Meltzoff, A. N., & Kuhl, P. K. (1994). Faces and speech: Intermodal processing of biologically-relevant signals in infants and adults. In D. J. Lewkowicz & R. Lickliter (Eds.), The development of intersensory perception: Comparative perspectives (pp.335-369). Hillsdale, NJ: Erlbaum.

Faces and Speech: Intermodal Processing of Biologically Relevant Signals in Infants and Adults

Andrew N. Meltzoff
Patricia K. Kuhl
University of Washington

The human face stands out as the single most important stimulus that we must recognize in the visual domain. In the auditory domain the human voice is the most important biological signal. Our faces and voices specify us as uniquely human, and a challenge in neuro- and cognitive science has been to understand how we recognize and process these two biologically relevant signals.

In both domains the conventional view is that the signals are at first recognized through unimodal mechanisms. Faces are thought to be visual objects and voices to be the province of audition. We intend to show that these stimuli are analyzed and represented through more than a single modality in both infancy and adulthood. Speech information can be perceived through the visual modality, and faces through proprioception. Indeed, visual information about speech is such a fundamental part of the speech code that it cannot be ignored by a listener. What listeners report "hearing" is not solely auditory, but a unified percept that is derived from auditory and visual sources. Faces and voices are thoroughly intermodal objects of perception.

Recent experiments have discovered that infants code faces and speech as intermodal objects of perception very early in life. We focus on these intermodal mappings, and explore the mechanism by which intermodal information is linked. Faces and speech can be used to examine central issues in theories of intermodal perception. How does information from two different sensory modalities mix? Is the input from separate modalities translated into a "common code?" If so, what is the nature of the code?

One phenomenon we discuss is infants' imitation of facial gestures. Infants can see the other person's facial movements but they cannot see their own

movements. If they are young enough, they have never seen their own face in a mirror. How do infants link up the gestures they can see but not feel with those that they can feel but not see? We show how this phenomenon illuminates models of intermodal development. New data indicate that infants correct their behavior so as to converge on the visual target through a series of approximations. Correction suggests that infants are guiding their unseen motor behavior to bring it into register with the seen target. In this sense, early imitation provides a key example of intermodal guidance in the execution of skilled action. Other new data reveal an ability to imitate from memory and imitation of novel gestures. Memory-based facial imitation is informative because infants are using information picked up from one modality (vision) to control nonvisual actions at a later point in time, after the visual target has been withdrawn. This suggests that infant intermodal functioning can be mediated by stored supramodal representations of absent events, a concept that is developed in some detail.

Intermodal speech perception involves auditory–visual mappings between the sound of speech and its visual instantiation on the lips of the talker. During typical conversations we see the talker's face, and watch the facial movements that are concomitant by-products of the speech event. As a stimulus in the real world, speech is both auditory and visual. But is speech truly an intermodal event for the listener/observer? To what extent are the visual events that accompany the auditory signal taken into account in determining the identity of the unit?

Adults benefit from watching a talker's mouth movements, especially in noise. People commonly look at the mouth of a talker during a noisy cocktail party because it feels like vision helps us to "hear" the talker. The second author of this chapter has been known to say to the first: "Hold on, let me get my glasses so I can hear you better." These are not examples of superstitious behavior. Research shows that watching the oral movements of a talker is equivalent to about a 20-dB boost in the auditory signal (Sumby & Pollack, 1954). A gain of 20-dB is substantial. It is equivalent to the difference in level between normal conversational speech (65 dB SPL) and shouting (85 dB SPL).

Here we also discuss new research on auditory–visual "illusions" showing that visual information about speech virtually cannot be ignored by the listener/observer. The development of the multimodal speech code and the necessary and sufficient stimulus that allows adults and infants to detect intermodal speech matches are explored. Finally, we hypothesize that infant babbling contributes to the intermodal organization of speech by consolidating auditory–articulatory links, yielding a kind of intermodal map for speech. Understanding the multimodal nature of the speech code is a new and complex issue. The mapping between physical cues and phonetic percepts goes beyond the realm of the single modality, audition, typically

associated with it. This chapter reveals the rather surprising extent to which speech, both its perception and its production, is a thoroughly intermodal event both for young infants and for adults.

