Speech and the Problem of Perceptual Constancy*

Donald Shankweiler,[+] Winifred Strange,[++] and Robert Verbrugge[+++]

ABSTRACT

Speech signals are intrinsically variable for many reasons. In this paper we consider the implications of variability for a theory of vowel perception. Current theories of the vowel emphasize the relational nature of the acoustic cues since no absolute values of formant frequencies could unambiguously distinguish vowels produced by different talkers and in different phonetic contexts. It has been assumed that the perceptual process of vowel identification includes a normalization stage whereby the listener calibrates his perceptual apparatus for each talker, according to some reference derived from preceding utterances by that talker. We have been unable to obtain evidence for such a perceptual mechanism. Theories of vowel perception have failed to give due weight to the richness of the natural speech signal. We attempt to show why the invariant acoustic information that specifies a vowel cannot be found in a temporal cross section, but can only be specified over time.

*Chapter prepared for Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc., in press).

[+] Also University of Connecticut, Storrs.

[++] University of Minnesota, Minneapolis.

[+++] Human Performance Center, University of Michigan, Ann Arbor.

Acknowledgment: The authors' research reported in this paper was begun during the academic year 1972-73, while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The early portion of the work was supported in part by a grant from the National Institute of Child Health and Human Development (NICHD) to the Center, and in part by grants awarded to Shankweiler and to J. J. Jenkins by the National Institute of Mental Health. The later portion of the research and the preparation of this paper was supported by grants from NICHD to the Center and to Haskins Laboratories.

We are grateful to T. Edman for his substantial assistance with all phases of the experimental work, and to J. J. Jenkins for advice and encouragement from the project's inception. Other colleagues who commented on earlier drafts of this paper are: G. M. Kuhn, T. Nearey, and M. Studdert-Kennedy. It is a pleasure to acknowledge their help.

Speech perception is not ordinarily included in the body of phenomena and theory that convention defines as the psychology of perception. Yet the problem of how the perceptual categories of speech are specified in the acoustic signal is a primary example of the problem of perceptual constancy. In spite of its neglect by psychology, speech and its signal have been intensively studied by members of other disciplines. We think that some of the results and puzzles generated by this research are relevant to the concerns of our colleagues whose primary interests are in other facets of human cognition.

Speech perception at any level involves classification. The classificatory step is assumed whenever we move beyond a purely physical (acoustic) description of speech to a psychophysical description in terms of perceptual units. Unlike certain problems in traditional psychophysics in which the choice of units may be arbitrary, there is wide consensus about what the units of perception are in the case of speech. This consensus is the product of centuries of linguistic investigation, during which many attempts have been made to isolate the various levels and units that constitute our perception of speech. Viewed in terms of structure, speech is a hierarchical system that manifests what Hockett (1958) called a "duality of patterning": it employs both meaningful and meaningless units. Morphemes (or, roughly speaking, words) are the smallest of the meaningful units. In all languages morphemes have an internal structure composed of smaller meaningless segments, the phonemes (Bloomfield, 1933). Since the communication of meanings ultimately rests on a foundation of phonemic structure, a basic part of the task of understanding how speech is perceived is to discover the conditions for the perception of phonemic categories.[1] For present purposes, we shall ignore meaning and concern ourselves only with the phonemic message--that is, with the perception of syllables and their phoneme segments, familiar to us as the consonants and vowels.

In speech, as in handwriting, no two "signatures" are alike. In generating the "same" phoneme, different speakers do not produce sounds that are acoustically the same. Indeed, the same signal is never exactly repeated by the same speaker. In perceiving speech, as in identifying objects, we ordinarily regard only those distinctions that are critical, ignoring those that are merely incidental. While no one would deny that speech signals are intrinsically variable for many reasons, the implications of this variability for perception have not been widely appreciated. In brief, they constitute a major problem in <u>perceptual constancy</u>.

[1] The phoneme is the minimal unit by which perceivers differentiate utterances. For example, the word <u>bad</u> has three phonemic segments, /b/, /æ/, and /d/, that differentiate it from such words as <u>dad</u>, <u>bed</u>, and <u>bat</u>. In different utterances, a phoneme may be realized acoustically in different ways; linguists call these variants "phones." For example, the final /t/ in <u>bat</u> might be either released (acoustically, a pause followed by a burst) or unreleased (no burst). The class of phones is potentially infinite, and it is arguable whether phones (however defined) are natural perceptual units. Our emphasis in this paper is on how the identity of a phoneme is perceived despite variations in its acoustic form.

THE PERCEIVING MACHINE AND THE ONE-TO-ONE PROBLEM

Although our concern is with how the human perceptual system works, it may help us to bring this problem into focus if we consider how a machine might proceed in recovering a string of phonemic segments. Consider, for example, the problems to be solved in designing a voice-operated typewriter. The goal of such an automatic speech recognition device is to type out the appropriate string of phonemic symbols (or perhaps standard orthographic symbols) in response to any speech input. In the simplest case, the only information available to the device will be the acoustic waveform itself. A human listener, of course, can usually take advantage of other sources of information, including both the linguistic and the situational context of the utterance.[2] While acknowledging the importance of context, we should not overlook the fact that listeners can identify arbitrarily chosen words and nonsense syllables with high accuracy when listening conditions are favorable. In other words, we are not posing an unrealistic problem for our hypothetical device; human listeners can do remarkably well when little contextual information is available.

Many attempts have been made in recent decades to design a voice-operated typewriter, but the problem has so far proved elusive. Despite a degree of success with severely restricted vocabularies when words are spoken by a trained talker, a generally useful speech recognizer continues to be unattainable. As Hyde (1972:399) notes, "there are still no devices which can perform even moderately well on normal (conversational) speech in normal (noisy) environments by a normal range of talkers."

It is worth considering for a moment how the operation of a speech recognition system has typically been conceived. As in many other automatic pattern recognition devices, the procedure involves two stages. In the first stage, the basic units or segments are located. For example, in automatic reading of print, this would correspond to isolating individual letters. In the second stage, each segment is identified as an instance of one of a fixed set of objects. In the case of print reading, this would correspond to identifying a segment as a particular letter of the alphabet. Thus a successful voice-operated typewriter would have to be able to perform two operations on the acoustic waveform of any speech signal. First, it would have to divide the waveform into acoustic segments that have a one-to-one correspondence through time with the sequence of phonemes in the utterance. Second, it would have to detect the presence or absence of acoustic features that are critical for identifying particular phonemes. This second stage is often conceived of as requiring a set of filters, each filter being tuned to a critical acoustic property (defined along dimensions such as frequency, intensity, and duration).

This strategy implies certain widespread assumptions that only in recent years have been successfully challenged. For example, we find that in the

[2] The importance of context to human listeners has been elegantly demonstrated by Miller, Heise, and Lichten (1951), who found a remarkably predictable relationship between the amount of acoustic distortion that yields a given level of intelligibility and the informational redundancy of the message. For a given signal-to-noise ratio, intelligibility was greater for words heard in sentence context than for words heard in isolation.

standard accounts of speech acquisition, it is tacitly assumed that speech consists of a collection of elementary sounds "transparent" to the infant, such that he automatically recognizes a parent's utterance of /d/ as "the same sound" as his own utterance of /d/ (Allport, 1924; Watson, 1924). Similarly, taxonomic linguists working in the tradition of Bloomfield (1933) supposed that all languages sampled from a common inventory of sounds (phones). Working from phonetic transcription of a large number of utterances as a base, these linguists developed highly successful procedures for determining which sound contrasts played a role in any particular language. It was believed that the great practical success of transcription as a tool for language description rested on a narrow physical base, and that, in principle, an acoustic definition could be given for each phone. In this view, speech was conceived as a kind of sound alphabet, in which each phone is conveyed by a discrete package of sound with a characteristic spectral composition. The pervasiveness of this assumption has been noted by Denes (1963:892):

> The basic premise of [most speech-recognition] work has always been that a one-to-one relationship existed between the acoustic event and the phoneme. Although it was recognized that the sound waves associated with the same phoneme would change according to circumstances, there was a deep-seated belief that if only the right way of examining the acoustic signal was found, then the much sought-after one-to-one relationship would come to light.

In fact, the perceptual skills that underlie phonetic transcription have never been explained well enough that an algorithm could be written to permit a machine to do the job. From our present perspective it is clear why no one has been able to develop a voice-operated typewriter based on the startegy outlined above. First, there are no clearly bounded segments in the acoustic waveform of roughly phonemic size; that is, there are no acoustic units available for setting up a correspondence with phonemes. Second, even if boundaries are arbitrarily imposed on the continuous signal, the segments corresponding to a particular phoneme often vary considerably in their acoustic composition. Moreover, any one of those acoustic segments, transferred to a different phonemic environment, might be heard as a different phoneme altogether. Not only do the physical attributes specifying a particular phoneme vary markedly, but the same physical attribute can specify different phonemes depending on the context.

THE CONTINUOUS SIGNAL DOES NOT REVEAL THE SEGMENTATION OF THE PHONEMIC MESSAGE

The conclusions stated above are the results of three lines of investigation begun in the mid 1940s and continuing to the present. We turn now to review briefly the nature of the evidence.

