The Perception of Speech

Michael Studdert-Kennedy
Haskins Laboratories, New Haven

"In short, for the study of speech sounds on any level whatsoever, their linguistic function is decisive." Jakobson, Fant, and Halle (1963:11)

INTRODUCTION

To perceive speech is to extract a message, coded according to the rules of a natural language, from an acoustic signal. The signal is a more or less continuously varying acoustic wave, usually generated by the speech organs of a human or by some device (such as a telephone, synthesizer, or mynah bird) that has been constrained to imitate human speech. The message is a string of lexical and grammatical items that may be transcribed as an appropriately marked sequence of discrete phonemic symbols. How analog signal is transformed to digital message has been studied intensively for no more than forty or fifty years.

Early work was largely guided by the demands of telephonic communication: its aim was to estimate, for example, how much distortion by frequency band-width compression, channel noise, or amplitude peak-clipping could be imposed on the speech wave without seriously reducing its intelligibility. Knowledge accumulated during this period has been reviewed by Licklider and Miller (1951) and by Miller (1951). Recent related work concerned with the effects of noise on speech intelligibility has been reviewed by Webster (1969). Among general conclusions that may be drawn, three are of particular interest to the present discussion. First, the frequency band contributing most to the intelligibility of speech is balanced around 1900 Hz and comprises the four or five octaves between about 200 Hz and 4000 Hz, the region of greatest sensitivity in the human auditory threshold curve. This hints at a not unexpected, biologically determined match between speech signal and perception that research has only recently begun to explore (e.g., Stevens, in press; Lieberman, 1970). Second, speech is highly resistant to distortion: even infinite peak clipping (which reduces the wave to no more than a pattern of zero-crossings) may have surprisingly small effects on intelligibility. Evidently, the speech signal is highly redundant, and the human listener is an adept of impletion, able to supply from within information that is lacking without. Again, only recently has research begun to track down the sources of the listener's information in his linguistic capacity.


+ Also, Graduate Center and Queens College, City University of New York.
A third conclusion of this early work (as peak clipping studies suggest) is that the key perceptual dimensions are not those of the waveform (amplitude and time) but those of its time-varying Fourier transform, as displayed in a spectrogram (frequency, intensity, time). Development of the sound spectrograph (Koenig et al., 1946), followed by publication of a work on "Visible Speech" by Potter et al. (1947) and of a monograph on "Acoustic Phonetics" by Joos (1948), paved the way for the first important task of any perceptual study: definition of the stimulus. In this undertaking, research has been increasingly guided by developments in linguistic theory concerning the structure of the message.

STAGES OF THE PERCEPTUAL PROCESS

Before considering this research, it will be useful to lay out, for purposes of exposition, a rough model of the transformation from signal to message. The process entails, conceptually, at least these stages of analysis: 1) auditory, 2) phonetic, 3) phonological, 4) lexical, syntactic, and semantic. The stages form a conceptual hierarchy but in a working model must be both successive and simultaneous: tentative results from higher levels feed back to lower levels, not only to permit correction of earlier decisions in light of later contradictions but also to permit partial determination of phonetic shape by phonological, syntactic, and semantic rules and decisions.

Only the first stage is based directly on the physical input. It is automatic, that is, beyond voluntary control; it transforms the acoustic waveform that impinges on the ear to some time-varying pattern of neurological events of which the spectrogram is, at present, our closest symbolic representation. What further transformation the signal may undergo is an area of active study (see, for example, Mattingly, this volume; Stevens 1967, 1968a, in press). The process requires at least partially independent, neurological systems for extraction of spectral structure, fundamental frequency, intensity, and duration. These interact and give rise to the auditory (psychological) dimensions of quality (timbre), pitch, loudness, and length. Whether acoustic-psychological transformation already involves neural mechanisms peculiar to speech we will consider below.

All stages beyond the first are abstract: they entail recognition of properties that do not inhere in the signal. Together with our knowledge of social context, they represent the set of expectations, some learned, some probably innate, by which we can (and without which we could not) perceive the signal as speech, speech as language. Training may separate the stages to some degree, and we may demonstrate their psychological reality, initially inferred by linguistic analysis, in the laboratory. But in normal conversation, we are unaware of their contributions to what we perceive. No fixed weights can be assigned to the stages. Indeed, their weights undoubtedly vary with the conditions of listening. Phonetic perception in a high wind is governed as much by situational and higher linguistic factors as by the acoustic signal.

Nonetheless, phonetic perception, the second stage, does bear a peculiarly intimate relation to the first. Though we may, without too much difficulty,

---

1 See Fry (1956) for this observation and for a discussion of stages in which a slightly different position than the one adopted here is taken.
attach phonetic properties to nonspeech, denying speech its phonetic properties is not so easy. There is a directness in perception that makes it difficult to hear the sounds, even of a totally foreign language, as purely auditory events. We hear them, instead, phonetically. That is to say, we hear them as sounds generated by the vocal organs of a human. The nature of this intermediate, phonetic representation is not known. For, despite the term "phonetic," this level is no longer one of sound but rather of some intricate, abstract derivative from the initial auditory analysis. Perhaps it is not without import that, in struggling to interpret the sounds of an unfamiliar language, we (like children who watch a parent or like ourselves straining for meaning through a glass door) often seek order by articulatory imitation: extraction of phonetic information may be closely tied to production mechanisms. Certainly, traditional phonetics, by its easy switches among terms that purport to describe the sounds and terms that unequivocally describe their presumed antecedents in production, hints at some peculiar relation between audition and articulation. But for the moment, we shelve the question. We take the output of this stage to be isomorphic, with a narrow phonetic transcription or, following the formulation of Chomsky and Halle (1968), with a phonetic matrix: the columns, headed by phonetic symbols, segment the phonetic features of the rows. By this point, segments and categories are already present: the sounds have become speech, if not language. But the message is still redundant and much allophonic variation remains to be resolved.

The third stage of the conceptual hierarchy is phonological: phonetic segment is converted to systematic phoneme (Chomsky, 1966). The stage corresponds to the lowest level of Chomsky and Halle's (1968) generative phonological component, a level at which, according to Chomsky and Miller (1963:fn9), the "phonemes" appear. During this stage, the listener applies phonological rules and determines the status of the perceived "sound" sequence within (or without) his language: he "hears through" the features of the phonetic matrix to the distinctive features of the underlying phonemic matrix. Phonetic analysis will have established, for example, the nasalized medial vowel of [k ǝ t] in many American English dialects. It remains for phonological analysis to reallocate the nasality from the phonetic column for [ǝ] to a new column for a following segment and so to arrive at recognition of /k ǝ nt/"can't" (Malécot, 1956). In this stage, also, the listener may dismiss phonetic information that serves no distinctive purpose in his language, treating, in English, both the initial stop of [tʰap] and the unreleased final stop of [pʰat] as instances of /t/. And in this stage, he applies the phonotactic rules of his language to derive an acceptable interpretation of the phonetic information.

We can separate the level experimentally from higher levels by calling for perceptual judgments of nonsense syllables. In cross-language studies, listeners reflect the phonological categories of their native languages by their classification of phonetic segments (Lotz et al., 1960; Abramsdn and Lisker, 1965; Chistovich et al., 1966; Stevens et al., 1969; Lisker and Abramson 1970). Operation of phonotactic rules within speakers of a single language has also been demonstrated (Brown and Hildum, 1956; Greenberg and Jenkins, 1964). We should note, incidentally, that, for the untrained listener, relations between phonologic and phonetic levels are as close as between phonetic and auditory: under normal conditions, he instantly hears speech according to the phonological categories of his native language.
The fourth and last stage (lexical, syntactic, and semantic) represents a complex of interrelated processes that we wish to exclude from the main line of argument. But there are several points to be made. First, we distinguish between direct and indirect perceptual effects of this stage. By direct effects, we intend those grounded in observable acoustic parameters of the signal. For example, lexical items are marked by the acoustic correlates of stress: variations in duration, intensity, and fundamental frequency (Fry, 1955; Hadding-Koch, 1961; Bolinger, 1958; Fry, 1968). If, in difficult listening conditions, segmental features are partly lost, stress patterns may help delimit the sampling space of the lexical items (Skinner, 1936; Savin, 1963; Kozhevnikov and Chistovich, 1965:238ff.). At the same time, perceived stress does not depend on its acoustic correlates alone. Listeners to a synthetically modified utterance with all vowels displaced to [a], but with timing, frequency, and amplitude variations retained, do a poor job of judging its stress pattern (Lieberman, 1965). They require syntactic or, as Klatt (1969) has argued in discussing Lieberman's results, at least segmental information to make reliable stress judgments (and vice versa: a neat instance of parallel processing). Similar results have been obtained for fundamental frequency contours conveying intonation patterns (Lieberman, 1965). In short, acoustic correlates of higher-order linguistic structures may be directly perceived, although relations between signal and message are relatively loose and we have, as yet, no detailed account of the underlying, interactive process.²

