Journal of Phonetics (1986) **14**, 99–104

Two cheers for direct realism

548

Michael Studdert-Kennedy

Queens College and Graduate Center, City University of New York, and Center for Advanced Study in the Behavioral Sciences, Stanford, CA, U.S.A.

"Beware Procrustes, bearing Occam's razor."
Lise Menn

I am very much in sympathy with Fowler's approach (henceforth, CAF) because it is grounded in a functionalist, biological view of language. No doubt the approach will be faulted, despite its disclaimers, for narrowly focussing on phonetic structure. Yet what is new in CAF is precisely its scope: the range of phonetic fact for which it takes responsibility. Basic research in speech perception and basic research in speech production (no less than applied research in speech synthesis and machine recognition) have tended to follow parallel lines. Perception research typically manipulates acoustic variables with little regard for articulatory constraints, while production research typically studies the actions of individual muscles or articulators with little concern for how they are coordinated to yield a perceptually coherent acoustic signal. By adopting a single abstract unit (corresponding to the phoneme-sized phonetic segment) as the presumed functional element of both production and perception, CAF lays the ground for a program of research responsible to both. Nor is it coincidence that the selected unit is potentially alphabetic. For CAF thus acknowledges that our accounts of speaking and listening must be consistent with the facts of writing and reading.

A signal virtue of CAF, then, is that it accepts responsibility for the segmental structure of all four modes of language action: like any good theory, it proposed to unify (eventually) related classes of fact that are commonly treated as separate. The faults of CAF largely stem, I believe, from a somewhat too zealous attempt to impose a framework, devised to handle an animal's traffic with the physical world, on a communication system with a quite different evolutionary history and function.

CAF includes three assumptions that need to be modified or, at least explicated: (1) perception is "unmediated by cognitive processes of inferencing or hypothesis testing"; (2) listeners "extract information about articulation from the acoustic speech signal"; (3) "it matters little through what sense we realize what speech event has occurred." My comments follow.

1. Unmediated perception

A corollary of this assumption seems to be that the phonetic segment should not be constructed, in either perception or production, from smaller units. Accordingly CAF, invoking speech error data to support the choice of unit, implicitly dismisses "feature" errors as unimportant. Yet such errors do occur, with some low frequency, and have to be accounted for. Voicing metathesis seems to be the most common (e.g. *clear blue sky* → *glear plue sky*: Fromkin, 1971), but place metathesis also occurs (e.g.

0095–4470/86/010099 + 06 $03.00/0

pedestrian → *tebestrian*: Fromkin, 1971; *wild goose chase* → *wild juice case*: Robert Remez, pers. commun.). These errors are interesting because they reflect a level of organization below the segment.

The possibility of such errors is implicit in CAF's definition of a phonetic segment as a "set of coordinated gestures". Elsewhere, Fowler and her colleagues (Fowler, Rubin, Remez & Turvey, 1980) treat the phonetic segment as a set of nested, or embedded, coordinative structures that arise as functional groupings of muscles, marshalled for moment-to-moment control of speech. The coordinative structures of Fowler *et al.* evidently correspond to the gestures of CAF. Similarly, Kelso, Saltzman & Tuller (this issue, henceforth KST) discuss the task-specific grouping of muscles to execute a gesture, nested within the CV syllable. CAF, quite properly in my view, regards these gestures as non-linguistic (or non-phonetic): "lip closure *per se* is not an articulatory speech event." Lip closure only becomes phonetic (i.e. only performs a linguistic function) by virtue of its coordination with other non-phonetic gestures in an appropriate linguistic context.

A speaker, then, is engaged in moment-to-moment marshalling of intrinsically functionless muscle systems to fulfil a phonetic function—much as a tennis player marshalls muscles to execute a tennis stroke. A skilled speaker has a repertoire of routinized processes that assemble non-phonetic gestures into phonetic segments. Errors in gestural assemblage may then be rare because the process occurs with very high frequency, so that a given gesture is called into a phonetic segment even more frequently than a phonetic segment is called into a syllable. Errors in the process may also be rare due to tight anatomical and physiological constraints on gestural coupling: Voicing metathesis is perhaps the most common error because voicing is relatively loosely coupled to supralaryngeal action. In any event, by this account, a gestural error is motoric, a segmental error phonetic.