Of special importance from our standpoint are a series of tape-cutting and tape-splicing experiments that had the effect of shaking the general confidence that phonemes are conveyed by isolable bits of sound. These experiments failed to find any way to divide the signal on the time axis to yield segments of phoneme size. For example, given a consonant-vowel syllable such as go, there is no way to cut the piece of magnetic tape so as to produce the consonant /g/ alone. Some vowel quality always remains. Moreover, if a consonant-vowel syllable is cut at some point, the consonantal portion may not be heard as the same phoneme when spliced to a recording of a different vowel. Schatz (1954), for example, found that the consonantal portion of an utterance of /pi/ was heard as /k/ when it was joined to the vowel /ɑ/. Harris (1953) and Peterson, Wang, and Sivertsen (1958) independently concluded that assembled speech made by splicing together prerecorded segments is not generally intelligible when the units are smaller than roughly a half-syllable.[3] Some investigators (e.g., Cole and Scott, 1974a, 1974b) continue to argue that speech perception may be based in large part on the detection of acoustic invariants for phonemes. They have claimed that a one-to-one correspondence _can_ often be found, that spliced segments _can_ preserve their identity when transferred to new phonemic contexts. However, much of this apparent "invariance" disappears when one is careful to cut the initial segment sufficiently short that no trace of the subsequent vowel remains (cf. Kuhl, 1974).

A second important development is the study of spectrographic displays of speech. This work was made possible by the invention of the sound spectrograph (Koenig, Dunn, and Lacey, 1946) during World War II and by the general availability of such devices for research during the postwar years. The spectrograph displays, in graphical form, the time variations of the spectrum of the speech wave. This representation of the sound patterns of speech is valuable for the information it gives about articulation. The energy in speech sounds is concentrated in a small number of frequency regions that appear on a spectrogram as horizontal bands (called "formants"). The location of the formants on the frequency scale reflects the primary resonances of a talker's vocal tract (Fant, 1960). Since the shape of the vocal cavity changes at the joining of successive consonants and vowels, the formant frequencies may be seen to modulate up and down as one scans a spectrogram along the time axis. However, efforts to locate discrete information-bearing units along the time axis have met with repeated failure. The phonemic and syllabic segments, which are so clear perceptually, have no obvious correlates in a spectrogram, as evidenced by the fact that spectrograms are very difficult to read, even after much experience (Fant, 1962; but see also Kuhn and McGuire, 1975).[4]

Figure 1 shows spectrograms of two syllables, _bib_ and _bub_. Note that the formant frequencies are nonoverlapping for the entire duration of the syllables, not just in the middle portion. Although the syllables differ phonemically (i.e., perceptually) in only one segment (the medial vowel), acoustically they differ throughout.

Failure to find obvious acoustic cues in spectrograms led to a third line of investigation: a variety of experiments with synthetic speech, produced by devices that place acoustic parameters under the experimenter's direct control. Early work by researchers at Haskins Laboratories made use of hand-painted

[3] It does not follow from this result that the minimal perceptual unit is larger than the phoneme. The inference we would draw is that decisions about phoneme identity are made with regard to information distributed over the whole syllable (and sometimes, perhaps, over a number of syllables).

[4] In remarking on the difficulty of reading spectrograms our point is not that the spectrogram does not represent the relevant phonemic information, but rather that the ear has readier access to the brain's phonemic decoder than the eye.

patterns resembling spectrograms that were converted into sound by a photoelectric device, the Pattern Playback (Cooper, Liberman, and Borst, 1951). The Pattern Playback and subsequent computer-controlled electronic synthesizers have made it possible to do analytic studies in which one parameter is varied at a time, to determine which parameters were critical for particular phonemes. Only through systematic psychophysical experimentation of this sort has it been possible to locate the linguistically relevant information in the speech spectrum (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952; Liberman, 1957; Liberman, Harris, Hoffman, and Griffith, 1957; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). A major conclusion of this research is that, in general, there is no simple one-to-one correspondence between perceptual units and the acoustic structure of the signal. To be successful, synthetic speech must encode the information for phonemes into acoustic patterns at least a half-syllable or full-syllable in length.

These findings make it possible to understand why the design of a voice-operated typewriter proved so difficult. Phonemes are not merely joined acoustically; they overlap so that two or more are represented simultaneously on the same stretch of sound. Conversely, segmentation is impossible because information for one phoneme is usually spread over wide stretches of the signal. Even if segmentation were attempted, the cues isolated would be radically different in another phonemic environment. As we saw in Figure 1, all four segments corresponding to /b/ would differ markedly in slope and formant frequency range.

In sum, the radically context-dependent structure of speech dooms to failure the kind of pattern recognition procedure outlined above. A procedure that combines a prior stage of segmentation with an analysis of segments by a set of tuned-filter "detectors," operating independently and in parallel, will be insufficient as a job description of an automatic recognition machine (and insufficient as a model of human speech perception as well). One contemporary approach to the recognition problem (Mermelstein, 1974) is explicit on this point, acknowledging that a speech recognizer would have to extract information about component phonemes over longer stretches of the signal than a syllable and would have to incorporate rules about how that information is distributed. The system described by Mermelstein does <u>not</u> assume that the segmentation and labeling problems are independent.

SYLLABLE NUCLEI AS TARGETS

A commonly suggested strategy for speech recognition (Fant, 1970) is to classify first those phonemes that are most "transparent" in the signal, and then to use that information as a basis for determining what the more contextually variable phonemes are. For example, the research described earlier found that the acoustic form of many consonants is heavily dependent on the coarticulated vowel. If the vowel were easily detected in speech signals, then it could be identified first and used as a basis for disambiguating the neighboring consonant.

There are several reasons for supposing that vowels could be extracted readily by a routine based on a filter bank, though it would not be possible, in general, for consonants. In productions of sustained vowels, the positions of the formants on the frequency axis serve roughly to distinguish the vowels of a particular talker. For a high front vowel such as /i/, the first and second

FIGURE 1

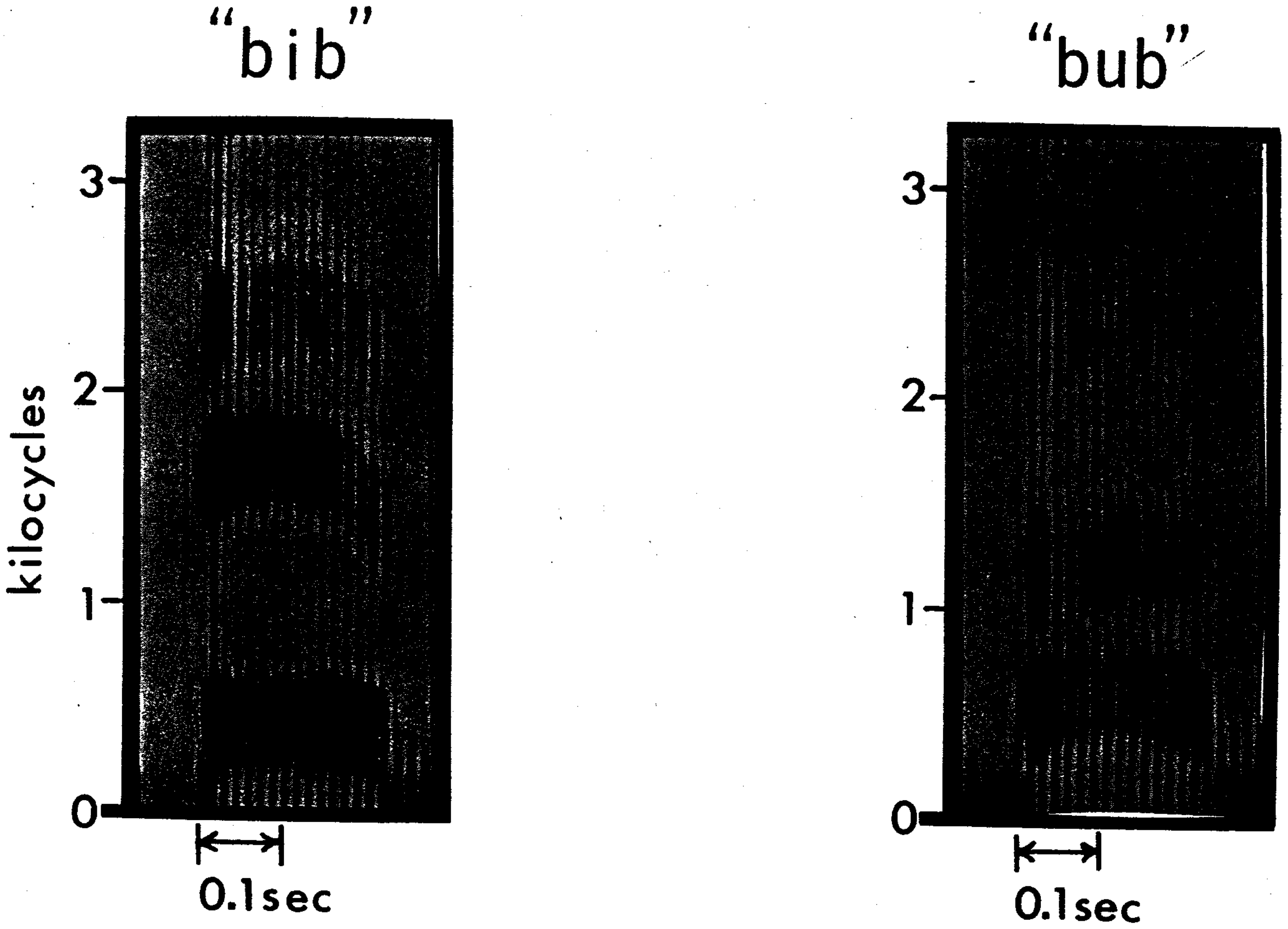

Figure 1: Spectrograms of tokens of two syllables that differ in the vowel. Note that the formant pattern of "bib" does not coincide with that of "bub" in any portion.

formants are spaced far apart (on the frequency axis) while the second and third formants are spaced close together. For a high back vowel /u/, the pattern is just reversed. These patterns can be synthesized with a combination of steady-state resonances that simulate the formants found in spectrograms of natural vowels. Synthetic stimuli are readily labeled as vowels by listeners, and it is possible to generate the full complement of English vowels with two or three formants (Delattre, Liberman, Cooper, and Gerstman, 1952).

