THE COGNITIVE REPRESENTATION OF SPEECH
T. Myers, J. Laver, J. Anderson (editors)

Perceiving Phonetic Segments

MICHAEL STUDDERT-KENNEDY
City University of New York and Haskins Laboratories

Why do we study speech perception? Is the speech signal merely a complex acoustic structure in which we are interested because we happen to use it for communication? Certainly, speech belongs to a natural class of acoustic patterns about which we know very little, namely, patterns structured over time by mechanical events—such as a footstep on gravel, a hand crumpling paper, a glass bottle bouncing or breaking (Warren, in preparation). If we knew more about the temporally and spectrally distributed acoustic properties of such dynamic events, and how we recognize them, would we then understand how we perceive speech? In other words, is speech merely a distinctive set of sounds, shaped, as no other sounds are, by movements of human articulators modulating a characteristic vocal source? Perhaps.

But there are reasons to believe that speech may also be distinctive in some deeper sense than this, and that it may engage distinctive perceptual mechanisms. First, the speech signal is the primary carrier of a language: it is rich in information not only about the abstract phonological structure of an utterance, but also about its syntactic and semantic structure. To perceive speech is to apprehend this structure, an act that can only be accomplished by a listener who knows a language. Just what the listener knows is, of course, very much the question. But, he must, at least, know the abstract linguistic units that compose the message—whether words, morphemes or phonemes—and he must know how such units are realized in the signal. No other natural sound, so far as we know, conveys so complex an abstract message.

A second and simpler reason for suspecting that the speech signal may be uniquely related to the human perceptual apparatus is that each speaker of a particular language can readily reproduce sounds uttered by another. Moreover, he can repeat the sounds so rapidly that we may reasonably hypothesize a privileged relation between elements of perception and elements of production. That the repetition may be far from an acoustically exact imitation, and yet be perceived as a repetition, only strengthens the hypothesis by emphasizing that more than mere acoustic identity is involved.

We should not lightly dismiss the ability to repeat or imitate. Imitation is, in fact, a remarkable skill that requires an animal to parse the behavior of another into components, and then activate its own corresponding motor controls to reproduce the behavior. A general capacity to imitate has evolved only in social animals—for the most part, in certain primates: its adaptive advantages are obvious in, for example, the well-known sweet potato washing of Japanese macaques (Wilson, 1975, pp. 170-171) or the nest-building of young chimpanzees (van Lawick-Goodall, 1971). Vocal imitation has evolved only in certain species of birds and in man: its adaptive advantages can only lie in communication. Moreover, transformation from sensory input to corresponding motor control is less direct for an acoustic

signal, shaped by movements within the internal space of a vocal tract, than for an optical signal, shaped by movements in the external space common to observer and observed. Perhaps we should not be surprised that the human infant begins its imitative attempts by tracking the visually available movements of its caretaker's mouth in "prespeech" oral play (Trevarthen, Hubley and Sheeran, 1975). In any event, the human capacity to imitate the speech of another implies a specialized sensorimotor device (perhaps analogous to that being discovered in the canary (Nottebohm, 1977) and suggests that phonetic structure may be represented in the brain in a form sufficiently abstract for there to be ready interchange between listening and speaking. Whether this representation is necessarily elicited during normal speech perception is, of course, an open question.

From this introduction, I wish to turn to the question of how listeners parse an utterance into its component linguistic segments. Whether the segments are words, morphs, syllables or phones need not concern us, for the moment, since once we have accepted the onus of describing the relation between linguistic segments and the acoustic signal, the problem is essentially the same. However, I shall concentrate on the phone, and its constituent cues or features, for several reasons, more fully discussed elsewhere (Liberman and Studdert-Kennedy, 1978). First, I assume the abstract elements of phoneme or feature that underlie the phonetic string to be psychologically real, structural units of lexical and grammatical morphemes, essential to the dual structure that makes linguistic communication possible. Second, a solution to the segmentation problem at the phonetic level should lead toward a solution at higher levels. In fact, we already have some evidence that English word juncture is signalled by allophonic variations in the immediate vicinity of the juncture, that these variations occur at onset rather than offset of words (Nakatani and Dukes, 1977) and that acoustic-phonetic structure at word onset is particularly important for lexical access in fluent speech (Marslen-Wilson and Welsh, 1978). Finally, the speed and efficiency of speech as a medium of communication rests, in large part, on coarticulation among phonetic segments, and this coarticulation is the primary source of the segmentation problem (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967).