More important to the present discussion are the indirect effects on perception of this fourth stage. Here, we intend those provisional syntactic and semantic decisions that may resolve phonetic doubt. As we shall see, even in citation forms, there is frequently no simple, invariant correspondence between spectral structure and phonetic shape; in continuous speech the lack of invariance is even more marked (Shearmee and Holmes, 1962). Pickett and Pollack (1963) and Lieberman (1963) found that words excised from sentences and presented to listeners in isolation, without syntactic and semantic context, were poorly recognized. Other studies have shown that words are perceived more accurately in a sentence than on a list (Miller et al., 1951) and have separated the contributions of syntax and meaning to the perceptual outcome (Miller and Isard, 1963).³

One may wonder whether these effects of higher-level factors are truly perceptual. A recent study speaks to this question. Warren (1970) demonstrated an effect that he termed "phonemic restoration." Listeners heard a tape-recorded sentence: "The state governors met with their respective legislatures convening in the capital city," with a 120 msec segment deleted and replaced by a cough (and, on another occasion, by a burst of 1000 Hz tone) of the same duration. The missing segment corresponded to the first "s" of...

²Lehiste (1970) reviews the experimental phonetic literature on prosodic features and evaluates its meaning for linguistic theory.

³There is also a sizeable literature examining the effects of surface structure on perceptual segmentation of an utterance (Fodor and Bever, 1965; Garrett et al., 1966; Johnson, 1966; Reber and Anderson, 1970) and of deep structure on immediate recall of sentences (Mehler, 1963; Miller, 1964; Savin and Perchonock, 1965; Blumenthal, 1967).
"legislatures," "together with portions of the adjacent phonemes which might provide transitional cues to the missing sound." Of twenty subjects listening to this sentence, nineteen reported that all speech sounds were present, and one reported a missing phoneme but the wrong one. Factors above the phonological may contribute to this illusion: Sherman (cited by Warren, 1970) "found that when a short cough was followed immediately by the sounds corresponding to 'ite,' so that the word fragment could have been derived from several words, such as 'kite' or 'bite,' the listener used other words in the sentence to determine the phonemic restoration; when the preceding and following context indicated that the incomplete word was a verb referring to the activity of snarling dogs, the ambiguous fragment was perceived quite clearly as either 'bite' or 'fight'" (Warren, 1970:393).

The important point is that "the ambiguous fragment was perceived quite clearly": the effect is an illusion and resists manipulation. Warren remarks that other listeners, "despite knowledge of the actual stimulus, still perceived the missing phoneme as distinctly as the clearly pronounced sounds actually present" (p. 392). The study not only demonstrates the abstract nature of phonetic perception, but it is also consistent with "a somewhat novel theory of speech perception" (Chomsky and Miller, 1963:311). This "novel theory" has been summarized by Chomsky and Halle (1968:24):

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

This account circumvents the failure of the acoustic signal to meet the "linearity condition" and the "invariance condition" of segmentation into fixed phonetic units (Chomsky and Miller, 1963), problems to which we turn in the following section. However, an adequate theoretical account must not only define the "certain cues" that the hearer uses but also explain how he makes use of them. Like the speaker of Jakobson and Halle (1956:5,6) who, if context and syntax cannot be trusted to take up the slack of slovenly speech, may deploy the full resources of his code to produce "an explicit form which... is apprehended by the listener in all its explicitness," so the listener, unaided by syntax or context, may deploy the code at his command to extract from an intrinsically slovenly acoustic signal an explicit phonetic message. Most of what follows is directed toward an understanding of this act of "primary recognition" (Fry, 1956).

---

4 Skinner (1936) and Miller (1956) describe similar illusions.

5 There are counterparts in other areas of perception. Neisser (1966) and Kolers (1968a) review similar problems in visual pattern recognition that have also invited abstract, "constructive" theories of perception.
DEFINITION OF THE STIMULUS

The first task of any perceptual study is to define the stimulus. For this, the sound spectrograph has been the principle instrument, and two complementary methods have been used: analysis and synthesis. Spectrographic measurements first provide data concerning the cues that seem likely to be important for perception (formant patterns, temporal relations, noise bandwidths, and so on). Synthesis is then used to verify or adjust the preliminary conclusions of analysis.

Synthesis was first used in this way at Haskins Laboratories in New York. Cooper (1950) (see also Borst, 1956) developed the Pattern Playback as a research tool for reconverting spectrographic patterns into sound. The patterns, painted on a moving acetate belt, reflect frequency-modulated light to a photo-electric cell that drives a speaker. Portions of the pattern can be systematically emphasized, pruned, deleted until minimal cues for the perception of a particular utterance have been determined (Liberman, 1957). With this device, and with its electronic successors at Haskins and elsewhere, a body of knowledge has been built up concerning acoustic cues for speech, sufficient for synthesis by rule of relatively high-quality speech (Liberman, 1957; Fant, 1960, 1968; Mattingly, 1966, this volume; Flanagan, 1965; Stevens and House, 1970).

Implications of this knowledge have been considered by Liberman et al. (1967b). They draw two pertinent conclusions. First, there are, for the most part, no segments in the acoustic signal that correspond to perceived segments of the message. Certainly the sound stream may be segmented and, as we shall see, these segments may be crucial to perceptual reconstruction of the message. But, whatever level of message unit we look for—distinctive feature, phoneme, or syllable—we frequently find no single sound segment corresponding to it and it alone. There are exceptions: fricatives or stressed vowels, for example, may be isolated in slow speech. But, in general, a single segment of sound contains information concerning several neighboring segments of the message, and a single segment of the message may draw information from several neighboring segments of the sound (see also Fant, 1962, 1968). Since perceptual segmentation not only occurs but is essential to the "duality of patterning" on which human language rests (Hockett, 1958), lack of one-to-one relations between signal segments and message segments constitutes a problem for both psychologists and linguists.

A second, closely related, conclusion of Liberman et al. (1967b) is that acoustically distinct signals (separated by differences that in nonspeech would be well above threshold) are frequently perceived as identical, while acoustically identical signals are frequently perceived as distinct. There is thus, for speech, a lack of isomorphism between sign and percept analogous to that in other areas of perception.

Here, we distinguish between two types of anisomorphism in the perceptual process: one can be observed by the unaided human listener, the other cannot. The first includes "extrinsic" allophonic variations peculiar to a particular language or dialect (Wang and Fillmore, 1961; Ladefoged, 1966) and constitutes a problem in the relations between phonetic and phonological segments (stage 3 of the model outlined above). Thus, in the formulation of Chomsky and Halle (1968), neither columns nor rows of the distinctive feature matrix that
serves as input to the generative phonological component are necessarily isomorphic with those of the phonetic feature matrix at output. The inputs /rayt + r/ and /rayd + r/, for example, (see Chomsky and Miller, 1963) emerge as [rayDr] and [ra.yDr]. Here, columnar segmentation of the phonemic input has been lost by transformation of a distinction in the fourth column (voiced/ unvoiced) into a distinction in the second (long/short vowel); also, phonologically distinct segments, /t/ and /d/, have become phonetically identical as an alveolar flap, [D], while phonetically distinct diphthongs, [ay] and [a.y], have emerged from the single phonological diphthong /ay/. These transformations may be generated by the ordered application of two phonetic rules. And, in general, the system underlying extrinsic allophonic variations may be inferred and stated in a set of rules relating phonetic and morphophonemic levels. In the present discussion, we shall not further consider this type of variation.

The second type of anisomorphism [and the one to which Liberman et al. (1967) address themselves] was discovered only when it became possible to substitute a suitable analyzing instrument (the sound spectrograph) for the human listener. There then appeared the discrepancies between acoustic signal and phonetic percept referred to above: a lack of one-to-one correspondence between acoustic and phonetic segments and a host of "intrinsic" allophonic variations, such that the acoustic signal clearly could not meet the linearity and invariance conditions imposed by traditional concepts of speech as a sequence of discrete phonetic segments. Anisomorphism of this type constitutes a problem in the relations between acoustic pattern and phonetic segments (stages 1 and/or 2 of the model outlined above).

Attention to these discrepancies was first drawn by perceptual experiments with synthetic speech. For example, Liberman et al. (1952) showed that a brief burst of energy centered at 1440 Hz and followed by a two-formant vowel pattern was sufficient cue for perception of [p] if the vowel was [i] or [u], of [k] if the vowel was [a]. In other words, perception of [p] or [k] was determined not by the frequency position of the stop burst but by the relation of this position to the following vowel. Here, a single acoustic cue controlled two distinct percepts. The authors concluded that, for these stops, "the irreducible acoustic stimulus is the sound pattern corresponding to the consonant vowel syllable." Schatz (1954) confirmed their results in a tape-cutting experiment with natural speech.