Consider now the child learning to speak. Its task is to discover how to marshall its repertoire of non-linguistic babbling gestures for linguistic use. Its first linguistic (functionally communicative) segments are words or formulaic phrases. The child evidently perceives these units as constructed from non-linguistic gestures. For example, Ferguson and Farwell (1975) report the following attempts by a 15-month-old child to say the word *pen*:

[mãᵊ, ˇ̃ʌ, dɛᵈᵑ, hɪn, ᵐbõ, pʰɪn, tʰn̩tʰn̩tʰn̩, baʰ, ɖʰauᴺ, buã].

In these attempts, we find all the gestures required to utter *pen*: lip closure, linguaalveolar closure, tongue raising and fronting, velum raising and lowering, glottal narrowing and spreading. The gestures are misordered and mistimed, but it is evident that the acoustic structure of the word did specify for the child the gestures that compose it.

As the child develops, it will come to recognize recurrent gestural groupings as functional elements in speaking: the phonetic segment will emerge as the interface between non-linguistic gesture and linguistic word. Will the child thereby lose its capacity to perceive gestures? It would seem not. The speech error data demonstrate that the adult may produce gestures separately from the segmental structure in which they are normally embedded. If the perspectives of speakers and listeners are "interchangeable", as CAF proposes, listeners must assemble segments from non-phonetic, auditory markers in the signal no less than speakers assemble them from a non-phonetic gestural repertoire. This may not call for "inferencing or hypothesis testing" in perception, but it does call for some process less immediate than the word "direct" would seem to imply.

2. Extracting information about articulation

Direct realism presses CAF into "defining speech event interchangeably from the perspectives of talkers and listeners". For the definition to hold we must assume that the problem of functional equivalence among diverse motor patterns, in general, or of the many-to-one relation between articulation and acoustics, in particular, has been solved (cf. KST, this issue). We could then be confident that articulation and acoustics are, at some abstract level of description, fully isomorphic: to every acoustic pattern of change in frequency and time there exactly corresponds an articulatory pattern of movement in space and time, and vice versa.

Ironically, this assumption renders ambiguous much of the evidence cited to support it. To show that listeners extract information about articulation from the speech signal, CAF cites several studies in which listeners' perceptual judgements seemed to be in better agreement with the articulatory pattern than with the acoustic. Such findings are anomalous, if articulatory and acoustic patterns are isomorphic. For the "P-center" studies CAF resolves the anomaly by arguing that it arose from an error in the conventional acoustic measurements of vowel onset. Once the error was corrected, acoustics, articulation and perception fell into line.

An equivalent move in the /slɪt–/splɪt/ "trading relations" phenomenon would require systematic measurement of the articulatory correlates of acoustic silent interval (stop closure) and formant transitions (stop release). Such measurements have never been reported, so far as I know, and in the cited study they could not be appropriately made because the experiments were done with synthetic speech. Articulatory equivalence (or non-equivalence) was therefore inferred, with some circularity, from perceptual equivalence (or non-equivalence). However, if the appropriate measurements were done on natural speech, articulation, acoustics and perception would, by the hypothesis of CAF, again fall into line.

In short, if acoustics and articulation are fully isomorphic, they are merely notational variants. Whether we describe the listener as perceiving sound patterns or as perceiving articulatory patterns, is then a matter of theoretical taste. Direct perception of articulation becomes merely an axiom of a direct-realist theory.

