Modularity and The Motor Theory of Speech Perception

Edited by Ignatius G. Mattingly and Michael Studdert-Kennedy

Lawrence Erlbaum Associates, 1991

Robert Allen Fox

Division of Speech and Hearing Science, The Ohio State University, U.S.A.

This book represents the proceedings of a conference entitled "Modularity and the Motor Theory of Speech Perception" held in New Haven, CT June 5-8, 1988, honoring Alvin Liberman and his contributions to speech research. This volume includes original contributions, commentaries and panel discussions and is a rich source of information about the nature of speech perception, in general, and the strengths (and weaknesses) of the Motor Theory (developed by Liberman and his colleagues at Haskins Laboratories), in particular. The Motor Theory has been (and remains) the most influential model of speech perception over the course of the past three decades and Alvin Liberman's contributions to our understanding of speech perception has been unprecedented. A thorough review of so many excellent papers is quite impossible in the short space available. However, my aim here will be to summarize the most important points brought up by contributors, restricting myself to those papers relating most directly relevant to the Motor Theory.

Following brief introductory remarks by Franklin Cooper to the conference, Bjorn Lindblom examines evidence challenging two central assumptions of the Motor Theory, namely, modularity and gestural invariance. Although agreeing that the distal object of speech perception is the speaker's intentions, Lindblom argues that they need not be motorically invariant. Rather, one can observe variation in gestures in compensatory articulations (e.g., in bite-block conditions, in loud speech and in clear speech) and that gestures must only have sufficient perceptual contrast to allow a listener to identify the more primary ends of speech acts such lexical access, message comprehension and social interaction. However attractive this view might seem (after all, in normal conversation we "hear" words, messages, etc., and must train ourselves to discern phonetic variations), Lindblom provides little more than a skeletal framework for explaining how such an approach might operate. In terms of modularity (i.e.,

assumption of a specialized speech processor), he claims that the evolution of phonetic inventories in languages of the world may be seen as adaptations to motoric and perceptual constraints that are language-independent are in no way special to speech. Lindblom suggests that appeals to a biologically specialized processor at this point in our investigations of the production and perception of speech will make it methodologically more difficult to understand the role of non-special cognitive mechanisms and may bias our understanding of the phenomenon. Osamu Fujimura, in commenting on Lindblom's remarks, notes that although segmental aspects of speech are important, we need to be at least as concerned about the principles of temporal organization in speech.

The next two papers concern the theory of direct-realism, as applied to speech perception, which represents a recent, and quite formidable challenger to the Motor Theory. As described by Carol Fowler & Lawrence Rosenblum, the direct-realist theory of speech perception concerns itself with characteristics of the distal event (the phonetic gestures) rather than the proximal stimuli (the acoustic characteristics of the speech wave). In this view (described by Fowler in several earlier publications and familiar to most readers), the vocal tract produces phonetic gestures which, in turn, lawfully structures the acoustic signal. Listeners then use this structure to directly recover the phonetic gestures without identifying the acoustical stimuli directly. Fowler & Rosenblum challenge several claims of the revised Motor Theory (Liberman & Mattingly, 1985; Mattingly & Liberman, 1988). First, Fowler & Rosenblum argue that the direct-realist position does not require a listener to have direct access to his/her speech-motor system. Rather, the phonetic gesture is automatically recovered by the perceptual system in the same way that all perceptual systems directly recover the distal event from information in the proximal display. Second, Liberman & Mattingly have claimed that auditory perception is "homomorphic" in the sense that listeners hear acoustic stimulus as ordinary sound. Speech perception, on the other hand, is "heteromorphic" in the sense that the acoustic stimulus is heard as speech (and thus is not mapped onto the stimuli in a one-to-one fashion). Fowler & Rosenblum claim that most perception is, in fact, heteromorphic and that what is heard in regular auditory perception is the distal event producing the sound (e.g., a ringing bell) rather than the acoustic signal itself. Speech perception, thus, is not "special." Finally, Fowler and

Rosenblum provide an alternative explanation of duplex perception phenomenon suggesting that a similar effect will occur whenever two parts of a signal together can specify a sound-producing distal event. It is useful to remember that the Motor Theory and direct-realism represent very different theoretical perspectives. Whereas the Motor Theory attempts to describe the processes that might be responsible for converting an auditory stimulus to a cognitive percept (using some form of analysis-by-synthesis), direct realism provides no information whatsoever about what cognitive processes might be involved. Rather, it concentrates solely on the nature of the distal stimulus and how it might structure the stimulation to the sense organ (here, the ear). In this sense, direct-realism by its very nature cannot directly answer some of the questions being addressed by the Motor Theory.