Later experiments demonstrated the importance of second formant transitions as cues for distinguishing among labial, alveolar, velar stop, and nasal consonants (Liberman et al., 1954) and went on to show that a sufficient acoustic cue for a given phonetic segment may prove not only different in different contexts but so different that there seems to be no physical basis for perceptual generalization between the tokens (Delattre et al., 1955). In fact, if sufficient cues for [d] in different phonetic contexts (a rapidly rising F2 transition for [di], a rapidly falling transition for [du]) are synthesized in isolation, their perceptual identity is lost, and they are heard as different "chirps," one rising in pitch, the other falling (Mattingly et al., in press). These and other examples from perceptual studies with synthetic speech are reviewed by Liberman (1957) and by Liberman et al. (1967b).

Studies of natural speech have confirmed that there is enormous variability in the acoustic correlates of a given phonetic segment as a function
of phonetic context, stress, and speaking rate (Shearme and Holmes, 1962; Lindblom, 1963; Stevens and House, 1963; Kozhevnikov and Chistovich, 1965; Stevens et al., 1966; Öhman, 1966; Menon et al., 1969). Öhman (1966), for example, collected data from spectrograms. He traced the paths of the first three formants in spectrograms of intervocalic [g], followed and preceded by all possible combinations of five Swedish vowels. He found large variations in the formant transitions on either side of the [g] occlusion as a function of the vowel on the opposite side of the stop. He concluded that "the perception of the intervocalic stop must be based on an auditory analysis of the entire VCV pattern rather than on any constant formant-frequency cue" (p. 167). Thus Öhman implies, as do Liberman et al. (1952), that we reduce acoustic variance to phonetic invariance by analyzing relations between portions of the auditory pattern over sections of roughly syllable length. This process is examined in a later section (Syllables, Segments, Features).

One other invariance problem deserves mention, if only because it fits neither of the types (extrinsic, intrinsic) discussed above: that arising from speaker-dependent variations in vowel formant frequencies. Center frequencies of the first two or three formants as principal acoustic determinants of vowel color have been known for many years (Delattre, et al., 1952; Peterson, 1952, 1959, 1961; Peterson and Barney, 1952; Ladefoged, 1967). Also known since the early work of Fant (1947, reported in Fant 1966) and of Peterson and Barney (1952) is that formant frequencies vary widely enough to produce considerable acoustic overlap between phonetically distinct vowels spoken by different classes of speakers (men, women, children) and by different individuals within a class. Formant frequencies of a vowel spoken by a woman tend to be some 10-20 percent higher than those of a phonetically identical vowel spoken by a man, and for children the shift is even greater. What are the grounds of the perceived identity?

The hypothesis that vowels are "normalized" at some point during initial auditory analysis by application of a simple scale factor, inversely proportional to vocal tract length, is probably not tenable (Peterson, 1961). Fant (1966) has brought the problem into relief by showing that the male-female shift for Swedish and American English vowels is not constant for all formant frequency regions (largely due to a greater ratio of male to female tract length in the pharynx than in the mouth cavity), so that the putative normalization factor would differ for rounded back, open unrounded, and close front vowels. Fujisaki and Nakamura (1969) have developed an algorithm that separates the five vowels of Japanese (spoken by men, women, and children) with more than 90 percent accuracy on the basis of their first two formants, but it is not known whether their method could resolve a richer system or the severely reduced vowels of running speech (Lindblom, 1963). Under these conditions, invariance may be derived at a later stage of the perceptual process.

In fact, evidence exists that listeners adjust their phonetic decision criteria to the speaker's vocal tract characteristics (inferred from F1 and F3, involving stage 2 of the perceptual process) (Kasuya et al., 1967; Fujisaki and Kawashima, 1968; Fourcin, 1968) and/or to his vowel quadrilateral.

---

6See also the work of Pols et al. (1969) on Dutch vowels.
(inferred with aid of situational and linguistic constraints, involving stages 3 and 4) (Joos, 1948; Ladefoged and Broadbent, 1957; Ladefoged, 1967; Gerstman, 1968). However, no solution is yet generally agreed upon, and speaker-dependent vowel variation therefore takes its place beside other invariance problems previously mentioned.

We conclude, then, that while study of the speech signal and its perception has led to an understanding of acoustic cues sufficient for rather successful speech synthesis, it has also revealed that the signal is an intricate pattern of highly variable overlapping acoustic segments, anisomorphic with the perceived message. It has thus raised more problems for speech perception than it has solved. We may bring these problems into focus if we briefly turn our attention to production and ask how "intrinsic" acoustic variability may be presumed to arise.

Over the past dozen years, there has been a number of studies of muscular activity in speech production. Initial impetus for much of the work was given by the notion of the Haskins group that an invariance lacking in the acoustic correlates of phonetic segments might be found among their articulatory correlates. Early electromyographic (EMG) studies (e.g., Harris et al., 1962; MacNeilage, 1963; Harris et al., 1965) offered some support for this hypothesis, but more recent work has not (e.g., Fromkin, 1966; Tatham and Morton, 1968; MacNeilage and DeClerk, 1969), and MacNeilage (1970:184) has remarked: "Paradoxically, the main result of the attempt to demonstrate invariance at the EMG level has been...to demonstrate the ubiquity of variability."?

The presumed invariance must therefore lie at some neurological level, presently inaccessible. Some theorists (e.g., Ladefoged, 1966; Ohman, 1965, 1966; Ohman et al., 1967; Liberman et al., 1967) have taken this to be the level of "motor commands" and have assumed muscular variability (and consequent "intrinsic" allophonic variations) to arise from mechanical constraints, neuromuscular inertia, and temporal overlap of successive commands. Others (e.g., Fromkin, 1966; MacNeilage, 1970) have suggested that variability may result from controlled, contextually adapted responses to a set of invariant "go-to" or target commands. Both approaches have been elaborated. Ohman (1967) has developed a mathematical model by which vocal tract shape, area function, and speech wave may be computed from a linear combination of fixed commands and coarticulation functions. MacNeilage (1970) has explored the possibility of controlled, variable responses to fixed target commands in light of current neurological theory. Both describe derivation of a more or less continuous, variable signal from discrete, invariant commands.

If, now, we take these commands to be isomorphic with, if not identical to, some articulatory phonetic feature matrix, such as that proposed by Chomsky and Halle (1968) (cf., Stevens and Halle, 1967; Chistovich et al., 1968; Mattingly, this volume), we have, at least, some way of conceptualizing the nature of the phonetic matrix and of its output relation to the acoustic signal. The problem for perceptual theory is that it has, at present, no firm grip on

---

7 This should not be taken to imply that production is unruly. As MacNeilage's paper makes clear, EMG studies are advancing our understanding of its laws. See also Harris (this volume).
the reverse relation between acoustic signal and perceptual phonetic matrix. Among the reasons for this are, first, that the processes relating these levels are even less accessible to observation that the corresponding processes on the production side; second, that these levels are so tightly connected perceptually that it is difficult to separate them in behavior; third, that we have no clear concept of, and no terminology to describe, the phonetic matrix at the output of stage 2. Our task is, therefore, to define this abstract, phonetic matrix and its relation to auditory parameters of the acoustic signal.

CLASSIFYING SPEECH SOUNDS

Let us begin by considering how we classify speech sounds. Experimental evidence reinforces our intuitive recognition that we do so rapidly and involuntarily (see, for example, Kozhevnikov and Chistovich 1965:222 ff.). In this, speech is sui generis. Walking through the woods, we instantly recognize the sound of a waterfall, but with little difficulty, we may choose to hear it as a senseless rumble. Similarly, we have little difficulty in suspending our cognition of a speaker's meaning. But we do find it difficult not to recognize his sounds as speech: recognition is automatic, instantaneous. In what follows, we attempt to disentangle auditory from phonetic (and phonological) stages and to estimate their roles in perception.

Listeners can certainly make purely auditory judgments of speech signals. Flanagan (1965: Ch. VII) has reviewed studies carried out with the general intent of setting upper and lower limits on the discriminability of acoustic dimensions known to be important in speech (vowel formant frequencies, formant amplitudes, formant bandwidths, fundamental frequency, fricative noise bandwidth, and so on). The results, where comparable, give values of the same order as those reported in nonspeech auditory psychophysical studies. But for this, two conditions are necessary: first, the signals must be relatively sustained; second, the listener must be instructed either explicitly or implicitly, by the nature of the experimental task, to listen to the signals as though they were not speech. If the signal is presented in a word or phrase, or among a set of phonetically opposed sounds, a distinctive speech mode of response tends to appear.

