WHAT TELLS US THAT SPEECH IS SPEECH?

Quentin Summerfield, MRC Institute of Hearing Research, University Medical School, Nottingham, UK, and Peter J. Bailey, Department of Psychology, University of York, York, UK.

Acoustic analysis and perceptual experimentation have suggested that speech sounds are special and distinct from other sounds. First, no obvious one-to-one isomorphism exists between acoustic and phonetic segments, and the latter have been said to be encoded in the former (e.g., Liberman et al., 1967). Secondly, phonetic perception is apparently not fully rationalised by known psychoacoustic properties of the auditory system. One illustration of this is provided by the experiments described by Dr Dorman (this symposium; see also Bailey, Summerfield and Dorman, 1977), in which the perception of sinewave analogues of speech is shown to depend upon listeners' interpretation of the signals as speechlike or non-speechlike. There is thus some support for the argument that speech perception entails a special decoding process (Liberman & Studdert-Kennedy, 1977). In what follows we shall explore two related questions: what is the nature of the information which might activate such a process, and to what extent should a specification of this information constrain a formulation of the process?

One simple hypothesis could be that speechlikeness is marked by acoustical attributes which, if detected in an initial stage of auditory analysis, direct the signal to a subsequent stage of special phonetic processing. Such attributes would have to be properties of all utterances, and, of necessity, would have to be unencoded, unlike the contextually mutable segments whose decoding they would trigger. Possible candidates have been considered to be rapid spectral changes (Haggard, 1971) and the onset of periodic excitation (Allen & Haggard, 1977), but their role as 'trigger features' has not been empirically demonstrable. Furthermore, even if elaborated by a variable criterion for the acceptability of a trigger feature, this hypothesis cannot account for the perceptual duality of sinewave analogues of CV syllables. Here the putative trigger features would have to be intrinsic to the information specifying phonetic identity, and so in failing to meet the criterion of invariance would exceed the categorising capabilities of purely auditory analysis. Paradoxically, to detect such trigger features successfully, the putatively auditory processor would require the properties of a special phonetic decoder.

These considerations call for a re-appraisal of the model in which signals are routed to one type of processor or another on the basis of prior detection of simple acoustical attributes. They suggest that an alternative solution to the problem of distinguishing speech from non-speech sounds could be that phonetic and generalised auditory processing are accorded in parallel to all acoustic inputs. Phenomenal perception would correspond to whichever process achieved a satisfactory analysis. In proposing such a solution, Liberman, Mattingly and Turvey (1972) suggested that a sound is recognised as speech if a phonetic processor succeeds in extracting phonetic features. Thus the acoustic specification of signals as speechlike is conceived as being isomorphic with the acoustic specification of the cues to phonetic elements, and a characterisation of the former would follow inevitably from a characterisation of the latter. We have already noted that speech is considered an intractable perceptual problem, as a result of the non-invariant relationship between acoustic cues and phonetic elements. This contrasts with the more straightforward relationship existing between phonetic elements and articulatory dynamics, which has led to the suggestion that acoustic cues are interpreted with respect to an internalised knowledge of vocal tract behaviour (Stevens & House, 1972; Liberman et al., 1967). A perceptual model of this kind would seem to involve at least two stages: in the first, a sequence of acoustic elements must be segregated and detected; in the second, these elements must be interpreted, presumably to reconstruct the information encoded in the sequential properties of the signal. Knowledge of vocal tract behaviour may assist the first stage, but it governs the second stage. While we have no doubt that speech perception is inextricably tied to the origin of the signal in a vocal tract, we wonder whether a process of fractionation followed by reintegration would best capture the information endowed in the signal by the continuous articulatory flow of a dynamic vocal tract (see also, Bailey & Summerfield, 1978).