Perhaps all this is sophistry. We know, after all, that listeners do extract information about articulation. How otherwise would every normal child come to speak the dialect of its peers? We know too from studies of "lip-reading" that acoustic and optic information about speech may combine in perception. These studies suggest that we are able to imitate or repeat the utterance of another because perception extracts an amodal pattern of information, isomorphic with the pattern that controls articulation—just as CAF claims. What seems to be at issue then is not whether listeners can extract information about articulation, but whether they always do, and whether perception is direct, in the sense that the medium structured by articulation is transparent and a matter of indifference to the perceiver.

3. The medium of amodality

Each species of animal has a unique combination of perceptual and motoric capacities. Characteristic motor systems have evolved for locomotion, predation, consumption, mating. Matching perceptual systems have evolved to guide the animal in these activites. The selection pressures shaping each species' perceptuomotor capacities have come, in the first instance, from physical properties of the world.

By contrast, these perceptuomotor capacities themselves must have played a crucial role in shaping the form of a social species' communication system. The general point was made by Huxley (1914) when he remarked that the elaborate courtship rituals of the great-crested grebe must have evolved by selection of perceptually salient patterns from the bird's repertoire of motorically possible actions. Certainly, specialized neuro-anatomical signaling devices have often evolved, but they have typically done so by modifying pre-existing structures just enough for them to perform their new function without appreciable loss of their old. The cricket stridulates with its wing, the grasshopper with its legs; birds and mammals vocalize with their eating and breathing apparatus. The quality and range of possible signals is thus limited by the structure and function of the co-opted mechanism.

A further constraint on signal form must come from the perceptual system to which the signals are addressed. Here again specialized devices (e.g. feature detecting systems, templates) have certainly evolved, presumably by some minimal modification of a pre-existing perceptual system. Typically, such specialized devices, in the auditory realm, seem to have evolved in animals with little or no parental care and therefore little opportunity to learn their species' call: bullfrogs, treefogs, certain species of bird, and so on. We have no evidence for such devices in the human.

We are not then surprised that the main speech frequencies are spread over the three octaves (500–4000 Hz) to which the human auditory system is most sensitive, and that (as the quality of deaf speech attests) speech sounds have evolved to be heard, not seen. Thus, the differences in degree of constriction among high vowels, intra-oral fricatives and stops are highly salient auditorily; but the same differences in, say, finger to thumb distance, would be scarcely detectable if they were incorporated in a visual sign language. Similarly, the abrupt acoustic changes at the onset of many CV syllables may have been favored, in part, because the mammalian auditory system is particularly sensitive to such discontinuities (Delgutte, 1982; Kiang, 1980; Stevens, 1981). The resulting auditory contrast perhaps facilitates the listener's perceptual segmentation both of the syllable from its context and of the consonant from its following vowel.

On the other hand, the signs of American Sign Language have evolved (over the past 170 years) to be seen, not heard. Accordingly, signs formed at the center of the signing space (that is, in the foveal region of the viewer) tend to use smaller movements and smaller handshape contrasts than signs formed at the periphery (Siple, 1978).

In short, even if the sense that informs us about our environment "matters little" in the farmyard (itself a dubious claim), it seems not to "matter little" for communication. Language has evolved within the constraints of pre-existing perceptual and motor systems. We surrender much of our power to understand that evolution, if we disregard the properties of those systems. And indeed, CAF concedes as much by citing with approval Lindblom's work on the emergence of phonetic structure. The success of that work, particularly for vowel systems, rests on an acoustic description of speech sounds, weighted according to a model of the auditory system, and on the use of an auditory distance metric to assess their perceptual distinctiveness.

How, then, are we to square the auditory properties of the speech signal with the evident amodality of the speech percept? We must, I think, question CAF's definition of speech events as "a talker's phonetically structured articulations." A speech event is not simply articulation, however structured, any more than a tennis serve is simply the server's swing. A speech event, even narrowly conceived as phonetic action, only occurs when a speaker executes, and a listener apprehends, a phonetic function. Elsewhere,

Fowler (1980) has termed this function the talker's phonetic "intent" (cf. Liberman, 1982). "Intent" seems to correspond, at least in level of abstraction, to task (or goal), the level at which KST (this issue) define a single function from which different, but equivalent, articulations may arise. Surely, this too must be the level—free of adventitious articulatory variation and its acoustic consequences—at which the listener's percept might properly be termed amodal.