For example, Lane et al. (1961) (see also Lane, 1965) asked subjects to judge the loudness of the vowel [a], produced in isolation, and determined a loudness function exponent of 0.7, a value close to that usually found in experiments with nonspeech sounds. Ladefoged and his colleagues (see Ladefoged, 1967:35-41, for a summary of their work) asked subjects to assess the relative loudness of two words in a constant carrier sentence. They found a loudness function exponent of 1.2. This was exactly equal to the exponent of their function relating loudness to the rate of work done upon air in phonation. Ladefoged and his colleagues concluded that their results reflected a distinctive speech mode by which loudness of sound is judged in terms of the physiological effort required to produce it. Lehiste and Peterson (1959) reached a similar conclusion in a study of the loudness of a set of nine steady-state vowels.

Analogous results have been reported in studies of intonation contours. Hadding-Koch and Studdert-Kennedy (1963, 1964) varied the extent and direction of the terminal glide of a fundamental frequency contour imposed synthetically on a vocoded carrier word. They asked subjects, under one experimental
condition, to judge the glide as either rising or falling, under another condition, to judge the word as a question or statement. Subjects' psychophysical judgments were influenced by their linguistic judgments: they tend to judge falling glides of words they considered questions as rising, and rising glides of words they considered statements as falling. In an extension and replication of this study (Studdert-Kennedy and Haddad, in preparation) the authors compared psychophysical judgments of contours imposed on a word with those of matched modulated sine-waves. The previously observed effects were much reduced in the sine-wave judgments. Lieberman (1967), in a theoretical account of these results, argues that listeners perceive intonation contours in terms of the subglottal pressure changes and laryngeal maneuvers required to produce them.

In short, if speech sounds are isolated and of fairly long duration, listeners will make reliable auditory judgments of the same order as they make for comparable nonspeech sounds. But if signals are presented in a context that encourages the listener to deploy his linguistic resources, a characteristic mode of perception appears: unable to separate auditory from phonetic, the listener bases supposedly auditory judgments on phonetic or linguistic decisions. By the same token, if experimental conditions permit auditory judgment, listeners may supplement phonetic skills with auditory, provided the signal is of sufficiently long duration. This appears to be a principal basis of differences observed in the discrimination of consonants and vowels.

A typical experiment goes as follows. Two or more phonetic segments are selected for study. Among reasons for selecting the segments is that they are distinguished by acoustic differences lying along a continuum, such as direction of a formant transition, duration of a silent interval, or center frequencies of formants. One of the selected segments is synthesized, usually within a nonsense word or syllable, sometimes (if a fricative or vowel) in isolation. The relevant acoustic cue is then varied systematically, in steps large enough to be psychophysically detectable in nonspeech, small enough for there to be several steps within phonetic categories. The result is a series of a dozen or so acoustic patterns that range in phonetic type from, say, [ba] through [da] to [ga], or from [ta] to [da], or from [i] through [I] to [ɛ]. Several tokens of each pattern are recorded and gathered into random test orders.

Listeners are then asked to identify and to discriminate between acoustic patterns. Identification is usually by forced choice: listeners assign each token to one of the designated phonetic categories. For discrimination, listeners usually hear tokens in triads, of which two are identical, the other different by one or more steps along the continuum, and are asked either to pick out the odd one or to indicate whether the third token is the same as the first or second.

Under these conditions a listener's performance typically approaches one or other of two ideal modes of perception, termed "categorical" and "continuous".

---

8 As a matter of fact, experiments customarily prescribed phonological categories and, therefore, engage phonological perception. But since our interest here is to separate auditory from phonetic, we disregard the phonological component in what follows.
(Liberman et al., 1957; Liberman et al., 1961b; Fry et al., 1962; Eimas, 1962; Studdert-Kennedy et al., 1970b). By "categorical" perception is intended a mode in which each acoustic pattern, whatever its context, is always and only perceived as a token of a particular phonetic type. Asked to discriminate between two acoustic patterns, the listener can do so if he assigns them to different phonetic categories but not if he assigns them to the same phonetic category. In other words, he can find no auditory basis for discrimination and so must rely on category assignments. By "continuous perception" is intended a mode in which a listener may, if asked, group different patterns into a single category, but his categories are not clearcut (due to context effects), and he is still able to discriminate between patterns that he assigns to the same category. In other words, discrimination is independent of category assignment. (For a fuller account, see Studdert-Kennedy et al., 1970).

Listeners have approached these two modes of perception in many studies. In general, they tend to perceive a continuum categorically if the acoustic variations separate stop consonant categories (e.g., [b, d, g], [p, b, [t, d]), continuously if identical variations are carried by nonspeech signals with no phonetic significance\(^9\) or if the acoustic variations separate sustained vowels.\(^11\)

The categorical/continuous distinction between speech and nonspeech is fundamental. But the same distinction between consonants and vowels is more troublesome. Early interpretations (e.g., Fry et al., 1962; Liberman, et al., 1967a) proposed two distinct perceptual mechanisms: a motor reference mechanism for the categorical stop consonants, paralleling their articulatory discontinuities, an auditory mechanism for the continuously graded vowels. There are many reasons why this is not satisfactory, not least, the difficulty of believing that the syllable, an articulatory and perceptual integer, compounded of consonant and vowel, is analyzed by two distinct mechanisms. Furthermore, this account has been superseded.

Recent work has demonstrated that continuous perception of vowels is a function of their duration and of the experimental method used to study them. Stevens (1968b) has shown that medial vowels in CVC syllables are more categorically perceived than the same vowels sustained in isolation. Fujisaki and Kawashima (1969) have shown that listeners' reliance on category assignment for discrimination increases as the duration of synthetic vowels is reduced from 6 to 3 to 1 glottal pulse. Sachs (1969) has demonstrated a similar effect of duration for vowels in isolation and in word context.

Fujisaki and Kawashima (1969) have also developed a quantitative model of the listener's behavior in discrimination studies. Briefly, the model

\(^9\)The term "categorical" is here preferred to "categorial," since it carries, in addition to the meaning "of or involving a category," shared by both words, the sense "absolute, unqualified." (Webster's Third New International Dictionary, 1965).

\(^10\)See Liberman et al., 1957; Liberman et al., 1961 a,b; Bastian et al., 1961; Eimas, 1963; Fujisaki and Kawashima, 1969; Abramson and Lisker, 1970; Mattingly et al., in press.

\(^11\)See Abramson, 1961; Fry et al., 1962; Eimas, 1963; Stevens et al., 1969.
states that the degree of categorical perception depends on whether auditory or phonetic short-term memory is summoned for the decision process during discrimination. The reliability of auditory short-term memory is less for the brief acoustic events that signal stop consonants than for the relatively sustained events that signal steady-state vowels. The listener has recourse to auditory discrimination whenever he is asked to distinguish between two identical phonetic types; in this, his vowel auditory memory serves him better than his consonant auditory memory, and his vowel discrimination is accordingly superior. The model makes quantitative predictions that have been repeatedly confirmed in experimental tests. The authors conclude that vowels may be perceived either categorically or continuously depending on experimental conditions.

Chistovich and her colleagues reached the same conclusion by a different route (Chistovich et al., 1966a, 1968). Chistovich has explicitly addressed herself to problems of phonetic classification. She has questioned the value of studying speech discrimination on grounds that the procedure invites a listener to search the signal for auditory qualities that he would not normally detect and so to hear it as nonspeech (1968:34, 35). She has confined her own studies to absolute identification, asking subjects to shadow, mimic, or transcribe natural or synthetic speech sounds. We will not describe the methods here (see below: Syllables, Segments, Features). But in an important paper (Chistovich et al., 1966a; see also Fant, 1968) she has demonstrated that even isolated, steady-state vowels may be perceived categorically, if the experimental method forces the listener's attention to phonetic, rather than auditory, qualities.

We should not be misled into supposing that there are no important differences between consonants and vowels: their functional opposition within the syllable is fundamental to both production and perception of speech. Vowels are acoustically more variable and phonetically more subject to the effects of context than are consonants; they carry a lighter segmental load and virtually all the auditory load of prosodic and indexical features. Their perceptual passage therefore leaves an auditory residue that the listener may put to nonphonetic use: his judgments are then continuous. If the residue is reduced by rapid speech, or if the listener's attention is diverted from it by some resolutely phonetic task, his judgments are categorical. In short, he may perceive vowels both auditorily and phonetically; consonants he perceives phonetically. The distinction to be drawn is not, therefore, between consonants and vowels, but between continuous auditory and categorical phonetic perception, the first typical of nonspeech, the second peculiar to speech.

