized when presented in a grammatical sentence than when spoken in isolation. Although I cannot prove this definitely, it seems extremely unlikely that this effect could be mediated by the grammatical sentential context biasing a particular set of phonemes or syllables or other structural elements of the correct words. The number of possible words that could be presented in most sentential contexts is generally very large, and all structural elements (phonemes, syllables, etc.) must be represented many, many times in these different words. Rather it seems necessary to assume that the effect is mediated by a contextually induced bias (set) for concept or word representatives that can appear in the particular context of other words in some sentence. Miller and Isard (1963) obtained an intelligibility advantage for words spoken in either grammatical sentences or partially grammatical (anomalous) sentences versus words presented in ungrammatical strings. The number of possible alternatives for each position in the anomalous strings is even greater than in the perfectly grammatical strings, making it seem very improbable that the effect could be mediated by setting oneself for particular syllables, phonemes, features, etc. Similarly, Bruce (1955) and Rubenstein and Pollack (1963) found that the intelligibility of a word was enhanced when it was known to come from a particular conceptually defined set, such as the set of all foods, rather than being any possible word in the language. Again, since the reduction in the set of alternatives is based on semantic criteria, rather than on phonetic criteria, it seems very unlikely that the effect could be obtained unless there were concept representatives in some semantic system, in addition to phonetic representatives of the components of words.

Finally, the word frequency effect (e.g., Brown & Rubenstein, 1961; Howes, 1957; Pollack, Rubenstein, Decker, 1959; Rosenzweig & Postman, 1957), in which more frequently occurring words in a language are more easily recognized than less frequently occurring words, seems difficult to explain without assuming that there are concept or word representatives and that we are biased (set) to either perceive or emit more familiar words, as compared with less familiar words. It seems unlikely that very much of this effect could be attributed to systematic differences at the structural (phonetic) level between frequently occurring and infrequently occurring words. However, Fredericksen (1971) has presented some limited evidence that there may be some systematic phonetic differences between frequently and infrequently occurring words, so perhaps the issue cannot be considered closed. Nevertheless, there is an enormous amount of converging evidence for the reality of concept representatives, in addition to the possibility of segmental and feature representatives, at lower levels of the verbal modality.

D. Concept Representatives versus Word Representatives

Finally, there is evidence to indicate that these semantic units are concept representatives that encode one meaning of a word as postulated by Wickelgren (1969b), not word representatives that encode all the meanings of a word, as postulated by Katz and Fodor (1963).

When words to be learned were originally presented in a sentence context that biased a particular meaning, rather than other meanings, for each word, false recognition rates were elevated for words related to the correct meaning but not for words related to the other meanings of each presented word (Perfetti & Goodman, 1970). Light and Carter-Sobell (1970) and Tulving and Thomson (1971) have demonstrated that correct recognition memory for a word is decreased by changing from learning to testing the verbal context in which it is presented, in a manner that presumably changes in some cases the meaning the subjects attach to the word.

Also, MacKay (personal communication) found that associative responses to ambiguous words (words with multiple meanings, such as *sound*) took longer than responses to "unambiguous" words (words with one primary meaning or closely related meaning such as *clock*). MacKay controlled for length, frequency, emotional tone, and "other factors" that he thought might possibly confound the comparison. As MacKay concludes, this indicates that associations are not between word representatives but between concept representatives.

IV. PHRASES, CLAUSES, AND SENTENCES

Although I know of no relevant evidence, it is reasonable to conjecture that familiar (frequently occurring) phrases may be represented by single units at the concept level, just as words are. That is to say, the phrase *black bird* may be represented as a single concept, just as is the single word *blackbird*.

Since human beings are capable of consistently identifying the phrases, clauses, and sentences in a long utterance, there is reason to believe that these constitute "units" in *some* sense in the nervous system. However, I know of no evidence to indicate that each distinct phrase, clause, or sentence is represented by a single unit in the nervous system in the sense used for concept representatives. Considering that most of the phrases, clauses, and sentences that we speak are novel (never before uttered by the speaker), it seems extremely unlikely that there are single representatives for each distinct phrase, clause, or sentence that we can produce. Rather, what seems more likely is that we have representatives for a variety of types of phrases, clauses, sentences, etc. and also for the relationships that can obtain between phrases, clauses, sentences, etc. These more general concepts probably play an important role in our understanding and production of speech, but further discussion of these matters is beyond the scope of the present chapter.

V. FEATURES

A. Articulatory Features

The production of the sounds of speech depends upon the movement of air through the vocal tract in a manner analogous to the production of a musical sound by the passage of air through a complex pipe. The vocal tract has two channels (pipes) through which the air may pass, namely, the oral tract and the nasal tract. The vocal tract begins above the larynx (glottis, voice box) with a single cavity (pipe) known as the pharynx, which branches to form two cavities, the nasal cavity (through the nose) and the oral cavity (through the mouth).

1. Chest

The production of the airstream essential for speech takes place primarily by operation of the chest muscles, which force air from the lungs, through the glottis, and then into the vocal tract. Alternate mechanisms for the production of an airstream are used in some languages.

Once an airstream has been produced, the acoustic characteristics of the sound one hears depend on the configuration of the vocal tract. In contrast to the usual pipe or system of pipes, however complex, the vocal tract is capable of a modification of its characteristics at a large number of places by means of muscular activity controlled by motor neurons.

2. Larynx

In the larynx, the vocal cords may be in a variety of states. One state is that used in the production of voiced sounds, including all English vowels and the consonants /b/, /d/, /g/, /z/, /m/, /n/, etc. In the case of all voiced sounds, the vocal cords are in vibration, producing a characteristic pitch (fundamental frequency) that can be heard in association with these sounds. A second state of the larynx is that characteristic of voiceless aspirated sounds in English, such as the consonants /p/, /t/, and /k/, when these sounds occur in the initial or terminal position of syllables. A third state of the larynx is the voiceless unaspirated state. This state is used in producing such sounds as /p/, /t/, and /k/ in many interior positions of syllables, such as /p/ in the word *spot,* /t/ in the word *stove,* or /k/ in the word *ski.* It is quite easy to tell for yourself whether a sound such as /p/ is aspirated or unaspirated. Put your hand in front of your mouth during the pronunciation of a word that contains the sound /p/ and feel whether or not a puff of air hits the palm of your hand during the pronunciation of /p/. Aspirated /p/, as in *pit,* is accompanied by a puff of air;

unaspirated /p/, as in *spit,* is not. A fourth state of the larynx is that characteristic of whispered voice.

3. Velum

Another modifiable aspect of the vocal tract is whether the soft palate (velum) is up or down. When the velum is up, the nasal cavity is shut off from the vocal tract and the airstream passes only through the oral cavity. In English, most sounds are strictly oral, including such consonants as /p/, /b/, /t/, /d/, /k/, /g/, /s/, /z/, /r/, /l/, /w/, etc. When the velum is down, the airstream passes through both the nose and the mouth. Such sounds are known as nasals and, in English, include only the phonemes /m/, /n/, /ŋ/, where /ŋ/ is the terminal phoneme in /ing,/ /ang/, /eng/, /ong/, etc.

4. Jaw

Another modifiable aspect of the vocal tract is the jaw, which partly controls the degree of openness of the vocal tract (principally the volume of the oral cavity). Other factors, including principally lip and tongue position, also control the openness of the vocal tract, and it is not clear to what extent jaw position is modified to control openness.

5. Lips

Lip position, including whether the lips are closed or opened and also whether the open position of the lips is spread or rounded, is another important characteristic of the vocal tract. For example, the consonant phonemes /p/, /b/, and /m/ are all produced by a complete closure of the lips. The consonant (semivowel) /w/, is characterized by a rounded, open lip position.

6. Tongue

Finally, the most important modifiable aspect of the vocal tract is the position of the tongue. Muscles in the tongue exert considerable control in determining the openness of the vocal tract. In addition, the tongue can be elevated at many different places, markedly changing the relative configuration of the oral and pharyngeal cavities. The point at which the vocal tract is maximally constricted is known as the place of articulation. Place of articulation is an important feature dimension of both vowels and consonants, and the tongue is the principal articulator determining place.

Vowels and consonants differ in place of articulation, but whether the places of articulation for the vowels in any way correspond to the places of articulation for consonants is unknown. The question is difficult, since the degree of openness (largely determined by the degree of elevation of the tongue) for vowels is much greater than that for consonants. Similarly, semivowels, including the sounds /w/, /r/, /l/, /y/, and possibly /h/, are generally characterized by a degree of openness of the vocal tract intermediate between that for vowels and consonants. Once again, although it is possible to make some type of identification of the places of articulation of semivowels with corresponding places for consonants, this identification is extremely tentative, at present.