In a commentary to Fowler & Rosenblum's chapter, Peter MacNeilage suggests that the notion of a phonetic gesture (as used both by motor theorists and direct-realists) has not been adequately defined; in particular, it has not been defined so as to allow a conclusive test of either theoretical approach. Motor theorists (Liberman & Mattingly, 1985) have claimed that motor acts are governed by invariants of some sort (the gestures), but have provided no detailed discussion of the nature of the underlying gestural control signal. MacNeilage is also concerned about the assumption that these gestures are innately specified, especially in light of the fact that neonates do not show the ability to produce all possible sounds and normally do not successfully produce a limited set of sounds appropriate to their native language until 3.5 years (nor are all possible "gestures" found in human languages produced during early babbling). In terms of direct-realism, MacNeilage criticizes Fowler & Rosenblum's failure to provide evidence for the assumption that there is no distinction between underlying and surface levels of gestures. In an earlier publication Fowler (1986), appealed to the "task dynamic" approach to speech production to support her claim that a distinction between a surface level (that would be affected by context) and underlying level (closer to neural control structures) is unnecessary. However, MacNeilage suggests that the task dynamic position involves a distinction between goal specification (underlying level) and a set of realization procedures (producing a surface level) that is not supportive of Fowler's view. MacNeilage also claims that Fowler's viewpoint is not adequate to account for those cases in which English listeners

recover only a single percept (e.g., a rhotic approximate) from two distinct gestures (bunched tongue [r] and retroflex [r]). A similar argument could be constructed with regard to the perception of vowels.

The next four papers turn to the question of how phonetic gestures might be acquired by young children. Marilyn Vihman provides evidence that each child discovers the relevant phonetic gestures in his/her native language on the basis of his/her own production patterns during babbling as well as on the perception of the speech of others. She contrasts this view with that suggested by Liberman & Mattingly (1985) in which gestures are said to be genetically specified and special to speech. Vihman's observational and experimental evidence falls into one of three basic categories in terms of differences between first-language learners: First, canonical babbling in deaf infants is qualitatively and quantitatively different from that of hearing infants—this underscores the importance of language input in the development of segmental vocalization (particularly of consonants). Second, differences among children learning the same language are greatest in the prelinguistic stage and the phonetic composition of the child's early words very similar to that of contemporaneous babbling. Since the earliest words reflect the closest (perceptual) match with the child's preexisting motor scheme "this would provide a plausible developmental origin for the perception-production link" (p. 76). Finally, Vihman outlines evidence of prosodic and phonatory differences among children learning different languages (there is little convincing evidence of cross-linguistic consonantal differences among prelinguistic infants). Vihman briefly discusses the possible implications of the age-related changes in perceptual ability (seen in work of Janet Werker and several others) to the perception-production link, but this is more speculative and less convincing than the first three points.

Michael Studdert-Kennedy continues with a discussion of the relationship between perceptual development and development of articulatory gestures. He states that the viewpoints of Vihman, on one hand, and Liberman and Mattingly, on the other, are not "entirely incompatible." In particular, the perceptual-motor link is not arbitrary (i.e., the speech signal is a lawful consequence articulatory movements), the link is biologically based (humans only), and it requires experience with other's speech. Studdert-Kennedy offers a thumbnail sketch of

the history of our knowledge about babbling beginning with Jakobson and suggests that current evidence consistently reveals a large gap between what a child recognizes (i.e., phonetic distinctions and lexical items he/she can discriminate and/or identify) and what a child can produce. This makes it more difficult for the Motor Theory's claims regarding the close identification between perceptual representations and internal, innately specified articulatory representations. In general, Studdert-Kennedy supports the notion of an early separation between comprehension and production, but cautions that "we should not let slip the central insight of Liberman & Mattingly ... that the structure of the speech signal is broadly isomorphic with the articulation that produced it ... [which permits] the child to discover the correspondences between its babbling repertoire and the speech it hears" (p. 88).

