Abstract. To explain the unique efficiency of speech as an acoustic carrier of linguistic information and to resolve the paradox that units corresponding to phonetic segments are not to be found in the signal, consonants and vowels were said to be "encoded" into syllabic units. This approach stimulated a decade of research into the nature of the speech code and of its presumably specialized perceptual decoding mechanisms, but began to lose force as its implicit circularity became apparent. An alternative resolution of the paradox proposes that the signal carries no message; it carries information concerning its source. The message, that is, the phonetic structure, emerges from the peculiar relation between the source and the listener, as a human and as a speaker of a particular language. This approach, like its predecessor and like much recent work in child phonology and phonetic theory, takes the study of speech to be a promising entry into the biology of language.

The earliest claim for the special status of speech as an acoustic signal sprang from the difficulty of devising an effective alternative code to use in reading machines for the blind. Many years of sporadic, occasionally concentrated effort have still yielded no acoustic system by which blind (or sighted) users can follow a text much more quickly than the 35 words a minute of skilled Morse code operators. Given the very high rates at which we handle an optical transform of language, in reading and writing, this failure with acoustic codes is particularly striking. Evidently, the advantage of speech lies not in the modality itself, but in the particular way it exploits the modality. What acoustic properties set speech in this privileged relation to language?

The concept of "encodedness" was an early attempt to answer this question (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Liberman and his colleagues embraced the paradox that, although speech carries a linguistic message, units corresponding to those of the message are not to be found in the signal. They proposed that speech should be viewed not as a cipher on linguistic structure, offering the listener a signal isomorphic, unit for unit, with the message, but as a code. The code collapsed the phonemic segments (consonants and vowels) into acoustic syllables, so that cues to the...
component segments were subtly interleaved. The function of the code was to finesse the limited temporal resolving power of the ear. We typically speak and comfortably understand speech at a rate of 10-15 phonemes/second, close to the rate at which discrete elements merge into a buzz. By packaging consonants and vowels into syllabic units, the argument went, we reduce this rate by a factor of two or three and so bring the signal within the resolving range of the ear.

This complex code called for specialized decoding mechanisms. More than a decade of research was devoted to establishing the existence of a specialized phonetic decoding device in the left cerebral hemisphere and to isolating the perceptual stages by which the supposed device analyzed the syllable into its phonetic components. This information-processing approach to speech perception exploited a variety of experimental paradigms that had seemed valuable in visual research (see Darwin, 1976, and Studdert-Kennedy, 1976, 1980, for reviews), but led eventually to a dead end, as it gradually became apparent that the undertaking was mired in tautology. A prime example was the proposal to "explain" sensitivity to features, whether phonetic or acoustic, as due to feature-detecting devices, and to look for evidence of such mechanisms in infants.

Current research has drawn back and is now moving along two different, though not necessarily divergent paths. The first bypasses the problems of segmental phonetic perception and focuses on what some believe to be the more realistic problem of describing the contributions of prosody, syntax, and pragmatics to understanding speech. The second path, with which I am concerned, reverses the procedure of the earlier encoding approach. Instead of assuming that linguistic units should somehow be represented as segments in the signal and then attempting to circumvent the paradox of their absence by tailoring a perceptual mechanism for their extraction, the new approach simply asks: What information does the speech signal, in fact, convey? If we could answer this question, we might be in a position not to assume and impose linguistic structure, but to describe how it emerges.

Consider the lexicon of an average middle-class American child of six years. The child has a lexicon of about 13,000 words (Miller, 1977), most of them learned over the previous four years at a rate of 7 or 8 a day. What makes this feat possible? Of course, the child must want to talk, and the meanings of the words she learns must match her experience: cat and funny, say, are more likely to be remembered than trepan and surd. But logically prior to the meaning of a word is its physical manifestation as a unit of neuromuscular action in the speaker and as an auditory event in the listener. Since the listening child readily becomes a speaker, even of words that she does not understand, the sound of a word must, at the very least, carry information on how to speak it. More exactly, the sound reflects a pattern of changes in laryngeal posture and in the supralaryngeal cavities of the vocal tract. The minimal endowment of the child is therefore a capacity to reproduce a functionally equivalent motor pattern with her own apparatus. What properties of the speech signal guide the child's reproduction?

We do not know the answer to this question. We do not even know the appropriate dimensions of description. But several lines of evidence suggest
that the properties may be more dynamic and more abstract than customary
descriptions of spectral sections and spectral change. For example, some half
dozens studies have demonstrated "trading relations" among acoustically inocom-
mensurate portions of the signal (e.g., Liberman & Pisoni, 1977; Repp,
Liberman, Eocardt, & Pesetsky, 1978; Fitch, Halwes, Erickson, & Liberman,
1980). Perhaps the most familiar example is the relation between onset
frequency of first formant transition and delay in voicing at the onset of a
stop consonant-vowel syllable: reciprocal variations in spectral structure
and duration of delay produce equivalent phonetic percepts (Summerfield &
Haggard, 1977). Presumably, the grounds of this and other such equivalences
lie in the articulatory dynamics of natural speech, of which we do not yet
have an adequate account. (For a review of studies of this type, see Repp,
1981)

A second line of evidence comes from studies of sine-wave speech
synthesis. Remez, Rubin, Pisoni, and Carrell (1981) have shown that much, if
not all, of the information for the perception of a novel utterance is
preserved if the acoustic pattern, stripped of variations in overall amplitude
and in the relative energy of formants, is reduced to a pattern of modulated
sine waves following the approximate center frequencies of the three lowest
formants. Here, it seems, nothing of the original signal is preserved other
than changes, and derivatives of changes, in the frequency positions of the
main peaks of the vocal tract transfer function (cf. Kuhn, 1975).

