668

ECOLOGICAL PSYCHOLOGY, 1(2), 181–193

Cues to What? A Comment on Diehl and Kluender "On the Objects of Speech Perception"

Michael Studdert-Kennedy
University of Connecticut, Yale University, and Haskins Laboratories

The principle of auditory enhancement is a valuable extension of the theory of adaptive dispersion. However, neither the principle nor the theory has any bearing on what we perceive in speech. All perceptual objects, including those of speech, are abstract, amodal structures made available to cognition through one or more sensory modalities. A focus on the modalities or media of information transfer in speech perception diverts attention from two central and related aspects of speech function: perceptuomotor functional equivalence and imitation. Arguments and evidence from studies of lipreading, short-term memory, and infant behavior are put forward to support the postulation of an output unit, the phonetic gesture, at a level in the communication chain corresponding on the input side to the acoustic cue. The object of speech perception is then taken to be the amodal phonetic segment, a cohesive set of direct mappings between sound and gesture.

MOSTLY THEORY

Ethological studies leave little doubt that animal communicative displays typically derive from prior noncommunicative motor repertoires and are then elaborated in evolution to enhance their perceptual salience and communicative effect (Hinde, 1970, chapters 27 and 28). Therefore, I am in full agreement with the proposal of Diehl and Kluender that language communities tend to select their phonological segment inventories from ". . . signals that are highly valued on both auditory and articulatory grounds" (cf. Lindblom, in press-a, in press-b; Lindblom, MacNeilage, & Studdert-Kennedy, 1983, 1989), and I view their principle of auditory enhancement as a valuable extension of Lindblom's theory

Requests for reprints should be sent to Michael Studdert-Kennedy, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695.

of adaptive dispersion. To be sure, Diehl and Kluender are sometimes a trifle extreme in their eagerness to upgrade auditory factors and downgrade articulatory factors in shaping phonologies. But, insofar as linguists have indeed tended to account for phonetic and phonological universals by appeal solely to ". . . physical and physiological constraints on speech production (rather than to auditory principles)," Diehl and Kluender are justified in their attempt to redress the balance. A phonetic universal is, after all, a universal of both production and perception.

However, when Diehl and Kluender turn to perceptual process and to the "objects of speech perception," their auditory bias leads them to altogether too narrow a view of speech function. Certainly we listen in order to hear in order to understand. But children, learning their native language, also listen in order to hear in order to speak. Both these facts must be accounted for in an adequate theory of speech perception. Children are able to learn to speak because they are heirs to a specialized capacity for imitation, and because every spoken utterance specifies (in a sense to be (somewhat) clarified later) the articulation necessary to reproduce it.

The general drift of my argument then is this. The "objects of speech perception" are neither auditory nor articulatory. All perceptual objects or events, including those of speech, are abstract, amodal structures in the world made available to cognition through one or more sensory modalities. Perception is a process of informational transfer by which structure in the world is transformed into structure within an organism: events in the world set up topologically correspondent patterns of neural activity in the perceiver. If the events are the subject's own actions, then the patterns are isomorphic with the neural patterns that controlled the action. (This, at least, is the necessary assumption of feedback theories of motor control.) The percept is then a pattern of relations that constitutes this isomorphism between input and output: an abstract, amodal, perceptuomotor structure.

If the perceived events are not the subject's own actions but those of a conspecific, then the pattern is again isomorphic with the pattern that controlled the action, because the bodily structures of perceiver and actor are isomorphic. Although not a sufficient condition for a capacity to imitate, this isomorphism is clearly necessary, because imitation of an act requires that the perceptual process induce in the imitator a neuromotor control structure isomorphic with that which produced it. That is why (with a few exceptions) imitation is restricted to copying the acts of conspecifics. Implicit in this formulation is an assumption that places reciprocal constraints on theories of both the production and the perception of acts learned by imitation: units of perception map directly onto units of production. We therefore need, for speech, an observable unit of production at a finer grain of analysis and at a lower level in the communicative chain than the linguistic segments that Diehl and Kluender evidently take to be

the objects of perception. A likely candidate, set at a level in production corresponding to "cues" in perception, is the phonetic gesture, discussed later.

Before this, notice that imitation is not normally mimicry: The copy is neither exact nor necessarily executed by exactly the same patterns of muscular action as those of the model. We do not expect a child learning to eat with a fork to copy the precise movements of the model. We expect only that the child should execute certain acts—loading the fork, transferring food to the mouth, withdrawing the fork—in a broad spatiotemporal pattern functionally equivalent to that of the model.