This acoustic characterization of vowels as steady-state entities is reinforced by articulatory considerations. The flow of speech is marked by a rhythmic pattern of syllables; each syllable contains a vowel "nucleus" that is usually coarticulated with one or more consonants. It is usual to think of consonants as the dynamic component of speech, since they are generally produced by movement of the articulators, and to regard vowels as the static component, since they may be produced with a stationary vocal-tract configuration and sustained indefinitely.

This contrast is emphasized by the concept of an idealized vowel as a prolonged, static entity defined (acoustically) by the frequencies of the first two or three formants, i.e., by the primary resonances of the stationary vocal tract. At least for individual talkers, then, there should be a distinctive set of frequency values associated with each vowel, in contrast to the variable values associated with the talker's consonants. In that case, vowels should be retrievable by a simple two-stage recognition procedure: it would be a straightforward matter to detect the presence of steady states electronically (thereby isolating vowel segments for analysis), and a filter bank could then determine which set of critical frequencies the vowel sound best fits.

Unfortunately for this approach, the apparent simplicity of vowels is largely an illusion. Vowels in natural continuous speech, unlike the artificially prolonged vowels of the phonetics laboratory, are not generally specified by steady states at all. Let us see why this is so.

First, as a result of coarticulation, vowels are encoded into the structure of a full syllable. The imprint of the vowel is not localized but is smeared throughout the entire temporal course of the syllable. Thus, information about a vowel is available in the transitions as well as in the steady-state portion (if, indeed, a steady state is even attained). This was clear in the earlier example of the syllables bib and bub; in both cases, the vowel affected the spectral pattern of the entire syllable. Moreover, the acoustic properties of the vowel nucleus may be affected by coarticulated consonants. Measurements by House and Fairbanks (1953) and Stevens and House (1963), for example, indicated consistent changes in the duration, fundamental frequency, and formant frequencies and intensities of vowels, depending on consonantal context. It should also be noted that consonants can affect the structure of neighboring vowels if (by the phonological rules of the dialect) a distinction between two consonants is actually manifested by a difference in the neighboring vowels. For example, the /d/ in rider is distinguished from the /t/ in writer by the increased duration of the vowel that precedes it. Similarly, in some dialects of English, a nasal phoneme, such as the /n/ in pants, is realized by a nasalization of the preceding vowel; thus the spectral structure of the vowel /æ/ will vary markedly depending on whether pats or pants is spoken. Because the coarticulation effects between consonants and vowels do not operate in one direction alone, but are two-way effects, there is no obvious acoustic invariant that characterizes a vowel in all consonantal contexts.

A second major source of variance in the acoustic structure of vowels is the tempo of articulation. During rapid rates of speech, steady-state configurations may never be attained at all. Acoustic analysis of rapid speech supports the hypothesis of articulatory "undershoot," since syllable nuclei often do not reach the steady-state formant frequency values characteristic of vowels in slowly articulated syllables (Lindblom, 1963; Stevens and House, 1963). Lindblom and Studdert-Kennedy (1967) found that listeners showed a shift in the acoustic criteria that they adopted for vowels (i.e., there was a shift in the phoneme boundary between them) as a function of perceived rate of utterance. Apparently, human listeners compensated for this simulated articulatory undershoot by perceptual overshoot. These data show that formant transitions, which are generally understood to carry consonantal information, may also aid in specifying the vowel. Thus, in ordinary speech, vowels, like consonants, are dynamic entities that are scaled by the pace of speaking.

A third source of acoustic variation in vowels is associated with the individual characteristics of the talker. We perceive this variation directly when we identify _persons_ on the basis of voice quality. On the other hand, such individual variation is irrelevant and becomes "noise" when our intent is to recover the linguistic message. Inasmuch as formant frequencies reflect vocal-tract dimensions, it is obvious that the absolute positions of the formants will not be the same for a child as they are for an adult. The extent of the problem is suggested by Joos (1948:64).

> The acoustic discrepancies which an adult has to adjust for when listening to a child speaker are nothing short of enormous--they commonly are as much as seven semitones or a frequency ratio of 3 to 2, about the distance from /ɛ/ to /ʊ/.

Somehow, in spite of this, we manage to understand small children's speech reasonably well and they ours. This is especially remarkable given that the difference between children and adults cannot be described by a simple scale factor. The vocal tract not only increases in size but changes in shape, and the consequent changes in the acoustic output are correspondingly complicated. Indeed, Fant (1966) has argued that the assumption of an invariant relation between formants one and two for a given vowel is just as untenable as the assumption that the _absolute_ formant frequencies of the vowel are invariant for all talkers. Thus, the relation between utterances of a syllable by an adult and a small child is not the multiplicative relation that obtains between versions of a melody played in different keys.

Having discussed variation based on physical differences in the sound-production apparatus, we should also mention differences that are social in origin, reflecting local variations of dialect within the larger language community. These variations are associated, of course, with geographical region, ethnic group, and socioeconomic class. Additional sources of talker-related variation are idiosyncratic speech mannerisms, emotional state, and fatigue. These sources, in addition to those we have discussed above, pose enormous difficulties to the design of an automatic recognition device.[5]

No one, to our knowledge, has seriously considered how an automatic speech recognition routine would adjust its criteria to compensate for the variations associated with coarticulation, tempo, and talker. When we consider the magnitude and variety of variations that we take in our stride as perceivers, we begin to realize something of the complexity of the relations between the signal and the phonemic message. The difficulties encountered by the task of machine recognition command a new respect for the subtlety and versatility of the human perceptual apparatus and lead us to a new appreciation of the abstract nature of speech perception.

WHAT SPECIFIES A VOWEL?

The idea that vowels can be defined as fixed sets of steady-state values is an oversimplification that bears little relation to the structure of natural speech. We have found it necessary to reopen the question of what specifies a vowel, and we wish to introduce some recent findings as a case study in the problem of perceptual constancy.

Vowels, as we noted earlier, are traditionally defined by formants. Vowel quality is associated with concentrations of acoustic energy in a few relatively narrow portions of the frequency spectrum; energy in the regions between these bands is generally weak and has little perceptible effect on vowel quality. In distinguishing among vowels, the lowest two formants are traditionally thought to be the most significant; the contribution to perception of the third and higher formants is problematical. For this reason, vowels are customarily represented as points located in a two-dimensional space defined by the first and second formants. As a result of variations among talkers, the points in this acoustic vowel space are actually regions. A critical question for perceptual theory is: How much or how little do these regions overlap?

A thorough assessment of this question was made by Peterson and his colleagues (Peterson, 1951; Peterson and Barney, 1952), who obtained spectrographic measurements of tokens of 10 American English vowels produced by 76 talkers (including men, women, and children). Figure 2, which is redrawn from Peterson and Barney (1952), shows the vowel space defined by measurements of formants one and two (F_1 and F_2). We note that there is considerable overlap in some regions. In running speech we might expect a comparable analysis to show still more overlap. The findings showed not only lack of invariance in the position of the formants in children and adults, but also considerable average differences between men and women and considerable variation among talkers of the same age group and sex.

In his pioneering monograph on acoustic phonetics, Joos (1948) had discussed the dilemma that such variation poses for theories of speech perception. If

[5] Reflect, too, on the variety of transformations of the signal that might be produced by the commonplace feats of talking with food in the mouth, with a cigar between the lips, or with teeth firmly clamped on a pencil (cf. Nooteboom and Slis, 1970).