The trouble began with the spectrogram and with the recognition that, as the cliché has it, phones are not strung together like beads on a string, but rather, as Hockett (1958) wrote, like a line of eggs passed between rollers. Historically, there have been two broad responses to the paradox. The first has been to accept the segments of linguistic analysis—whether the features of distinctive feature theory or the static targets of traditional articulatory phonetics—as psychologically real, and then either to attempt an acoustic formulation of the signal isomorphic with those segments or to posit some specialized mechanism for linking signal and message. The second response has been frankly pragmatic, proposing to finesse the linguist's units as little more than analytically useful fictions (perhaps encouraged by the historical "accident" of the alphabet) and to go for supposedly simpler groupings such as phoneme dyads, syllables, or even words.

The first line of response has been the more varied and complex of the two, and will be the focus of most of what I have to say. An important contributor here was Fant who early recognized the disparity between signal and message, remarking that a single segment of sound may convey information concerning several segments of the message, and a single segment of the message may draw information from several segments of the sound (Fant, 1962). Moreover, despite the claim of distinctive feature theory that correlates of the features are to be found at every level of the speech process (articulatory, acoustic, auditory), Fant observed 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, p. 94). He and his colleagues therefore set about describing the actual segments of sound observable in a spectrogram, the presumed elements with which the perceptual system has to work. They developed a terminology for describing these segments as manner or segment type features (e.g., voicing, plosive release, frication, nasal resonance and so on) and place features specified in terms of formant pattern and energy distribution over frequency and time (Fant, 1968). On two general points, central to Fant's position, there now seems to be a growing consensus: first, one-to-one correspondence does not obtain between features of the signal and features of the message (cf. Parker, 1977, Ganong, in press); second, the signal takes effect through its entire spectral array rather than through the pattern of individual formants. However, Fant's approach has not yielded a solution to the segmentation puzzle.

An emphasis on the entire spectral array has also characterized the recent work of Stevens (e.g. 1972, 1975). Stevens has confronted the problem head-on by undertaking to reformulate the acoustic description of the signal to make it isomorphic with the message (that is, the distinctive feature). He has adopted an explicitly evolutionary approach to the presumed link between production and perception. According to Stevens' (1972) quantal theory, phonetic categories have come to occupy those acoustic spaces where (by calculations from a vocal tract model) large articulatory variations have little acoustic effect, and to be bounded by those regions where small articulatory changes have a large acoustic effect. As a simple example, consider the articulatory-acoustic series that carries the speaker-listener from a high front vowel [i], through an alveolar fricative [s], to an alveolar stop [d]. However, most of Stevens' recent empirical work has concentrated on stop consonants. For example, Blumstein and Stevens (1979) derive three invariant spectral templates, determined by integration over a 26 msec. window at the onset of the three voiced stops, /b,d,g/, spoken by two male speakers before five representative vowels. Their terminology explicitly recalls distinctive feature theory: "diffuse-rising" for alveolar, "diffuse-falling" for labial and "compact" for velar. To match these acoustic properties, and as a step toward solving both the invariance and the segmentation problem, Stevens (1975) posits innate property detecting devices by means of which the infant is presumed to latch onto the speech system. We should note that these devices are conceived by Stevens as general to the mammalian auditory system rather than as tuned specifically to speech; in this they differ from the feature detectors to be discussed below.

Nonetheless, there are difficulties with Stevens' approach. First, the proposed stop-consonant templates invoke a fixed "locus," incompatible with the findings of Delattre, Liberman and Cooper (1955). Thus, the supposedly invariant properties are static rather than dynamic and the role of their matching detectors is essentially to filter out "irrelevant" variation. Yet there is a mass of empirical data to demonstrate that the perceptual system is not only sensitive to the very information that these detectors are designed to exclude, but also uses it to reach phonetic decisions (see Studdert-Kennedy, 1976 for a review). Second, the proposed property detectors are currently confined to consonant place of articulation, and there is reason to believe that an attempt to specify analogous detectors for, say, consonant voicing or vowel features would run foul of the many acoustic variations induced by phonetic context, rate and speaker. Finally, no integrative mechanism is proposed for aligning the features in the rows of the phonetic matrix within their appropriate segmental columns.

