Journal of Phonetics (1986) **14**, 69–74

A new perspective and old problems

Dominic W. Massaro

Program in Experimental Psychology, University of California, Santa Cruz, CA 95064, U.S.A.

The value of approaching the problem of perception from a direct–realist perspective is becoming apparent, and it is valuable to explore this approach in the study of speech perception. Fowler has offered a thorough and stimulating framework for how such a perspective might be instantiated. Although there is much that is positive in the paper, my charge as a commentator is perceived as being critical. One goal of the commentary is to make the case that the direct–realist perspective, as articulated by Fowler, bears more similarities with alternative or previous perspectives than differences. New neologisms for old beliefs do not a Kuhnian paradigm shift make. In fact, it is almost distressing that the author converged on the consensus opinion for many of the traditional problems in the area.

Within the domain of communication, only the linguistic utterance is addressed and not other aspects, such as the talker's gestures. No-one should criticize an attempt to constrain the phenomena to be explained, since it is a necessary aspect of scientific endeavor. However, such a constraint is highly unfortunate when the phenomenon placed outside the purview of the analysis has important implications for the analysis itself. For Fowler, the distal event in speech perception is "the articulating vocal tract", and supposedly it is recovered by the perceiver. We might inquire how this recovery interacts with nonarticulatory information, such as gestures of the speaker. What does a motor theory of speech perception imply about such a contribution? It is reasonable to assume that activities of the vocal tract might be recovered from the acoustic speech signal, since the latter was shaped by the former. A similar case does not seem possible for gestural signals, since they are not shaped by the vocal tract.

McNeill (1985) makes a strong case that gestures are not non-verbal, but bear striking resemblance to other linguistic modalities such as speech. As an example, some gestures might be characterized by distinctive features analogous to those characterizing speech sounds. Communicators appear to compute speech and gesture in parallel and make both sources of information available to the perceiver. As perceptual psychologists, we might ask when the gestural source of information is psychologically real. For the direct realist who wants the perceiver to recover articulation of the vocal tract, it is necessary to address how gesture would interact with this function. A contribution of gesture information might constrain significantly the type of recovery process that is postulated.

It is thus an important question if and how gestural information contributes to speech perception. Thompson & Massaro (1985) addressed this issue by independently varying both gestural and acoustic sources of information in a categorization task. Subjects were given either acoustic, gestural, or both sources of information and asked to identify the intended reference. The task involved discriminating a *ball* from a *doll*. The acoustic information was manipulated by creating a /ba/ to /da/ speech continuum and presenting

0095–4470/86/010069 + 06 $03.00/0

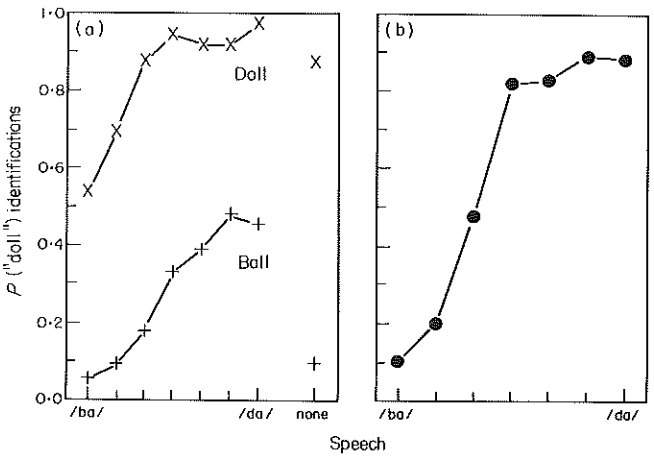

Figure 1. Observed proportion of "doll" responses as a function of the level of speech and gesture for five-year-olds. (a) Conditions containing a gesture; (b) speech-alone condition. x, Doll gesture; +, ball gesture. Data from Thompson & Massaro (1985).

one of the sounds along this continuum. The gestural information involved pointing to either a ball or a doll. On some trials, there was no acoustic signal; on other trials, there was no gesture. On the remaining trials, there were both acoustic and gestural signals which could be compatible, ambiguous, or in conflict with respect to each other. Pre-school children and adults were tested.