INFANT FACIAL IMITATION AS AN INSTANCE OF INTERMODAL FUNCTIONING

There is broad consensus among developmentalists that young infants are highly imitative. However, all imitation is not created equal: Some is more relevant to intermodal theory than others. For example, infants can see the hand movements of others, and can also see their own hands. In principle, infants could imitate by visually matching their own hands to those of another. This would require visually guided responses and visual categorization (the infant's hand is smaller and seen from a different orientation than the adult's), but it would not put much demand on the intermodal system per se.

What is intriguing for students of intermodal functioning is that facial imitation cannot, even in principle, rely on such *intra*modal matching. Infants can see the facial movements of others, but not their own faces. They can feel their own movements, but not the movements of others. How can the infant relate the seen but unfelt other to the felt but unseen self? What bridges the gap between the visible and the invisible? The answer proposed by Meltzoff and Moore (1977, 1983, 1992, 1993; Meltzoff, 1993) is intermodal perception.

It has been known for 50 years that 1-year-old infants imitate facial gestures (e.g., Piaget, 1945/1962). It came as rather more of a surprise to developmentalists when Meltzoff and Moore (1977) reported facial imitation in 2- to 3-week-old infants and later showed that newborns as young as 42 min old could imitate (Meltzoff & Moore, 1983, 1989). The reason for this surprise is instructive. It is not because imitation demands a sensory-motor connection from young infants: There are many infant reflexes that exhibit such a connection. The surprise was engendered by Meltzoff and Moore's hypothesis that early facial imitation was a manifestation of active intermodal mapping (the AIM hypothesis), in which infants used the visual stimulus as a target against which they actively compared their motor output.

At the raw behavioral level, the basic phenomenon of early imitation has now been replicated and extended by many independent investigators. Findings of early imitation have been reported from infants across multiple cultures and ethnic backgrounds: United States (Abravanel & DeYong, 1991; Abravanel & Sigafoos, 1984; Field et al., 1983; Field, Goldstein, Vaga-Lahr, & Porter, 1986; Field, Woodson, Greenberg, & Cohen, 1982; Jacobson, 1979), Sweden (Heimann, Nelson, & Schaller, 1989; Heimann, & Schaller, 1985), Israel (Kaitz, Meschulach-Sarfaty, Auerbach, & Eidelman, 1988), Canada (Legerstee, 1991), Switzerland (Vinter, 1986), Greece (Kugiumutzakis, 1985),

France (Fontaine, 1984), and Nepal (Reissland, 1988). Collectively, these studies report imitation of a range of movements including mouths, tongues, and hands. It is safe to conclude that certain elementary gestures performed by adults elicit matching behavior by infants. The discussion in the field has now turned to the thornier question of the basis of early imitation: Does the AIM hypothesis provide the right general framework, or might there be some more primitive explanation, wholly independent of intermodal functioning? Apparently infants poke out their tongues when adults do so, but what mechanism mediates this behavior?

Imitation Versus Arousal

One hypothesis Meltzoff and Moore explored before suggesting AIM was that early matching might simply be due to a general arousal of facial movements with no processing of the intermodal correspondence. Studies were designed to test whether imitation could be distinguished from a more global arousal response by assessing the specificity of the matching (Meltzoff & Moore, 1977, 1989). It was reasoned that the sight of human faces might arouse infants. It might also be true that increased facial movements are a concomitant of arousal in babies. If so, then infants might produce more facial movements when they saw a human face than when they saw no face at all. This would not implicate an intermodal matching to target.