Similarly, we do not expect a child learning to speak to imitate a model exactly (although, as the transmission of dialect attests, the child may eventually come close to this). And we certainly do not expect the child—with a vocal tract quite differently proportioned than that of the adult model—to accomplish its phonetic ends by exactly the same spatiotemporal pattern of articulatory movements as the adult. In fact, to repeat a statement for which Diehl and Kluender take me to task (footnote 2), ". . . the claim that an utterance specifies its articulation cannot mean that it specifies precisely which articulators are to be engaged, and when. Rather, it must mean that the utterance specifies a range of functionally equivalent articulatory actions" (Studdert-Kennedy, 1987, p. 71).

Diehl and Kluender find this statement ". . . exceedingly unparsimonious inasmuch as the equivalence class of perceptuomotor objects it refers to is virtually unbounded in size." Moreover, they object that the statement ". . . obscures the real basis of functional equivalence. The only thing that all members of a functional equivalence class of gestures have in common . . . is their similar acoustic/auditory effects." If we change "acoustic/auditory effects" to "perceptual effects" I entirely agree with the last sentence. In fact, a few lines beyond the quoted passage I observed that ". . . the arbiter of [functional] equivalence [in speech] . . . is a listener's judgment." In other words, speech is listener-oriented, and the goal of a speaker is to produce an acoustic signal that a listener will understand, despite its variability across contexts (cf. Lindblom, in press-a).

However, the first objection is puzzling. First, because an equivalence class of perceptuomotor objects is exactly as "unbounded" as the "equivalence classes of acoustically diverse . . . speech tokens" that Diehl and Kluender taught their quail. Second, and more importantly, the objection is puzzling because it seems blind to a commonplace of animal behavior studies, namely, that many different motor patterns can be marshalled to accomplish the same end. Even the stereotyped (and phylogenetically ancient) patterns of rodent self-grooming, for example, can be executed by novel combinations of a mouse's muscle systems, if the animal is prevented from following its habitual routines (Fentress, 1981). "Unparsimonious" or not, the summoning of variable motor patterns drawn from an unbounded set to achieve a relatively small set of invariant goals is the

foundation of the capacity of living organisms to adapt their behavior to a changing environment. As Lashley (1951) and many others recognized, the neurological basis for this biological universal of motor equivalence is one of the deepest and most difficult problems in neurobiology.

The only serious attempt to address the issue, so far as I know, has been through the work of Bernstein (1967) and his associates in the Soviet Union and of their colleagues in this country (e.g., Greene, 1972; Kelso, Holt, Kugler, & Turvey, 1980; Turvey, 1977). These students have cast the problem as one of regulating the motor system's internal degrees of freedom in the execution of its functions. They model the synergistic actions of many muscles in terms of "coordinative structures," that is, groupings of muscles temporarily marshalled to act as a single unit in the performance of a particular function. Implicit in the model is the notion that different groups of muscles, or differently weighted patterns of action in the same group of muscles, may be marshalled to execute the same function on different occasions. Thus, the approach contributes to our understanding not only of how a motor system regulates its degrees of freedom, but also of how an animal adapts its behavior to achieve the same end in different contexts.

Application of these principles to speech has entailed the definition of *phonetic coordinative structures*, or synergistic functions, termed *phonetic gestures* (e.g., Browman & Goldstein, 1986; Fowler, 1986; Fowler, Remez, Rubin, & Turvey, 1980; Saltzman, 1986). A gesture is, by dictionary definition, an intentional communicative act. A phonetic gesture is an act that effects or contributes to the formation of a vocal tract configuration or sequence of configurations specifying a particular phonetic element. This element corresponds to some abstract linguistic unit, such as a feature, phoneme, syllable, word, or prosodic phrase.