The argument of the last few pages has brought us no closer to separating the auditory from the phonetic stages of speech perception. But it does suggest that, insofar as listeners can achieve this separation, they are judging auditory aspects of the signal irrelevant to phonetic perception, and that, insofar as they perceive phonetically, they cannot achieve the separation. In other words, the output of stage 1 cannot be brought into consciousness under

---

12 Abercrombie (1967: Ch. 1) distinguishes between linguistic and indexical (dialectal, personal) features of the acoustic signal.
normal listening conditions. We might even suppose (although this is not a necessary conclusion) that phonetic classification of speech already begins during the initial auditory analysis of stage 1.

Such a position is implicit in the "immanent approach" of distinctive feature theory. Theorists emphasize that correlates of the features are to be found at every level of the speech process (articulatory, acoustic, auditory) and that the invariance to be sought in the signal is "relational" rather than absolute (Jakobson et al., 1963; Jakobson and Halle, 1956; Chomsky and Miller 1963). They stress the perceptual importance of the entire spectral pattern rather than of band-limited cues, such as a single formant transition (Fant, 1964). The relational concept is difficult, since reference points for spectral relations must vary with speaker, dialect, phonetic context, stress pattern, and speaking rate. Further, Fant has remarked that "statements of the acoustic correlates to distinctive features have been condensed to an extent where they retain merely a generalized abstraction insufficient as a basis for the quantitative operations needed for practical applications" (Fant, 1962), and no one has attempted to use the acoustic specifications of distinctive features to synthesize speech.

Stevens, in his recent work (1967, 1968a, in press) undertakes to remedy this situation by showing that there is "some justification on a purely physical basis for a characterization of phonemes in terms of discrete properties or features" (Stevens, in press). His general procedure is to compute from an idealized vocal tract model the spectral poles and zeros associated with, for example, a particular point of closure or constriction. For certain points of constriction, there appears a significant concentration of spectral energy; the frequency position of this concentration proves relatively insensitive to small shifts in position of the constriction. These "quantal places of articulation...are optimal from the point of view of sound generation" (Stevens, 1968:200), since they permit relatively imprecise articulation without serious perturbation of the signal. Furthermore, they tend to correspond to places of articulation used in many languages (e.g., velar, postalveolar (retroflex), postdental). Similar computations for [i, a, u], the pivots of most vowel systems, provide spectral correlates of their distinctive feature definitions, in terms of F1-F2-F3 positions (Stevens, in press).

We note, incidentally, that Chistovich et al., (1966b) have reported related perceptual data for vowels. They used a handmaneuvered version of the Stockholm Royal Institute of Technology's OVE I b. The instrument permits a subject to trace any selected path through the F1-F2 plane (with F0, F3, and F4 set at appropriate values) and to judge the resulting sounds. Subjects indicated whenever the continuously changing vowel crossed a boundary into a region of altered phonetic quality. Over a hundred such boundary points were determined by four subjects, marked on an F1-F2 plot and connected by best-fitting straight lines. On this plot most boundaries were either vertical (fixed F1) or horizontal (fixed F2), suggesting that "extremely simple rules employing critical boundary values of formant frequencies operate in vowel perception" (Chistovich et al., 1966b; see also Fant, 1968).

Day (1968, 1969, 1970) and Day and Cutting (1970a, b) report the results of work with dichotic and other experimental methods that may serve to separate the stages behaviorally.

13
Implicit in Stevens' work is the assumption that there should prove to be a biologically comfortable match between articulatory and auditory capacities (cf., Halle, 1964; Stevens and Halle, 1967; Stevens et al., 1969). Lieberman (1970) has developed this position more fully, arguing that phonological features may have been selected through a combination of articulatory constraints and "best matches" to specific neural acoustic detectors. Recent work in neurophysiology has demonstrated the existence of relatively complex property, or feature, detectors in cat (Whitfield and Evans, 1965) and frog (Frishkopf and Goldstein, 1963; Capranica, 1965). It is not unreasonable to suppose that comparable detectors, tuned to features of speech, may exist in man.

In short, there are arguments and some evidence to suggest that linguistically relevant features of the acoustic signal may be extracted during initial auditory analysis. How fixed pattern detectors could resolve intrinsic allophonic variations and the incipient entropy of running speech is not clear. But let us suppose that sets of property detectors are, indeed, neatly sprung by the flow of speech. There would then remain the deeper task of grouping the outputs of these detectors into phonetic segments. For this, more than an auditory analysis is required.

**SYLLABLES, SEGMENTS, FEATURES**

Problems of segmentation have bedeviled speech research since its inception.14 Here, particularly, research has relied on linguistic theory for definition of the perceptual terminus and has sought to validate postulated theoretical units empirically. In this, students have had the support of linguists. Jakobson and Halle, for example, emphasized the "immanent approach which locates the distinctive features and their bundles within the speech sounds, be it on their motor, acoustical, or auditory level" (1956:8). More recently, Halle (1964) has implied that universal phonetic features may be grounded in man's innate auditory capacities, while Chomsky and Halle take phonetic features to be "identical with the set of properties that can in principle be controlled in speech" (1968:295) and assume each feature to have (presumably discoverable) "acoustical and perceptual correlates" (1968:299).

We are not, however, entirely at the mercy of theory, nor even of possibly illusory perception, in our choice of perceptual units. If we are willing to make the assumption that perceptual units are isomorphic with production units, we have in errors of speech a natural body of materials from which to infer segments. Any unit subject to errors of metathesis (Spoonerism), substitution, or omission must be under some degree of independent control in production. Fromkin (1970) has analyzed six hundred errors collected by herself and her colleagues over three years. The observed units of error pertinent to this discussion were: syllables (e.g., "butterpillar and catterfly"), phone-length segments (e.g., "the niper is jarrow"), and features (e.g., "tebestrian" for "pedestrian"). (Interestingly, she found no metathesis of consonants and vowels: phone-length metathesis across syllable boundaries always involved exchange of segments having similar functions within the syllable.) Fromkin

---

14 For experiments on and discussions of segmentation, see Harris (1953); Lisker (1957); Wang and Peterson (1958); Peterson et al. (1958); Lisker et al. (1962); Ladefoged (1967).
concludes that an adequate model of speech performance must include mechanisms for producing such errors. We may say the same, mutatis mutandis, of an adequate model of speech perception.

Each unit mentioned has been validated in perception. The feature is the least intuitively obvious segment and has received most experimental attention. For the syllable, we have already cited Liberman et al. (1952) and Ohman (1966) (p. 22). We may add, from among many, the series of experiments reported by Kozhevnikov and Chistovich (1965: Ch. VI) and an experiment of Huggins (1964), in which he alternated speech rapidly between ears and found that the rate most disruptive to speech perception was close to the syllable rate. For the phonetic segment, Pike (1943: Ch. VII) provided cogent arguments, observing, for example, that phonetic transcriptions of experts, and even of those with little training, generally agree on the total number of segments in an utterance. Fry, also, remarked that "the existence and widespread use of alphabetic writing are an indication that a phonemic system and segmentation into phonemic units are features which find a ready response in speakers and listeners" (Fry, 1964:60). We may add the experimental evidence of Kozhevnikov and Chistovich (1965: 217 ff.), who found that mean reaction time for transcribing the consonant from a spoken CV syllable could be as much as 100 msec less than for transcribing the vowel from the same syllable. The same result was reported by Savin and Bever (1970). There is also evidence for the phoneme as an encoding unit in short-term memory (Conrad, 1964; Wickelgren, 1966b; Sperling and Speelman, 1970).

Savin and Bever (1970), as Warren (in press), made another interesting observation: subjects responded consistently faster to syllable targets than to phoneme targets.\footnote{Kolers (1968b) reports similar results for tachistoscopically exposed words and the letters that compose them.} They concluded that "phonemes are identified only after some larger linguistic sequence of which they are parts" (p. 300).\footnote{Ladefoged (1967:147) has developed a similar argument, drawing an unfortunate analogy with typing. He points out that skilled typists type by the word, not by the letter. Certainly, skilled typing may be governed by an hierarchical system of temporo-spatial coordination that includes integrating commands for sequences of letters as units [MacNeilage (1964) has an elegant discussion of these matters]. But it is evident that the typist does type letter by letter and that his behavior, not to mention the typewriter, would quickly jam, if he did not.} They explain that phonemes are primarily "neither perceptual nor articulatory," but rather "psychological entities of a non-sensory, non-motor kind... in short, phonemes are abstract" (p. 301). Without entering into discussion of the boundaries between sensation and perception, we may agree with their last remark since, as earlier observed, even phonetic perception is abstract. But if this distinction is to explain the longer latency for phoneme than for syllable identification, we must infer that syllables are not abstract. Certainly, as we argue below, syllables (unlike phonetic segments and features) may exist as articulatory, acoustic, and auditory units. But, insofar as they are phonetic units, they too are abstract. How an entity (whether concrete or abstract) of which the existence and form are determined by discrete components can be perceived without prior extraction of at least some of those components is hard to imagine. But it is not hard to imagine that the extraction of these
components is normally so rapid, automatic, and unconscious that their conscious
recovery is slow. In fact, this may also be true of syllables in running
speech: one would not be surprised to learn that recognition of syllables
took longer than recognition of the words that they compose. Differences in
recovery time for the several phonemes of Savin and Bever remain, of course,
to be explained. In any event, their study has added evidence for the psycho-
logical and, in our view, the perceptual reality of the phonetic segment.