It has been the general conclusion of students of perception that distal events and proximal stimulation relate equivocally, and the traditional response to this problem has been to assert that perception is a constructive process mediated by abstract internal

knowledge (see, for instance, Neisser, 1967). This view of perception is currently coming under increasing scrutiny, urging a re-examination of the peculiarities of phonetic perception. Theoretical appraisal (e.g., Turvey, 1977; Shaw & Bransford, 1977) and empirical analysis (e.g., Lee, 1974; Blumstein & Stevens, 1978) suggest that distal events may have a more veridical, if complex, representation in perceptual data than has generally been supposed. Thus it may be profitable to explore the notion that phonetic percepts are not constructed from discrete acoustic elements by the mediation of articulatory knowledge, but rather that they are specified in acoustic dynamics structured by a speech-specific organisation of the vocal apparatus (e.g., Krmpotic, 1959; Fowler et al., in press). The acoustic signal must remain the focus of our concern, given that an unequivocal reconstruction of articulatory dynamics from the acoustic signal is not possible (e.g., Atal et al., 1978).

Implicit in the foregoing is the assumption that information for speechlikeness can be specified at a single level of analysis, for which the most promising popular candidate has been the level of phonetic processing. This is a necessary view, given that listeners can describe as speechlike even highly schematic analogues of speech sounds, provided they permit a phonetic interpretation. However, the notion that speechlikeness is specified only in the information for phonetic elements is insufficient to account for the certainty and immediacy with which naive listeners can identify utterances in an unfamiliar language as human speech. In recognising as speech snatches of foreign languages heard, for instance, when tuning a radio receiver, we are presumably attending to information of a different kind from that which specifies a sinewave analogue of a CV syllable as speechlike. A particular suggestion by Stevens and House (1972) is that natural speech sounds are characterised by 'certain dynamic or time-varying properties, among which are syllabic intensity fluctuations such as are associated with one of the most fundamental attributes of speech - the vowel-consonant dichotomy' (p. 13). Recent reformulations of the processes underlying speech production (e.g., Fowler et al., in press) provide a means of rationalising the multiplicity of information in a speech signal that specifies it as such. In this view, the speaker progressively organises his articulatory musculature such

that moment-to-moment control need only be exercised over the minimum number of muscle groups during the act of speech production. It is suggested that speech is the concomitant of a set of functionally nested constraints upon the organisation of the vocal apparatus as a whole, so that short-term events like consonantal articulations are nested within longer-term events like the reconfiguration of the vocal tract for successive vowels; these are themselves nested within events of even longer life-spans, such as the speech-specific respiratory synergism (e.g., Lenneberg, 1967). All of these articulatory events are characteristic of speech production, and all endow the speech signal with distinctive dynamic properties to which listeners may be sensitive.

This conceptualisation of speech production, and the type of perceptual attunement it implies, are consistent with a broader view of the development of sensitivity to sound in general. In the natural world, sounds result from the participation of three-dimensional structures in events that occur over time. It is held that the evolution in organisms of sensitivity to vibration in the media that surround them progressed as a developing facility in identifying not just vibration or sound per se, but ecologically relevant events whose concomitants are sounds (see, for instance, Masterson & Diamond, 1973). To a greater or lesser degree, a natural sound is specific to (though not necessarily completely descriptive of) both its particular source, and the particular event in which the source is participating.

Following Turvey and Prindle (1978), therefore, we suggest that the distinction typically made in the laboratory between perception of natural (or even synthetic) speech sounds, and perception of non-natural waveforms like isolated pure or complex tones, should be recast as the distinction between the perception of events and the perception of non-events. In terms of this categorisation, speech perception is a particular instance of event perception, and a general description of the auditory perception of natural events should throw light on the specific problem of perceiving articulatory events. A tentative description could be that the perception of events depends upon the registration of the coherence of information specific to a source and information specific to the transformation wrought upon that source. (See Shaw & Pittenger, 1977.) Thus a preliminary answer to the question of