Finally, several recent audio-visual studies have shown that phonetic
judgments of a spoken syllable can be modified if the listener simultaneously
watches a video presentation of a face mouthing a different syllable: for
example, a face uttering [ga] on video, while a loudspeaker presents [ba], is
usually judged to be saying [da] (McGurk & MacDonald, 1976; Summerfield,
1979). The phonetic percept, in such a case, evidently derives from some
combination of abstract, dynamic properties that characterize both auditory
and visual patterns.

Moreover, infants are sensitive to dynamic correspondences between speech
heard and speech seen. Three-month-old infants look longer at the face of a
woman reading nursery rhymes if auditory and visual displays are synchronized,
than if the auditory pattern is delayed by 400 milliseconds (Dodd, 1979).
This finding evidently reflects more than a general preference for audiovisual
synchrony, since six-month-old infants also look longer at the video display
of a face repeating a disyllable that they hear (e.g., [lulu]) than at the
synchronized display of a face repeating a different disyllable (e.g., [mama])
(MacKain, Studdert-Kennedy, Spieker, & Stern, Note 1).

The point here is not the cross-modal transfer of a pattern, which can be
demonstrated readily in lower animals. Rather, it is the inference from this
cross-modal transfer, and from the other evidence cited, that the speech
signal conveys information about articulation by means of an abstract (and
therefore modality-free) dynamic pattern. The infant studies hint further
that the infant learns to speak by discovering its capacity to transpose that
pattern into an organizing scheme for control of its own vocal apparatus.

Here we should note that, while the capacity to imitate general motor
behavior may be quite common across animal species, a capacity for vocal
imitation is rare. We should also distinguish social facilitation and general observational learning from the detailed processes of imitation, evidenced by the cultural phenomenon of dialects among whales, seals, certain songbirds, and humans. Finally, we should note that speech (like musical performance and, perhaps, dance) has the peculiarity of being organized, at one level of execution, in terms of a relatively small number of recurrent and, within limits, interchangeable gestures. Salient among these gestures are those that correspond to the processes of closing and opening the vocal tract, that is, to the onsets (or offsets) and to the nuclei of syllables.

We do not have to suppose that the child must analyze adult speech into features, segments, syllables, or even words, before she can set about imitating what she has heard. To suppose this would be to posit for speech a mode of development that precisely reverses the normal (phylogenetic and ontogenetic) process of differentiation. And, in fact, the earliest utterances used for symbolic or communicative ends seem to be prosodic patterns, which retain their unity across a wide variety of segmental realizations (Menn, 1976). Moreover, the early words also seem to be indivisible: for example, the child commonly pronounces certain sounds correctly in some words, but not in others (Menyuk & Menn, 1979). This implies that the child's first pass at the adult model of a word is an unsegmented sweep, a rough, analog copy of the unsegmented syllable. And there is no reason to believe that the child's percept is very much more differentiated than her production. Differentiation begins perhaps, when, with the growth of vocabulary, recurrent patterns emerge in the child's motor repertoire. Words intersect, and similar control patterns coalesce into more or less invariant segments. The segmental organization is then revealed to the listener by the child's distortions. Menn (1978, 1980) describes these distortions as the result of systematic constraints on the child's output: the execution of one segment of a word is distorted as a function of the properties of another. She classifies these constraints in terms of consonant harmony (e.g., [gәk] for duck), consonant sequence (e.g., [nos] for snow), relative position (e.g., [dәge] for 'gator), and absolute position (e.g., [ri] for fish).

Here we touch on deep issues concerning the origin and nature of phonological rules. But the descriptive insights of Menn and others working in child phonology are important to the present argument because they seem to justify a view of the phonetic segment as emerging from recurrent motor patterns in the execution of syllables rather than as imposed by a specialized perceptual device. As motor differentiation proceeds, these recurrent patterns form classes, defined by their shared motor components—shared, in part, because the vocal tract has relatively few independently movable parts. These components are, of course, the motor origins of phonetic features (cf. Studdert-Kennedy & Lane, 1980). Some such formulation is necessary to resolve the paradox of a quasi-continuous signal carrying a segmented linguistic message. The signal carries no message: it carries information concerning its source. The message lies in the peculiar relation between the source and the listener, as a human and as a speaker of a particular language.

Readers familiar with the work of Turvey and Shaw (e.g., 1979) will recognize that the present sketch of a new approach to speech perception owes much to their ecological perspective (as also to Fowler, Rubin, Remez, &
Turvey, 1980). What may not be generally realized is that this perspective is highly compatible with much recent work in natural phonology (e.g., Stampe, 1979), child phonology (e.g., Menn, 1980), and phonetic theory (e.g., Lindblom, 1980; MacNeilage & Ladefoged, 1976; Ohala, in press). For example, Lindblom and his colleagues have, for several years, been developing principles by which the feature structure of the sound systems of different languages might be derived from perceptual and articulatory constraints. More generally, Lindblom (1980) has stressed that explanatory theory must refer "...to principles that are independent of the domain of the observations themselves" (p. 18) and has urged that phonetic theory "...move [its] search for basic explanatory principles into the physics and physiology of the brain, nervous system and speech organs..." (p. 18). In short, if language is a window on the mind, speech is the thin end of an experimental wedge that will pry the window open. The next ten years may finally see the first steps toward a genuine biology of language.

REFERENCE NOTE

REFERENCES