B. Auditory Features

1. Narrow-Band Frequency Cues

There is considerable neurophysiological evidence to support the existence of neurons at many levels of the nervous system that respond selectively to certain frequencies or frequency components of complex sounds. These neurons typically have a "tuning curve," responding maximally to a tone of a particular frequency and responding progressively less well to frequencies farther away from the "ideal" frequency for that neuron. Thus, there are frequency detectors to represent the fundamental frequency and also those relatively narrow-band resonances of the vocal tract known as formants, which are considered to be distinctive features for the perception of vowels (Delattre, Liberman, Cooper, & Gerstman, 1952; Peterson & Barney, 1952).

2. Frequency Transitions

In addition, some neurons have been discovered at higher levels of the auditory nervous system that seem to respond to various frequency transitions (Suga, 1965, 1968; Whitfield & Evans, 1965), raising the possibility that there is unitary internal representation for each type of frequency transition. For example, there might be a set of neurons that respond selectively to a transition from 400 to 700 Hz, occurring over a 20–100 msec interval, a different set of neurons responding primarily to a transition from 400 to 900 Hz, still another group of analyzers responding selectively to transitions from 1400 to 1000 Hz, etc. Since changes in the formant frequencies (formant transitions) are known to be good cues for identifying consonants that precede or follow a vowel (see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967, for a review), single units representing these features would be very helpful. To my knowledge, there is no direct evidence that formant transitions have a unitary feature representation in the human nervous system. Virtually all of the work that has been done on the auditory nervous system has been done on animals other than man, for obvious reasons.

3. Broad-Band Formants

The formants and formant transitions that provide the important acoustic features for the perception of stop consonants, semivowels, and vowels are relatively narrow-band concentrations of energy around maxima at certain characteristic frequencies. By contrast, fricatives and affricates (f, v, θ, ð, s, z, š, ž, etc.) are characterized by extremely broad-band concentrations of energy in different frequency regions of the spectrum (Heinz & Stevens, 1961; Hughes & Halle, 1956; Jassem, 1965). The exact cues that distinguish fricatives and affricates are less well established than the cues for other English sounds. However, it does appear that concentrations of energy over frequency regions of many thousands of hertz are more characteristic cues for fricatives (especially s, z, š, and ž, see Harris, 1958) than local maxima (formants) with bands on the order of tens or hundreds of hertz. In part, this may be a consequence of the fact that most of the friction noise present in fricatives and affricates is at much higher frequencies than the fundamental frequency or the first-, second-, and third-formant frequencies characteristic of other speech sounds. The friction noise in fricatives and affricates is to a very large extent contained between 2000 and 12,000 Hz. It is well established that a given degree of discriminability of pure tones requires a progressively greater absolute (but not relative or percentage) separation in frequency as the frequency of the tones increases (e.g., Harris, 1952). Along this line, it is reasonable to suppose that frequency detecting units have broader bands at higher frequencies, measured in absolute frequency units (hertz).

While the friction noises characteristic of fricatives and affricates can be prolonged, there are other brief noise cues, called noise bursts, that provide useful cues in the recognition of stop consonants (Liberman *et al.,* 1967).

4. Silence

When a stop consonant, such as /p/ or /t/, is articulated, there is a short period of silence during which the vocal tract is completely closed and no sound can be heard. This period of silence is known to be an important cue for the perception of the stop consonants (Liberman *et al.,* 1967). Of course, the period of silence occurs at a different point in time than the formant transitions to or from the following or preceding vowel. Thus, in recognizing a phoneme we must assume that the features used in recognition do not all occur at exactly the same points in time but, instead, are spread over some small region in time on the order of 100 to 500 msec.

5. Duration

The duration of different periods of silence, noise bursts, broad-band formants, narrow-band formants, formant transitions, etc. may be cues of some importance in speech recognition at a given rate of speaking. Studies of the duration of different phonemes or other classes of speech segments invariably indicate systematic differences between different phonemes or other classes of speech sounds (e.g., Kozhevnikov & Chistovich, 1965; Lehiste, 1970b). However, it is often difficult or impossible to determine any particular point in time where one phoneme ends and the next phoneme begins. The cues for a given phoneme are intermixed in time to a large extent with the cues for adjacent phonemes. This being the case, it is not clear from peripheral articulatory or acoustic measurements what underlying psychological duration should be assigned to each segment (phoneme).

Although segmental duration may be a relatively meaningless concept in speech recognition, feature duration is known to be an important cue in discriminating between certain phonemes. Lisker (1957) has shown quite conclusively that the duration of stop consonant closure is a critically important cue for the distinction between voiced and unvoiced consonants /p/ versus /b/. The acoustic cue for the duration of stop-consonant closure is, of course, a period of relative silence, with voiceless consonants such as /p/ having closing durations in the Lisker study ranging from 90 to 140 msec (averaging 120 msec) and closure durations for the voiced consonant /b/ varying from 65 to 90 msec (averaging 75 msec).

Furthermore, Bolinger and Gerstman (1957) have shown that silence duration is an important cue for juncture (in this case, word segmentation) in distinguishing such phrases as *lighthouse keeper* and *light housekeeper*. The results are that the tendency to put *lighthouse* together is enhanced by having a short silent period between *light* and *house,* relative to the silent period between *house* and *keeper*. Conversely, the tendency to hear *housekeeper* as a single word is enhanced by making the silent period between *light* and *house* long, in relation to the silent period between *house* and *keeper*.

Given that rates of talking are variable, it is probably necessary for the listener to take speech rate into account in some way to make maximal use of duration cues. Relative durations of some kind are probably more critical than absolute durations.

6. Intensity

Integrating over all audible frequencies, phonemes differ systematically in their acoustic intensity (Lehiste, 1970b). Of course, we have already mentioned that an important cue for distinguishing different phonemes or allophones is the intensity in different (narrow or broad) regions of the frequency spectrum. Whether any type of integration over the entire audible frequency spectrum is an independently significant cue for speech recognition is not known. To establish this, we would need to know how intensity in different frequency regions is weighted by the "loudness" units in the nervous system assessing overall intensity. Furthermore, as with the duration cue, absolute intensities (air pressure levels) differ very considerably with different degrees of stress and different emotional states, even for the same speaker, to say nothing of differences across different speakers.

It was once thought that intensity was the primary cue for the suprasegmental stress feature, but the status of intensity as a cue to stress is now quite debatable (Bolinger, 1957–1958; Lehiste, 1970b; Lieberman, 1967; Morton & Jassem, 1965). It now seems generally agreed that fundamental frequency is the primary cue to stress (frequency changes of several types cueing stress; Bolinger, 1958), with duration being of some significance (greater duration indicates greater stress). Intensity, if it is a cue at all, is far less important than either fundamental frequency (intonation) or duration. It seems conceivable that the function of intensity variation in speech is largely a matter of achieving the appropriate signal-to-noise ratio for speech to be heard under different conditions of background noise. Overall intensity may be of no significance whatever, at either a segmental or suprasegmental level, as a cue to the meaning of the utterance.

C. Abstract Linguistic Features

Phonetics would be greatly simplified, if there were a simple relation between each articulatory feature and some auditory feature.

In some cases, there is a relatively simple relation. For example, the voicing state of the larynx is correlated with the presence of a fundamental frequency component in the acoustic signal. For steady-state vowels, there is a fairly simple relation between the openness and the frequency of the first formant and between place of articulation and the frequency of the second formant (e.g., Peterson & Barney, 1952). However, it should be noted that openness of the vocal tract and place of articulation are not simple articulatory feature dimensions to the same extent as is voicing.

At the other extreme, place of articulation for consonants is determined by a number of different muscles in both the tongue and the lips (and possibly the jaw), and the acoustic features of the "same" place differ for different phonemes and for the same phoneme in different phonemic environments. For example, the acoustic cues for the place of articulation of the /d/ phoneme differ greatly depending upon whether it is followed (or preceded) by the vowel /ɪ/ or the vowel /ɑ/. (See Liberman *et al.*, 1967, for a more complete discussion of this issue.)

At present, there is no accepted theory of the articulatory and auditory features of speech sounds and the relation between them. Instead of having two feature analyses, one auditory and one articulatory, what we have are a variety of somewhat similar proposals for a single abstract (linguistic) feature analysis of those abstract classes of speech sounds known as phonemes (e.g., /p/, /z/, /o/, etc.).

Phonological arguments can be given for one distinctive feature system as opposed to another, but these are beyond the scope of the present chapter, and the psychological reality for articulation and recognition of much of phonology is in doubt (see Ladefoged, 1970).