The specificity of the imitative behavior was demonstrated because infants responded differentially to two types of lip movements (mouth opening vs. lip protrusion) and two types of protrusion actions (lip protrusion vs. tongue protrusion). The results showed that when the body part was controlled—when lips were used to perform two different movements—infants responded differentially. Likewise, when the same general movement pattern was demonstrated (protrusion) but with two different body parts (lip vs. tongue), they also responded differentially. The response was not a global arousal reaction to a human face, because the same face at the same distance moving at the same rate was used in all of these conditions. Yet the infants responded differentially.

Memory in Imitation and Intermodal Mapping

The temporal constraints on the linkage between perception and action was also investigated. It seemed possible that infants might imitate if and only if they could respond immediately, wherein the motor system was entrained by the visual movement pattern. To use a rough analogy, it would be as if infants seeing swaying began swaying themselves, but could not reenact this act from a stored memory of the visual scene. In perceptual psychology the term *resonance* is sometimes used to describe tight perception–action

couplings of this type (e.g., J. J. Gibson, 1966, 1979; or Gestalt psychology). The analogy that is popular (though perhaps a bit too mechanistic) is that of tuning forks: "Information" is directly transferred from one tuning fork to another with no mediation, memory, or processing of the signal. Of course, if one tuning fork were held immobile while the other sounded, it would not resonate at a later point in time. If early imitation were due to some kind of perceptual-motor resonance or to a simple, hard-wired reflex, it might fall to chance if a delay was inserted between stimulus and response.

Two studies were directed to assessing this point: one using a pacifier and short delays (Meltzoff & Moore, 1977) and the other using much longer delays of 24 hours (Meltzoff & Moore, 1994). In the 1977 study, a pacifier was put in 3-week-old infants' mouths as they watched the display so that they could observe the adult demonstration but not duplicate the gestures on-line. The pacifier was effective in disrupting imitation while the adult was demonstrating; the neonatal sucking reflex was activated and infants did not tend to push the pacifier out with their tongues or open their mouths and let it fall out. However, when the pacifier was removed and the adult presented only a passive face, the infants initiated imitation.

The notion that infants could match remembered targets was further explored in a recent study that lengthened the memory interval from seconds to hours (Meltzoff & Moore, 1994). In this study 6-week-old infants were shown facial acts on three days in a row. The novel part of the design was that infants on day 2 and day 3 were used to test memory of the display shown 24 hr earlier. When the infants returned to the laboratory, they were shown the adult with a passive-face pose. This constituted a test of cued-recall memory. The results showed that they succeeded on this imitation-from-memory task. Infants differentially imitated the gesture they had seen the day before. This could hardly be called resonance or a reflexive automatically triggered response, because the actual target display that the infants were imitating was not perceptually present; it was stored in the infant's mind. The passive face was a cue to producing the motor response based on memory. This case is interesting for intermodal theory because infants are matching a nonvisible target (yesterday's act) with a response that cannot be visually monitored (their own facial movements).

Novel Behaviors and Response Correction

Another question concerned whether infants were confined to imitating familiar, well-practiced acts or whether they could construct novel responses based on visual targets. Older children can use visual targets to fashion novel body actions (Meltzoff, 1988); it is not that the visual stimulus simply acts as a "releaser" of an already-formed motor packet. Response novelty was investigated by using a tongue protrusion to the side (TP_{side}) display as one of the

stimuli in the 3-day experiment (Meltzoff & Moore, 1994). For this act, the adult protruded and withdrew his tongue on a slant from the corner of his mouth instead of the usual, straight-tongue protrusion from midline. The results showed that infants imitated this display, and the overall organization and topography of the response helped to illuminate the underlying mechanism. It appeared that the infants were correcting their imitative responses.

Infant tongue protrusion responses were subdivided into four different levels that bore an ordinal relationship according to their fidelity to the TP_{side} display. Time sequential analyses showed that over the 3-day study there was a progression from level 1 to level 4 behavior for those infants who had seen the TP_{side} display. This was not the result of a general arousal, because infants in control groups, including a group exposed only to a tongue protrusion from midline, did not show any such progression. These findings of infants homing in on the target fit with Meltzoff and Moore's AIM hypothesis. The core notion is that early imitation is a matching-to-target process. The gradual correction in the infant's response supports this idea of an active matching to target. The "target" was picked up visually by watching the adult. The infants respond with an approximation (they usually get the body part correct and activate their tongue immediately) and then use proprioceptive information from their own self-produced movements as feedback for homing in on the target.