Ideally, we would derive the set of phonetic gestures used in any particular language by selection from the pool of all possible articulations and their auditory consequences. The selection pressures would be such as to assure communicatively sufficient auditory and sensorimotor discriminability across the entire phonological system (Lindblom, in press-a, in press-b). However, for immediate practical use in modeling articulation, we must have recourse to speech patterns that we actually observe. At least two ways of estimating the set of phonetic gestures necessary to describe a particular language have been proposed. One, going from speech sounds to vocal tract shapes, uses factor analysis of formant frequencies to ". . . make an empirical estimate of the degrees of freedom of the motion of the tongue" (Ladefoged, Harshman, Goldstein, & Rice, 1978, p. 1028; see also Ladefoged, 1980). The other draws on the analysis of articulator trajectories (Browman & Goldstein, 1986) and, where possible, on articulatory synthesis (Saltzman, Rubin, Goldstein, & Browman, 1987). Both paths lead to gestural inventories that include elements more or less familiar from traditional introspective phonetics, such as bilabial closure, glottal open-

ing, velic opening, tongue front raising, and so on. Note that, although some phonetic gestures may be executed by movements of a single articulator (e.g., velic opening), others normally require the coordinated movements of several more or less independent articulators (e.g., tongue raising, by movements of jaw and tongue; bilabial closure by movements of upper lip, lower lip, and jaw.)

The distinction between movements and gestures is critical, because most of the variability in speaking (and, therefore, in the acoustic signal) arises at the level of movements not of gestures. All speakers close the lips to form /p,b,m/, raise the front of the tongue to form /i/, lower the back of the tongue to form /a/, and so on. Equivalence classes of phonetic gestures (e.g., larynx lowering and lip protrusion to form /u/, cited both by Diehl and Kluender and by myself in the paragraph from which Diehl and Kluender quote) are, therefore, rare, whereas equivalence classes of movements are common. One task for a theory of speech production is then to explain how listeners vary the timing and amplitude of articulator movements as a function of phonetic context, rate, and speaking style to assure a constant phonetic percept. The reciprocal task of perceptual theory is to explain how listeners exploit the systematic acoustic variability, introduced by variation across an equivalence class of movements, to recover the speaker's phonetic goal. Notice that both these tasks are special cases of what we would require of a general theory of imitation.

In one respect, Diehl and Kluender's approach is strikingly similar to that just outlined. They believe that ". . . the separate gestural components or acoustic correlates . . . are . . . more or less independently controllable by talkers." In other words, to every cue there is a gesture and vice versa, thus assuring one-to-one correspondence between units of production and units of perception. The difference between the two approaches lies in the definition of and in the procedures for determining the presumed "gestural components." Diehl and Kluender apparently equate gestures with articulator movements; they isolate gestures not by analysis of articulation or the acoustic signal, but by inference from perceptual studies of "cues." They thus shut themselves off from what I take to be a central issue in speech research—the problem of perceptuomotor functional equivalence.

Moreover, as Diehl and Kluender themselves acknowledge, ". . . often two cues that are mutually enhancing are the product of the same gesture and are thus not independently controllable." In other words, inference from cue to gesture is unreliable. One reason for the uncertainty is that cues are typically isolated by manipulation of an articulatorily unconstrained terminal analog synthesizer. Thus, although they are independently controlled by an experimenter, they may not be independently controlled by a talker. Their perceptual effect is then simply due to their lawful covariation in the patterns we are accustomed to hear as speech.

Diehl and Kluender seem to acknowledge this difficulty in their selection of cues from Lisker's (1978) catalogue of 16 acoustic features associated with

utterance-medial /b/ and /p/. Three of the four cues they selected do seem to be, at least potentially, under independent control of the speaker (lip-closure duration, preceding vowel duration, presence/absence of glottal vibration). But many others that they disregarded are not, and this indeed is the central point of Lisker's article, which concludes with these words:

> The ensemble of features [i.e., cues], spread over two syllables, shows a degree of disparity at the purely acoustic level that seems strange, given that they all affect the same phonetic judgment. However, they can all be referred back to a single crucial articulatory difference in the management of the larynx. (Lisker, 1978, p. 132)

In another article, primarily intended to clear up the widespread misunderstanding of voice onset time (VOT) as an acoustic rather than an articulatory dimension (a misunderstanding shared by Diehl & Kluender who refer to VOT as a "major voicing cue for initial stop consonants"), Abramson (1977) wrote:

> The timing of the valvular action of the larynx may be said to be a physiological mechanism that underlies such acoustic phonetic features as the onset and offset of voice pulsing, intensity of plosive release, amount of aspiration noise, attenuation of the first formant, onset of voice-excited formant transitions, and perturbations of fundamental frequency. These features intersect in various combinations to furnish the phonetic basis of phonologically relevant voicing and aspiration. . . . (p. 295)

And later, introducing a discussion that extends the conceptual framework to medial and final consonants, Abramson (1977) continued:

> A more appropriate concept is simply that of voice timing—i.e., laryngeal timing—which subsumes voicing as a special case . . . VOT is a laryngeal dimension with a complex set of intersecting, overlapping or even discrete acoustic cues. . . . (pp. 297–298)

Diehl and Kluender are, of course, free to question whether all, or even any, of the acoustic features, listed by Lisker and Abramson, are indeed merely incidental consequences of laryngeal timing. And this they have done for four cues, as indicated in the target article. Yet it surely behooves them to acknowledge that their endeavor is far from complete, lacking both the scope and parsimony of Lisker and Abramson's unified account of the articulatory origins of diverse voicing cues in utterance initial, medial, and final positions. Such an acknowledgment, extended to other dimensions as well as voicing, would entail placing at least as much emphasis on redundancy in the speech signal that arises automatically from articulatory mechanics and aerodynamics, outside a speaker's control, as on the controlled redundancy that Diehl and Kluender promote.

The functional value of both forms of redundancy lies in specifying phonetic structure at a level below that of the linguistic segment, namely, that of the phonetic gesture. Particular gestures, and particular combinations of gestures, may then have been selected for deployment in the world's languages precisely because they offer a rich conspiracy of correlated cues specifying an economical, easily achieved gestural pattern. This view complements rather than contradicts Diehl and Kluender's theory of auditory enhancement.

MOSTLY EXPERIMENT

Lip Reading: Adults

In their brief treatment of the McGurk–MacDonald effect, Diehl and Kluender propose that lipreading is the outcome of "perceptual learning." In this, they are at least partially correct, because individuals vary in their ability to lipread (children tend to be worse than adults) and can improve with practice (Massaro, Thompson, Barron, & Laren, 1986). However, Diehl and Kluender imply that the learning comes about by simple "association" between the optic and acoustic properties of speech. The phrase, "learned association," implies a link between arbitrary patterns, perhaps analogous to the link between an orthography and the spoken language it represents. Yet, studies of lipreading under various conditions (not only those of conflicting acoustic–optic information) have demonstrated that optic and acoustic information ". . . is integrated before phonetic or lexical categorization takes place; the two streams are analogue at their conflux" (Summerfield, 1987, p. 16). This conclusion argues that acoustic and optic streams share the same amodal structure; learning to lipread is a matter not of associating two arbitrary patterns, but of discovering their structural equivalence.

Further evidence for rapid transformation of the speech input into an amodal representation comes from studies of short-term memory for list of words. In their well-known article, "Precategorical Acoustic Storage (PAS)," Crowder and Morton (1969) showed that listeners recall words from the end of an auditorily presented list better than words from the middle (recency effect) of the list. They also showed that the effect was significantly reduced if the words were presented graphically (modality effect) or if a spoken word was appended to the list, not for recall, but simply as a signal to begin recall (suffix effect). Evidently, the suffix somehow interferes with the representation of recent items. That this interference is at some relatively low (precategorical) level is argued by the facts that the effect: (a) is unaffected by degree of semantic similarity between suffix and list, (b) is reduced if suffix and list are presented to opposite ears, or are spoken by different voices, and (c) does not occur if the suffix is a tone or burst of noise.

Campbell and Dodd (1980) used this paradigm to test recall of digits, either

lipread (seen, but not heard) or presented graphically, with and without a spoken suffix (heard, but not seen). They found significant recency and suffix effects for the lipread lists, but not for the graphic lists (thus undermining the claim for an auditory/visual modality effect). In a complementary study, Spoehr and Corin (1978) demonstrated that a lipread suffix (seen, but not heard) reduced recall of auditorily presented lists. Evidently, speech seen, but not heard, and speech heard, but not seen, share a common representation. Moreover, because Campbell and Dodd (1980) found no suffix effect for the graphically presented lists, the shared representation is not at some abstract, phonological level where spoken and written language converge. Rather, these studies, like those reviewed by Summerfield (1987), suggest an analog, amodal representation in some form common to the light reflected and the sound radiated from mouth and lips. (For further work along these lines, see Crowder, 1983; Greene & Crowder, 1984).