One way to justify assuming an abstract feature level is if the sensory and motor aspects of speech unite at the feature level, as opposed to the segmental level or the concept level. According to such a theory, there should be a definable relation between articulatory and auditory features, so that the same feature representatives receive sensory input during perception, and control motor output during articulation. The last 20 years of research on speech perception and production make it seem very unlikely that such a relation can be defined (Liberman *et al.,* 1967). The theory defended in subsequent sections is that the sensory and motor aspects of speech unite at a segmental level, assuming that the segmental units are allophones (auditory and articulatory variants) of phonemes.

Conventional linguistic distinctive feature systems (excluding such purely phonological systems as that of Chomsky and Halle, 1968) may only be a convenient way of representing the average similarity of the phoneme classes of allophones at a segmental level. This average similarity might be jointly determined by both auditory and articulatory feature similarity in a manner too complex for present determination.

There is no doubt that some kinds of features have psychological reality, since articulatory and auditory features are needed for the speech production and perception processes, respectively. Furthermore, Hintzman (1967) and Wickelgren (1965, 1966) have demonstrated that errors in short-term memory for single phonemes are nonrandom in ways that can be described by distinctive feature systems, though these data do not permit a decision regarding the auditory, articulatory, or abstract nature of the features (Wickelgren, 1969d). However, there is no compelling evidence for the psychological reality of units representing abstract features, and existing feature systems include dimensions that may not represent either real articulatory or real auditory feature dimensions.

These latter dimensions are place and manner (openness) of articulation, which sound like articulatory dimensions (and may be at some central level), but which do not appear to be single dimensions at the peripheral motor neuron level.

Nevertheless, conventional linguistic feature systems serve the very useful function of representing, in at least an approximate way, the similarity of different phoneme classes of allophones. Furthermore, it is conceivable that such systems have articulatory or abstract linguistic reality (though it is very unlikely that they have auditory reality). For these reasons, a distinctive feature analysis of vowels is presented in Table 1 and a distinctive feature analysis of consonants is presented in Table 2. Arguments for something like this feature analysis can be found in Denes and Pinson (1963), Flanagan (1965), Gleason (1955), Ladefoged (1971), Pike (1947), Wickelgren (1965, 1966), and many other places.

Basically three or four feature dimensions are used to analyze the vowels. First, the major distinction is between simple vowels and complex vowels. Simple vowels refer in essence to a single configuration of the vocal tract, which is maintained in some sense throughout the vowel phoneme (though, of course, there are transitions from the preceding phoneme and transitions to the succeeding phoneme). In the case of complex vowels (complex syllable nuclei, diphthongs), the place of articulation or degree of openness or both change during the articulation of the vowel. In some cases, this change is extensive, as in the case of the complex vowel /ī/, where the initial portion of /ī/ is similar to the vowel /ɑ/, as in *hot,* but the terminal portion of /ī/ is similar to the semivowel /y/. In other instances, notably /ē/ or /o͞o/, and possibly /ō/, the degree of alteration of the vocal tract configuration during the complex vowel is less dramatic,

TABLE I

DISTINCTIVE FEATURES OF ENGLISH VOWELS

	Place		
Openness	Front	Middle	Back
		Simple Vowels	
Narrow (High)	ɪ (h*i*t)	ɨ (b*i*rd)	ʊ (f*oo*t)
Medium (middle)	ɛ (s*e*t)	ʌ (h*u*t)	o ([a])
Wide (Low)	æ (h*a*t)	ɑ (h*o*t)	ɔ (all)
		Complex Vowels	
Narrow	ē or i or ɪy (h*ea*t)	er or ɨr (b*ir*d)	o͞o or u or Uw (b*oo*t)
Medium	ā or e or ɛy (h*a*te)		ō or ow (b*oa*t)
Wide		ī or ɑy (*i*ce) and ou or ɑw (c*ou*ch)	oi or ɔy (*oi*l)

[a] To my knowledge, short "o" does not appear in my English dialect, except as a part of the long vowel "ow."

TABLE II

DISTINCTIVE FEATURES OF ENGLISH CONSONANTS

	Place							
Openness (Manner)	0	1	2	3	4	5	6	7
Stops								
UV	p			t			k	
V	b			d			g	
N	m			n			ŋ	
Affricates								
UV		f	θ			č		
V		v	ð			j		
Fricatives					s	š		
V					z	ž		
Semivowels								
UV								h
V	w			r	l	y		

NOTE: /θ/ = /th/ in "*th*ink", /ð/ = /th/ in "*th*e", /č/ = /ch/ in "*ch*eck" /š/ = /sh/ in "*sh*ook", /ž/ = /zh/ in "azure", /ŋ/ = /ng/ in "bri*ng*" UV = unvoiced, V = voiced, N = voiced nasal

to the point where some phoneticians have considered these vowels to have a single vocal tract configuration. (For a further discussion of this unresolved issue regarding which vowels are complex and the general relationship between various complex vowels and "corresponding" simple vowels and semivowels, see Lehiste, 1964, and Ladefoged, 1965.)

The other two distinctive feature dimensions for vowels, namely, the degree of openness and the place of articulation, are reasonably well established by comparison to the simple versus complex dimension. However, there are some disagreements regarding the classification of place and openness shown in Table 1. One question is whether /ɑ/, as in *hot,* should be classified as a middle-wide vowel (as in Table 1) or as a back vowel with a wider degree of openness than /ɔ/.

In general, for a given speaker there is no guarantee that the vowels found in the same row have the same degree of openness. Nor is there any guarantee that vowels found in the same column have the same place of articulation. The values of openness and place may be only relative to other entries within the same row or column, respectively. That is to say, within the front simple vowels, /ɪ/ is narrower than /ɛ/, which is narrower than /æ/, but /ɪ/ and /ᴜ/ do not necessarily have the same degrees of openness of the vocal tract. It is not even clear from an articulatory standpoint, given the different shape of the tongue for front versus back vowels, what it means to have the same degree of "openness." Nevertheless, this type of relative classification with regard to the values of place and openness has proved to be of substantial value in understanding many of the differences between vowels.

Disregarding such features as aspiration, which are not critical in a logical sense for distinguishing any pair of words in English, English consonants can be analyzed in terms of four distinct feature dimensions: openness, place of articulation, voicing, and nasality.

The openness dimension in Table 2 refers to how much the vocal tract is constricted at the place of maximum constriction. This varies all the way from complete closure for stop consonants to a rather open vocal tract in the case of semivowels. Fricatives have a sufficiently narrowed vocal tract so that the air passes through a very narrow passage producing considerable friction noise. Affricates are somewhat intermediate between stops and fricatives, and their exact status in relation to stops and fricatives is not definitely decided.

Place of articulation varies in a more extreme manner for consonants than for vowels, namely, there are some consonants, such as /p/, /b/, /m/, and the semivowel /w/, that have their place of articulation at the lips. In the case of /p/, /b/, and /m/ this is by means of complete lip closure. Other consonants are produced by putting the lower lip against the upper teeth, as in the case of /f/ and /v/, or by placing the tip of the tongue against the upper teeth, as in /θ/ and /ð/, all the way back to /h/, which is made by pushing the root of the tongue against the back wall of the pharynx.

The number of different values of the place dimension could be reduced from the number used in Table 2 by collapsing place values 0 and 1, place values 2 and 3, and place values 6 and 7, since the more detailed analysis shown in Table 2 is not logically necessary for assigning a distinct set of feature values to each phoneme. However, since the places of articulation are reasonably different, a relatively arbitrary decision was made to adopt the more finely differentiated set of values on the place dimension. On the other hand, the feature analysis shown in Table 2 ignores many additional important characteristics of each consonant. For example, nothing in Table 2 indicates that /w/ has lip rounding, rather than the spread lip configuration. Nothing in Table 2 indicates some of the distinctive characteristics of the shape of the tongue for the phonemes /r/ and /l/. Aspiration is ignored, and so on.

Voicing and nasality have been represented in Table 2 as if they were a single dimension with three values. This reduction in the complexity of the table is possible for English, since all nasal consonants are voiced. However, logically speaking, voicing refers to a state of the vocal cords, while nasality refers to whether the velum is up or down.

VI. SEGMENTS

A. Phonemes

Although this analysis is currently being challenged within transformational generative phonology (e.g., Chomsky & Halle, 1968), linguists traditionally analyze a word or phrase that represents a concept into a sequence of structural units called phonemes. Thus, the English word *struck* would be analyzed to consist of the following ordered set of five phonemes /s/, /t/, /r/, /u/, /k/. The six letters used to spell the word *struck* in English orthography are often referred to as graphemes. The spelling of a word in written English is an ordered set of graphemes (letters). Although there is a rough correspondence in English between graphemes and phonemes, the relationship is very far from one-to-one.