Although this analysis highlights error detection and correction in the motor control of early imitation, Meltzoff and Moore did not rule out visual-motor mapping of basic acts on "first effort," without the need for feedback. It seems likely that there is a small set of elementary acts (midline tongue protrusion?) that can be achieved relatively directly, whereas other more complex acts involve the computation of transformations on these primitives (e.g., TP_{side}) and more proprioceptive monitoring about current tongue position and the nature of the "miss." Infants cannot have innately specified templates for each of the numerous transformations that different body parts may be put through. There has to be some more generative process involved in imitation. It is therefore informative that infants did not immediately produce imitations of the novel TP_{side} behavior; they needed to correct their behavior to achieve it. Such correction deeply implicates intermodal functioning in imitation.

Development and the Role of Experience

It has been reported that neonatal imitation exists, but then disappears or "drops out" at approximately 2–3 months of age (Abravanel & Sigafoos, 1984; Fontaine, 1984; Maratos, 1982). The two most common interpretations are that newborn imitation is based on simple reflexes that are inhibited with a cortical take over of motor actions, or that the neonatal period entails a brief period

of perceptual unity that is followed by a differentiation of the modalities, and therefore a loss of neonatal sensory-motor coordinations (including imitation), until they can be reconstituted under more intentional control (Bower, 1982, 1989). The reflexive and the modality-differentiation views emphasize an inevitable, maturationally-based drop out of facial imitation. Meltzoff and Moore (1992) recently presented a third view. They argued that learned expectations about face-to-face encounters play a more central role in the previously-reported disappearance of imitation. This may not be as exciting as the notion that a completely amodal perceptual system differentiates at 2–3 months of age, but it better accounts for the results we recently obtained.

Meltzoff and Moore (1992) conducted a multitrial, repeated-measures experiment involving 16 infants between 2 and 3 months of age, the heart of the drop-out period. The overall results yielded strong evidence for imitation at this age; however, these infants gave no sign of imitating the adult gestures in the first trials alone. The same children who did not imitate on first encounter successfully imitated when measured across the entire repeated-measures experiment. This is hardly compatible with a drop-out due to modality differentiation; it is more suggestive of motivational or performance factors that can be reversed.

Further analysis suggested that the previously reported decline in imitation is attributable to infants' growing expectations about social interactions with people. When these older infants first encountered the adult, they initiated social overtures as if to engage in a nonverbal interchange—cooing, smiling, trying out familiar games. This behavior supplanted any first-trial imitation effects. After the initial social gestures failed to elicit a response (by experimental design because our *E* did not respond contingently to the infant), infants settled down and engaged in imitation.

It thus appears that development indeed affects early imitation. Imitation is a primitive way of interacting with people that exists prior to other social responses such as cooing, smiling, and so on. Once these other responses take hold, they become the first line of action in the presence of a friendly person. Hence the *apparent* "loss" of imitation. If the typical designs are modified, however, this is reversible and there is quite robust imitation among older infants. What develops are social games that are higher on the response hierarchy than is simple imitation, but there is no fundamental drop out of competence.

Converging Evidence

Returning to the mechanism question, Meltzoff and Moore have proposed that information about facial acts is fed into the same representational code regardless of whether those body transformations are seen or felt. There is a "supramodal" network that unites body acts within a common framework.

Imitation is seen as being tied to a network of skills, particularly to speech–motor phenomena, which also involve early perception-production links involving oral-facial movements (Meltzoff, Kuhl, & Moore, 1991). The development and neural bases for such an intermodal representation of the face are a pressing issue for developmental neuroscience (Damasio, Tranel, & Damasio, 1990; de Schonen & Mathivet, 1989; Stein & Meredith, 1993).