Finally, for the perceptual reality of features below the level of the
phonetic segment, a large body of experimental evidence exists. First, virtu-
ally all studies of synthetic speech continua (many were cited in the preceding
section) in which a "phoneme boundary" is observed may be regarded as studies of
feature boundaries, since segments on either side of the boundary differ by
a single articulatory feature. Second, perceptual confusions among consonants
or vowels heard under difficult listening conditions (through noise, through
filters, or under dichotic competition) group themselves systematically: the
more feature values two segments have in common, the more likely they are to
be confused. This has been shown for consonants (Miller and Nicely, 1955;
Singh, 1966, 1969; Studdert-Kennedy and Shankweiler, 1970) and for vowels
(Miller, 1956; Singh and Woods, 1970). In several of these experiments, fea-
tures were shown to be approximately additive (independent). A third line of
evidence comes from scaling studies. Hanson (1967), for example, used multi-
dimensional scaling techniques to place nine synthetic Swedish vowels in a
psychological space which proved to have two dimensions corresponding to the
tonal features (grave/acute; compact/diffuse) of Jakobson et al. (1963) and
a third dimension corresponding to no defined feature [cf., the three dimensions
of Pols, van der Kamp and Plomp (1969) in their study of Dutch vowels]. For
scaling the six stop consonants of English, Greenberg and Jenkins (1964) used
magnitude estimation and found, among other results, that sounds differing
on one feature were judged to be closer than sounds differing on two. Peters
(1963) found that manner, voicing, and place of articulation (in this order)
were the main determinants of similarity judgments among consonants. There
is also evidence for encoding of both consonants and vowels in short-term
memory according to phonological features (Wickelgren, 1965, 1966a, 1969;
Sales et al., 1969, and four previous papers by these authors, cited therein).
Finally, we note that several of these studies evaluated different sets of
features and their definitional terms (articulatory, auditory). Klatt (1968)
has presented a quantitative method for evaluating binary features from con-
fusion matrices and for estimating their independence. For the present dis-
cussion, however, it is enough to know that some set of features functions in
perception.

Let us now return to the theme by recalling that, despite their perceptual
reality, neither phonetic segments nor their component features have acoustic

---

17 For consonants, filtering typically tends to damage place of articulation;
reverberation or echo effects tend to damage manner; noise damages both
manner and place (Fant et al., 1966).

18 Ladefoged (1969) was able to predict, with almost perfect accuracy, speaker
judgments of articulatory similarities among thirty words in each of twenty
Bantu languages by counting numbers of shared feature values on an ad hoc
set of binary features.
reality. Our task is therefore to understand how these abstract (physically nonexistent) entities achieve psychological reality. The syllable has a different status, since we may define it not only linguistically but also in articulatory and acoustic terms. Whatever the difficulties of defining its boundaries acoustically [see Malmberg (1955) for one of the few attempts], its general function in production as a carrier of phonetic information is fairly clear. The syllable is the unit of consonant/vowel coarticulation: it arises from imposition of a precisely timed and coordinated pattern of articulatory gestures (the multiple physical manifestations of phonetic features) upon a pulse of air. As the word itself indicates, the speaker collapses discrete muscular movements into a pattern of overlap that forms a larger unit. For perception, the new unit has a double function. First, it reduces the number of auditory segments emitted per unit time below the number of phonetic segments and so brings segment repetition rate within the temporal resolving power of the ear. This function has been discussed elsewhere (Studdert-Kennedy and Liberman, 1963; Studdert-Kennedy and Cooper, 1966; Liberman et al., 1967b). Second, and we dwell on this here, its function is to contrast, and so to permit the listener to detect, segments of sound.

Fant and his colleagues (Fant and Lindblom, 1961; Fant, 1962) are among the few researchers to recognize the importance of sound contrast for phonetic perception. Dissatisfied by the abstract nature of distinctive features, and by the difficulty of specifying their acoustic correlates, Fant undertook to work upwards from signal to message rather than downwards from message to signal. He has developed a system for describing spectrograms in terms of sound segments with boundaries determined by switching events in the speech production mechanism. He decomposes sound segments into sound features specifying production and speech wave characteristics for each. He carefully reiterates that neither sound segments nor sound features are isomorphic with phonetic segments or features. For a recent account, see Fant (1968:235-241).

The importance of Fant's work is that, by systematic analysis of the acoustic signal, it provides an objective account of the factors with which the perceptual mechanism has to work. His system permits precise division of the spectrographic pattern in frequency and in time and invites exploration of the perceptual importance of its segments by filtering and gating techniques (e.g., Öhman 1961a, b; Fant, et al., 1966). Such studies may correct or corroborate work with synthetic speech, which, Fant (1964) believes, tends to overemphasize single cues at the expense of the entire auditory pattern.

Returning now to the syllable as vehicle of sound contrast, we may illustrate with part of a study by Bondarko (1969). Her intent was to examine "the means by which [phonemes] are contrasted within the syllable" (1969:2),

---

19 We are aware that the work of Ladefoged and his colleagues (Ladefoged 1967: Ch. I) forces modification of Stetson's (1951) chest pulse theory of the syllable. But the disagreement is over the physiological control mechanism not over the function of the syllable in production and perception.

20 The linguistic function of the syllable, as carrier of stress, is incidental to its coarticulatory origin and can be as well performed by an isolated vowel or, in some circumstances, consonant.
in other words, to search for the acoustic basis of distinctive feature oppositions. But we need not accept the theoretical framework to be interested by her study. Adopting an approach reminiscent of Fant's, she defined five types of contrast that may occur between sound segments in Russian CV syllables: contrasts in fundamental frequency, duration, formant structure, intensity, and "locus" (F2 transition). She then examined spectrographically twenty-five syllable types [five consonant classes—voiced and voiceless stops and fricatives, and sonants (lateralis, nasals)—followed by five vowels]. She classified each syllable type according to how many of her contrast types it carried: the results ranged from all five for voiceless stops to one (or two) for laterals and nasals. The recorded syllables were gathered into random test orders and presented to twenty-five subjects for identification. Many details are omitted from the present account (in particular, the different values of stress used), but the general outcome was clear: probability of correct identification declined as intrasyllable contrast declined.

The portion of the syllable most likely to be missed with decline in contrast was the vowel. This is not unexpected, since we know vowel recognition to be heavily dependent on acoustic context, although most studies have given their attention to effects over signal stretches longer than the syllable (e.g., Ladefoged and Broadbent, 1957; Fry et al., 1962; Hiki et al., 1968). Among studies of effects within the syllable is one by Fujimura and Ochiai (1963), who compared identification of Japanese vowels spoken in syllabic context with identification of 50 msec segments gated out of the vowel centers: identifications shifted in the absence of surrounding formant movements. In a related study with synthetic vowels, Lindblom and Studdert-Kennedy (1967) showed that identification of a particular vowel pattern varied as a function of the rate and direction of its surrounding formant transitions.

An attempt not simply to demonstrate, but to watch the development of, contrast within the syllable is made by the shadowing studies of Kozhevnikov and Chistovich (1965). An experimenter reads over a microphone a list of VCV (or CV) patterns with V fixed and C a stop consonant that varies from trial to trial. A listener, in another room, repeats each utterance as rapidly as possible. Contact electrodes (lips, artificial palate) and throat microphones provide oscillographic records of the two speakers' utterances. If VCV patterns are being used, it is found that soon after the first speaker releases the consonant, the shadower's articulators constrict. The place of constriction may be more or less random, but, as the speaker continues, the shadower adjusts the point of constriction, if necessary, and completes his gesture. The latency of the shadower's release is 100-150 msec from the time of the speaker's release or from the time that the shadower's articulators assume the correct point of constriction. This delay is far too short for a normal choice reaction time. Kozhevnikov and Chistovich argue that shadowing shortcuts higher-level processes to reveal the normal, involuntary sequence of states in phonetic perception. Each state is said to be a function of both the preceding one and of some change in the external signal (p. 231). Since the first nonrandom state must be a function of the external signal, the entire sequence of phonetic states is a function of the external sequence. This sequence takes phonetic effect through acoustic contrast between its segments.