what is a speech sound could be this: a pattern of sound may be perceived as speech if it cospecifies its source as a human vocal tract participating in a physiologically and phonologically permissable act of articulation. The registration of coherence is analogous to perceiving the solutions to a set of simultaneous equations: the equations provide structure and coherence for the solutions, but no one solution necessarily mediates the attainment of any other. What we understand by coherence may be illustrated further with a visual analogy. When a man runs, he structures light in such a way that both his identity as a man and his act of running are specified optically. When we perceive him running, we detect the coherence of these conjoint specifications; we do not first perceive the actor in order that we may interpret the elements of his act. (For a particularly succinct demonstration of the registration of coherence in the perception of such events, see Johansson, 1974.)

It will be apparent that we lack a formal means of characterising the coherence in speech sounds. Nevertheless, the notion provides us with an appealing informal account of the perceptual strategies adopted by listeners in the experiments on sinewave analogues of speech. When sinewaves were heard as non-speech sounds, we suppose that listeners attended to the elements in the acoustic array but not to their potential organisation. In hearing them as speechlike, on the other hand, they attended both to the acoustic elements and to their organisation, which together specify, albeit in a highly reduced form, a vocal tract undergoing a phonologically permissable act of articulation. Those familiar with R.C. James' photograph, reproduced in Lindsay and Norman (1972, p. 8) will recognise that the foregoing analogously describes the initial perception of the picture as a random array of dark and light areas, and the subsequent perception of a Dalmatian dog walking in dappled sunlight. Both hearing sinewaves as speechlike and seeing the dog are compelling perceptions. It may be that the search for coherence in stimulus information is a general goal of perceptual systems, guided and rewarded by the attainment of clarity (Woodworth, 1947; Gibson, 1969). We note that when listeners began to hear sinewaves as speechlike, their identification functions became more consistent and more categorical.

In summary, we are suggesting that the achievement of speech articulation is to present the information for speech perception unequivocally in the surrounding media. The acoustic signal is clearly the most important vehicle for speech, but we acknowledge also the perceptual importance of the speech-specific optical concomitants of articulatory events (e.g., Miller & Nicely, 1955; see Erber, 1975, for a review). Progress beyond the phenomenological interest of demonstrations such as the perceptual duality of sinewave analogues of CV syllables requires the development of a vocabulary with which to describe how articulatory events structure sound and light in perceptually accessible ways. The mathematics of this description will be complex. Nevertheless, we are encouraged that optical invariants supporting the visual perception of aspects of one human activity, locomotion, have been formally described (Lee, 1974; Cutting et al., 1978). The rebirth of articulatory synthesis for perceptual experimentation (Mermelstein & Rubin, 1978; cf., Haggard, in press) is one precursor of the attainment of a similar specification of the optical and acoustical invariants supporting the perception of speech articulation: that is, to specify what it is that tells us that speech is speech.

References

Allen, J. and M.P. Haggard (1977): "Perception of voicing and place features in whispered speech: a dichotic choice analysis", Perc. Psych. 21, 315-322.

Atal, B.S., J.J. Chang, M.V. Mathews, and J.W. Turkey (1978): "Inversion of articulatory to acoustic transformation in the vocal tract by a computer sorting technique", JASA 63, 1535-1555.

Bailey, P.J. and A.Q. Summerfield (1978): "Some observations on the perception of [s]+stop clusters", Haskins Laboratories Status Report on Speech Research SR53 (2), 25-60.

Bailey, P.J., A.Q. Summerfield, and M.F. Dorman (1977): "On the identification of sine-wave analogues of certain speech sounds", Haskins Laboratories Status Report on Speech Research SR51-52, 1-25.

Blumstein, S.E. and K.N. Stevens (1977): "Acoustic invariance for place of articulation in stops and nasals across syllable contexts", JASA 62, S26(A).

Cutting, J.E., D.R. Proffitt, and L.T. Kozlowski (1978): "A biomechanical invariant for gait perception", J. Exp. Psych: HPP 4, 357-372.