That neonates can relate information across modalities is no longer the surprise it was in 1977. There has been an outpouring of findings that are compatible with this view, although the ages, tasks, and intermodal information have varied widely (e.g., Bahrick, 1983, 1987, 1988; Bahrick & Watson, 1985; Bower, 1982; Bushnell & Winberger, 1987; Butterworth, 1981, 1983, 1990; Dodd, 1979; Lewkowicz, 1985, 1986, 1992; Meltzoff, 1990; Rose, 1990; Rose & Ruff, 1987; Spelke, 1981, 1987; Walker, 1982; Walker-Andrews, 1986, 1988). One example from our own laboratory is particularly relevant, because it involves neonates of about the same age as in the studies of imitation and involved vision and touch. Meltzoff and Borton (1979) provided infants tactual experience by molding a small shape and fitting it on a pacifier (Fig. 14.1). The infants orally explored the shape but were not permitted to see it. The shape was withdrawn from their mouths, and they were given a choice between two shapes, one that matched the shape they had tactually explored and one that did not (tactual and visual shapes were appropriately counterbalanced). The results of two studies showed that 29-day-old infants systematically looked longer at the shape that they had tactually explored. The finding of cross-modal perception in 1-month-olds was replicated and cleverly broadened in an experiment by E. J. Gibson and Walker (1984), who used soft versus rigid visual and tactual objects (instead of shape/texture information), and by Pêcheux, Lepecq, and Salzarulo (1988), who used Meltzoff and Borton's shapes and examined the degree of tactual familiarization necessary to recognize information across modalities. More recently the oral–visual cross-

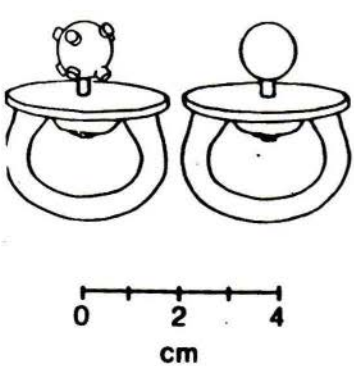

FIG. 14.1. Shapes used to assess tactual–visual matching. The pacifiers were inserted in the infants mouths without them seeing them. After a 90-sec familiarization period the shape was withdrawn, and a visual test was administered to investigate whether the tactual exposure influenced visual preference. From Meltzoff and Borton (1979). Reprinted by permission.

modal matching effect was extended to a newborn population by Kaye (1993), who found visual recognition of differently shaped rubber nipples that were explored by mouth. Streri (1987; Streri & Milhet, 1988; Streri & Spelke, 1988) conducted tactual–visual studies in 2- to 4-month-old infants and demonstrated cross-modal matching of shapes from manual touch to vision. Finally, Gunderson (1983) used Meltzoff and Borton's shapes mounted on pacifiers and replicated the same effect in 1-month-old monkeys, indicating that cross-modal matching is not specific to neonatal humans.

We have been especially interested in pursuing infant intermodal perception of biologically relevant stimuli. Toward that end we have investigated other phenomena involving faces. In particular, we have found that young infants recognize the correspondence between facial movements and speech sounds. This line of work affords a particularly detailed look at the nature of the information that is "shared" across modalities.

SPEECH PERCEPTION AS AN INSTANCE OF INTERMODAL FUNCTIONING

Speech perception has classically been considered an auditory process. What we perceived was thought to be based solely on the auditory information that reached our ears. This belief has been deeply shaken by data showing that speech perception is an intermodal phenomenon in which vision (and even touch) plays a role in determining what a subject reports hearing. Visual information contributes to speech perception in the absence of a hearing impairment and even when the auditory signal is perfectly intelligible. In fact, it appears that when it is available, visual information cannot be ignored by the listener; it is automatically combined with the auditory information to derive the percept.