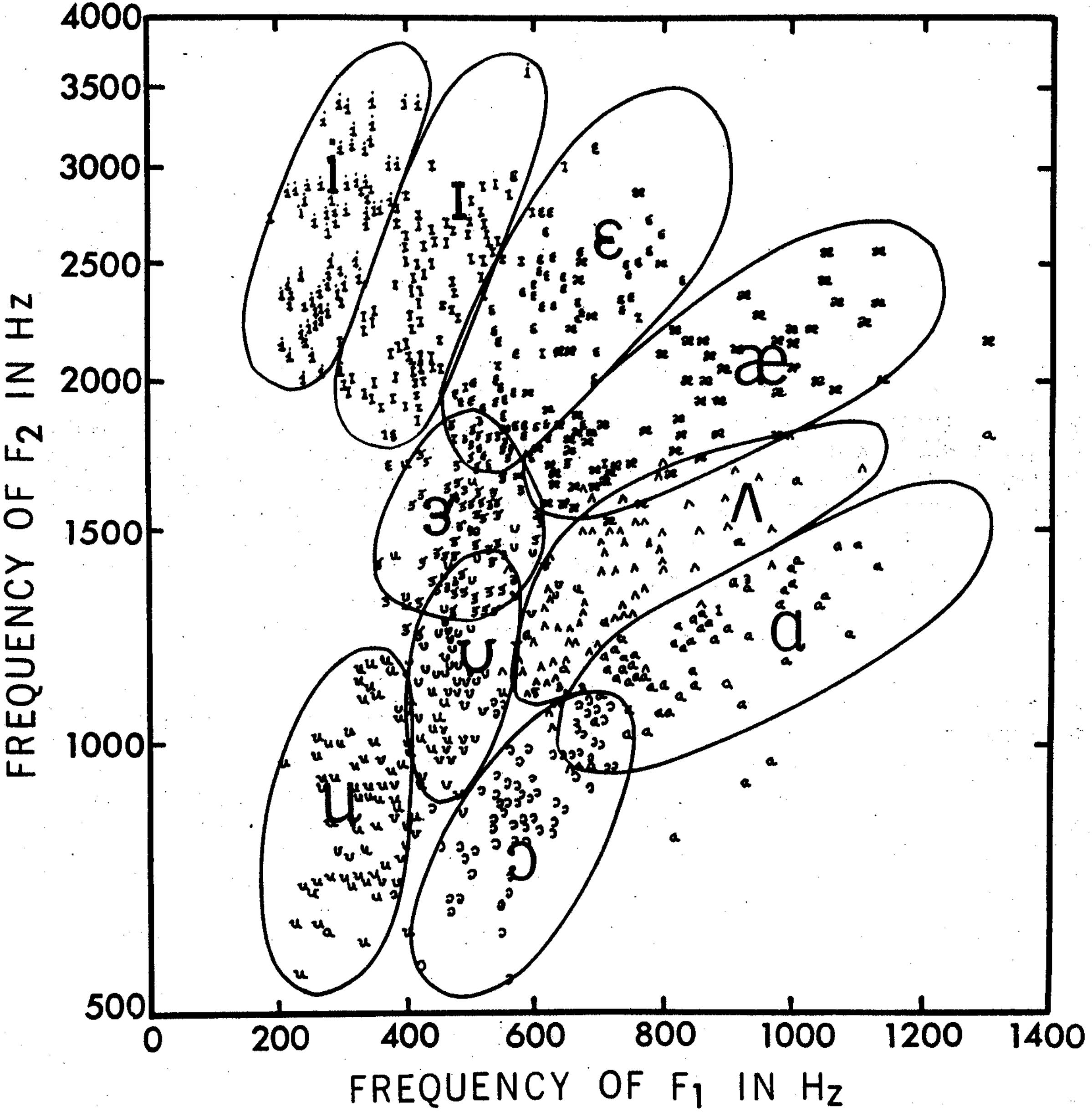

Figure 2: First- and second-formant frequencies of American-English vowels for a sample of 76 adult men, adult women, and children. The closed loops enclose 90 percent of the data points for each vowel category (redrawn from Peterson and Barney, 1952).

different spectra are heard by listeners as the same vowel, on what does the judgment of sameness depend? It cannot, he concludes, be due to any evidence in the sound:

> Therefore the identification is based on outside evidence....If this (outside evidence) were merely the memory of what the same phoneme sounded like a little earlier in the conversation, the task of interpreting rapid speech would presumably be vastly more difficult than it is. What seems to happen, rather, is this. On first meeting a person, the listener hears a few vowel phones and on the basis of this small but apparently sufficient evidence he swiftly constructs a fairly complete vowel pattern to serve as background (coordinate system) upon which he correctly locates new phones as fast as he hears them....(p. 61).

Thus, in Joos's view, the listener calibrates the talker's vowel space on the basis of a small subset of sample utterances. The listener needs some reference points to define the range and distribution of the talker's vowels. These reference signals, Joos suggests, could be supplied by extreme articulations (in terms of tongue height and point of tongue contact). Thus, Joos (1948) leaves no doubt that the coordinate system he has in mind is based at least in part on a model of the vocal tract:

> The process of correctly identifying heard vowel colors doubtless in some way involves articulation. A person who is listening to the sounds of a language that he can speak is not restricted to merely acoustic evidence for the interpretation of what he hears, but can and probably does profit from everything he "knows," including of course his own way of articulating the same phones.

Since the publication of Joos's work, a normalization step in speech perception has been assumed by virtually everyone who has written on the subject, whether or not the writer accepted Joos's version of the motor theory of speech perception (or, indeed, any other version of motor theory). The idea that a listener makes use of reference vowels for calibration of a talker's vowel space has also persisted. Gerstman (1968) and Lieberman (1973)--whose contributions we shall discuss presently--each have taken up the reference vowel idea and defended it. What is surprising, given this overwhelming consensus, is how few attempts have been made to measure the ambiguity in perception of vowels that is directly attributable to talker variation.

Joos's (1948) statement of the constancy problem implies that, for an unknown talker, isolated syllables should be highly ambiguous from the standpoint of the perceiver. It is perplexing, then, that two experiments that directly measured the perceptual ambiguity of natural speech found little support for this prediction. Peterson and Barney (1952), to whom we are indebted for systematic acoustic measurements of individual differences in vowel formants, also attempted to assess the perceptual consequences of the variation they discovered in production. The same recorded utterances used in making the spectrographic measurements were also assembled into listening tests. Listeners had to identify tokens of 10 vowels in /h-d/ consonantal environment; the set consisted of *heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who'd*, *hud*, and *heard*. Syllables produced by groups of 10 talkers (men, women, and children) were randomly mixed on each listening test, ensuring that opportunities for normalization would be slight. Although plots of the first two formants show that the regions occupied by these vowels overlap considerably, perceptual judgments were remarkably accurate, with 94 percent of the words perceived correctly. A similarly low error rate for perception of a larger set of vowels, which included diphthongs, was reported by Abramson and Cooper (1959).

These perceptual data do not support the notion that a single syllable, spoken in isolation, is necessarily ambiguous. Apparently, the information contained within a single syllable is usually sufficient to allow whatever adjustment for individual talker characteristics might be required. The history of the research following Peterson and Barney's (1952) study shows that this conclusion was not generally drawn.

The work of Ladefoged and Broadbent (1957) and Ladefoged (1967) is widely cited as evidence that the listener has relative criteria for vowel identification and that the identity of a vowel depends on the relationship between the formant frequencies for that vowel and the formant frequencies of other vowels produced by the same talker. As a result, vowel space must presumably be rescaled for each voice a listener encounters. The Ladefoged and Broadbent (1957) study was designed to find out whether subjects could be influenced in their identifications of a test word by variations in an introductory sentence preceding it. Synthetic speech was used in order to gain precise control over the acoustic parameters. A set of test syllables of the form /b-vowel-t/ was prepared on the synthesizer. Listening tests were made up in which the test words were presented following a standard sentence: _Please say what this word is_. Variants of this sentence were produced by shifting the frequencies of the first or second formants up or down. Each was intelligible despite wide acoustic differences. The results showed that the same test word was identified as _bit_ when preceded by one version of the test sentence (i.e., one "voice") and as _bet_ when preceded by a second version (another "voice") in which the first formant varied over a lower range. The authors conclude from perceptual shifts such as this that the identification of a vowel is determined by its relation to a prior sample of the talker's speech (provided here by the test sentence).

Ladefoged (1967) interprets these findings within the framework of Helson's (1948) adaptation level theory. This theory attempts to account for the extraordinary efficiency of the compensatory mechanisms that achieve constancies, such as color constancy under changing illumination, by supposing that the perceiver scales his responses not to the absolute properties of each stimulus, but according to the weighted mean of a set of stimuli distributed over time. The introductory sentence, in the Ladefoged and Broadbent experiment (1957), is understood as providing the standard or anchor, thus creating an internal adaptation level to which the test words are referred. We shall return to the adaptation level hypothesis presently, after we have introduced some relevant findings of our own.

If it is true that a listener needs a sample of speech in order to fix the coordinates of a talker's vowel system, we need to know how large a sample is required and whether particular vowels are more effective than others as "anchors." As we noted earlier, Joos (1948) believed that the best reference signal would be one that allows the listener to determine the major dimensions of the talker's vocal tract. He therefore suggested that the "point vowels"

/i/, /ɑ/, and /u/ might be the primary calibrators of vowel space, since they are the vowels associated with the extremes of articulation. Lieberman and colleagues (Lieberman, Crelin, and Klatt, 1972; Lieberman, 1973) agree that these vowels probably play an important role in disambiguating syllables produced by a novel talker. They note that the point vowels are exceptional in several ways: they represent extremes in acoustic and articulatory vowel space, they are acoustically stable for small changes in articulation (Stevens, 1972), and they are the only vowels in which an acoustic pattern can be related to a unique vocal-tract area function (Lindblom and Sundberg, 1969; Stevens, 1972).

Gerstman (1968) has made one of the few direct attempts to test the idea that a subset of vowels can serve to calibrate a talker's vowel space and reduce errors in recognition of subsequently occurring vowels. Gerstman developed a computer algorithm that correctly classified an average of 97 percent of the vowels in the syllables produced by the Peterson and Barney panel of 76 talkers. For each talker's set of 10 utterances, the program rescaled the first- and second-formant frequency values of each medial vowel, taking the extreme values in the set as the endpoints. Since these extreme formant values are typically associated with /i/, /ɑ/, and /u/, the procedure corresponds to a normalization of each talker's vowel space with reference to his own utterances of the point vowels. The classification system was essentially a filter bank that classified the vowels according to the _scaled_ values for the first two formants, and the sums and differences of these values. By inserting the normalization stage between segmentation and classification, Gerstman's program succeeded in reducing by half the errors in classification made by human perceivers [recall that Peterson and Barney's (1952) listeners made 6 percent errors]. We must keep in mind however, that a successful algorithm is not a perceptual strategy, but only a possible strategy. Although it is of interest that such an algorithm can, in principle, serve as the basis for categorization of the signals, we are aware of no evidence that the human perceptual apparatus functions analogously. For example, it would be necessary to demonstrate that humans scale individual formants and calculate sums and differences between them.

It seemed to us that speculation had far outstripped the data bearing on the vowel constancy problem. In fact, the few studies of perceivers' recognition of _natural_ (as distinguished from _synthetic_) speech indicated that isolated syllables spoken by novel talkers are remarkably intelligible. Therefore, it seemed important to verify Ladefoged and Broadbent's (1957) demonstration with natural speech in which all the potential sources of information that are ordinarily available to perceivers are present in the signal. Similarly, Gerstman's (1968) success in machine recognition using the three point vowels as calibrating signals needed to be evaluated against the performance of human listeners. Accordingly, we designed experiments to determine the size of the perceptual problem posed for the listener when the speaker is unknown. This involved a comparison of perceptual errors, under matched conditions, when the test words were spoken by many talkers and when all were produced by only one talker. We also sought to discover whether certain vowels (e.g., the point vowels) have a special role in specifying the coordinates of a given talker's vowel space.