A third approach within this line of response has been that of researchers at Haskins Laboratories. They too have taken the goal of research in speech perception to be specification of relations between signal structure and linguistic units—but the units of traditional articulatory phonetics rather than of distinctive feature theory.

They too early recognized the problem of acoustic segmentation, remarking at the conclusion of a study demonstrating contextual dependencies among plosive release bursts and following vowel: ". . .the irreducible acoustic stimulus is the sound pattern corresponding to the consonant-vowel syllable" (Liberman, Delattre and Cooper, 1952, 516). However, they have been more atomistic in their approach, searching the signal for "minimal cues" (Liberman, Ingemann, Lisker, Delattre and Cooper, 1959) to phonetic contrasts rather than characterizing the entire spectrum.

Two broad lines of research bearing on the segmentation issue have stemmed from this approach: one, deriving from a seemingly plausible account of categorical perception, has issued in recent work on feature detectors; the other, pursuing multiple acoustic cues, has sought to explain how they are integrated within the phonetic segment. Let us consider each in turn.

Early work with speech synthesizers showed that a useful procedure for defining the acoustic properties of a phoneme was to construct tokens of opponent categories, distinguished on a single phonetic feature, by varying a single acoustic parameter along a continuum (e.g., /ba/ to /da/, /da/ to /ta/, etc.). If listeners were asked to identify these tokens, they tended to identify any particular stimulus in the same way every time they heard it: there were few ambiguous tokens. Moreover, when asked to discriminate between neighboring tokens, listeners tended to do badly, if they assigned them to the same phonetic category, well if they assigned them to different categories, even though the acoustic distance between tokens was identical in the two cases. This phenomenon was dubbed "categorical perception" (Liberman, Harris, Hoffman and Griffith, 1957). Recent work has demonstrated that the phenomenon is not purely psychoacoustic, but, in some degree, a function of the listener's language (for a review, see Strange and Jenkins, 1977). Here, however, I want to pursue its possible psychoacoustic implications for segmentation—implications exploited, incidentally, by Stevens in his quantal theory (Stevens, 1972).

These implications have been elaborated (although with an eye to the problem of invariance rather than of segmentation) by recent work in selective adaptation. Following the lead of Eimas and Corbit (1973), several dozen studies over the past five years have demonstrated that listeners, asked to identify tokens along a synthetic speech continuum before and after repeated exposure to (that is, adaptation with) a good category exemplar from one end of the continuum, report fewer instances from the adapting category after than before adaptation. Since this effect is observed on a labial voice onset time (VOT) continuum after adaptation with a syllable drawn from an alveolar continuum, and vice versa, adaptation is clearly neither of the syllable as a whole nor of the unanlyzed phoneme, but of a feature within the syllable. Eimas and Corbit therefore termed the adaptation "selective" and attributed their results to the "fatigue" of specialized detectors and to the relative "sensitization" of opponent detectors. Later studies have replicated the results for VOT and extended them to other featural continua, such as those for place and manner of articulation.

Unfortunately, several lines of evidence and argument undermine the hypothesis that selective adaptation reflects the operation of feature detecting mechanisms, specialized for speech. First, is the fact that adaptation has been shown to depend on spectral overlap between adaptor and test continuum: adaptation of consonantal features is specific to following vowel, syllable position and even fundamental frequency (see Ades, 1976, for a review) demonstrating that the supposed detectors are certainly not context-free (as would be required, if we wished to identify these mechanisms with Stevens' (1975) property detectors). Second, the grounds for arguing that, say, variations in voice onset time or place of articulation are mediated by opponent detectors, analogous to those posited for color perception, are not