The fact that speech can be perceived by the eye is increasingly playing a role in theories of both adult and infant speech perception (Fowler, 1986; Kuhl, 1992, 1993a; Liberman & Mattingly, 1985; Massaro, 1987a; Studdert-Kennedy, 1986, 1993; Summerfield, 1987). This change in how we think about speech results from two sets of recent findings. First, studies show that visual speech information profoundly affects the perception of speech in adults (Dodd & Campbell, 1987; Grant, Ardell, Kuhl, & Sparks, 1985; Green & Kuhl, 1989, 1991; Green, Kuhl, Meltzoff, & Stevens, 1991; Massaro, 1987a, 1987b; Massaro & Cohen, 1990; McGurk & MacDonald, 1976; Summerfield, 1979, 1987). Second, even young infants are sensitive to the correspondence between speech information presented by eye and by ear (Kuhl & Meltzoff, 1982, 1984; Kuhl, Williams, & Meltzoff, 1991; MacKain, Studdert-Kennedy, Speiker, & Stern, 1983; Walton & Bower, 1993). The work on infants and adults is discussed in turn.

Auditory–Visual Speech Perception in Infants

Our work on the auditory–visual perception of speech began with the discovery of infants' abilities to relate auditory and visual speech information (Kuhl & Meltzoff, 1982). A baby-appropriate lipreading problem was posed (Fig. 14.2). Four-month-old infants were shown two filmed images side by side of a talker articulating two different vowel sounds. The soundtrack corresponding to one of the two faces was played from a loudspeaker located midway between the two facial images, thus eliminating spatial clues concerning which of the two faces produced the sound. The auditory and visual stimuli were aligned such that the temporal synchronization was equally good for both the "matched" and "mismatched" face–voice pairs, thus eliminating any temporal clues (Kuhl & Meltzoff, 1984). The only way that infants could detect a match between auditory and visual instantiations of speech was to recognize what individual speech sounds looked like on the face of a talker.

Our hypothesis was that infants would look longer at the face that matched the sound rather than at the mismatched face. The results of the study were in accordance with the prediction. They showed that 18- to 20-week-old infants recognized that particular sound patterns emanate from mouths moving in particular ways. In effect the data suggested the possibility that infants recognized that a sound like /i/ is produced with retracted lips and a tongue-high posture, whereas an /a/ is produced using a lips-open, tongue-lowered posture. That speech was coded in a polymodal fashion at such a young age—a code that includes both its auditory and visual specifications—was quite surprising. It was neither predicted nor expected by the then existing models of speech perception.

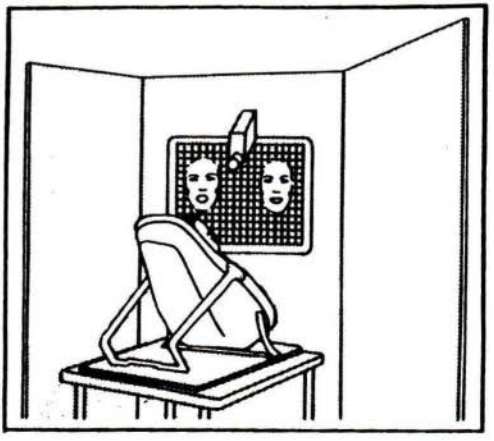

FIG. 14.2. Experimental arrangement used to test cross-modal speech perception in infants. The infants watched a film of two faces and heard speech played from a central loudspeaker. From Kuhl and Meltzoff (1982). Reprinted by permission.

Generality of Infant Auditory–Visual Matching for Speech

The generality of infants' abilities to detect auditory-visual correspondence was examined by testing a new vowel pair, /i/ and /u/ (Kuhl & Meltzoff, 1988). Using the /u/ vowel was based on speech theory because the /i/, /a/, and /u/ vowels constitute the "point" vowels. Acoustically and articulatorily, they represent the extreme points in vowel space. They are more discriminable, both auditorially and visually, than any other vowel combinations and are also linguistically universal. The results of the /i/–/u/ study confirmed infants' abilities to detect auditory–visual correspondence for this vowel pair. Two independent teams of investigators have replicated and extended the cross-modal speech results in interesting ways. MacKain et al. (1983) demonstrated that 5- to 6-month-old infants detected auditory–visual correspondences for disyllables such as /bebi/ and /zuzi/ and argued that such matching was mediated by left hemisphere functioning. More recently, Walton and Bower (1993) showed cross-modal speech matching for both native and foreign phonetic units in 4.5-month-olds.