Whereas words stand for concepts that have meaning, phonemes stand for structural components of words and have, in general, no meaning at all. This is a general characteristic of the segmental level of analysis, namely that the segments themselves are purely structural units with no meaning. Communication of concepts from one individual to another virtually requires representation of words by a succession of segments, since the number of unique, hearable, positions of the vocal tract is many orders of magnitude less than the number of different concepts that human beings possess. Since the human vocal system is essentially a one-sound-at-a-time system, it is necessary to construct sequences of sounds (vocal tract positions) in order to represent all of the concepts human beings possess.

There are usually considered to be approximately 40 consonant and vowel phonemes in the English language (for my listing of 41 consonant and vowel phonemes see Tables 1 and 2). Although there is some disagreement regarding the exact inventory of phonemes characteristic of English (particularly concerning whether complex vowels should be considered phonemes), there would be little disagreement concerning how any particular word ought to be spelled phonemically, within any given system of phonemes.

The primary theoretical justification for phonemes is that they constitute a theory of the segmental representation of words at some central level in the nervous system. Three properties of this representation bear explicit mention. First, the inventory of segments is relatively small in comparison to the number of words in English. Second, the segments are nonoverlapping. Knowledge that a word contains a given segment does not logically restrict what other segments might be used in the spelling of the word. [In regard to this second point it must be mentioned that, if we have statistical data on the spelling of words in the language, then we can make statements concerning conditional probabilities that certain phonemes will be found preceding or following or otherwise occurring in the same word with other phonemes. Furthermore, if we have some additional restrictions added to the phonemic theory regarding what phonemes are permitted to appear in association with what other phonemes in words (namely, a theory of phonology), then we also have a basis for making predictions. However, for the moment I am assuming that we lack both statistical and phonological information. My point is that spelling words in terms of ordered sets of phonemes does not logically require that there be nonrandom orderings of the segments across the dictionary of words in a language.] Third, the spelling of a word is by an *ordered set* of segments (phonemes) not by an *unordered set* or a two-dimensional tree structure, etc.

Note that, thus far, nothing has been said regarding the articulatory or acoustic representation of phonemes, though clearly the fundamental purpose of a segmental representation of words (concepts) is to permit communication by a sequential output device (the vocal tract) to a sequential input device (the ear).

The simplest relation between abstract phonemes and peripheral articulatory and acoustic events would be for a sequence of phonemes to translate into a discrete sequence of vocal tract configurations and, therefore, a discrete sequence of acoustic patterns characteristic of each phoneme. In this simple system, if there were 40 phonemes, there would be 40 vocal tract configurations (or 40 short sequences of vocal tract configurations) and 40 different acoustical events, one for each phoneme. Any given phoneme would have the same vocal tract configuration and the same acoustical properties independent of its context. That is to say, the articulatory and acoustic features of a phoneme would be independent of what other phonemes occurred next to it or in other positions in the same word. Furthermore, the vocal tract configurations and acoustic events characteristic of each phoneme would be nonoverlapping in time. That is, the articulatory and acoustic features of each phoneme would be presented during a period of time that in no way overlapped with the articulatory and acoustic features of any other phoneme. If speech segments had these two properties, we might say they were "context free" and temporally segmentable.

In actual fact, at an articulatory and acoustic level, speech does not consist of context-free segments, nor is it temporally segmentable. The vocal tract configuration and associated acoustic representation of a given phoneme vary considerably with the context, primarily the immediate left and right phonemic environment. That is to say, the phoneme /d/ will be very different in both articulatory and acoustic features depending on whether it is preceded or followed by the phoneme /ɛ/ or the phoneme

/ɑ/. Furthermore, there is not even a core of invariant acoustic features for each phoneme across all different contexts. (For a review of the extensive literature supporting these conclusions, see Liberman *et al.*, 1967; MacNeilage, 1970; Wickelgren, 1969a.)

There are two good reasons for this context-conditioned variation (coarticulation), which are inherent in the properties of the speech musculature: (*a*) inertia (the vocal tract cannot move instantly from one configuration to another) and (*b*) starting position (different muscular contractions are required to achieve the same terminal position of an articulator from different starting positions). Given these properties of the speech apparatus, one could only realize a sequence of invariant articulatory and acoustic segments by ignoring the transition regions between adjacent speech sounds and concentrating only on some invariant central period of time during which the vocal tract configuration and acoustic cues might be invariant for a given phoneme in all different contexts. This later, somewhat weaker type of context-free, temporally segmentable, phonemic coding does not obtain either. Given the neural and mechanical limitations of speech articulation, speech, if it were required to have these context-free, temporally segmentable properties, would be incapable of occurring at as rapid a rate as it occurs (Liberman *et al.*, 1967). Instead, the acoustic cues characteristic of the transitions between target positions for successive phonemes constitute extremely important cues for the recognition of both the prior and the subsequent phonemes (Liberman *et al.*, 1967). Under these circumstances, it is often rather arbitrary and absurd to try to decide some particular point in time at which one phoneme ends and the next phoneme begins.

It is possible to subdivide speech into a number of discrete segments, but these segments have no simple relationship to the phonemes. For example, a stop-consonant plus vowel utterance contains at least four different distinguishable segments: a silent period during the complete occlusion characteristic of the stop consonant, the fricative burst characteristic of the very short period of time following opening of the consonant (during which there is only a very narrow opening in the vocal tract), the "vowel-like" period of the transition to the target vowel position, and finally, if the speech rate is slow, the steady-state vowel period. The acoustical cues inherent in the transition to the vowel contain information regarding both the consonant and the vowel, so there is no particular point in time at which one could decide that the consonant portion has ended and the vowel portion has begun. Furthermore, in rapidly articulated speech, steady-state target positions for vowels may never be reached, which further complicates analysis.

What ability we do have to segment speech acoustically derives from the fact that certain underlying features of speech are relatively discrete, in the sense of having only two or a few states, with relatively rapid transitions from one state to the other. Examples of such discrete features include voicing (whether the vocal cords are vibrating or not), nasality (whether the nasal cavity is open to the passage of sound or not), whether the vocal tract is completely occluded or not, and, to a lesser extent, the presence versus absence of fricative noise. The very important features of tongue position and the degree of openness of the vocal tract, for any period other than the complete occlusion phase of stop consonants, are relatively continuous features with a large number of different states and relatively gradual transitions between the states. Since some of the continuous feature cues are among the most important for the identification of speech segments, the impossibility of segmenting the values on these feature dimensions effectively precludes the segmenting of speech into discrete phonemes. Even the discrete features are of no value in segmenting speech when adjacent phonemes have the same value on that discrete feature dimension, such as when adjacent phonemes are voiced. In such cases, the voicing is simply maintained continuously, without any break to indicate a separation between the two phonemes. Such discrete features as the voicelessness of the preceding consonant may extend well into the transition to a following vowel, and the anticipatory presence of features characteristic of subsequent phonemes is also observed. For example, sometimes a vowel adjacent to a nasal consonant will be itself nasalized throughout its entire duration. This is just one example of what is meant by saying that the features characteristic of phonemes at a peripheral articulatory and acoustic level are not independent of the nature of the adjacent phonemic context. Note that this also contributes to the impossibility of segmenting speech into nonoverlapping phonemes at a peripheral articulatory or acoustic level. (See Fant, 1962, for a further discussion of the segmentability of certain speech features and the lack of segmentability of other features and of phonemes.)

Consistent with the lack of segmentability and the context-conditioned variation in the acoustic cues for phonemes, it is known to be impossible to achieve highly intelligible speech by cutting "phoneme-size" segments from recording tape and splicing them together in new combinations to form new words. Such techniques have been tried several times and do not work (Harris, 1953). The smallest size units that can be cut and spliced from recorded utterances to produce intelligible speech are roughly a half syllable in length, and these half syllables must mesh properly in order to produce intelligible speech (Peterson, Wang, & Sivertsen, 1958). That is to say, an /ni/ half syllable must mesh with an /ic/ half syllable to produce natural and intelligible speech. Thus, speech at the vocal tract and acoustic cue level is not composed of successive, context-free segments. That is to say, speech at these levels is not phonemic.

Of course, this in no way implies that speech is not phonemic at some higher level of the nervous system. However, in subsequent sections, arguments will be presented to suggest that phonemic coding is not used at any level of the nervous system in normal adult speech recognition and articulation.