neurophysiological. On the contrary, the notion of binary opposition is drawn from distinctive feature theory, despite the fact that this theory holds such oppositions to be abstract relations within a phonological system, often realized phonetically by multiple, context-dependent values (cf. Parker, 1977). Finally, the proposed feature detectors lack biological warrant. The communication system of an animal reflects the pressures of its environment and life-history. Animals such as bullfrogs (Capranica,, 1965) or certain songbirds (Marler, 1963) require innate feature detectors or templates, because they must identify their species accurately, and must learn to do so (when learning is involved) within a relatively brief (or non-existent) period of parental care. The message conveyed by their species-specific signals is simple and largely confined to matters associated with reproduction. This is not the case for human infants. The patterns of mother-infant interaction (e.g., Stern, Jaffe, Beebe and Bennett, 1975; Freedle and Lewis, 1977) and of early infant vocalization (e.g., Huxley and Ingram, 1971, 162 ff.; Menn, 1979) suggest a lengthy, epigenetically guided search for sound structure rather than an innate response to species-specific calls.

I take the preceding arguments and evidence to rule out feature or property detecting mechanisms specialized for speech, but to say nothing about property detecting mechanisms in the general auditory system. In fact, selective adaptation studies bear on the segmentation issue precisely by reaffirming the familiar fact that listeners can engage distinct channels of analysis to perceive contrasts between properties of acoustic signals and that, in speech, these properties may be the many-to-one or one-to-many correlates of phonetic features.

Whether or how listeners normally use these channels in listening to speech is both unknown and a separate issue, one that returns us to the second line of research that has developed out of the Haskins work—the search for cues. This work can be seen as, in a sense, coordinate with Stevens' research on spectral properties. Just as Stevens' work raises the question of how diverse spectral properties, said to correspond to distinctive features, are organized and aligned into phonemes and syllables, so the Haskins work raises the issue of how diverse acoustic cues are integrated perceptually into spectral and temporal properties corresponding to features or phonemes.

The puzzle arises because each phonetic distinction is susceptible to manipulation by many acoustic cues. For example, Lisker (1978) has listed sixteen different cues (not all of them tested, it is true) that may serve to distinguish the medial stops of *rapid* and *rabid*. Similarly, Bailey and Summerfield (1978) have shown that perceived place of articulation of a stop consonant (/p/, /t/ or /k/), consequent upon the introduction of a brief silence between /s/ and a following vowel, depends in English on the duration of the silent closure, on spectral properties at the offset of /s/ and on the relation between these properties and the following vowel.

Typically, cues seem to form a hierarchy with one or more cues dominating the others. Thus, presence of voicing during medial closure forces perception of *rabid* rather than *rapid*, but its absence permits other cues to come into play. These cues may then engage in trading relations, so that equivalent percepts result from reciprocal increases and decreases in their values. A well-worked example is provided by the distinction between *slit* and *split*. Here, equivalent percepts of *split* are yielded either by a relatively brief silence (stop closure) after /s/, followed by second and third formant transitions appropriate to /p/, or by a longer silence followed by steady-state vowel formants without transitions (Fitch, Halwes, Erickson and Liberman, in press). (For a review of such work, see Liberman and Studdert-Kennedy, 1978).

At first glance, multiple cue equivalences, or trading relations, seem common -place in light of the familiar intensity-time relations of audition and vision.

However, while constant energy functions are readily rationalized in terms of basilar mechanics or retinal photochemistry, there is no obvious account of why cues as diverse as, say, silence and formant transitions should be perceptually equivalent. No less puzzling is the function of such spectrally diverse and temporally distributed cues in normal perception. Given the multiplicity of cues to any given phonetic distinction, it hardly seems plausible that each cue be extracted by a separate channel and then combined with other cues to yield a phonetic feature—which must then be combined with other phonetic features to form a phone or syllable (Pisoni and Sawusch, 1975). What principle would rationalize these successive integrations?