In the third paper, Janet Werker reviews the literature (to which she has consistently made important contributions) on developmental changes in speech perception ability. She notes that adults normally have difficulty in discriminating among non-native phonemic contrasts that young children (particular those in the first year of life) make easily. However, as Werker & Tees (1984) have shown, under special testing conditions, adults can be shown to be able to make these same non-native discriminations. On the basis of these and related data, Werker argues that this developmental change represents a type of perceptual reorganization rather than loss of discrimination ability. This reorganization happens by the end of the first year and older infants (11-13 months) can be shown to be less sensitive to at least some non-native contrasts than are younger infants (6-8 months). Werker offers five different hypotheses to account for this reorganization: (1) as an effect of experience hearing native language contrasts during perceptual development which allows the listener to maintain certain innate distinctions; (2) as a result of parameter resetting in the phonetic module (as suggested by Liberman & Mattingly); (3) as mediated through articulatory repertoire in canonical babbling (note Vihman above); (4) as a function of applying more general cognitive abilities to the domain of phonetic categorization; and (5) as a result of the development of a phonological system in the receptive lexicon. The first hypothesis is very general, with no indication on what aspect of auditory/linguistic processing the experience is having an effect. It is thus incomplete in that it does not account for the fact that listeners can make nonnative distinctions under some task

conditions (which suggests that the effect of experience is limited to linguistic categories and not general auditory abilities). In addition, some of these accounts are incompatible (e.g., 2 and 4), but Werker cannot reject any of them strictly on the basis of the data currently available (although she and her colleagues prefer an explanation based on the emergence of phonemic categories and systematic phonological system).

In a comment to Werker's paper, Peter Eimas shares her view that there is a set of innate phonetic categories that allow an infant to match acoustic signals with categorical representations and that during language development, listeners lose the ability to make phonetic/phonemic contrasts that do not occur in the native language they are learning (but do not lose auditory sensitivity). However, Eimas expands on the view (elaborated by Best and her colleagues) that not all nonnative phonetic distinctions are the same. Best, et al. (1988) suggested four different types of nonnative contrasts (although Eimas only talks about three of them): (1) those in which both nonnative categories correspond to a single native category; (2) those in which the member of the nonnative categories corresponds to contrasting native categories; (3) those in which one member corresponds to an existing category, but the other does not; and (4) those in which neither member of the nonnative contrast corresponds to a native language category. Eimas argues in the case of (1) that one could image that the infant would eventually ignore a phonetic distinction not used to contrast meaning in the language he/she is learning. Thus when information is contained in the acoustic signal that corresponds to such a linguistic contrast, it is no longer discriminated (since there is no relevant phonetic contrast on which to base the response). Eimas speculated that the purely acoustic difference would be ignored because "higher levels of linguistic representation takes precedence over lower levels, and lower levels of linguistic representation ..[phonetic]..take precedence over auditory representations." Eimas argues that in the case of (4) since the nonnative categories do not easily correspond to any native category, that discrimination is likely done at the auditory level alone. Eimas provides the prediction that the nonnative contrasts corresponding to (1) will be lost as phonetic categories before those in (4), with those in (2) and (3) lost at some intermediate point. However, a recent study by Polka (1992) provides data suggesting that this view may be too simple and that other factors (including individual differences) may affect a

listener's response to nonnative distinctions. Eimas concludes by suggesting that much of this data is compatible with the Revised Motor Theory's modularity hypothesis. In particular, in the reorganization of the perceptual system, although the module is penetrable this penetration involves linguistic information only. The auditory processing system is not affected.

Quentin Summerfield discusses the visual perception of phonetic gestures (e.g., lipreading) and provides a very useful synopsis of the relevant experimental literature. He first discusses the properties that the perceptual system must have in order to lipread or to process speech audio-visually (i.e., simultaneously); next he speculates as to whether these properties are consistent with the requirements of perceptual modules. His basic thesis is that the ability to lipread and its close relationship to the perception of speech audio-visually (as demonstrated in the McGurk effect, etc.) would suggest that "lipreading is not distinct from the linguistic-processing module but part of it" (p. 126). Summerfield speculates that the integration of visual and auditory cues takes place before phonetic or lexical categorization with the internal pre-categorical representation taking the form of invariant visual cues (which he does not define in greater detail).

In a comment on Summerfield's paper, Joanne Miller agrees with Summerfield in that if the speech perception module is specific to phonetic structure then it should utilize all relevant stimuli (across different modalities, including both auditory and visual) as, indeed, it seems to do. However, she notes that it is difficult to reconcile the fact that although there is little individual variation in perceiving speech on the basis of acoustic information, there is significant individual variation in the utilization of visual cues (note Summerfield's suggestion that good lipreaders are "born and not made"). In terms of the innateness issue (as required by the Motor Theory), although several studies have demonstrated that very young children show sensitivity to the acoustic and optic matches in articulated speech, we do not know if infants process linguistically relevant facial movements differently from other movements.