B. Context-Sensitive Allophones

A number of problems in speech recognition and articulation can be solved by making an assumption regarding the segmental encoding of words that is an alternative to the phonemic coding assumption (Wickelgren 1969a,c). Instead of assuming that a word such as *struck* is encoded as an ordered set of context-free phonemes /s/, /t/, /r/, /u/, /k/, one could assume that a word such as *struck* was encoded by an unordered set of context-sensitive allophones: $/{}_{\#}s_{t}/$, $/{}_{s}t_{r}/$, $/{}_{t}r_{u}/$, $/{}_{r}u_{k}/$, $/{}_{u}k_{\#}/$. Each of the context-sensitive elements in this code essentially contains a little bit of local information concerning the ordering of this element in relation to other elements in the set.

If there were a unit in the nervous system for each such context-sensitive allophone, then the representation of the word could be by an unordered set of such units whenever this was convenient for speech recognition and articulation, since the information concerning the ordering of these elements can be derived from the unordered set in all cases for single English words. For example, one first looks for the only element that has # as its initial element, then this initial allophone provides the information that the next allophone is of the form $/{}_{s}t_{-}/$, where "–" stands for an unknown phoneme, and so on until all the allophones in a word are correctly ordered. It is much like assembling a linear jigsaw puzzle. With phonemic coding, the same unordered set of phonemes can often be ordered to form two or more words. However, with context-sensitive allophonic coding it is a remarkable fact that there is, at most, one way of making an English word out of any unordered set of context-sensitive allophonic symbols.

Even if one were to take rather long phrases consisting of many words and scramble their context-sensitive allophonic segments, it will almost always be possible to reorder the symbols to form a unique reconstruction of the ordering of the allophones to form words in the phrase. This is especially true, if the terminal segment of the *i*th word is sensitive to the initial segment of the $(i+1)$st word and the initial segment of the $(i+1)$st word is context-sensitive to the terminal segment of the *i*th word. Such a context-sensitive coding would be said to "cross" word boundaries. For such an encoding of the phrase *Jim struck,* the terminal segment of *Jim* would be $/{}_{i}m_{s}/$ and the initial segment of struck would be $/{}_{m}s_{t}/$.

Other mechanisms for achieving the correct ordering of words (sets of allophones) without context-sensitive coding crossing word boundaries are presented in Wickelgren (1969a) and in the section on suprasegmental components later in the present chapter.

Even for a completely artificial language with random selection of symbols for the spelling of "words" with ordered sets of symbols from a relatively small vocabulary of letters or phonemes (on the order of 20 or more), it is very improbable that the information concerning the ordering of these letters could not be recovered from the overlapping-triple type of context-sensitive coding proposed here. It came as a considerable surprise to me to realize how much of the information concerning the ordering of very, very long sets of elements can be communicated by this type of extremely local information concerning the relative order of adjacent elements. (For a further discussion of the mathematical properties of context-sensitive coding see Wickelgren, 1969a,c.)

The number of context-sensitive allophones (phoneme triples) that would be required for English can be determined by a count of the number of different phoneme triples that occur in the language. One such count is that of Hultzén, Allen, and Miron (1964) who found a total of 3,083 different phoneme triples in their sample of 20,032 consectuvie phonemes. This analysis considered accented and unaccented vowels to be different phonemes, and this two-level lexical accent system is probably completely adequate. Nevertheless, these figures probably underestimate by a factor of two or three (surely less than ten) the number of different triples occurring in English. This is because the sample size is too small. However, it seems likely from these data that there are no more than about 10 or 20 thousand phoneme triples in English. If syllable boundaries are marked in the manner suggested in the next section, then less than 5 or 10 thousand context-sensitive allophone representatives would probably be required. While such a number is large in comparison to 40–100 phonemes as the segmental units of the language, it is small in relation to the number of neurones in the nervous system (about 10^{10}), so that there is no need to be concerned about the number of segmental representatives that must be assumed by this theory.

The assumption that context-sensitive allophones are the primary internal representatives used at the segmental level in speech recognition and articulation solves the two previously discussed problems with the assumption of phonemic segmental coding. First, context-sensitive coding solves the problem of segmenting speech at a peripheral articulatory and acoustic level. Since the segments are now overlapping, rather than nonoverlapping, by definition, there is no need to segment speech into nonoverlapping portions of time. Second, there obviously can be context-conditioned variation in the articulatory and auditory features of the allophones within each phoneme class. No invariant core of articulatory or auditory features is required to define a phoneme, because the units are allophones of each phoneme, which may be different in these features for every different immediate left and right phonemic environment. There should be a core of articulatory and auditory features for each allophone that are invariant over variations in more remote phonemic context. This latter requirement has not been adequately tested. However, remote context-conditioned variation is known to be far smaller than immediate context-conditioned variation, so this prediction is probably valid.

In the control of speech articulation, there could be activation of a sequence of context-sensitive allophone representatives at some central level of the nervous system that would translate into a smoothly flowing sequence of articulatory movements at a peripheral level. Since each segmental control unit is sensitive to ("knows") the target position for the previous and the succeeding segments, the instructions to the articulatory muscles can take prior and subsequent target positions into account in trying to achieve approximately the same target position for the currently articulated allophone as for any other allophone within the same phoneme class. For example, the motor intructions (features) for the /${}_{\upsilon}d_{\varepsilon}$/ allophone can be different from the motor instructions for the /${}_{\alpha}d_{\text{I}}$/ allophone in just the manner necessary to achieve approximately the same target position for both allophones of the /d/phoneme. According to the context-sensitive coding theory, the peripheral articulatory features of each phoneme class of allophones should exhibit context-conditioned variation. Also, it should be difficult or impossible to segment the stream of motor commands or vocal tract configurations, since the (allophonic) control elements are essentially overlapping phoneme-triples. Furthermore, the turning-off of one allophonic control unit and the turning-on of subsequent allophonic control units could be temporally overlapping to some extent. Thus, context-sensitive coding is consistent with these basic facts concerning speech articulation.

In speech recognition, all context-sensitive allophone representatives could be operating in parallel "looking" for the set of features needed to activate them with the requirement that features all be presented over some maximum unit in time. That maximum unit of time might well be variable for different perceived rates of talking. With such a parallel speech recognition scheme, there would be no need for prior segmentation of the speech stream. Also, with context-sensitive allophone detectors, the presence of context-condition variation in the acoustic cues for different "segments" would be an aid to speech recognition by increasing redundancy, rather than a hindrance to speech recognition.

It bears mentioning again that the results of Peterson *et al.* (1958), that the smallest segments that can be used for speech synthesis by the cutting and splicing technique are properly meshed half syllables, is precisely what one would expect on the view that the principal segmental representatives are context-sensitive allophones.

C. Syllables

The syllable has sometimes been proposed as an important unit in the segmental analysis of words. Thus, a word like *construct* is analyzable into two syllables: /con/, /strukt/. There are rather serious definitional problems in the case of segmental analysis of words into syllables, due to the fact that intuitions are far more variable regarding the syllabic analysis of words than the phonemic (and allophonic) analysis of words. A word such as *syllabic* might be subdivided into syllables by one person as /sil/, /lab/, /ik/; another might analyze it as /sil/, /ab/, /ik/; still another might analyze it as /sil/, /lab/, /ik/; another might analyze it as /si/, /lab/, /ik/; yet another might analyze it as /sil/, /ab/, /ik/; still another might analyze it as /sil/, /la/, /bik/, and so on. More sophisticated higher-level linguistic arguments can sometimes be given for a particular mode of syllabic analysis as opposed to another, but there is considerable disagreement at this level also regarding what, if any, syllabic analysis of words is useful. Personally, my intuition regarding *syllabic* is that there are no definite syllable boundaries in this word, though there clearly are three vowels and a definite accent pattern.

Context-sensitive allophonic coding is in a sense a type of overlapping syllabic coding, which would assert that there was a reality to all the different ways of analyzing a word such as *syllabic* into syllables. By contrast, syllabic coding requires analysis of a word into nonoverlapping segments. Syllabic coding provides no information regarding the ordering of syllables within a word. Furthermore, it is not at all clear how the assumption of syllabic coding could provide information regarding the ordering of the phonemic (allophonic) constituents of a syllable. So, although there are some superficial similarities between context-sensitive allophonic coding and syllabic coding, there are a greater number of differences.

Syllabic coding has sometimes been suggested as a way to provide invariant segmental units for auditory recognition, to get around the enormous effects of the immediate left and right phonemic context on the acoustic cues for the single phonemes. However, to the extent that initial and terminal phonemic (allophonic) segments of syllables are sensitive to terminal and initial phonemic segments of adjacent syllables, this solution is not satisfactory. Nevertheless syllabic coding would to some extent get around the problem of context-conditioned variation in the acoustic cues for each phoneme (Mattingly & Liberman, 1969). Segmentability problems would be reduced (but not eliminated), since there would be a need to segment only at syllable boundaries.