The quandary was recognized and a rationale for its solution proposed a number of years ago by Lisker and Abramson (1964, 1971). They pointed out that the diverse array of cues which separate so-called voiced and voiceless initial stop consonants in many languages—plosive release energy, aspiration energy, first formant onset frequency—were all consequences of variations in timing of the onset of laryngeal vibration with respect to plosive release—in other words, of voice onset time (VOT). Moreover, they proposed VOT as no more than an instance of a general articulatory variable, timing of laryngeal action, from which the multiple acoustic cues to consonant voicing in all contexts (initial, medial, final/stressed, unstressed) might be derived (Abramson, 1977; cf. Fant, 1960, p. 225). Unfortunately, VOT has often been narrowly interpreted as an acoustic variable, comparable with supposed non-speech analogs, such as the relative onset time of noise-buzz sequences (Stevens and Klatt, 1974; Miller, Wier, Pastore, Kelly and Dooling, 1976) or of two tones (Pisoni, 1977). Such studies have diverted attention from the deeper issue which Lisker and Abramson were addressing, namely, the origins of acoustic cue diversity.

If we extend their account to perception, we have to say that the perceptual counterpart of the unitary articulatory gesture is not an arbitrary collection of cues, but an integral auditory array. In apprehending this array, we perceive the event that shaped it. In other words, we perceive the gesture by means of its radiated sound pattern, just as we perceive the movement of a hand by reflected light—or the articulated gestures of speech by cineradiography. This is not a new notion. Both Paget in his book on human speech (1930) and Dudley (1940) some years later pointed out that the sounds of speech could be regarded as the carriers of articulated gesture.

Yet, attractive as this view may be and important as an account of the origins of speech cue diversity, it still does not explain how we divide the gesture into its underlying linguistic segments. On the contrary, if we regard the consonant-vowel syllable as the product of an integrated, ballistic gesture (Stetson, (1952) and if, further, we regard this gesture as the essential perceptual object which underlies the diverse acoustic cues, we are led to the paradoxical conclusion that the atomistic approach of the Haskins researchers returns us to a view that anchors perception in the entire spectral array rather than in individual cues—a view, moreover, that comes close to that of the more pragmatically directed line of response to the segmentation problem, mentioned at the beginning. This approach makes no attempt to align segments of the signal with the abstract phonetic segments that shape it.

Let us turn briefly to this second line of response. In the first instance, the approach was adopted in order to synthesize or compile speech with "building blocks" (Harris, 1953), formed from half syllables or "phoneme dyads" (Peterson, Wang and Sivertsen, 1958). Given our ignorance of precisley how the consonant gesture merges with the vowel gesture to yield the *motoric* consonant-vowel syllable, this was an eminently practical approach to *acoustic* synthesis, and one that is still viable (e.g., Fujimura, 1975; Mattingly, 1976). Recently, Klatt (1979) has developed a detailed model in which elements larger than the phone are also the elements of

perception (cf. Fujimura, 1975). Klatt proposes a store of a few hundred spectral templates corresponding to phone-pairs, as sufficient to bypass application of phonological rules during perception and to enable recognition of a sizeable lexicon. The point of interest to the present discussion is that Klatt's phone-pairs are, in principle, no different than the acoustic counterparts of syllable gestures. He has taken seriously the familiar claim that ". . . phones are not directly perceived, but must rather be derived from a running analysis of the signal over stretches of at least syllable length" (Liberman and Studdert-Kennedy, 1978, p. 153), but has elected to sidestep their analytic derivation. What we must ask is whether this brute force solution is our only recourse.

I believe that it is not and that the solution to the problem lies in recognizing that the speech signal can indeed be segmented into acoustic groups corresponding to phonetic segments, but only by an organism that already knows that phonetic segments are there to be found. The point is clearly made by the results of recent work on reading spectrograms (Cole, Rudnicky, Reddy and Zue, 1979). The subject, VZ, is a skilled acoustic phonetician who has devoted some 2500-3000 hours to learning to read spectrograms. What is of interest here is that while VZ's performance on correctly labeling segments seems to hover around 85% (a vast improvement, incidentally, on previous reported work), he identifies the existence of segments with an accuracy close to 97%. What is the basis of this remarkable performance?

A moment's reflection tells us that the performance must rest on two crucial facts: first, VZ knows that the segments are there to be found; second, the spectrographic display must represent sufficient acoustic information for the task to be accomplished. Consider each in turn.