Figure 1 gives the results for the pre-school children. As can be seen in the figure, both sources of information influenced categorization, and the interaction was such that the contribution of one source was greater to the extent the other source was ambiguous. How do these results address the issue of recovering articulation from the acoustic signal? There are two general approaches that might be taken. In the first, it could be assumed that gestural information aids in the recovery of the articulatory event. If this tack is adopted, the direct–realist perspective will have to be expanded to include exactly how gestural information is evaluated and how it is integrated with the acoustic signal. This approach would seem to be incompatible with the direct–realist perspective because of the loans on processing that are required.

The second approach would be to assume that recovery of articulation occurs via the acoustic signal only, and that gestural information does not influence this recovery. In this case, speech and gesture would interact after phonetic identification (i.e. recovery of articulation, since Fowler implies that recovery of the latter is recovery of the phonetic level). In this manner, the direct realist can explain the interaction of the two sources after the phonetic identification of each source. Thus, recovery of articulation could proceed by extracting invariants from the acoustic signal, and the gesture could be identified by some other means. For identification, the perceiver could utilize one source or the other. This process is analogous to that originally proposed by McGurk & MacDonald (1976) and MacDonald & McGurk (1978) except that the two sources were lip movements and spoken speech, and the former determined phonetic identification of place and the latter determined phonetic identification of voicing and manner.

This analysis can be described mathematically in a straightforward manner. Let a_i correspond to probability of a /da/ identification given speech and g_j represent the

probability of a /da/ identification given the gesture. The subscripts i and j refer to the level of the speech and gesture information, respectively, and indicate that identification of speech depends only on speech and analogously for gesture. Two covert identifications occur on each trial, and the subject must follow one or the other or both if they agree (Massaro & Cohen, 1983). In this case, the probability of a /da/ identification, given a bimodal linguistic event A_iG_j, is equal to

$$P(/\text{da}/|A_iG_j) = pa_i + (1 - p)g_j \quad (1)$$

where p is the probability of following covert identification of the auditory source. This seems to be a very reasonable explanation from the direct–realist perspective, but it is clearly contradicted by the results. The predicted $P(/\text{da}/|A_iG_j)$ is some compromise between the identification of the speech and gesture identifications (i.e. it is a weighted average), and therefore it cannot be more extreme than identification of either source presented alone. Consider the seventh level of auditory information and the "doll" pointing gesture in Figure 1. When these sources are presented alone, the probability of a child's /da/ judgment is about 0.90. When they are presented in combination, $P(/\text{da}/)$ is close to 1. Given a_i and g_j values of 0.09 in Equation 1, $P(/\text{da}/|A_iG_j)$ cannot be greater than 0.90 regardless of the value of p. In general, the prediction is that $P(/\text{da}/|A_iG_j)$ cannot exceed either $P(/\text{da}/|A_i)$ or $P(/\text{da}/|G_j)$. Thus we have a nice disproof of the model.

The evidence seems to demand that gesture and speech information are integrated prior to phonetic identification of each source. This evidence seems difficult to square with the recovery of articulatory information. Recovery of articulatory information from the acoustic stream is justified by those who find the concept appealing because the acoustic stream is shaped by articulation. Not so for gesture, but yet gesture information contributes to phonetic identification, which for Fowler means recovery of articulatory information. In my view, the result greatly weakens her thesis and related motor theories (Liberman & Mattingly, 1985).