Often it has been suggested that the syllable is an important unit in the timing of speech. At one time it was thought that each syllable was the result of a separate chest pulse of air, but this is now known to be false (Ladefoged, 1971).

Recently, a rather complex statistical analysis of temporal compensation in the pronunciation of words and phrases has been used to argue for the reality of the syllables (and the higher word and phrase levels) as being important units in the timing of speech (Kozhevnikov & Chistovich, 1965; Lehiste, 1970a; Shockey, Gregorski, & Lehiste, 1971). The method for determining the existence of timing units in articulation consists of looking for negative correlation between the durations of different segments within a word or a phrase. The notion is that if a syllable is a timing unit in the nervous system, then a single syllable such as /kon/ should be pronounced in a certain amount of time, given a particular rate of talking. If, in repeated pronunciation of the syllable /kon/, the speaker prolongs the /k/ phoneme more at some times than at other times, and if the syllable is a unit of timing in speech articulation, then the longer /k/ phoneme should be compensated for by a shorter /o/ phoneme or a shorter /n/ phoneme within that particular utterance of the syllable. The previously mentioned studies have found many examples of this negative correlation or temporal compensation within syllables, words, and even phrases.

However, all of the previously mentioned studies, except Lehiste (1970a), used a completely inappropriate method of data analysis. They selected from all of the utterances those that maintained the closest approximation to some particular time for the entire utterance. This, for statistical reasons, guarantees negative correlation between segmental components of the entire utterance. Since this negative correlation is guaranteed by selecting utterances of approximately the same total length, the finding of negative correlation gives no evidence for the psychological reality of any unit.

When such statistical selection was not used, as in Lehiste (1970a), a greater mixture of positive and negative correlation was found, though there were some consistent negative correlations, such as between a vowel and a subsequent consonant. However, even such findings are questionable, since speakers were instructed to maintain a constant rate of talking. Whatever the causes of differences in rate of speaking in normal conversation, instructing a subject to maintain a constant rate of talking may induce artificial types of temporal compensation that ordinarily never take place. For example, in these experimental tasks, where the subject is to produce the same short utterance tens or hundreds of times at a constant rate of talking, it is possible for subjects to maintain a constant rate by controlling either rate of talking or the total duration of their utterances. Subjects may attempt to maintain a constant duration of talking in short utterances by *changing* their rate of talking at various points during the utterance. It may be quite erroneous to interpret constant durations of short utterances to indicate constant rates of talking. It is conceivable that constant rates of talking may be more achievable under natural conversational conditions with longer utterances. The foregoing studies may have guaranteed negative correlation simply by the instructions to the subject, with the results indicating nothing regarding the functioning units in ordinary speech production. I do not see how to solve this problem within the context of this type of study. Until it is satisfactorily resolved, it seems to me we can place no confidence in the results that have been obtained.

Allen (1972) presents similar criticisms of these studies and reports also that he and Ohala have failed to replicate many of the negative correlations. In addition, Allen (1972) and Kozhevnikov and Chistovich (1965) point out that any measurement error in locating segment boundaries will automatically produce negative correlation between adjacent segments. Given the difficulties of segmenting speech at an acoustic level, this is a serious problem.

Further reason for rejecting these findings as evidence for the psychological reality of syllables comes from studies by Huggins (1968a,b), who found no evidence for the perceptual reality of temporal compensation. Huggins found that, when an adjacent vowel was lengthened or shortened, there was no compensatory change in the subject's preferences for the duration of an accompanying consonant in the same syllable.

Savin and Bever (1970) and Warren (1971) found, surprisingly enough, that it took less time to identify either a monosyllabic word or a nonsense syllable target than it took to identify single phoneme targets. This provides further evidence of concept (word) representatives in the nervous system. The nonsense syllable versus phoneme comparison is also evidence of some units more extensive than a single phoneme being primary in the speech recognition process. However, these units could either be syllables or context-sensitive allophones (phoneme triples). In Warren's study, CV bigrams were recognized almost as quickly as entire nonsense syllables, suggesting that phoneme triples would be recognized just as quickly as entire syllables.

D. Syllable Juncture

The strongest evidence known to me for the psychological reality of the syllable comes from studies of errors in speech articulation. There has been repeated confirmation of the law stated by Boomer and Laver (1968), that "segmental slips obey a natural law with regard to syllable-place; that is, initial segments in the origin syllable replace initial segments in the target syllable, nuclear replace nuclear, and final replace final [p. 7]." Besides Boomer and Laver, this law has been supported by Fromkin (1968, 1970), MacKay (1970a,b), and Nooteboom (1967, 1968).

Another finding reported by MacKay (1970a,b) is that reversals (transpositions, spoonerisms) of phonemes more frequently involve the initial consonants of different syllables, whether the reversals are within a word or between words. This latter effect was extremely large, indicating that the transition from the terminal consonant of one syllable to the initial consonant of the following syllable is the weakest link in the ordering of phonemes in an utterance.

While the above findings regarding speech errors indicate that *syllable boundaries* must be taken into account in representing the phonetic encoding of words and phrases, these data do not show that the *syllable* as an ordered set of phonemes is a basic segmental unit in speech production. Occasionally one observes a transposition or substitution of an entire syllable, but "subsyllabic" (phonemes or clusters of phonemes) substitutions and transpositions account for almost all speech errors at a segmental level.

As an example of how one might mark syllable boundaries within the context-sensitive allophonic coding theory described earlier, consider the following analysis of the word *segment:* $/_{\#}s_{e}/$, $/_{s}e_{g}/$, $/_{e}g_{+}/$, $/_{+}m_{e}/$, $/_{m}e_{n}/$, $/_{e}n_{t}/$, $/_{n}t_{\#}/$. This type of context-sensitive analysis marks the initial, medial, and terminal allophones in syllables in a manner that explicitly indicates similarity to other allophones that occupy initial, medial, or terminal syllable positions. Thus, errors should tend to preserve syllable position, as reported by the above investigators. In addition, it should be obvious that the encoding of order information is weakest across syllable boundaries, according to the above analysis. This is consistent with MacKay's finding that reversal errors frequently involve syllable-initial position. An additional assumption should probably be made that word juncture (#) has high similarity to syllable juncture (+).

Marking syllable boundaries explains the previous speech error findings without requiring the syllable to be a segmental unit. However, the legimate question can be raised as to what purpose is served for the organism by the marking of syllable boundaries. It has just been observed that marking a syllable boundary reduces the integrity of the order information across that boundary, linking the last allophone of one syllable less uniquely to the initial allophone of the following syllable. This hardly seems like a desirable property for a speech production system. The only functional reason I can think of for marking syllable boundaries in this way is that it would substantially reduce the number of context-sensitive allophones required to be learned and represented in the system. There are few restrictions concerning what phonemes can follow what phonemes across syllable boundaries in English, but there are very substantial restrictions on the phoneme sequences within syllables.

Another way to look at somewhat the same point is to argue that phoneme sequences that cross syllable boundaries are not coarticulated to the same extent as phoneme sequences within syllables, and that this fact ought to be represented in the coding theory at a phonetic level. Along this line, one would argue that the /dwi/ phoneme triple in *sandwich* (assuming one does pronounce all of the phonemes indicated in the spelling of this word) and the /dwi/ sequence in *dwindle* are somewhat different in their allophonic coding. Namely, the /w/ allophone in *sandwich* is $/{}_{+}w_{i}/$ and the /w/ allophone in *dwindle* is $/{}_{d}w_{i}/$. According to this hypothesis, different acoustic cues would be associated with the two different /w/ allophones. The relatively small number of intrasyllable consonant clusters that we have in English are acquired developmentally later than the consonant vowel clusters. The pronunciation of consonant clusters may well be more difficult, due to mechanical properties of the vocal tract, and/or the perception of consonant clusters might be more difficult. This could mean that it is dysfunctional for a language to require coarticulated representations of all the possible consonant transitions. In any event, it seems in accord with both data and intuition to mark syllable boundaries in the manner indicated.