Looked at in this way, articulation becomes as much a medium of speech, structured by the talker's goals, as the acoustic signal, structured by the talker's articulations, and as its heard counterpart, structured by the listener's (suitably "attuned") auditory system. Each medium is then subject to its own characteristic type of variability.

One happy side-effect of setting speaker and listener (articulation and audition) on equal footing is that we can rationalize perceptual error more simply than does CAF. The likelihood of an error is a function of its cost. Collisions between swallows, swarming in hundreds through a cloud of insects, or between pelicans flocking and diving into a school of fish, are rare (though, *pace* direct realism, they do occur!). Natural selection prunes the error-prone from the species, honing the perceptuomotor systems of the survivors to a fine precision. By contrast, errors in speaking and listening carry essentially no penalty. Moreover, if phonetic form has been shaped by compromise between the articulatory capacities of a speaker and the perceptual capacities of a listener we might expect some instability in phonetic execution, some slight oscillation between the opacity comfortable for a speaker, the transparency called for by a listener (cf. Slobin, 1980). We may view a conversation as a microcosm of evolution: the speaker balances a desire to be understood against articulatory ease, the listener a desire to understand against the costs of attention (Lindblom, 1983). Given these conflicting demands and the modest penalties for error, we might even be surprised that errors are not more frequent than they are. In this regard, while no one, so far as I know, has studied the social contexts in which perceptual errors occur, they are probably rare when the speaker is, say, delivering instructions for a parachute jump.

In conclusion, the fact that we hear speech is no less important and no more accidental than the fact that we articulate it. Many of the longstanding problems of speech research, including normalization, segmentation and even the lack of invariance, may be illuminated by an understanding of audition. Even if the information we extract is amodal, just what information we extract and the precision with which we extract it depend on our auditory sensitivity.

This comment was written while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, CA. My thanks go to the Spencer Foundation for financial support, and to Björn Lindblom and Peter MacNeilage for discussion and comments.

References

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In: Carlson, R. & Granström, B., *The representation of speech in the peripheral auditory system*, pp. 131–149. New York: Elsevier.

Ferguson, C. A. & Farwell, C. B. (1975). Words and sounds in early language acquisition: English initial consonants in the first fifty words, *Language*, **51**, 419–430.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing, *Journal of Phonetics*, **8**, 113–133.

Fowler, C. A., Rubin, P., Remez, R. E. & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In: Butterworth, B. (ed.), *Language production*, Vol. 1, pp. 373–420. New York: Academic Press.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances, *Language*, **47**, 27–52.

Huxley, J. S. (1914). The courtship habits of the great crested grebe (*Podiceps cristatus*), with an addition on the theory of sexual selection, *Proceedings of Zoological Society* (London), **XXXV**, 491–562.
Kiang, N. Y. S. (1980). Processing of speech by the auditory nervous system, *Journal of the Acoustical Society of America*, **68**, 830–835.
Liberman, A. M. (1982). On finding that speech is special, *American Psychologist*, **37**, 148–167.
Lindblom, B. (1983). Economy of speech gestures. In: MacNeilage, P. F. (ed.), *The production of speech*, pp. 217–245. New York: Springer-Verlag.
Siple, P. (1978). Visual constraints for sign language communication, *Sign Language Studies*, **19**, 97–112.
Slobin, D. I. (1980). The repeated path between transparency and opacity in language. In: Bellugi, U. & Studdert-Kennedy, M. (eds.), *Signed and spoken language: biological constraints on linguistic form*, pp. 229–243. Deerfield Beach, FL: Verlag Chemie.
Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: data from phonology, acoustics and psychoacoustics. In: Myers, T., Laver, J. & Anderson, J. (eds.), *The cognitive representation of speech*, pp. 61–74. New York: North-Holland.