In an examination of a different type of visual perception relevant to language communication, Howard Poizner, Ursula Bellugi, and Edward Klima present data from a group of six congenitally deaf signers, three of whom had lesions in the left hemisphere (normally specialized for speech) and three in the right hemisphere (normally specialized for

visuospatial processing). These results indicate that significant disruption in the use of sign language (a visual-gesture communication system) occurs with left hemisphere damage, but not, necessarily, with right hemisphere damage. Thus hearing and speech are not necessary prerequisites for left hemisphere specialization for language. In addition, even though sign language is conveyed via visuospatial means, deaf signers can develop separate functional specialization for language vs. nonlanguage processing. One can assume that these data would support the assumption of innate predisposition for speech processing (although in a different modality) by the Motor Theory.

The next four papers are directed at understanding possible distinctions between auditory processing, on the one hand, and phonetic processing on the other. Albert Bregman describes a very different approach to the explanation of the speech research data. In particular, Bregman suggest that speech perception results from schema-based recognition, rather than the special properties of a phonetic module. These schemas represent "epistemic systems...whose job it is to understand the world" and are characteristic of a set of mechanisms that cut across different mental capacities and sensory systems. A schema is a knowledge "specialist" that deals with regularities that the entity encounters at some level of abstraction. These hierarchically organized schemas are found in sensory (perceptual) systems as well as behavioral (production) systems. He suggests that the phenomenon of duplex perception arises whenever the same (sensory) evidence can be shared by different schemas. Unfortunately, Bregman only briefly mentions auditory scene analysis (which Darwin discusses more extensively in a later paper)—to which he has made many contributions—in terms of whether or not it occurs prior to phonetic processing. In general, this schema approach (as outlined by Bregman) seems too powerful and unconstrained (in the sense of early generative grammar) and it is unclear just what type of data would serve to falsify his theory.

In a comment to Bregman's paper, David Pisoni does not address the issue of schema-based recognition directly, but rather explores the distinction between speech and nonspeech modes of processing. Pisoni first provides a brief historical synopsis of experiments addressing differences between the perception of speech and nonspeech stimuli. He notes that

although these studies consistently demonstrated that a given speech signal was processed differently than its nonspeech counterpart, that these early studies often failed to ensure that the two sets of stimuli were truly equivalent in terms of the psychoacoustic properties being manipulated or that subjects were properly familiarized with the nonspeech signals. However, more recent experiments controlling these factors (many carried out by Pisoni and his colleagues) have shown significant dissociation between auditory and phonetic responses on the same set of stimuli. These data support the conclusion that speech signals elicit a "qualitatively different mode of processing, a speech mode, that appears to be quite different from the way other nonspeech auditory signals are responded to" (p. 236). However, as some psychoacousticians have pointed out (e.g., Larry Feth, personal communication), the experiments demonstrating categorical perception with speech stimuli (as opposed to nonspeech stimuli) seldom provide extensive practice to their subjects (compared to psychoacoustic methods employing a small n with large numbers of repetitions) and often fail to separate contributions of sensitivity and bias. It could be, as Johnson and Ralston (1990) have suggested, that speech categories differ from nonspeech categories only because listeners have much greater experience in categorizing speech sounds.

In a discussion of the relationship between auditory processing and speech perception, C.J. Darwin argues that although there may be a special phonetic processor it does not "preempt" auditory processes (the accepted Haskins' explanation for duplex perception) but, rather, utilizes the organized (grouped) auditory information provided by earlier processing. Darwin summarizes a variety of studies that indicate whether or not a given sound component (e.g., a harmonic) is integrated with other acoustic components into a single speech percept depends on factors such as the pitch of the harmonic or its onset time—i.e., this integration depends upon how the auditory grouping or scene analysis segregates the auditory components into one or more acoustic sources. The speech module creates a percept based on these auditory groupings that are consistent with articulatory constraints.

In a short commentary, Bruno Repp agrees that auditory scene analysis may operate prior to more abstract (phonetic) processing. These arguments buttress the claim for a separation between an auditory level of processing and a more specialized phonetic level of processing.

He also briefly discusses the nature of dichotic integration in duplex perception suggesting that the relevant experimental data suggest the importance of relationships between various auditory factors and their relevance to "natural auditory events".

The next two papers (the first by Helen Neville, the second by Masakazu Konishi) concentrate on neurophysiological aspects of speech perception. Neville reports behavioral and electrophysiological data demonstrating left hemispheric specialization for English or ASL in both deaf and hearing subjects while Konishi provides an overview of the neurological substrates of spatial localization in the barn owl. Together these studies show that portions of the brain can, in fact, be specialized for very specific functions (e.g., spatial localization) as well as more general functions (e.g., language processing). However, beyond demonstrating that such specialization can occur, they provide little or no support for the existence of a specialized "phonetic module" and few details on how its neural substrate would be organized.