Analogous to focusing on the acoustic signal in the recovery of articulatory events, Fowler also chooses to focus on speech consisting of phonetically structured syllables. This choice is again traditional in the study of linguistic phenomena in which cross-talk across levels is assumed to be minimal. Fowler's choice again seems to stand in marked contradiction of what we know about top-down processes in speech perception (Ganong, 1980; Marslen-Wilson & Welsh, 1978). Consider an experiment carried out by Isenberg, Walker & Ryder (1980). A speech continuum was made between the function words *the* and *to*, and each of these test words was placed in two different sentence contexts. One context was appropriate for *the* and the other for *to*.

The speech continuum was created by beginning with a natural utterance of the word *to* and attenuating the onset energy between 14 and 36 dB in steps of 2 dB. With little attenuation, the word is heard consistently as *to*; with a lot of attenuation, the word is heard as *the*. Intermediate levels of attenuation give more ambiguous percepts. The test word was placed as the initial word in one of two sentence contexts

To/the go is essential.
To/the gold is essential.

The only difference between the sentences is whether *go* or *gold* follows the test word. Subjects were presented with the 12 versions of the test word in the two sentence contexts and were asked to identify the test words as *the* or *to*.

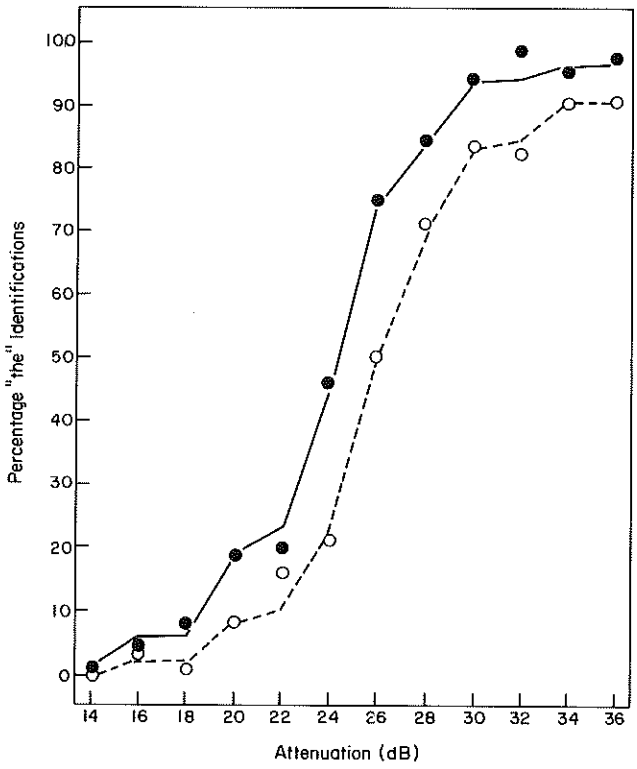

Figure 2. Percentage of "the" identification as a function of the onset attenuation of the test stimulus and the sentential context. Content /gold/: ●, observed; ——, predicted. The predicted results are derived from a fuzzy logical model of speech perception (Massaro & Cohen, 1983). Content /go/: o, observed: – – –, predicted. Data from Isenberg *et al.* (1980).

Figure 2 gives the results. As expected, the percentage of *the* identifications increased systematically with increases in the attenuation of the onset of the test word. Sentential context also had an effect, especially at the intermediate levels of attenuation of the test word. These results indicate that phonetic and lexical levels of analysis cannot be described independently of one another. Recovery of articulation followed by higher-order analyses simply will not work as an appropriate description. Rather, it appears that listeners have continuous information from multiple sources (both bottom-up and top-down), and these are integrated to achieve recognition at the highest level of analysis possible. A more global critique of Fowler's theme might be: why is recovery of articulation so crucial if the action is at a much higher level.

I must admit I do not understand how Fowler comes to grips with top-down effects, and in my mind they represent a real barrier to a direct–realist view. To say that perceivers might conserve effort by utilizing top-down information where possible implies effort is involved in perception, something which seems incompatible with the neo-Gibsonian framework. Also, it seems odd to give priority to top-down over bottom-up processing when bottom-up putatively costs so little for just "picking it up".