Lipreading: Infants

Infants are also sensitive to structural correspondences between acoustic and optic specifications of phonetic events. Dodd (1979) showed that 4-month-old infants watched the face of a woman reading nursery rhymes more attentively when her voice was synchronized with her facial movements than when it was delayed by 400 msec. However, more than mere synchrony is involved. Kuhl and Meltzoff (1982) showed that 4- to 5-month-old infants looked longer at the face of a woman repeatedly articulating the vowel they were hearing (either [i] or [a]) than at the same face articulating the other vowel in synchrony. The preference disappeared when the signals were pure tones matched in amplitude and duration to the vowels; the infant preference was evidently for a match between a mouth shape and a particular spectral structure. Similarly, MacKain, Studdert-Kennedy, Spieker, and Stern (1983) showed that 5- to 6-month-old infants preferred to look at the face of a woman repeating the disyllable they were hearing rather than the synchronized face of the same woman repeating another disyllable.

The interest of all these lipreading studies, in our context, is that they take us a step closer toward understanding the transformation from sensory input to motor output, the mechanism by which light or sound "gets into a muscle." Presumably, acoustic and optic representations share a common amodal form because the optic signal that we read from the lips arises from exactly the same physical source as the acoustic signal that we hear, namely, the speaker's articulations.

Hemispheric Specialization for Lipreading in Adults and Infants

Further evidence of a perceptuomotor link comes from studies of hemispheric specialization. If lipreading is a visuospatial skill, established by associative

learning, then we might expect it to be lateralized to the right hemisphere. If it is a language skill closely related to speech production, then we would expect it to be lateralized to the left hemisphere. In an extensive study too complicated for me to describe here in any detail, Campbell, Landis, and Regard (1986) tested two women, one with a right occipitotemporal lesion, the other with a left occipitotemporal lesion. Neither woman was aphasic.

The woman with the right occipitotemporal lesion could neither recognize familiar faces nor correctly classify facial expressions. Yet her performance on a variety of lipreading tasks (including the McGurk–MacDonald blend illusion) was entirely normal; she was fast and accurate in classifying photographs of faces according to the speech sound being spoken, but could not classify the same faces by individual or even by sex. The woman with the left occipitotemporal lesion suffered severe alexia, being able to read only one letter at a time, but was unimpaired in the recognition of faces and expressions; she was also able to lipread spoken digits. But she was unable to sort faces by the speech sound being spoken, nor was she susceptible to the audiovisual blend illusion. Obviously, the cases are complex, but what is remarkable is the double dissociation between sites of lesion, facial recognition, and aspects of lipreading skill.

We also have evidence for localization of lipreading to the left hemisphere in 6-month-old infants. In the study by MacKain et al. (1983), infant preferences for a match between the facial movements they were watching and the speech sounds they were hearing were statistically significant only when the infants were looking to their right sides. The result is consistent with studies by Lempert and Kinsbourne (1982) and Kinsbourne (1972), showing that attention to one side of the body may facilitate processes for which the contralateral hemisphere is specialized. From this we may infer that infants with a preference for matches only on their right sides were revealing a left hemisphere capacity for recognizing acoustic–optic correspondences in speech.

Prelinguistic Phonetic Imitation in Infants

For many years, established wisdom held that infants only learned to imitate facial expressions after being exposed to the sight of their own faces. Similarly, infants were said to be able to imitate speech sounds only after exposure to the sounds of their own voices in babbling. Although babbling probably does contribute to the shaping of speech sounds (Locke & Pearson, 1988), some capacity for speech sound imitation is present before the onset of babbling. Meltzoff and Moore (1977) showed that 12- to 21-day-old infants could imitate both arbitrary mouth movements, such as tongue protrusion and mouth opening, and arbitrary hand movements, such as opening and closing the hand by serially moving the fingeres. Subsequently, they demonstrated imitation of mouth movements in infants less than 72 hr old (Meltzoff & Moore, 1983). (Meltzoff & Borton, 1979, also demonstrated that 4-week-old infants could

recognize structural correspondences between objects presented visually and objects they had not seen, but had previously explored tactually with their mouths—A nice puzzle for the associationist!)

In the mouth movement studies, imitation was elicited without vocalization, but had vocalization occurred, its spectral structure would presumably have reflected the shape of the mouth. In fact, Kuhl and Meltzoff (1982) reported that 10 of the 32 4- to 5-month-old infants in the study just cited ". . . produced sounds that resembled the adult female's vowels . . . alternating their vocalizations with hers" (p. 1140). Whether these imitations were based on the optic or on the acoustic specification of the female's articulations or on an integrated amodal structure, we cannot tell. But it is difficult to resist the conclusion that these infants already had rudimentary control of the perceptuomotor link between their acoustic–optic percepts and articulation.