Furthermore, a slight modification of the above hypothesis would preserve considerably more of the order information across syllable boundaries in a manner that is still qualitatively consistent with the evidence for syllable boundaries derived from speech error data. The alternative hypothesis can most easily be explained by giving an alternative encoding of the word *segment:* $/{}_{\#}s_{e}/$, $/{}_{s}e_{g}/$, $/{}_{e}g_{+}/$, $/{}_{g}{+}_{m}/$, $/{}_{+}m_{e}/$, $/{}_{m}e_{n}/$, $/{}_{e}n_{t}/$, $/{}_{n}t_{\#}/$. By this analysis, syllable juncture is considered to be more than a conditioning factor on the syllable-terminal and syllable-initial allophones. Syllable juncture is also a context-sensitive allophonic segment in and of itself, conditioned by the terminal phoneme of the preceding syllable and the initial phoneme of the succeeding syllable. Such syllable-boundary segmental units would play an important role in preserving order information across syllable boundaries for speech articulation and recognition. In addition, if there are explicit cues for syllable boundaries, as there must be if the /w/ allophones in *sandwich* and *dwindle* are different, then there exist cues adequate for the acoustic definition of such syllable-juncture segments. In some words, with difficult transitions from the terminal phoneme of one syllable to the initial phoneme of the next syllable, the syllable juncture allophone might be associated with a short pause or transition region in speech. However, there is *no* need for the articulatory output of the syllable boundary segment to occupy any appreciable period of time. Within context-sensitive allophonic theory, it is simply not necessary to have each segment associated with some particular period of time that does not overlap with the period of time occupied by other segments.

Another point about the current theory of representing syllable juncture is that there is no need to assume that a word with *n* vowels has *n* syllables or $n - 1$ syllable junctures. Syllable juncture is not introduced to mark off nonoverlapping domains of different vowels, as necessary units of the speech production or recognition processes. Rather, syllable juncture is introduced to separate consonants that are not be coarticulated. In a word such as *syllabic,* there may be no syllable junctures at all in my dialect, though conceivably there might be one or more syllable junctures in someone else's dialect.

E. Suprasegmental Components

In addition to the traditional segmental components of speech, there is another class of articulatory features that are considered to be important phonetic components of speech. These factors are often grouped together and referred to as suprasegmental or prosodic features of speech. They include such linguistic notions as the intonation (pitch) contour of an utterance, stress, accent, rhythm, etc.

Intonation contours serve as important cues for the meaning of a message. For example, in English, a falling pitch contour at the end of an utterance is typical for statements, and a "not-falling" intonation contour is typical for questions. In addition, certain words in an utterance will be stressed more than other words. For example, articles, prepositions, and other grammatical "function" words are typically the most weakly stressed words in an utterance, and the other words in an utterance differ among themselves in judged degree of stress. In at least some utterances, linguistic intuition will be reasonably consistent in distinguishing three or four levels of what we shall here call "stress," which is the stress placed on one entire word, as opposed to another entire word, in the utterance.

Stress is thought to be primarily realized in conjunction with a particular syllable or vowel of the stressed word, the "accented" syllable, or the "accented" vowel. To reduce the possibility of confusion between word "stress" and segmental "accent," I have followed Bolinger (1958) in giving them different names. If the syllable has no reality, then "accent" should probably be assumed to be a distinctive feature of vowels. People make remarkably consistent judgments of vowel (syllable) accent, so there is little doubt concerning its psychological reality as a feature. For example, in the word "fundamental," the third vowel has the primary accent, while the first vowel has the secondary accent, and the second and fourth vowels are least accented. It is possible that there are only two distinctive levels of vowel accent, accented and unaccented.

Regarding the domain of the accent feature, it should be noted that considering accent to be a feature of vowels automatically conditions the (immediately) preceding and subsequent phonemes adjacent to the vowel. This could somewhat stretch out the articulatory and acoustic realization of the effects of different levels of accent. In addition, the anticipatory (priming) mechanism described earlier would further increase the articulatory and acoustic time domain of accent on an articulatory and acoustical feature level.

When a word is stressed, as for example when it is produced in isolation, there are measurable articulatory and acoustic features that differentiate different levels of vowel accent. However, when a word is relatively unstressed in continuous speech, there are frequently no observable acoustic differences between accented and unaccented vowels in such an unstressed word. Thus, the accent and stress systems are closely linked, with accent reflecting the different potential of the vowels in a word for exhibiting the effects of word stress.

Linguists, especially Chomsky and Halle (1968) have the intuition that every sentence has a normal word stress pattern. Accounting for this normal word stress pattern is part of the task of their theory of phonology. For example, in a sentence such as *Joe ran into a green door*, the word *door* would be judged by me to have primary stress, with *Joe, ran,* and *green* having secondary stress, and the function words *into* and *a* being unstressed. However, under appropriate circumstances, one could place primary stress on any word of the utterance, even the grammatical function words. Chomsky and Halle wish to account for "normal" stress patterns and leave these other stress patterns outside their theory, attributing them to different mechanisms. Beyond the linguistic intuition that there is one "most typical" stress pattern for virtually every sentence, I know of no other reasons to justify assuming different stress mechanisms in the two cases. The stress pattern judged "normal" for a sentence may simply be that which is appropriate for the most probable set of conditions for the utterance (intended meaning and assumed prior knowledge of the hearer). Thus, I do not find the evidence for assuming different mechanisms to be at all compelling. In any event, I am not concerned with accounting for any particular stress patterns for the words in an utterance. I shall take them to be given by the concept level as input for the phonetic level. My only concern is to ask how this word stress pattern is represented at the phonetic level and what its interaction is with the segmental aspects of speech.

We could characterize the stress pattern in an utterance as an ordered set of three or four stress levels. Thus, if 3 indicates the highest stress level, 2 the next highest, and 1 the lowest, we could characterize the stress sequence in *Joe ran to the green door* as (2,2,1,1,2,3). However, for reasons similar to the assumption of context-sensitive allophonic coding for segmental representatives, it seems preferable to assume that stress pattern is not represented at a phonetic level by an ordered set of context-free stress representatives but is, instead, represented by an unordered set of context-sensitive stress representatives. For example, in the above sentence, the context-sensitive stress coding would be $({}_{\#}2_{2})$, $({}_{2}2_{1})$, $({}_{2}1_{1})$, $({}_{1}1_{2})$, $({}_{1}2_{3})$, $({}_{2}3_{\#})$. Such a sequence of context-sensitive stress representatives could be input from the concept level to the phonetic level, either simultaneously or successively. In either case, the correct order of stress representatives would be activated by long-term associations between such context-sensitive stress representatives, even though the temporal ordering of stress representatives was not maintained at all times at the phonetic level. With only a small number of different stress representatives (probably only three or four), it is not clear whether such context-sensitive coding would always unambiguously encode the temporal ordering of stress levels (the intonation contour). However, input from the concept level to the phonetic level could be not simultaneous but successive, with each word and its associated stress level being output from the concept level to the phonetic level, in temporal order. In this case, the short-term associations among stress levels and between them and their associated words would help the stress representatives successfully traverse the ambiguous transitions. At present, the assumption that input to the phonetic level from the concept level is word-by-word successive seems slightly more reasonable than the assumption of parallel input. However, there is no real evidence on the point.

There is a reason for assuming that the ordering of stress representatives is at least partially (if not completely) independent of the ordering that might be induced by their associations with the successive words in an utterance. Evidence for this comes from Boomer and Laver (1968) and Fromkin (1970), who observed a number of speech errors in which subjects transposed words in an utterance but the stress patterns remained the same. That is to say, the sequence of stress representatives remained the same, though the words came out in the wrong order. This resulted in the "wrong" stress being assigned to words that appeared in the wrong positions in the utterance. If stresses do not have some representation of their ordering independent of their association with ordered word representatives, this should not happen. This constitutes some evidence for the validity of the context-sensitive associative-chain theory of serial ordering of stress representatives, as previously described, though of course other theories of the ordering of stress representatives could also account for this fact.

Although it seems desirable to assume that stress representatives have their own encoding of order to account for the above findings, it seems very desirable to assume that each stress representative has an association to its appropriate word (set of allophone representatives at the phonetic level). Although it is possible to assume that the representation of the order of words and the order of stress representatives are two separate systems that operate in parallel, this sort of system would probably lead to some coordination difficulties, unless there were interaction between the two sequences to keep them in proper phase. Assuming associations between stress representatives and sets of allophone representatives could probably be used to achieve this coordination. In addition, such a correlation would help to replicate the ordering of stress representatives and of the sets of allophone representatives for adjacent words, improving the accuracy of transitions from the terminal phoneme of a word to the initial phoneme of the following word.

References

Allen, G. D. Timing control in speech production: Some theoretical and methodological issues. Paper presented at Phonetics Symposium, Univ. of Essex Language Centre, January, 1972.

Anisfeld, M., & Knapp, M. Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology,* 1968, **77,** 171–179.

Bolinger, D. L. On intensity as a qualitative improvement of pitch accent. *Lingua,* 1957–58, **7,** 175–182.