HOW DOES A LISTENER MAP A TALKER'S VOWEL SPACE? AN EXPERIMENT TO DETERMINE THE SIZE OF THE PROBLEM

We (Verbrugge, Strange, and Shankweiler, 1974) first attempted to measure the degree of ambiguity in vowel perception attributable to lack of congruence of the vowel spaces of different talkers. To measure this, we presented listeners with unrelated words or nonsense syllables, so that broad linguistic context would make no contribution to the act of identification. Our studies of this problem were fashioned after Peterson and Barney (1952). We presented nine vowels in the consonantal environment /p-p/; thus, the set consisted of /pip/, /pɪp/, /pɛp/, /pæp/, /pɑp/, /pɔp/, /pʌp/, /pʊp/, and /pup/. In one listening test, the listeners heard tokens spoken by 15 different voices (5 men, 5 women, 5 children), arranged in random order. To determine what proportion of perceptual error is due to uncertainty about the talker, as opposed to other sources, a second listening test was employed in which a single talker uttered all the tokens on a given test. The talkers were given no special training. They were urged to recite the syllables briskly in order to bring about some undershoot of steady-state targets. Our objective was to achieve conditions as similar as possible to normal conversational speech.

Listeners misidentified[6] an average of 17 percent of the vowels when spoken by the panel of 15 talkers, and 9 percent when each of three tests was spoken by a single individual (a representative man, woman, and child from their respective groups). The difference between these two averages, 8 percent, is a measure of the error attributable to talker variation. Although this is a statistically significant difference [$\underline{t}$(50 df) = 5.14, $\underline{p}$ < .01], its absolute magnitude is surprisingly small. Less than half the total number of errors obtained in the variable-talker condition can be attributed to talker variation.

Listeners can identify vowels in consonant-vowel-consonant syllables with considerable accuracy even when they are spoken by an assortment of talkers deliberately chosen for vocal diversity. The intended vowel was identified on 83 percent of the tokens in a test designed to maximize ambiguity contributed by vocal-tract variation.[7] In a second study, listeners were asked to identify 15 vowels (monophthongs and diphthongs) spoken in /h-d/ context by 30 talkers. Here, the rate of identification errors was 13 percent overall and 17 percent for the nine monophthongs alone. The results of both studies are in essential agreement with earlier perceptual data reported by Peterson and Barney (1952)

[6] We have taken the intended vowel of the talker as the criterion of correct identification. That is, we have defined an identification error as a response by the listener that does not correspond to the phonemic category intended by the talker. It might be the case that errors so defined are as much due to mispronunciation as to misperception. No correction was given to talkers during recording other than to clarify orthographic confusions in a few instances. In the case of the youngest children, some coaching was required before they pronounced the nonsense syllables. However, no adult models were provided immediately prior to utterances that were included in the tests.

[7] Each of the 15 talkers spoke only three tokens containing different vowels. These tokens were separated in the test by no fewer than eight intervening tokens spoken by different talkers. Listeners were unable to judge how many talkers were included on a test.

and Abramson and Cooper (1959), although the error rates obtained by these investigators were even lower than those we obtained.[8]

These findings do not bear out the common assumption of a critical need for extended prior exposure to a talker's speech. The information contained within a single syllable appears to be sufficient in most cases to permit recognition of the intended vowel; familiarity with a voice seems to play a rather small role in the identification process.

ARE THE POINT VOWELS USED BY LISTENERS AS AIDS TO NORMALIZATION?

We (Verbrugge, Strange, and Shankweiler, 1974) next examined the possibility that an introductory set of syllables increased the likelihood that a succeeding vowel produced by the same talker was correctly recognized. Because we wished to test a specific hypothesis about the stimulus information required for normalization, we did not employ an introductory sentence as Ladefoged and Broadbent (1957) had done. Instead, we introduced each target syllable by three precursor syllables; this provided three samples of the talker's vowels and little else. In one condition of the experiment, the precursors were /hi/, /hɑ/, and /hu/; these syllables contain examples of the talker's point vowels. For a second condition, we chose /hɪ, hæ , hʌ/, a set of nonpoint vowels that (like the point vowels) are quite widely separated in the space defined by the first and second formants.

Neither set of precursor syllables brought about a systematic reduction of perceptual errors in identifying the target vowels. The errors in each precursor condition averaged 15 percent (compared to 17 percent in the earlier condition without precursors), but in neither case does this difference approach significance.[9] The principal effect of the /hi/, /hɑ/, /hu/ precursors was to shift the pattern of responses somewhat, some vowels showing improved identification, others showing poorer identification.

The idea that normalization is specifically aided by the point vowels--as suggested by Joos (1948), Gerstman (1968), and Lieberman (1973)--is not supported by these data. In fact, <u>no</u> precursor syllables that we tried were found to have a systematic effect. A single, isolated syllable is usually sufficient to

[8] We suspect that these studies made somewhat less severe demands on listeners' perceptual capacities than our own. In the Peterson and Barney (1952) study, listeners heard only 10 different talkers on a particular test. Each talker spoke two tokens of each of 10 /h-d/ syllables. The study yielded an overall error rate of 6 percent. The Abramson and Cooper (1959) study employed eight talkers, each of whom spoke one token of 15 /h-d/ syllables. The overall error rate in that study ranged from 4 to 6 percent. An additional source of perceptual difficulty in our tests is the fact that /ɑ/ and /ɔ/ are homophonous in the dialect of most of our talkers.

[9] This result was confirmed in a separate study (cf. Verbrugge, Strange, and Shankweiler, 1974) of 15 vowels in /h-d/ context. When no precursors were present, errors averaged 13 percent. When each /h-d/ syllable was preceded by the syllables, /kip/, /kɑp/, /kup/, 12 percent of the responses were errors. The difference was not significant.

specify a vowel; prior exposure to specific subsets of vowels could not be shown to supply additional information. It would seem unnecessary to invoke a psychophysical weighting function in order to establish an internal adaptation level (cf. Ladefoged, 1967). We may surmise that the isolated syllable is not so ambiguous an entity as is sometimes implied.

Because of the repetitive and stereotyped manner in which the precursors were presented, some readers might be inclined to doubt whether listeners made full use of the phonetic information potentially available and therefore to question whether these experiments are adequate to test the hypothesis. We can reply to this objection indirectly by referring to a further experiment in which the same precursor syllables did produce a measurable effect on perception of a subsequent target syllable. This experiment involved the same 15 talkers as the earlier experiment, but differed in that the test syllables were produced in a fixed sentence frame: The little /p-p/'s chair is red. Each talker was instructed to produce the sentence rapidly, placing heavy stress on the word chair. The unstressed, rapidly articulated /p-p/ syllables were excised from the tape recording and assembled into two new listening tests. In one condition, the /p-p/ targets were prefaced by the same tokens of the /hi, hα, hu/ precursors employed in the previous experiment. On this test, listeners made an average of 29 percent errors in identifying the vowels in the target syllables. In the other condition, no precursors were present. On this test, listeners misidentified 24 percent of the same vowels. Thus, misperception of target vowels occurred with significantly greater frequency when they were preceded by precursors [$t(35\ df) = 2.88$, $p < .01$]. We may suppose that the precursors impaired recognition of succeeding vowels in this instance because they specified a speaking rate slower than that at which the /p-p/ syllables were actually produced. Thus, whereas we failed to find evidence for effects of precursors on normalization of vocal-tract differences, we do find evidence for adjustment to a talker's tempo (as hypothesized by Lindblom and Studdert-Kennedy, 1967), on the basis of preceding segments of speech.

THE ROLE OF FORMANT TRANSITIONS IN VOWEL PERCEPTION

Our results suggest that the identity of a vowel in a syllable spoken by a new talker is likely to be specified by information within the syllable itself. The phonetic context supplied by preceding syllables apparently serves a function other than that of adjustment for a new set of vocal-tract parameters: it may enable the perceptual system to gauge the tempo of incoming speech and to set its criteria accordingly. We were encouraged by these preliminary findings to look for the sources of information that specify the vowel within the syllable, and to explore how that information is used by the perceiver in the process of vowel perception.

As we noted earlier, the formant transitions in a syllable vary systematically as a function of both the consonant and the vowel. Therefore, we might expect that the listener utilizes information contained in the transitions in recovering the identity of the medial vowel. Research on the identification of isolated steady-state vowels (i.e., vowels that are not coarticulated with consonants) indirectly supports this expectation. Perception of isolated vowels is notably unreliable. Fairbanks and Grubb (1961) presented nine isolated vowels produced by seven phonetically trained talkers to eight experienced listeners. The overall identification rate was only 74 percent; rates for individual vowels ranged from 53 to 92 percent. Slightly better identification of isolated vowels was obtained by Lehiste and Meltzer (1973) for three talkers, where, again, talkers and listeners were phonetically skilled. Fujimura and Ochiai (1963) directly compared the identifiability of vowels in consonantal context and in isolation. They found that the center portions of vowels, which had been gated out of CVC syllables, were less intelligible in isolation than in syllabic context.

Research bearing on this question has also been done with synthetic speech. Millar and Ainsworth (1972) reported that synthetically generated vowels were more reliably identified when embedded in an /h-d/ environment than when acoustically identical steady-state target values were presented in isolation. Finally, Lindblom and Studdert-Kennedy (1967) noted that listeners used different acoustic criteria to distinguish pairs of vowels depending on whether judgments were made on isolated vowels or on the same vowel targets embedded within a CVC environment.

