Cognition, 10 (1981) 301–306
© Elsevier Sequoia S.A., Lausanne – Printed in The Netherlands

The emergence of phonetic structure

MICHAEL STUDDERT-KENNEDY*

Queens College and Graduate Center
City University of New York
and Haskins Laboratories.

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 and 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

*I thank Alvin Liberman, Ignatius Mattingly and Bruno Repp for much fruitful discussion and advice. Preparation of this paper was supported in part by NICHD Grant HD 01994 to Haskins Laboratories. Reprint requests should be sent to Michael Studdert-Kennedy, Haskins Laboratories, 270 Crown Street, New Haven, Ct. 06510, U.S.A.

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 explain 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 some 12–15,000 words, 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 he learns must match his 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 he 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 his 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

dozen studies have demonstrated 'trading relations' among acoustically incommensurate portions of the signal (e.g., Fitch, Halwes, Erickson and Liberman, 1980; Liberman and Pisoni, 1977; Repp, Liberman, Eccardt and Pesetsky, 1978. Perhaps the most familiar example is the relation between extent of first formant transition and delay in voicing at the onset of a stop consonant-vowel syllable: reciprocal variations in spectral structure and onset delay produce equivalent phonetic percepts (Summerfield and 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 review of studies of this type, see Repp [in preparation])

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 and 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 audio-visual 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 and Stern, Reference 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, syllables, segments or even words, before he can set about imitating what he 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 and 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 his 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. [gʌk] for *duck*), consonant sequence (e.g., [nos] for *snow*), relative position (e.g. [dæge] for *'gator*) and absolute position (e.g., [Iʃ] 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 and 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 and 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, 1973), child phonology (e.g., Menn, 1980) and phonetic theory (e.g., Lindblom, 1980; MacNeilage and 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.

References

Darwin, C. J. (1976) The perception of speech. In Carterette, E. and Friedman, M. (eds.), *Handbook of Perception, Vol. VII.* New York, Academic Press, 176–226.

Dodd, B. (1979) Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cog. Psychol., 11,* 478–484.

Fitch, H. L., Halwes, T., Erickson, D. M. and Liberman, A. M. (1980) Perceptual equivalence of two acoustic cues for stop-consonant manner. *Percep. Psychophys., 27,* 343–350.

Fowler, C. A., Rubin, P., Remez, R. E. and Turvey, M. T. (1980) Implications for speech production of a general theory of action. In Butterworth, B. (ed.), *Language Production.* New York, Academic Press, 373–42.

Kuhn, C. M. (1975) On the front cavity resonance and its possible role in speech perception. *J. Acoust. Soc. Am., 58,* 428–433.

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

Liberman, A. M. and Pisoni, D. B. (1977) Evidence for a special speech-perceiving mechanism in the human. In Bullock, T. H. (ed.), *The Recognition of Complex Acoustic Signals.* Berlin, Dahlem Konferenzen, 59–76.

Lindblom, B. (1980) The goal of phonetics, its unification and application. *Phonetica, 37,* 7–26.

McGurk, H. and McDonald, J. (1976) Hearing lips and seeing voices. *Nature, 264,* 746–748.

MacNeilage, P. and Ladefoged, P. (1976) The production of speech and language. In Carterette, E. C. and Friedman, M. (eds.), *Handbook of Perception, Vol. VII.* New York, Academic Press, 75–120.

Menn, L. (1976) *Pattern, Control and Contrast in Beginning Speech: A Case Study in the Development of Word Form and Word Function.* Bloomington, Indiana University Linguistics Club.

Menn, L. (1978) Phonological units in beginning speech. In Bell, A. and Hooper, J. B. (eds.), *Syllables and Segments.* Amsterdam, North-Holland.

Menn, L. (1980) Phonological theory and child phonology. In Yeni-Komshian, G., Kavanagh, J. F. and Ferguson, C. A. (eds.), *Child Phonology: Perception and Production, Vol. I,* 23–41.

Menyuk, P. and Menn, L. (1979) Early strategies for the perception and production of words and sounds. In Fletcher, P. and Garman, M. (eds.), *Language Acquisition.* New York, Cambridge University Press, 49–70.

Ohala, J. (In press) The origin of sound patterns in vocal tract constraints. In MacNeilage, P. F. (ed.), *Speech Production.* New York, Springer-Verlag.

Remez, R. E., Rubin, P. E., Pisoni, D. B. and Carrell, T. D. (1981) Speech perception without traditional speech cues. *Science 212,* 947–950.

Repp, B. H. (In preparation) Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception.

Repp, B. H., Liberman, A. M., Eccardt, T. and Pesetsky, D. (1978) Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *J. Exper. Psychol.: Hum. Percep. Perf., 4,* 621–637.

Stampe, D. (In press) *A Dissertation on Natural Phonology.* New York, Garland.

Studdert-Kennedy, M. (1976) Speech perception. In Lass, N. J. (ed.), *Contemporary Issues in Experimental Phonetics.* New York, Academic Press, 243–293.

Studdert-Kennedy, M. (1980) Speech perception. *Lang. Sp., 23,* 45–66.

Studdert-Kennedy, M. and Lane, H. (1980) The structuring of language: Clues from the differences between signed and spoken language. In Bellugi, U. and Studdert-Kennedy, M. (eds.), *Signed Language and Spoken Language: Biological Constraints on Linguistic Form.* Deerfield Beach, Fl., Verlag Chemie, 29–40.

Summerfield, Q. (1979) Use of visual information for phonetic perception. *Phonetica, 36,* 314–331.

Summerfield, Q. and Haggard, M. (1977) On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *J. Acoust. Soc. Am., 62,* 436–448.

Turvey, M. T. and Shaw, R. E. (1979) The primacy of perceiving: An ecological reformulation of perception for understanding memory. In Nilsson, L.-G. (ed.), *Perspectives on Memory Research: Essays in Honor of Uppsala's 500th Anniversary.* Hillsdale, NJ, Erlbaum.

Reference Note

MacKain, K., Studdert-Kennedy, M., Spieker, S. and Stern, D. Cross-modal coordination in infants' perception of speech. Paper to be read at the Second International Conference on Child Psychology, Vancouver, B.C., August 1981.