Bolinger, D. W. A theory of pitch accent in English. *Word,* 1958, **14,** 111–149.

Bolinger, D. W., & Gerstman, L. J. Disjuncture as a cue to constructs. *Word,* 1957, **13,** 246–255.

Boomer, D. S., & Laver, J. D. M. Slips of the tongue. *British Journal of Disorders of Communication,* 1968, **3,** 2–12.

Brown, C., & Rubenstein, H. Test of response bias explanation of word-frequency effect. *Science,* 1961, **133,** 280–281.

Bruce, D. J. Effects of content upon intelligibility of heard speech. In C. Cherry (Ed.), *Information theory,* Third London Symposium: 1955. New York: Academic Press, 1956. Chapter 26, pp. 245–252.

Chomsky, N., & Halle, M. *The sound pattern of English.* New York: Harper, 1968.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word,* 1952, **8,** 195–210.

Denes, P. B., & Pinson, E. N. *The speech chain.* Bell Telephone Laboratories: Waverly Press, 1963.

Fant, C. G. M. Descriptive analysis of the acoustic aspects of speech. *Logos,* 1962, **5,** 3–17.

Flanagan, J. L. *Speech analysis, synthesis, and perception.* New York: Springer-Verlag, 1965.

Frederiksen, J. R. Statistical decision model for auditory word recognition. *Psychological Review,* 1971, **78,** 409–419.

Fromkin, V. A. Speculations on performance models. *Journal of Linguistics,* 1968, **4,** 47–68.

Fromkin, V. A. Tips of the slung—or—to err is human. *UCLA Working Papers in Phonetics,* 1970, No. **14,** March, 40–79.

Gleason, H. A. *An introduction to descriptive linguistics.* New York: Holt, 1955.

Grossman, L., & Eagle, M. Synonymity, antonymity, and association in false recognition responses. *Journal of Experimental Psychology,* 1970, **83,** 244–248.

Hall, J. L. Binaural interaction in the accessory superior-olivary nucleus of the cat. *Journal of the Acoustical Society of America,* 1965, **37,** 814–823.

Harris, C. M. A study of the building blocks in speech. *Journal of the Acoustical Society of America,* 1953, **25,** 962–969.

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

Harris, J. D. Pitch discrimination. *Journal of the Acoustical Society of America,* 1952, **24,** 750–755.

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

Hintzman, D. L. Articulatory coding in short-term memory. *Journal of Verbal Learning and Verbal Behavior,* 1967, **6,** 312–316.

Howes, D. On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America,* 1957, **29,** 296–305.

Hubel, D. H., & Wiesel, T. N. Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology,* 1963, **26,** 994–1002.

Huggins, A. W. F. The perception of timing in natural speech I: Compensation within the syllable. *Language and Speech,* 1968, **11,** 1–11. (a)

Huggins, A. W. F. How accurately must a speaker time his articulations? *IEEE Transactions on Audio and Electroacoustics,* 1968, **AU-16,** 112–117. (b)

Hughes, G. W., & Halle, M. Spectral properties of fricative consonants. *Journal of the Acoustical Society of America,* 1956, **28,** 303–310.

Hultzén, L. S., Allen, J. H. D., & Miron, M. S. *Tables of transitional frequencies of English phonemes.* Urbana: Univ. of Illinois Press, 1964.

Jassem, W. The formants of fricative consonants. *Language and Speech,* 1965, **8,** 1–16.

Katz, J. J., & Fodor, J. A. The structure of semantic theory. *Language,* 1963, **39,** 170–210.

Kimble, G. A. Mediating associations. *Journal of Experimental Psychology,* 1968, **76,** 263–266.

Klein, G. A. Temporal changes in acoustic and semantic confusion effects. *Journal of Experimental Psychology,* 1970, **86,** 236–240.

Kozhevnikov, V. A., & Chistovich, L. A. *Speech: Articulation and perception.* Translated by J.P.R.S., Washington, D.C., No. JPRS 30, 543. Moscow-Leningrad, 1965.

Ladefoged, P. Acoustical characteristics of selected English consonants. *Language,* 1965, **41,** 332–338.

Ladefoged, P. The phonetic framework of generative phonology. *UCLA Working Papers in Phonetics,* 1970, No. **14,** March, 25–32.

Ladefoged, P. Phonetics. *UCLA Working Papers in Phonetics,* 1971, No. **20,** March, 1–28.

Lehiste, I. Acoustical characteristics of selected English consonants. *International Journal of American Linguistics,* 1964, **30:**3, Part 4, 34, pp. xi–197. Bloomington: Indiana University.

Lehiste, I. Temporal organization of spoken language. *Working Papers in Linguistics,* 1970, No. 4, Ohio State Univ., 95–114. (a)

Lehiste, I. *Suprasegmentals.* Cambridge, Massachusetts: M.I.T. Press, 1970. (b)

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

Lieberman, P. *Intonation, perception, and language.* Cambridge, Massachusetts: M.I.T. Press, 1967.

Light, L. L., & Carter-Sobell, L. Effects of changed semantic context on recognition memory. *Journal of Verbal and Verbal Behavior,* 1970, **9,** 1–11.

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

MacKay, D. G. Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia,* 1970, **8,** 323–350. (a)

MacKay, D. G. Spoonerisms of children. *Neuropsychologia,* 1970, **8,** 315–322. (b)

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

Mattingly, I. G., & Liberman, A. M. The speech code and the physiology of language. In K. N. Leibovic (Ed.), *Information processing in the nervous system.* Springer Verlag, 1969. Pp. 97–114.

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

Miller, G. A., Heise, G. A., & 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, G. A., & Isard, S. Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 217–228.

Morton, J., & Jassem, W. Acoustic correlates of stress. *Language and Speech,* 1965, **8,** 148–158.

Nooteboom, S. G. Some regularities in phonemic speech errors. *IPO Annual Progress Report II* (Instituut Voor Perceptie Onderzoek), 1967.

Nooteboom, S. G. The tongue slips into patterns. *Nomen, linguistic and phonetic studies.* The Hague, Mouton, 1968.

Perfetti, C. A., & Goodman, D. Semantic constraint on the decoding of ambiguous words. *Journal of Experimental Psychology,* 1970, **86,** 420–427.

Peterson, G. E., & Barney, H. L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America,* 1952, **24,** 15–24.

Peterson, G. E., Wang, W. S.-Y., & Sivertsen, E. Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America,* 1958, **30,** 739–742.

Pike, K. L. *Phonemics: A technique for reducing languages to writing.* Ann Arbor: Univ. of Michigan Press, 1947.

Pollack, I., Rubenstein, H., & Decker, L. Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America,* 1959, **31,** 273–279.

Rosenzweig, M. R., & Postman, L. Intelligibility as a function of frequency of usage. *Journal of Experimental Psychology,* 1957, **54,** 412–422.

Rubenstein, H., & Pollack, I. Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 147–158.

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

Shockey, L., Gregorski, R., & Lehiste, I. Word unit temporal compensation. *Working Papers in Linguistics,* 1971, No. **9,** Ohio State Univ. 145–165.

Stowe, A. N., Harris, W. P., & Hampton, D. B. Signal and context components of word-recognition behavior. *Journal of the Acoustical Society of America,* 1963, **35,** 639–644.

Suga, N. Functional properties of auditory neurones in the cortex of echo-locating bats. *Journal of Physiology,* 1965, **181,** 671–700.

Suga, N. Analysis of frequency-modulated and complex sounds by single auditory neurones of bats. *Journal of Physiology,* 1968, **198,** 51–80.

Tulving, E., & Thomson, D. M. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology,* 1971, **87,** 116–124.

Underwood, B. J. False recognition produced by implicit verbal responses. *Journal of Experimental Psychology,* 1965, **70,** 122–129.

Warren, R. M. Identification times for phonemic components of graded complexity and for spelling of speech. *Perception and Psychophysics,* 1971, **9,** 345–349.

Whitfield, I. C., & Evans, E. F. Responses of auditory cortical neurons to stimuli of changing frequency. *Journal of Neurophysiology,* 1965, **28,** 655–672.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America,* 1965, **38,** 583–588.

Wickelgren, W. A. Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America,* 1966, **39,** 388–398.

Wickelgren, W. A. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review,* 1969, **76,** 1–15. (a)

Wickelgren, W. A. Learned specification of concept neurons. *Bulletin of Mathematical Biophysics,* 1969, **31,** 123–142. (b)

Wickelgren, W. A. Context-sensitive coding in speech recognition, articulation, and development. In K. N. Leivovic (Ed.), *Information processing in the nervous system.* New York: Springer-Verlag, 1969. Pp. 85–95. (c)

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