Although the evidence, cursorily reviewed under this and the three preceding headings, may not be conclusive, it clearly challenges Diehl and Kluender's claim that: "For all instances of acoustic or intermodal covariation subsumed under our second qualification, it is possible that perceivers simply learn [by association] to use informational correlates to identify a phonetic message, without actually detecting the (nonvisible) gestures or their underlying control structures by which that message was articulatorily realized."

Quail

I have three comments. First, about 10 years ago, Zoloth et al. (1979) trained two species of macaque (pig-tail and bonnet) and an African vervet to learn an arbitrary discriminative response to contrasting calls of the Japanese macaque. The Japanese macaques learned the arbitrary response significantly more rapidly than the other species. Moreover, the processes underlying the Japanese macaques' responses to their own calls are evidently localized in the left cerebral hemisphere, whereas those of the other two species of macaque are not (Petersen, Beecher, Zoloth, Moody, and Stebbins, 1978; cf. Heffner & Heffner, 1984). These studies demonstrate, not unexpectedly, that the information in a Japanese macaque's call is in the acoustic signal, and that even species lacking this macaque's hemispheric specialization can be trained to discover that information. But the species' differences in rate of learning the arbitrary response and in hemispheric specialization suggest that showing a particular discriminative task to be within the psychophysical competences of two different species is not necessarily to show that the two species' percepts are equivalent. Perhaps Diehl and Kluender's next undertaking should be to demonstrate that quail who have learned to discriminate between acoustic–phonetic categories can make the same discrimination by lipreading! We also may well suspect that had Kluender, Diehl, and Killeen (1987) conditioned quail to execute arbitrary discriminative responses to categories of quail vocalizations, the birds would

have learned the task in rather less than the 8,000 to 12,000 trials they needed to form categories of human speech sounds. We may even suspect that infants (in whom left hemisphere sensitivity to speech sounds is normally present at birth, Molfese & Betz, 1988) could be more rapidly trained to discriminate categories of speech sounds than of quail vocalizations. But both these studies remain to be done.

Second, teaching quail auditory categories that correspond to phonetic categories is only of interest if phonetic categories play a functional role in speech perception. I have no space to go into the issue here, but I believe it remains to be shown that they do. Phonetic categorization may be an epiphenomenal consequence rather than a functional precondition for perceiving speech (cf. Studdert-Kennedy, 1986).

Third, whatever the function of phonetic categories, we have no reason to believe that they are polymorphous, have no single invariant property, or are learned by conditioning. In fact, there are grounds for believing that phonetic categories form by self-organization according to shared phonetic function (Lindblom et al., 1983) in a fashion perhaps analogous to the formation of syntactic categories (e.g., Maratsos & Chalkley, 1980).

CONCLUSION

I applaud Diehl and Kluender's endeavor to redress the balance between auditory and articulatory accounts of the origins of speech sound patterns. However, in my view, although the primary modality by which speech patterns are conveyed is audition, the objects of segmental speech perception themselves are neither auditory nor articulatory. Rather, they are amodal structures, each segment a cohesive set of direct mappings between sound and gesture. Accordingly, I do not applaud Diehl and Kluender's relative disregard of speech production: In my view, we are unlikely to have an adequate theory of speech perception as long as we do not have an adequate theory of speech production. Nor, finally, do I applaud their disregard of speech acquisition. To quote some words that I wrote nearly 10 years ago:

> The infant is a listener, a very attentive one, because by learning to listen it learns to speak. In my opinion, only by carefully tracking the infant through its first two years of life shall we come to understand adult speech perception and, in particular, how speaking and listening establish their links at the base of the language system. (Studdert-Kennedy, 1980, p. 45)

ACKNOWLEDGMENT

Preparation of this article was supported in part by National Institute of Child Health and Development Grant No. 01994 to Haskins Laboratories.

REFERENCES

Abramson, A. (1977). Laryngeal timing in consonant distinctions. *Phonetica, 34,* 295–303.

Bernstein, N. (1967). *The coordination and regulation of movement.* London: Pergamon.

Browman, C., & Goldstein, L. (1986). Toward an articulatory phonology. In C. Ewan & J. Anderson (Eds.), *Phonology yearbook* (Vol. 3, pp. 219–254). Cambridge, UK: Cambridge University Press.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology, 32,* 85–99.

Campbell, R., Landis, T., & Regard, M. (1986). Face recognition and lipreading: A neurological dissociation. *Brain, 109,* 509–521.