In short, experiments support a view of the syllable as carrier of contrast. But the contrast is not, as distinctive feature theory might have it,
between linguistic features in their acoustic manifestations: we have seen that there are few, if any, acoustic segments isomorphic with linguistic features. We must, therefore, read the epigraph to this chapter in a sense different from that intended by its authors: "for the study of speech sounds ... their linguistic function is decisive" (Jakobson et al., 1963:11). One linguistic function of the syllable is to provide a rhythmic acoustic signal within the temporal resolving power of the ear and to facilitate, by its inherent acoustic contrasts, exercise of the listener's capacities for auditory discrimination. Without such a signal, linguistic communication would not be possible.

We conclude, then, that, while study of the acoustic signal may lend insight into its auditory function, it will not lead us appreciably closer to an understanding of how auditory patterns are related to phonetic matrix. For this, we must start from the known message and examine its manifestation in the signal.

THE PERCEPTUAL PHONETIC MATRIX

A sizeable body of knowledge exists concerning acoustic cues for phonetic segments and their features (see earlier citations, and Mattingly, this volume). The relations are not one-to-one and, from an acoustic point of view, seem arbitrary. Some features are signaled by more than one cue; some cues signal more than one feature. A brief explosion, for example, may indicate both voicing and place of articulation in final stops. In initial stops, voicing may be cued by explosion energy, degree of aspiration, and first formant intensity. Each cue may be emphasized in synthetic speech an used as the principal cue to voicing (Liberman et al., 1952, 1961b). In natural speech multiple cues, scattered within the syllable over time and frequency, combine, and according to synthesis experiments, their perceptual weights vary with phonetic context (Hoffman, 1958; Ainsworth, 1968). But we may make sense of their arbitrary nature and varying perceptual weights by applying the acoustic theory of speech production. The relations between source, vocal tract transfer function, and signal are well understood (Fant, 1960; Flanagan, 1965), and for any apparently arbitrary collocation of acoustic events, we may specify the conditions of production. The conditions of production themselves, however, remain unexplained until we can relate them to their underlying phonetic (articulatory) features.

For the feature of voicing, research is approaching this level of explanation. In 1960, Fant suggested that the main factor underlying the voiced/voiceless distinction for stop consonants in initial position was "the instant of time when the vocal cords close for the production of the following voiced sound" (Fant, 1960:225). Lisker and Abramson, by spectrographic analysis of stops in eleven languages and by perceptual experiments with synthetic speech in three, have explicated Fant's suggestion (Lisker and Abramson, 1964a, b, 1967, 1970; Abramson and Lisker, 1965, 1970; also Lisker et al., 1969). Their work suggests that the disparate acoustic features of explosion energy, aspiration, and first-formant intensity may all be derived from the single, underlying articulatory variable of voice onset time, that is, the relative timing of closure release and the onset of laryngeal vibration.20

20A fourth cue, rapid pitch changes at the onset of voicing, though probably trivial in natural speech, may be deliberately exaggerated in synthetic speech
They write:

Laryngeal vibration provides the periodic or quasi-periodic carrier that we call voicing. Voicing yields harmonic excitation of a low frequency band during closure, and of the full formant pattern after release of the stop. Should the onset of voicing be delayed until some time after the release, however, there will be an interval between release and voicing onset when the relatively unimpeded air rushing through the glottis will provide the turbulent excitation of a voiceless carrier commonly called aspiration. This aspiration is accompanied by considerable attenuation of the first formant, an effect presumably to be ascribed to the presence of the tracheal tube below the open glottis. Finally, the intensity of the burst, that is, the transient shock excitation of the oral cavity upon release of the stop, may vary depending on the pressures developed behind the stop closure where such pressures will in turn be affected by the phasing of laryngeal closure. Thus it seems reasonable to us to suppose that all these acoustic features, despite their physical dissimilarities, can be ascribed ultimately to actions of the laryngeal mechanisms. (Abramson and Lisker, 1965)

We have quoted this account at length because it provides a model for the reduction of an apparently incoherent set of acoustic cues to an underlying articulatory variable or phonetic feature. How far this approach may be carried with other features remains to be seen. Also open for the future is the degree to which this and other approaches in experimental phonetics may force a modification of the phonetic features posited by phonological theory, if that theory is to be given a physical base. Ladefoged (1966) argued that the list of distinctive features was already overextended at fifteen, largely because the specifications disregarded physiological constraints on their combination.

Our argument then is that only through their articulatory origin can the temporally scattered and contextually variable acoustic (and auditory) patterns of speech be understood. The listener, who begins life as a babbler, develops, by repeated association of articulatory controls with their auditory consequences, a "knowledge" of his phonetic capacity and of the use to which it is put in his native language (Weir, 1962). By imitation of the voices around him and through adult acceptance of his imitations, he learns the relation between his own acoustic output and that of others who have larger vocal tracts. Thus, he learns to "infer" from phonetically adventitious components of a speaker's signal (such as overall fundamental frequency or frequency position of the third formant) characteristics of the tract that produced it and the instructions required by his own tract for a "matching" signal. There are even grounds for suspecting that he may be born with some "knowledge" of phonetic capacity. Lisker and Abramson (1964a, b) found that nine languages, some with two, some with three stop

and will then serve as an effective cue (Haggard et al., 1970). The relatively small pitch changes of natural speech are probably also attributable to voice onset time.
categories, implemented only three values of voice onset time and suggested that physiological constraints may underlie this nonrandom distribution. Recently, Eimas and his colleagues (Eimas et al., 1971) tested discrimination of the labial voice onset time continuum in one-month-old infants by tracing adaptation and recovery of sucking responses. The infants showed significantly higher discrimination of a 20 msec difference in voice onset time that straddled two of Lisker and Abramson's phonetic categories than of the same difference within a category.

But whatever its source, "knowledge" of the relations between auditory patterns and articulatory features is available to every speaker/listener, and through this knowledge, we hypothesize, he is able to resolve an intricate pattern into its simple origin. In light of our earlier discussion, we must suppose the resolution to be automatic and, probably, beyond conscious recovery. Precisely how it may be accomplished we will not here speculate. But the general argument is not new. Similar positions have been adopted by Stevens and by Liberman and their colleagues (Halle and Stevens, 1962; Stevens and Halle, 1967; Stevens and House, 1970, Liberman et al., 1967b). Chistovich and her co-workers (Kozhevnikov and Chistovich, 1965; Chistovich et al., 1968) have applied the model experimentally. Their shadowing studies bear on perception only if we take stage 2 (phonetic) to be an automatic, running analysis, with its final output an assemblage of segments and features that may serve, in production, as instructions to the articulatory mechanism and, in perception, as input to the phonological component. In short, the perceptual matrix is identical with the generative. Its columns (phonetic segments) and rows (phonetic features), although determined by man's physiological capacities, are not themselves part of his auditory and articulatory systems. Rather, they are abstract linguistic entities uniting the complementary communicative functions of speaking and listening.

The model is broad, stripped of detail that might invite experimental test. Until our knowledge of perceptual processes and their physiological correlates has vastly increased, it is likely to remain so: less empirical than protreptic.

**NEURAL SPECIALIZATION FOR SPEECH PERCEPTION**

Cats can discriminate between speech sounds. Dewson (1964) trained five cats, by operant procedures, to discriminate between sustained [u] and [i], spoken by a man (F0 = 136 Hz), and to transfer their learning in one fifth the number of original trials to the same vowels spoken by a woman (F0 = 219 Hz). Warfield et al., (1966) trained ten cats to discriminate between the words "cat" and "bat," spoken by a woman, and demonstrated by a gating procedure that discrimination was based on the initial portion of each syllable. The cats did not transfer their learning to other words beginning with the same phonetic segments.

We may doubt that cats can perceive speech. This is not simply because they do not know a language, but also because they are not physiologically equipped to do so. For over a century, evidence has been accumulating that the human brain is asymmetrically organized for language functions. Recently, experiments have demonstrated that "dominance" of one or the other of the cerebral hemispheres (usually the left) extends to mechanisms for the perception of speech.
Kimura (1961a) discovered that listeners, presented with triads of competing digits in opposite ears (i.e., dichotically), were better able to recall those presented to the right ear than those presented to the left. She attributed the effect to functional prepotency of contralateral over ipsilateral auditory pathways and to cerebral dominance for language (Kimura, 1961b); her interpretation has since been supported by many studies. Shankweiler and Studdert-Kennedy (1966, 1967) showed that the right-ear advantage for speech did not depend on meaning, since it could be obtained if subjects were asked to identify contrasting consonants in pairs of dichotically presented, synthetic, CV nonsense syllables. There was no right-ear advantage if the competing stimuli were vowels. A later study with natural, CVC nonsense syllables (Studdert-Kennedy and Shankweiler, 1970) confirmed these results and showed, by error analysis, right-ear advantages for voicing and place of articulation in stop consonants, a result confirmed by Haggard (1970). Other work has demonstrated specialization of the nonlanguage hemisphere for recognition and discrimination of nonspeech auditory patterns (Miller, 1962; Kimura, 1964, 1967; Benton, 1965; Chaney and Webster, 1965; Shankweiler, 1966; Curry, 1967; Vignolo, 1969; Darwin, 1969).