There are at least two ways that the transitional portions of the acoustic signal might provide information for vowel identity. One possibility is that transitions play a role in specifying talker characteristics. Since the loci of formant transitions for a particular consonant vary with differences in vocal-tract dimensions (Fourcin, 1968; Rand, 1971), transitions might serve as calibration signals for normalization. Particularly when the phonemic identity of the consonants is fixed and known to the listener, the transitions might serve to reduce the ambiguity of the vowel by providing information about vocal-tract characteristics of the talker who produced the syllable.

We may also envision a second possibility that is at once more general and more parsimonious: the acoustic specification of vowels, like consonants, is carried in the dynamic configuration of the syllable. In other words, the syllable as a whole cospecifies both consonants and vowel. In this view, transitions may be regarded as belonging to the vowel no less than to the consonants. If this were true, we would expect that the perception of medial vowels would be aided by the presence of consonantal transitions regardless of whether the perceiver encounters many talkers on successive tokens or only one.

To make an experimental test of these possibilities, we (Strange, Verbrugge, and Shankweiler, 1974) constructed a new set of listening tests that contained a series of isolated vowels. In one condition the vowels were spoken by the same panel of 15 talkers described above. In a second condition, a single talker produced the full series of vowels. Together with the earlier tests with /p-p/ syllables, these materials allowed us to compare the relative effects on vowel identifiability of two major variables: presence or absence of consonantal environment and presence or absence of talker variation within a test. This also placed us in a position to evaluate the alternative hypotheses about how consonantal environment contributes to vowel perception.

According to either hypothesis, we would expect that the perception of isolated vowels would be less accurate than the perception of medial vowels on a listening test in which the tokens are produced by different talkers. However, the two alternative hypotheses generate different expectations concerning the error rate on isolated vowels and medial vowels when the talker does not vary within a test. If the advantage of consonantal environment is due to use of transition cues for normalization, we could expect to obtain no difference between performance on these two conditions, because in neither case is there a need for repeated calibration. Therefore, we would expect that vowel recognition would be as accurate for the isolated vowels as for the medial vowels. If, on the other hand, the consonantal environment provides critical information for the vowel independent of talker-related variation, we would expect a difference in consonantal environment to affect performance whether or not talkers vary within a test. Thus, we would expect identification of isolated vowels to be less accurate than medial vowels even for tests in which the talker did not vary.

The results for the isolated vowel tests support the latter hypothesis. The average error in the variable-talker condition was 42 percent (compared to 17 percent errors on the comparable test in which vowels were spoken in /p-p/ environment). This increase in errors is consistent with either hypothesis. However, the results for the single-talker condition also showed a large increase in errors when there was no consonantal environment. The average error in the single-talker conditions was 31 percent for the isolated vowels (compared to 9 percent errors on medial vowels). Moreover, a vowel-by-vowel comparison showed that for every vowel in both talker conditions, the error rate on the isolated vowel was greater than on the corresponding medial vowel. Both major variables (Consonants Present versus Absent and Talker Variation Present versus Absent) were shown to produce significant differences in overall errors [F(1,94 df) = 125.17 and 21.18, respectively, $\underline{p} < .01$]. The decrease in accuracy of vowel recognition due to the absence of consonantal environment was approximately the same whether talkers varied or not (i.e., the analysis showed no significant interaction between variables). We may surmise, therefore, that consonantal transitions do not aid in specifying a vowel by providing information for a normalization stage. On the contrary, these results indicate that the presence or absence of transitions is much more critical for accurate recognition than the degree of experience with a talker's vocal-tract parameters. Whereas the presence of within-test talker variation impairs recognition by only about 8 percent, the absence of a consonantal environment impairs performance by more than 20 percent.

The possibility cannot be overlooked, however, that the relatively poor perception of isolated vowels is attributable primarily to the talkers' inability to produce them reliably. Since isolated vowels do not occur in natural speech (with a few exceptions), talkers may produce them in peculiar ways, with formant frequencies uncharacteristic of the values found in natural syllables. Also, the characteristic relative durations of the vowels (Peterson and Lehiste, 1960) might not be preserved by talkers in their productions of isolated vowels.

To investigate these possibilities, we undertook spectrographic analysis of the tokens of isolated vowels and medial vowels used in our listening tests. Center frequencies of the first three formants and vowel duration were measured for all the tokens in the variable-talker tests, as well as for tokens of all nine vowels spoken in isolation by each of the 15 talkers. The data provided no evidence that the isolated vowels were produced in an aberrant manner. Average formant frequencies for men, women, and children correspond quite closely to those reported by Peterson and Barney (1952) (for vowels in /h-d/ environment), with the exception of /ɔ/.[10] When the formant frequencies of each talker's

[10] This deviation is due to a dialect difference between our group of talkers (predominantly natives of the upper Midwest) and Peterson and Barney's group.

isolated and medial vowels are compared, the values are found to be highly similar. Measurements of vowel duration also fail to account for the increased error rate for isolated vowels. Although the durations of these were for the most part longer than the vowels in /p-p/ environment, the relative durations of the nine isolated vowels were much the same as the relative durations of vowels in consonantal environment. We may suppose, therefore, that the higher error rate for isolated vowels compared to that for vowels in a fixed consonantal environment cannot be explained on the grounds that isolated vowels tend to be produced in an aberrant manner.

The message of these perceptual data is clear: isolated, sustained vowels, although they correspond well to the phonetician's idealized conception of a vowel,[11] are poorly specified targets from the standpoint of the listener. Lehiste and Peterson (1959) found that many hours of practice were needed by untrained listeners before they could identify isolated vowels accurately, even when the tokens were painstakingly produced by a single phonetically trained talker. The ability to identify these "ideal" vowels may be a highly specific skill with little relevance to the identification of vowels in natural speaking situations.

At this point the objection might still be raised that the tests used to measure the perceptual difficulty of medial vowels are unrepresentative of natural conditions. One possibility is that there may be an advantage associated with consonantal context if the context is known beforehand (/p-p/ in this case), but that this advantage would be largely eliminated if the identity of neighboring consonants were unknown (as is often the case in natural speech). To test this possibility, we constructed a listening test in which the target vowels were enclosed by a variable consonantal environment. A panel of 12 talkers (a subset of the original 15) spoke a series of consonant-vowel-consonant syllables. In each syllable, one of the six stop consonants (/b, d, g, p, t, k/) appeared before the vowel and one of the six appeared after the vowel; consonants were selected so that each occurred equally often in each position. One group of listeners was asked to identify only the vowel in each test token; a second group was asked to identify the two consonants as well as the vowel. The average error in identifying the vowels was 22 percent for the first group and 29 percent for the second. Both error rates are well below the 42 percent error rate obtained on the variable-talker test with isolated vowels. In other words, even when listeners do not know the identity of either the consonants or the vowel, recognition is significantly more accurate for medial vowels than for isolated vowels.

A second possible objection to the earlier tests with medial vowels might be that syllables spoken in isolation (in "citation form") are unrepresentative of the syllables found in rapid, connected speech. The medial vowels in rapidly spoken syllables might be at least as difficult to identify as isolated vowels, since the vocalic portions of such syllables often fail to reach the steady-

[11] The formant space is less compressed for isolated vowels than for medial vowels. Thus, if the values of static first and second formants were the primary carriers of vowel quality, isolated vowels should be better perceived because their acoustic values are more widely separated.

state values characteristic of syllables spoken in citation form. The study reported earlier in this paper bears directly on this question. When /p-p/ syllables spoken in unstressed position are excised from a carrier sentence and assembled into a listening test, listeners made an average of 24 percent errors in perceiving the medial vowels. This is not much greater than the 17 percent error rate for perception of /p-p/ syllables read from a list, but is substantially less than the 42 percent error rate for isolated vowels. One might have guessed that the brevity of the short /p-p/ syllables and their failure to reach steady-state values would make them more difficult to identify than isolated vowels, which are longer in duration and more stable acoustically. Apparently, the presence of a consonantal environment more than compensates for these difficulties.

CONCLUDING REMARKS ON THE PROBLEM OF VOWEL CONSTANCY

Let us consider what we have learned about how the perceiver might achieve constancy of vowel quality. In our studies of vowel perception, the objective was to isolate sources of vocalic information in the natural speech signal. We employed signals that presented as many characteristics as possible of normal conversational speech, including a representative range of signal variations that result from physical differences among talkers.

Each way of conceptualizing the vowel contains an implied solution to the problem of perceptual constancy. We first considered the assumption that the vowel can be characterized by a steady-state output of the vocal tract, and that, to a first approximation, fixed-formant loci are associated with each vowel quality of all speakers. To the extent that this assumption is correct, the constancy problem is trivial. Only minor adjustments for variation would be required.

We saw that this conception of the vowel as a simple acoustic event, segmented in time and in spectral frequency composition, was widely shared among students of speech, including those who initiated earlier attempts at automatic speech recognition. We have reviewed a number of findings that are incompatible with this view. First, steady states are the exception, not the rule, in continuous natural speech. As a result of coarticulation of vowels with preceding and following consonants, the syllable is not discretely partitioned, and the information for the vowel is smeared throughout the syllable. Moreover, the variability occasioned by the phonemic environment of a vowel is compounded by the changes that accompany different speaking rates and different vocal-tract sizes. In retrospect, it is easy to see why attempts to design a generally useful speech recognition machine have so far failed.