Crowder, R. G. (1983). The purity of auditory memory. *Philosophical Transactions of the Royal Society of London, B 302,* 251–265.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception and Psychophysics, 5,* 365–373.

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

Fentress, J. (1981). Order in ontogeny: Relational dynamics. In K. Immelman, G. Barlow, M. Main, & L. Petrinovich (Eds.), *Behavioral development* (pp. 338–371). Cambridge, UK: Cambridge University Press.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

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

Greene, P. H. (1972). Problems of organization of motor systems. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology.* New York: Academic.

Greene, R. L., & Crowder, R. G. (1984). Modality and suffix effects in the absence of auditory stimulation. *Journal of Verbal Learning and Verbal Behavior, 23,* 371–382.

Heffner, H. E., & Heffner, R. S. (1984). Temporal lobe lesions and perception of species-specific vocalizations by macaques. *Science, 226,* 75–76.

Hinde, R. A. (1970). *Animal behavior.* New York: McGraw-Hill.

Kelso, J. A. S., Holt, K. G., Kugler, P. N., & Turvey, M. T. (1980). On the concept of coordinative structures as dissipative structures: II. Empirical lines of convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 1–47). Amsterdam, The Netherlands: Elsevier/North Holland.

Kinsbourne, M. (1972). Eye and head turning indicates cerebral lateralization. *Science, 176,* 539–541.

Kluender, D. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237,* 1195–1197.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138–1141.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56,* 485–502.

Ladefoged, P., Harshman, R., Goldstein, L., & Rice, L. (1978). Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America, 64,* 1027–1035.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.

Lempert, H., & Kinsbourne, M. (1982). Effect of laterality of orientation on verbal memory. *Neuropsychologia, 20,* 211–214.

Lindblom, B. (in press-a). Models of phonetic variation and selection. In A. Piazza (Ed.), *Language change and biological evolution.* Stanford, CA: Stanford University Press.

Lindblom, B. (in press-b). The status of phonetic gestures. In I. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Lawrence Erlbaum

Associates, Inc.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations for language universals* (pp. 181–203). The Hague, The Netherlands: Mouton.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1988). *Evolution of spoken language.* Unpublished manuscript.

Lisker, L. (1978). *Rapid* vs. *rabid:* A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research, 54,* 127–132.

Locke, J., & Pearson, D. M. (1988). Linguistic significance of babbling: Evidence from a tracheostomized infant. *Report from the Neurolinguistics Laboratory, Massachusetts General Hospital* (Whole issue).

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left hemisphere function. *Science, 219,* 1347–1349.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (Vol. 2, pp. 127–214). New York: Gardner.

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41,* 93–113.

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature, 282,* 403–404.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75–78.

Meltzoff, A. N., & Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child Development, 54,* 702–709.

Molfese, D. H., & Betz, J. C. (1988). Electrophysiological indices of the early development of lateralization for language and cognition, and their implications for predicting later development. In D. H. Molfese & S. J. Segalowitz (Eds.), *Brain lateralization in children* (pp. 171–190). New York: Guilford.

Petersen, M. R., Beecher, M. D., Zoloth, S. R., Moody, D. B., & Stebbins, W. (1978). Neural lateralization of species-specific vocalizations by Japanese macaques *(Macaca fuscata). Science, 202,* 324–327.

Saltzman, E. (1986). Task-dynamic coordination of the articulators. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (pp. 129–144). New York: Springer-Verlag.

Saltzman, E., Rubin, P. E., Goldstein, L., & Browman, C. (1987). Task-dynamic modeling of interarticulator coordination. *Journal of the Acoustical Society of America, 82,* S15 (Abstract).

Spoehr, K., & Corin, W. J. (1978). The stimulus suffix effect as a memory coding phenomenon. *Memory and Cognition, 6,* 583–589.

Studdert-Kennedy, M. (1980). Speech perception. *Language and Speech, 23,* 45–65.

Studdert-Kennedy, M. (1986). Invariance: Functional or descriptive? In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 51–53). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production* (pp. 67–84). London: Academic.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In R. Campbell & B. Dodd (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing* (pp. 211–265). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zoloth, S. R., Peterson, M. R., Beecher, M. D., Green, S., Marler, P., Moody, D. B., & Stebbins, W. (1979). Species-specific perceptual processing of vocal sounds by monkeys. *Science, 204,* 870–873.