Erber, N.P. (1975): "Audio-visual perception of speech", JSHD 40, 481-492.

Fowler, C.A., P. Rubin, R.E. Remez, and M.T. Turvey (in press): "Implication for speech production of a general theory of action", in Language production, B. Butterworth (ed.), New York: Academic Press.

Gibson, E.J. (1969): Principles of perceptual learning and development, New York: Appleton.

Haggard, M.P. (1971): "Encoding and the REA for speech signals", Quart. J. Exp. Psych. 23, 34-45.

Haggard, M.P. (in press): "Experience and perspectives in articulatory synthesis", in Frontiers of speech communication research, B. Lindblom and S. Öhman (eds.), London: Academic Press.

Johansson, G. (1974): "Projective transformations as determining visual space perception", in Essays in honor of J.J. Gibson, R.B. MacLeod and H.L. Pick (eds.), 117-138, Ithaca: Cornell University Press.

Krmpotic, J. (1959): "Donnés anatomiques et histologiques relatives aux effecteurs laryngo-pharyngo-buccaux", Revue Lar. Otol. Rhinol. 80, 829-848.

Lee, D.N. (1974): "Visual information during locomotion", in Essays in honor of J.J. Gibson, R.B. MacLeod and H.L. Pick (eds.), 250-267, Ithaca: Cornell University Press.

Lenneberg, E.H. (1967): Biological foundations of language, New York: Wiley.

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

Liberman, A.M., I.G. Mattingly, and M.T. Turvey (1972): "Language codes and memory codes", in Coding processes in human memory, A.W. Melton and E. Martin (eds.), 307-334, New York: Winston.

Liberman, A.M. and M. Studdert-Kennedy (1977): "Phonetic perception", Haskins Laboratories Status Report on Speech Research SR50, 21-60. (To appear in Handbook of Sensory Physiology, Vol. VIII, "Perception", R. Held, H. Leibowitz, and H-L. Teuber (eds.), Heidelberg: Springer Verlag.

Lindsay, P.H. and D.A. Norman (1972): Human information processing: An introduction to psychology, New York: Academic Press.

Masterson, B. and I.T. Diamond (1973): "Hearing: central neural mechanisms", in Handbook of perception, Vol. III, Biology of Perceptual Systems, E.C. Carterette and M.P. Friedman (eds.), 408-448, New York: Academic Press.

Mermelstein, P. and P. Rubin (1978): "Articulatory synthesis - a tool for the perceptual evaluation of articulatory gestures", Haskins Laboratories Status Report on Speech Research SR53 (1), 1-11.

Miller, G.A. and P.E. Nicely (1955): "An analysis of perceptual confusions among some English consonants", JASA 27, 338-352.

Neisser, U. (1967): Cognitive psychology, New York: Appleton.

Shaw, R. and J. Pittenger (1977): "Perceiving the face of change in changing faces: implications for a theory of object perception", in Perceiving, acting and knowing, R. Shaw and J. Bransford (eds.), 103-132, Hillsdale, N.J.: Erlbaum.

Shaw, R. and J. Bransford (1977): "Psychological approaches to the problem of knowledge", in Perceiving, acting and knowing, R. Shaw and J. Bransford (eds.), 1-39, Hillsdale, N.J.: Erlbaum.

Stevens, K.N. and A.S. House (1972): "Speech perception", in Foundations of modern auditory theory, Vol. II, J.V. Tobias (ed.), 1-62, New York: Academic Press.

Turvey, M.T. (1977): "Contrasting orientations to the theory of visual information processing", Psych. Rev. 84, 67-88.

Turvey, M.T. and S.S. Prindle (1978): "Modes of perceiving: abstracts, comments and notes", in Modes of perceiving and processing information, H.L. Pick and E. Saltzman (eds.), 205-224, Hillsdale, N.J.: Erlbaum.

Woodworth, R.S. (1947): "Reinforcement of perception", AJPs 60, 119-124.