A more sophisticated conception of the vowel acknowledges the problem of variability but continues to assume that vowels, even in running speech, can be perceived with reference to a single set of acoustic values. This view proposes that tokens of the "same" vowel fall on a line in vowel space defined by the first and second formants. The formant frequencies of two talkers' vowels would then be constant multiples of one another. We noted that this relationship could not literally hold, because vocal tracts differ in shape as well as in size. This rules out an analogy to a melody played in a different key, or to a magnetic tape recording played back at a different speed. The failure of these analogies is revealed by Peterson and Barney's (1952) measurements of first- and second-formant frequencies in men, women, and children. The results of the measurements (displayed in Figure 2) showed wide dispersion of formant values for different speakers with considerable overlap of formants for neighboring vowels. Even when one considers only those tokens on which perfect agreement was obtained by listeners, much scatter among formant values is observed (cf. Peterson and Barney's Figure 9).

Failing to find the invariant relation preserved by linear scaling, investigators have sought a transformation that might yield a closer approximation. For example, it has become an accepted practice to plot units of frequency (Hz) on a scale of mels (Peterson, 1961; Ladefoged, 1967).[12] Transformation of formant frequencies to mels might be defended on the grounds that this unit reflects the response of the auditory system to frequency. However, we are skeptical that the constancy problem can be illuminated by a search for the right scale factor. Ladefoged (1967), who attempted to reduce variability by employing phoneticians as talkers, concluded that separation of all vowels cannot be attained by scaling the first- and second-formant frequencies, whether in linear fashion or nonlinearly, as on a scale of mels.

Although no one has succeeded in demonstrating a generally applicable scaling (normalizing) function, it is widely assumed that perceivers must apply such a function to each new talker they encounter. There has been speculation about the minimal stretch of speech required for calibration. Ladefoged's (1967) application of adaptation-level theory to the problem of speaker normalization reflects the common assumption that some extended sample of a new talker's utterances is required for determining the weights that enable the normalizing adjustments to be made. As we noted, Joos (1948) and Lieberman (1973) proposed that ambiguity of a new talker's utterances can be resolved by reference vowels that permit the perceiver to construct a model of the talker's vowel space, scaling the input according to parameters derived from these calibration vowels.

Listeners can apparently adjust their criteria for perception of synthetic vowels according to the formant ranges specified by a precursor sentence (Ladefoged and Broadbent, 1957). The successful performance of Gerstman's (1968) normalizing algorithm indicates that frequencies of the first and second formants could, in principle, suffice for this purpose. However, we doubt, as does Ladefoged (1967) himself, that first- and second-formant frequencies exhaust the sources of information that specify a vowel in the natural speech signal.[13] Moreover, the fact that listeners can perceive randomly ordered syllables accurately, indicates that there is little need for a mechanism that requires a sample of several syllables in order to construct a normalization schema. Finally, in our own experiments with natural speech, we failed to find that point vowel precursors, or another set of widely spaced vowels, brought about a systematic improvement in recognition of the following vowel.

Our results do not, therefore, support the view that vowels are relational values in a metric space that must be scaled according to other vowels produced

[12] A mel is a psychophysical unit reflecting equal sense distances of pitch and bearing an approximately logarithmic relation to frequency for frequencies above 1000 Hz (Stevens and Volkmann, 1940).

[13] This is also Peterson's (1961) conclusion, based on studies of filtering.

by the same talker. If that theory were correct, it is difficult to see how precursors could fail to improve recognition of an immediately following medial vowel. The presence of coarticulated consonants within the syllable proved far more useful for categorizing natural vowels than prior experience with a talker's utterance. In sum, our studies failed to provide supporting evidence for current conceptions of the normalization process. They force us to consider whether there is justification for a separate and preliminary normalization stage in speech perception.[14]

A major difficulty with all the proposals stated above is that they view the invariance problem in terms of the relation among formant frequencies of relatively sustained vowels. Even if such efforts to discover an algorithm were successful, they would not be sufficient to explain perception of vowels in natural conversational speech, because in such utterances a region of steady-state energy is rarely present in the signal. The presence of sustained acoustic energy at certain "target" frequencies is not essential for identification of a vowel under natural listening conditions. On the contrary, there is evidence that changing spectral patterns are much superior to sustained values as carriers of vowel quality. We cited earlier reports that isolated steady-state vowels are poorly perceived, even after listeners were given substantial training and when the target vowels were spoken by phonetically trained talkers (Fairbanks and Grubb, 1961; Lehiste and Meltzer, 1973). The results of our perceptual studies definitely confirm the perceptual difficulty of isolated vowels. Listeners misidentified 31 percent even when all items within a given test list were produced by the same talker. Moreover, vowels coarticulated with surrounding consonants, as is normal in running speech, were considerably more intelligible than isolated vowels spoken by the same talkers (e.g., 9 percent of vowels in /p-p/ environment were misidentified). It seems unlikely, therefore, that the perceptual system operates by throwing away information contained in formant transitions. Indeed, Lindblom and Studdert-Kennedy (1967), in studies with synthetic speech, demonstrated that listeners use this information directly in their placement of vowel phoneme boundaries. Vowel identifications varied with direction and rate of transitions even when the formant frequency values at the syllable centers were held constant. In short, it is futile to seek a solution to the constancy problem by analysis of any acoustic cross section taken at a single instant in time, and we must conclude that the vowel in natural speech is inescapably a dynamic entity.

Heuristic procedures for automatic recognition of consonants often begin by guessing the identity of the coarticulated vowel, since it is known that the specific shapes of the formant transitions are conditioned by the vowel. The vowel is assumed to be a stable reference point against which the identity of the consonant may be determined. But this, of course, presupposes that the vowel is more directly available than the consonant. We have now examined a number of indications that the problem of perceptual constancy may be no less abstract for the vowel than for the consonant. We found that isolated steady-state vowels

[14] But see Summerfield and Haggard (1973). These investigators measured an increase in reaction time to synthetic syllables from different (simulated) vocal tracts, which they interpret as reflecting extra processing time required for a normalization stage.

provided especially poor perceptual targets. Moreover, there were not substantially more errors in identifying the medial vowels of rapidly spoken syllables (where steady states were presumably not attained), compared to errors on medial vowels in syllables read from a list. It is clear that we can no longer think of defining vowels in terms of acoustic energy at characteristic target frequencies. A recognition device based on filters tuned to specific frequencies is as unworkable for vowels as it is for consonants. If the idea of the vowel target is to be retained, it must take account of the dynamic character of the syllable.

Lindblom's (1963) conception has this virtue. Formant contours are characterized as exponential functions that tend toward asymptotic "target" values associated with the vowel nucleus. Thus, a target can be defined acoustically even though it corresponds to no spectral cross section through the syllable. Our perceptual data presented here are compatible with this view that vowels are specified by contours of moving formants with certain invariant properties over stretches of approximately the length of a syllable.[15]

We may find a parallel to this conclusion in studies of speech production. Investigation of the manner in which phonemes are joined in the syllable reveals context-dependent relationships similar to those we have noted in the acoustic signal. Lindblom's (1963) dynamic theory of vowel articulation is, in fact, an attempt to explain how contextual influences on the acoustic and phonetic properties of vowels are produced. According to this view, undershoot in running speech is brought about by inertia in the response of the articulators to motor excitations occurring in rapid temporal succession. Invariant neural events corresponding to vowel targets thus fail to bring articulators to the positions they assume when the vowel is produced in a sustained manner.

Lindblom's (1963) inference of articulatory undershoot during rapid speech has been confirmed by cinefluorographic data (Gay, 1974). His account of the mechanism of undershoot has not gone unchallenged, however. MacNeilage (1970) presented a different view of the inherent variability of speech production. Basing his conclusions on extensive electromyographic studies of context effects in articulation, he argued that variability of muscle contraction is not to be understood merely as an unfortunate consequence of mechanical constraints on articulator motion, but as necessarily built into the system in order to permit attainment of relatively invariant target shapes. Gay (1974) cited his cinefluorographic findings in support of MacNeilage's hypothesis that variability of gesture must be regarded as a design characteristic. However, unlike MacNeilage,

[15] Our understanding of the vowel has been influenced by Gibson's (1966) approach to the problems of event constancy in visual perception, which is to seek regularities in the stimulus pattern that can only be defined over time. A similar approach is taken by Shaw, McIntyre, and Mace (1974), and by Shaw and Pittenger (in press). We tend to agree with these authors that the dynamic invariants specifying an event may be perceived directly by perceptual systems that are appropriately tuned. While Lindblom (1963) and Lindblom and Studdert-Kennedy (1967) offer a dynamic characterization of vowels, they apparently made the usual assumption that only temporal cross sections can be directly perceived, and they supposed that vowel perception is mediated by a process of analysis-by-synthesis in which the dynamic invariants are used to compute possible input patterns.

he concluded that there was variability not only of gesture, but of spatial target, since he failed to find invariance of vocal-tract shape for a central vowel /ɑ/ that held across speaking rates and consonantal environments. The same conclusion was drawn by Nooteboom (1970), who argued that the kinds of reorganization that occur in talking with the teeth clenched make it difficult to retain the idea of invariant spatial targets. Just as the attainment of a specific acoustic target value is not necessary for the successful perception of a vowel, it is probably the case that the attainment of a specific target shape is not necessary for its effective production. Thus, in production as in perception, it has become increasingly difficult to entertain the notion of an invariant target for each vowel, as long as the meaning of invariance is restricted to a specific vocal-tract shape and its resonances. It is likely that the units of production, like the units of perception, cannot be defined independently of the temporal dimension. Speech, viewed either as motor gesture or as acoustic signal, is not a succession of static states. Invariance in vowel production, then, can be discovered only in the context of the dynamic configuration of the syllable.