Catherine Browman and Louis Goldstein provide a brief description (described at greater length in several other publications) of their conception of articulatory gestures and then discuss how these gestures could form the basis for lexical representations and account for phonological alternations as well as historical changes. Browman and Goldstein base their view of the gesture on the concept of coordinative structures developed by Elliot Saltzman, Scott Kelso and other colleagues at Haskins. The "articulatory gesture" does not correspond to actual articulatory movements but represents the "coordinated movement of groups of articulators ... [that] can be characterized using dynamical equations" (p. 314). These equations stem from work done by Kelso and others modeling the control of other muscular systems (e.g., arm movement) and thus do not represent a domain of control unique to speech. This view of gestures is very attractive—at least for consonantal distinctions—but from the reviewer's point of view, these gestural "scores", as presently described, are inadequate for vowels (which may still be best defined acoustically/auditorally). From the point of view of the Motor Theory, Browman and Goldstein present a possible and well-considered characterization of what a speaker's phonetic "intentions" might be. However, little space is given to discussion about how a listener might make (in any direct manner, as would be required by the Motor Theory) the conversion from acoustic patterns to gestural patterns.

The next five chapters (including presentations by Ignatius Mattingly, Stephen Crain & Donald Shankweiler, Paul Bertelson & Beatrice de Gelder; commentaries by Daniel Holender, and Edward Klima; and two panel discussions) have a much broader focus than speech perception and the Motor Theory. In particular, the issues covered include speculations on the nature of reading (and reading disorders) in the language module; arguments as to whether this specialized module (as strictly defined by Fodor, 1983) can be appropriately expanded to including all of language processing and not just speech processing; and the nature of metalinguistic awareness. However interesting, these papers provide little direct insight into the Motor Theory itself, and I will not discuss them further (in this already lengthy review).

James Jenkins provides a useful summary of the conference. He concludes that there are at least six important claims associated with the Revised Motor Theory: (1) there is, to at least some extent, neural specialization for language systems; (2) this specialization begins very early in life; (3) that speech production is regularized at some point during the babbling stage; (4) that visual perception plays an important role in speech perception and the development of speech; (5) that speech sounds are "separated", in some fashion, from the other acoustic input that the listener hears; and (6) that speech sounds may tell us about the speech gestures that produced them. In terms of modularity, Jenkins agrees with many of the panel discussants in suggesting that although some subsystems of language may be modular (e.g., the phonetic module), that it is unlikely that the entire language systems is strictly modular. In addition, he notes several issues that have not been discussed adequately at the conference (that are relevant to theories of speech processing) including (1) the well-known plasticity of the nervous system, (2) whether putative modules are, in fact, "leaky" and not as impenetrable as a strict Fodor-type module would require, and (3) the extent to which speech behaviors are "soft-wired" (through years of practice and automatization) as opposed to "hard-wired" (innate). I would add that discussions concerning the phonetic module would have benefited from greater input from the psychoacoustic literature (particularly in term of practice effects and listener biases) as well as a discussion of context effects and whether they occur during or following phonetic categorization.

This book is an important contribution to the speech literature and should be required

reading for all graduate students (as well as established speech researchers) interested in perceptual issues. It is not a simple paean to the Motor Theory or to Alvin Liberman, but rather represents a serious discussion of the theory along with a presentation of formidable challenges to many assumptions which the theory explicitly or implicitly makes.

References

Best, C., M., Roberts, G. & Sithole, N. (1988). The phonological basis of perceptual loss of nonnative contrasts: Maintenance of discrimination among Zulu Clicks by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 245-260.

Fodor, J.A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Johnson, K. & Ralston, J.B. (1990). Automaticity in speech perception. *Research on Speech Perception*, No. 16, Bloomington, IN: Indiana University.

Liberman, A.K. & Mattingly, I.G. (1985). The Motor Theory of speech perception revised. *Cognition*, 21, 1-36.

Mattingly, I.G. & Liberman, A.K. (1988). Specialized systems for speech and other biologically significant sounds. In G. Edelman, W. Gall, & W. Cowan (eds.), *Auditory function: The Neurological Basis of Hearing*, (pp. 775-795). New York: Wiley.

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics*, 52, 37-52.

Werker, J. & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, 75, 1866-1878.