Evidence that Fowler presents to support her view often reflects confirmation bias. The problem is that the same evidence is consistent with just the opposite point of view. Although Popper (1959) and Platt (1964) might be evaluated as too rigid for today's

liberal science (Feyerabend, 1976), one cannot deny what is right. One simply cannot interpret evidence as supporting one thesis if it is equally consistent with an alternative one. Fowler, Liberman (1982), and Liberman & Mattingly (1985) cannot claim "trading relations" for their point of view since the relations are even more consistent with an alternative (Massaro, 1984; Massaro & Cohen, 1983). In fact, I originally developed the trading-relation paradigm of studying the evaluation and integration of multiple sources of information in speech perception because the experiments generated within the framework of motor theory and categorical perception were limited in value (Massaro & Cohen, 1976).

To illustrate the problem with confirmation bias and to argue for a more fine-grained analysis, consider the recent infatuation with the McGurk & MacDonald (1976) effect. Fowler uses the discovery that perceivers are influenced by both auditory and visual speech as evidence for her point of view. Since she does not acknowledge that other frameworks can provide a better account of the phenomenon, I must say that we have captured a broad range of results with a process model that might be considered as the antithesis of direct realism. Theories must be mapped into specific models, and this is the challenge to a direct perception viewpoint. Can a direct–realist model be developed to give a good quantitative description of the same results that they have used as evidence for their theory?

Fowler proposes a mediated theory of speech perception in which the mediation is articulatory. To complete the analogy to Gibson's theory, it needs to be demonstrated that the invariants are at this level and that this level is directly perceived. This has not been done. Finally, we should remind ourselves that Fowler has instantiated only one candidate for a direct perception theory. Her arguments and the critiques are most relevant to the single candidate rather than to the direct-perception approach in general.

The direct-perception framework might be more productive if it forgets about articulation and becomes less direct.

The writing of this paper and the research reported in the paper were supported, in part, by NINCDS Grant 20314 from the Public Health Service and Grant BNS-83-15192 from the National Science Foundation.

References

Feyerabend, P. K. (1975) *Against method.* London: NLB.

Ganong, W. F. III (1980) Phonetic categorization in auditory word perception, *Journal of Experimental Psychology: Human Perception and Performance,* **6**, 110–125.

Isenberg, D., Walker, E. C. T. & Ryder, J. M. (1980) A top-down effect on the identification of function words, *Journal of the Acoustical Society of America,* **68**, AA6 (abstract).

Liberman, A. M. (1982) On finding that speech is special, *American Psychologist,* **37**, 148–167.

Liberman, A. M. & Mattingly, I. G. (in press) The motor theory of speech perception revised, *Cognition* **21**, 1–36.

MacDonald, J. & McGurk, H. (1978) Visual influences on speech perception processes, *Perception & Psychophysics,* **24**, 253–257.

Marslen-Wilson, W. & Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech, *Cognitive Psychology,* **10**, 29–63.

Massaro, D. W. (1984) Children's perception of visual and auditory speech, *Child Development,* **55**, 1777–1788.

Massaro, D. W. & Cohen, M. M. (1976) The contribution of fundamental frequency and voice onset time to the /zi/–/si/ distinction, *Journal of the Acoustical Society of America,* **60**, 704–717.

Massaro, D. W. & Cohen, M. M. (1983) Evaluation and integration of visual and auditory information in speech perception, *Journal of Experimental Psychology: Human Perception and Performance,* **9**, 753–771.

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

McNeill, D. (1985) So you think gestures are nonverbal? *Psychological Review,* **92**, 350–371.

Platt, J. R. (1964) Strong inference, *Science*, **146**, 347–353.
Popper, K. (1959) *The logic of scientific discovery*. New York: Basic Books.
Thompson, L. A. & Massaro, D. W. (1985) Speech and gestures in the development of referential understanding. Report 37/1985 Perception and Action. Zentrum für interdisziplinäre Forschung. Universität Bielefeld.