The reader might wonder at this point whether the various productions and acoustic forms of a vowel are so heterogeneous that no coherent physical definition (however abstract) could be found that embraces them all. Perhaps the required invariance is not to be found in the acoustic signal at all. If it is not, a radical solution to the constancy problem is to suppose that the variants of a phoneme are physically unrelated, and to assume that the brain stores separately a prototype of each vowel and consonant for every phonemic environment. If we can extend to speech perception an argument made by Wickelgren (1969) concerning its production, then Wickelgren's hypothesis of "context-sensitive allophones" is such a proposal. However, the proposal has little to recommend it. Halwes and Jenkins (1971) find a number of flaws, two of which are critical. First, the proposal fails to capture the phonological relations that are known to be important in understanding both the production and perception of speech. Second, it ignores the "creativity" inherent in the production of speech that permits the reorganization of articulatory movements to maintain intelligibility even when normal speech movements are blocked, as when talking with the teeth clenched or with food in the mouth, or when under the influence of oral anesthesia.

In light of the evidence we have surveyed, it is obvious that attempts to understand the psychophysical constancy relations in speech have failed to discover transparent isomorphisms between signal and perception. This failure has led many to doubt whether a psychophysics of speech could ever illuminate the constancy problem. But certainly, it does not follow from the complexity of the psychophysical relation that the signal fails to specify the phonemic message uniquely.[16] The emphasis that current theories place on the relational nature

[16] We doubt that a "distinctive feature" description of speech would allow a simpler psychophysical relation to be stated. Phonemes are often characterized by a set of component features that are the basis for contrastive phoneme pairs. For example, /b/ and /d/ contrast in <u>place of articulation</u>, while /d/ and /t/ contrast in <u>voicing</u>. There is substantial evidence that such features are integral to the perceptual analysis of speech. We believe that the same arguments apply to the detection of distinctive features as apply to the

of the vowel is misleading because it underestimates the richness of the signal in natural speech, a richness that is attested by the great tolerance of the perceptual system for a degraded speech signal (as in noisy environments or after filtering). To abandon the search for acoustic invariants because the psychophysical relations are complex would surely be a backward step. It should be appreciated, however, that commitment to the principle of invariance does not bind us to a literal isomorphism between signal and percept. The weight of evidence conclusively opposes a one-to-one mapping of perceptual segments and their dimensions on physical segments and their dimensions. In the case of vowels, we have argued that the invariants cannot be found in a temporal cross section but can only be specified over time.[17] For vowels, as for other phonological segments, a major goal of research is to discover the appropriate time domains over which invariance might be found.

REFERENCES

Abramson, A. S. and F. S. Cooper. (1959) Perception of American English vowels in terms of a reference system. Haskins Laboratories Quarterly Progress Report QPR-32, Appendix 1.

Allport, F. H. (1924) Social Psychology. (Boston: Houghton Mifflin).

Bloomfield, L. (1933) Language. (New York: Henry Holt).

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

Cole, R. A. and B. Scott. (1974b) Toward a theory of speech perception. Psychol. Rev. 81, 348-374.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Amer. 24, 597-606.

Cooper, F. S., A. M. Liberman, and J. M. Borst. (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. In Proceedings of the National Academy of Sciences 37, 318-328.

perception of phonemes; namely, they are specified abstractly over stretches of speech of varying length. Thus, a recognition model that includes feature recognition as an early stage must meet the same tests that we have outlined for phoneme recognition in general.

[17] A derived invariant, defined in terms of relations, has been described for the voicing distinction in consonants. In spectrographic analyses of voicing in stop consonants in many languages, Lisker and Abramson (1971) discovered a unity among the apparently diverse and unrelated acoustic features that are correlates of the voiced-voiceless distinction. Their work suggests that aspiration, explosion energy accompanying stop release, and first-formant intensity may all be understood in terms of control of the time relations between stop-closure release and the onset of laryngeal vibration. The derived cue, voice onset time, is a relatively invariant property of the signal for a given overall speaking rate. However, voice onset time is not a simple property in that it is definable only in terms of a temporal relation between two events occurring within the syllable. See Lisker (1975) and MacNeilage (1972) for discussions bearing on the importance of timing in speech.

Delattre, P. C., A. M. Liberman, F. S. Cooper, and L. J. Gerstman. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. Word 8, 195-210.

Denes, P. B. (1963) On the statistics of spoken English. J. Acoust. Soc. Amer. 35, 892-904.

Fairbanks, G. and P. A. Grubb. (1961) A psychophysical investigation of vowel formants. J. Speech Hearing Res. 4, 203-219.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

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

Fant, C. G. M. (1966) A note on vocal-tract size factors and nonuniform F-pattern scalings. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-4, 22-30.

Fant, C. G. M. (1970) Automatic recognition and speech research. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-4, 16-31.

Fourcin, A. J. (1968) Speech source inference. IEEE Trans. Audio Electroacoust. AU-16, 65-67.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. J. Acoust. Soc. Amer. 35, 1889(A).

Gay, T. (1974) A cinefluorographic study of vowel production. J. Phonetics 2, 255-266.

Gerstman, L. H. (1968) Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. AU-16, 78-80.

Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).

Halwes, T. and J. J. Jenkins. (1971) Problem of serial order in behavior is not resolved by context-sensitive associative memory models. Psychol. Rev. 78, 122-129.

Harris, C. M. (1953) A study of the building blocks in speech. J. Acoust. Soc. Amer. 25, 962-969.

Helson, H. (1948) Adaptation level as a basis for a quantitative theory of frames of reference. Psychol. Rev. 55, 297-313.

Hockett, C. F. (1958) A Course in Modern Linguistics. (New York: MacMillan).

House, A. S. and G. Fairbanks. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust. Soc. Amer. 25, 105-113.

Hyde, S. R. (1972) Automatic speech recognition: A critical survey and discussion of the literature. In Human Communication: A Unified View, ed. by E. E. David, Jr., and P. B. Denes. (New York: McGraw-Hill).

Joos, M. A. (1948) Acoustic phonetics. Language, Suppl. 24, 1-136.

Koenig, W. H., H. K. Dunn, and L. Y. Lacey. (1946) The sound spectrograph. J. Acoust. Soc. Amer. 18, 19-49.

Kuhl, P. (1974) Acoustic invariance for stop consonants. J. Acoust. Soc. Amer., Suppl. 55, 55(A).

Kuhn, G. M. and R. McI. McGuire. (1974) Results of a spectrogram reading experiment. Haskins Laboratories Status Report on Speech Research SR-39/40, 67-79.

Ladefoged, P. (1967) Three Areas of Experimental Phonetics. (New York: Oxford University Press).

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. 29, 98-104.

Lehiste, I. and D. Meltzer. (1973) Vowel and speaker identification in natural and synthetic speech. Lang. Speech 16, 356-364.

Lehiste, I. and G. E. Peterson. (1959) The identification of filtered vowels. Phonetica 4, 161-177.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.

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

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.

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

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Amer. 42, 830-843.

Lindblom, B. E. F. and J. Sundberg. (1969) A quantitative model of vowel production and the distinctive features of Swedish vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-1, 14-32.

Lisker, L. (1975) On time and timing in speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton), vol. 12.

Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 767-785.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.

MacNeilage, P. F. (1972) Speech physiology. In Speech and Cortical Functioning, ed. by J. H. Gilbert. (New York: Academic Press).

Mermelstein, P. (1974) A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. Haskins Laboratories Status Report on Speech Research SR-37/38, 191-197.

Millar, J. B. and W. A. Ainsworth. (1972) Identification of synthetic isolated vowels and vowels in h-d context. Acoustica 27, 278-282.

Miller, G. A., G. A. Heise, and W. Lichten. (1951) The intelligibility of speech as a function of the context of the test materials. J. Acoust. Soc. Amer. 41, 329-335.

Nooteboom, S. G. (1970) The target theory of speech production. IPO Annual Progress Report (Institute for Perceptual Research, Eindhoven, Holland) 5, 51-55.

Nooteboom, S. G. and I. Slis. (1970) A note on the degree of opening and the duration of vowels in normal and "pipe" speech. IPO Annual Progress Report (Institute for Perceptual Research, Eindhoven, Holland) 5, 55-58.

Peterson, G. E. (1951) The phonetic value of vowels. Language 27, 541-553.

Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res. 4, 10-29.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in the study of the vowels. J. Acoust. Soc. Amer. 24, 175-184.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllabic nuclei in English. J. Acoust. Soc. Amer. 32, 693-703.

Peterson, G. E., W. S. T. Wang, and E. Sivertsen. (1958) Segmentation techniques in speech synthesis. J. Acoust. Soc. Amer. 30, 739-742.

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

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

Shaw, R., M. McIntyre, and W. Mace. (1974) The role of symmetry in event perception. In Perception: Essays in Honor of J. J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca, N. Y.: Cornell University Press).

Shaw, R., and J. Pittenger. (in press) Perceiving the face of change in changing faces. In Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

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

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.

Stevens, S. S. and J. Volkmann. (1940) The relation of pitch to frequency: A revised scale. Amer. J. Psychol. 53, 329-353.

Strange, W., R. Verbrugge, and D. Shankweiler. (1974) Consonant environment specifies vowel identity. Haskins Laboratories Status Report on Speech Research SR-37/38, 209-216.

Summerfield, A. W. and M. P. Haggard. (1973) Vocal tract normalization as demonstrated by reaction times. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University of Belfast) Series 2, no. 3, 1-26.

Verbrugge, R., W. Strange, and D. Shankweiler. (1974) What information enables a listener to map a talker's vowel space? Haskins Laboratories Status Report on Speech Research SR-37/38, 199-208.

Watson, J. B. (1924) Behaviorism. (New York: Norton).

Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychol. Rev. 76, 1-15.