Models of the neural mechanisms underlying these effects are still fluid, and no firm account will be offered here.21 We will, however, assume that ear advantages reflect a degree of functional cerebral asymmetry. Cerebral asymmetry, or dominance, for speech perception requires that some portion of the perceptual function be performed more efficiently by the dominant hemisphere. One aim of current dichotic speech research is to define that portion by determining the acoustic and psychological conditions of the right-ear advantage.

There are three broad possibilities. The dominant hemisphere may be specialized for: 1) response alone, 2) phonetic analysis and response, or 3) auditory analysis, phonetic analysis, and response. The first possibility was tentatively ruled out by Studdert-Kennedy and Shankweiler (1970); subjects' error patterns indicated that place of articulation and voicing features in stop consonants were independently extracted by a single center in the dominant hemisphere. Darwin (1971) also concluded that specialization was not simply for response.

Darwin pushed the analysis further by showing that, in synthetic fricatives, the right-ear advantage for place of articulation only occurs if the feature is signaled by a formant transition, while for voicing, the advantage occurs only if the fricative is followed by a vowel. This suggests that the dominant hemisphere may be superior at some stage of auditory analysis, being better equipped, perhaps, for detection of certain acoustic features of speech, such as rapid formant movement or voice onset.

An alternative, though not incompatible, interpretation is that the dominant hemisphere is skilled at extracting phonetic information under "difficult" conditions, such as those provided by the complex auditory pattern of coarticulated consonant and vowel. This interpretation meshes

---

with the lack of a reliable ear advantage for vowels and with the further fact that, if vowels are presented under conditions of general phonetic ambiguity, a right-ear advantage appears. Darwin (1971) demonstrated an advantage for dichotically competing, synthetic vowels, if the test included utterances apparently formed by different-sized vocal tracts: the listener was then in doubt as to which vocal tract would be presented to a particular ear on any trial. But if all vowels on the test sounded to have come from the same vocal tract, no right-ear advantage emerged.

There are parallels here with earlier studies. Continuous perception of vowels gives way to categorical perception, if attention is forced to phonetic qualities. The characteristic speech mode of judging loudness, in terms of physiological effort required for phonation, appears if signals are presented in a context that demands phonetic processing. The characteristic speech skill of the dominant hemisphere is evinced in perception of rapid or otherwise "difficult" auditory/phonetic patterns. In other words, the evidence is consistent with the view that the language-dominant hemisphere is superior to the minor hemisphere in its capacity to accomplish the phonetic analysis of stage 2.

However, the peculiarity of speech perception cannot be solely in the use to which we put an acoustic signal. What engages the phonetic processors? Most nonspeech sounds cannot be heard as speech. But to Tennyson's farmer, the pony's hooves sang "Property, property, property," and we hear the chaffinch call "chewink," despite his lack of formant structure (Thorpe, 1961). A nonspeech sound with rapid acoustic variations that suggest the fundamental consonant/vowel alternations of speech may, if other conditions dispose, engage the phonetic processors. But the specifications are vague. Research has only begun (Darwin, 1969) to exploit the double dissociation of left and right hemispheres, speech and nonspeech signals, as the thin end of an experimental wedge for separating signals that are, or can be, heard as speech from those which cannot and for defining their characteristics. The acoustic border beyond which phonetic, and perhaps specialized auditory, processors are involuntarily engaged is still undefined.

Finally, the discovery that phonetic perception is neurologically linked to language processes emphasizes the unity of language and its medium of expression. Despite the contrary views of some phoneticians and linguists, language is not independent of its medium. Language, as we know it, could not exist without speech, nor speech without sound. Mattingly and Liberman (1970) explore formal analogies between phonetic and syntactic structure and suggest that study of speech perception may throw light on more general linguistic processes. At the same time, it may ground man's characteristic mode of communication in his physiology: to study speech is to enlarge our understanding of the biological foundations of language.

REFERENCES


cluster. JAcS. 48.95(A).

Day, Ruth S. and James E. Cutting. 1970a. Levels of processing in speech
perception. Paper presented at 10th Annual Meeting of Psychonomic
Society, San Antonio.

speech and nonspeech. Paper presented at 80th meeting of Acoustical
Society of America, Houston.

Acoustic loci and transitional cues for consonants. JAcS. 27.769-773.

Delattre, Pierre C., Alvin M. Liberman, Franklin S. Cooper, and Louis J.
Gerstman. 1952. An experimental study of the acoustic determinants
of vowel color: Observations one- and two-formant vowels synthesized

555-556.

Eimas, Peter D. 1963. The relation between identification and discrimination

Eimas, Peter D., Einar R. Siqueland, Peter Jusczyk, and James Vigorito. 1971.

Mouton.

Fant, C. Gunnar M. 1962. Descriptive analysis of the acoustic aspects of
speech. Logos 5.3-17.

Laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR3.
16-20.

Fant, C. Gunnar M. 1966. A note on vocal tract size factors and non-uniform
F-pattern scalings. Speech Transmission Laboratory, Royal Institute of

Fant, C. Gunnar M. 1968. Analysis and synthesis of speech processes. In

Fant, C. Gunnar M. and Bjorn E.F. Lindblom. 1961. Studies of minimal speech
sound units. Speech Transmission Laboratory, Royal Institute of Technology,
Stockholm, Sweden. QPSR2.

Fant, C. Gunnar M., Bjorn E.F. Lindblom, and A. De Serpa-Leitão. 1966. Con-
sonant confusions in English and Swedish. Speech Transmission Laboratory,
Royal Institute of Technology, Stockholm, Sweden. QPSR4.


Podor, Jerry A. and Thomas G. Bever. 1965. The psychological reality of

Fourcin, Adrian J. 1968. Speech source inference. IEEE Transactions on Audio-

Frishkopf, Larry S. and Moise H. Goldstein, Jr. 1963. Responses to acoustic
stimuli from single units in the eighth nerve of the bullfrog. JAcS. 35.
1219-1228.


Fromkin, Victoria A. 1970. Tips of the tongue or to err is human. Univ. of

Fry, Dennis B. 1955. Duration and intensity as physical correlates of lin-
guistic stress. JAcS. 27.765-768.


Harris, Cyril M. 1953. A study of the building blocks of speech. JAcS. 25. 962-969.
Harris, Katherine S. Physiological aspects of articulatory behavior. See this volume.


Ladefoged, Peter and Donald E. Broadbent. 1957. Information conveyed by vowels. JAcS. 29.98-104.


Liberman, Alvin M. 1957. Some results of research on speech perception. JAcS. 29.117-123.


Liberman, Alvin M., Pierre C. Delattre, and Franklin S. Cooper. 1952. The role of selected stimulus variables in the perception of the unvoiced stop consonants. Amer. J. Psychol. 65.497-516.

Liberman, Alvin M., Pierre C. Delattre, Franklin S. Cooper, and Louis J. Gerstman. 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monograph. 68 (whole number).


Liberman, Alvin M., Katherine S. Harris, Howard S. Hoffman, and Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54.358-368.

Liberman, Alvin M., Katherine S. Harris, Joanne Kinney, and Harlan Lane. 1961b. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. J. Exp. Psychol. 61.379-388.


MacNeilage, Peter F. 1964. Typing errors as clues to serial order mechanisms in language behavior. L & S. 7:144-159.


Mattingly, Ignatius G. Speech synthesis for phonetic and phonological models. See this volume.


Miller, George A. 1964. Some psychological studies of grammar. Amer. Psychol. 17.748-762.


Pickett, James M. and Irwin Pollack. 1963. Adjacent context and the intelligibility of words excised from fluent speech. JAcS. 35.807(A).


Savin, Harris B. 1963. Word-frequency effect and errors in the perception of speech. JAcS. 35.200-206.
Schatz, Carol. 1954. The role of context in the perception of stops. Language 30.47-56.
Sparks, Robert and Norman Geschwind. 1968. Dichotic listening in man after section of neocortical commissures. Cortex 4.3-16.
Stevens, Kenneth N. 1968a. Acoustic correlates of place of articulation for stop and fricative consonants. QPR. 89. Research Laboratory of Electronics, M.I.T., 199-205.