That VZ's skill rests on his knowledge that phonetic segments exist becomes obvious as soon as we imagine a reader who lacks this knowledge and confronts a spectrogram as a cryptoanalyst, knowing nothing more than that it conveys a message. How would he proceed? Let us grant that, being human, the cryptographer would start by looking for units (very much like the epigraphist, confronted with Minoan Linear B (Chadwick, 1958). What he would find would be, of course, what Fant (1968) found, namely, a large number of clearly defined acoustic segments bearing (as we know, but the cryptographer does not) a one-to-many and many-to-one relation to the phonetic message. Since this appears to be precisely the condition of the human infant, we must ask how the infant acquires the knowledge that segments exist.

Before attempting to answer this, consider the second condition of VZ's performance—that the spectrographic display does represent enough information for the task to be done. What information does VZ, in fact, use? Apparently, the primary sources of information are spectral discontinuities (including, we may assume, the brief silences of stop closure, contrasts between friction noise and vowel periodicity, formant transitions, nasal resonances, plosive release bursts, and so on) and duration. These are the properties recommended by Fant (1968) in his prescription for spectrogram reading. They are also the properties described by Bondarko (1969) in her account of within-syllable segmentation as based on auditory contrast, a relational process, rather than on the extraction of absolute, context-free features.

Yet, as we have already argued, use of this contrast for segment recognition rests on the prior knowledge that phonetic segments exist. Only when the cryptographer possesses this key to the code, is he in a position to discover, by prolonged practice, the groupings and divisions of the acoustic stream that relate the signal to its phonetic message. How then does the human infant learn the code?

Perhaps the answer will emerge from a deeper understanding of the development of imitation. The process is epigenetic: it seems to rest on an innate, imitative response to the sounds of speech, gradually shaped by the language to which the infant is exposed. The newborn quickly learns to discriminate sound from silence (Friedlander, 1970), voices from other sounds (Alegria and Noirot, 1977), its mother's voice from a stranger's, intonation from monotone (Mehler, Bertoncini, Barriere and Jassik-Gershenfeld, 1978), and, in due course, syllable from intonation contour. Motorically, within weeks of birth, the infant is able to imitate the facial movements of its caretakers (Meltzoff and Moore, 1977), and soon engages in "prespeech" oral play, watching its mother's eyes and mimicking the movements of her mouth (Trevarthen, et al. 1977). Gradually, it shifts from cooing and crowing to intonated babble, until, before the end of its first year, syllables emerge.

Here, the process of differentiation ends—saving us from the paradoxical claim that the infant can imitate what it cannot perceive: the babbling infant imitates syllables, not phonetic segments. For, as we have seen, phonetic segments exist neither in the articulatory gesture nor (a fortiori) in the acoustic signal. Rather, phonetic segments are abstract control processes that emerge as the links between acoustic syllables and their corresponding gestures.

We do not have to suppose that development requires actual activation of the motor system, although this must surely facilitate the process. That activation is not necessary is shown by several well-attested cases of children who have learned to understand without being able to speak. An impressive case is that of Richard Boydell, a victim of congenital cerebral palsy, who, unable to speak and lacking all use of hands and arms, nonetheless learned by prolonged maternal tutelage to understand speech, to read, and even, when provided at the age of 30 with a foot-typewriter, to express his thoughts in highly literate English (Fourcin, 1975). If the difficulties of purely auditory segmentation are indeed as real as the evidence suggests, we must conclude that what remained unimpaired in Boydell was the sensori-motor center that links sound and gesture, and that permits the imitating infant to discover the abstract components of its language.

If this view has any merit and any import for future research, it lies in the implication that we cannot understand speech perception by simply charting relations between signal and percept. We have to go behind the signal to the gestures that shape the resonant volumes of the vocal tract. We have to understand how consonant movement and vowel movement merge to form the motor syllable. Perhaps we shall then conclude that the sounds of speech do indeed form a natural class of dynamic events—distinctive in that those who perceive them have learned how to speak.

Note

Acknowledgement I thank Alvin Liberman for shrewd comments. Preparations of this chapter was supported in part by NICHD grant HD-01994 to Haskins Laboratories, New Haven, Connecticut.