THE BASIS OF AUDITORY–VISUAL SPEECH PERCEPTION: PARAMETRIC VARIATIONS

From a theoretical standpoint the next most important issue was to determine how infants accomplished the intermodal speech task. As in all cases of intermodal perception, a key question is the means by which the information is related across modalities. One alternative is that perceivers recode the information from each of the two modalities into a set of basic common features that allows the information from the two streams to be matched or combined. The central idea is that complex forms are decomposed into elementary features and that this aids in intermodal recognition. We conducted a set of experiments to determine whether speech was broken down into its basic features during intermodal speech perception. In these studies we presented both infants and adults with tasks using nonspeech stimuli that captured critical features of the speech stimulus. The underlying rationale of the studies was to determine whether an isolated feature of the speech unit was sufficient to allow the detection of intermodal correspondence for speech.

Speech and Distinctive Feature Theory

The goal of these studies was to "take apart" the auditory stimulus. We wanted to identify features that were necessary and sufficient for the detection of the cross-modal match between a visual phonetic gesture and its concomitant

sound. Distinctive Feature Theory provides a list of the elemental features that make up speech sounds (Jakobson, Fant, & Halle, 1969). Speech events can be broken down into a set of basic features that describe the phonetic units. For example, one acoustic feature specifies the location of the main frequency components of the sound. It is called the *grave-acute* feature and distinguishes the sounds /a/ and /i/. In the vowel /a/, the main concentration of energy is low in frequency; in the vowel /i/ the main concentration of energy is high in frequency. The vowel /a/ is thus *grave* and the vowel /i/ *acute* in Distinctive Feature Theory (Jakobson et al., 1969). Features such as grave and acute can be duplicated with simple nonspeech sounds.

That speech features can be approximated with nonspeech sounds such as simple tones was recognized by Isaac Newton. His notebooks described how he created the impression of a series of vowels, beginning with low-pitched vowels like /a/ and /u/ up to high ones such as /i/, by slowly filling a deep pitcher with a constant stream of beer (see also Helmholtz, 1885/1954). When a small amount of beer was in the container, low-pitched sounds resembling /a/ were produced; when the pitcher of beer was filled, higher tones resembling /i/ were produced.

This experiment can also be conducted in a laboratory rather than a pub (although it is a bit more tedious). Psychoacoustic matching experiments have demonstrated that individual vowels are perceived to have "predominant pitches" corresponding to the grave-acute feature of Distinctive Feature Theory (Chiba & Kajiyama, 1958; Fant, 1973; Farnsworth, 1937). For example, Farnsworth (1937) presented subjects with 22 pure tones ranging from 375 to 2400 Hz. Subjects were instructed to label each of the tones as one of 12 vowels. The results showed that tones in the high frequency range tended to be labeled as vowels like /i/, while tones in the low to middle frequencies tended to be labeled as vowels like /a/ and /u/. These experiments establishing the psychological reality of features used solely unimodal tasks and all involved adults.

Is Intermodal Speech Perception Based on a Feature Analysis?

We tested whether the "predominant pitch" of the vowel (as captured in a nonspeech stimulus, a pure tone) was sufficient to produce the cross-modal matching effect observed in infants (Kuhl et al., 1991). We also examined this in adults by administering unimodal, cross-modal, and amodal tasks.

Two kinds of nonspeech stimuli were used, single isolated pure tones and three-tone complexes. The pure-tone signals varied from 750 to 4000 Hz. The three-tone complexes approximated the speech signals more closely in that they contained three tones, one located at each of the center frequencies of each of the first three formants of the vowels. (The three-tone

complexes provided additional features that matched those in the original vowels, such as the relationships between individual formant frequencies.) Neither of these two nonspeech signals sounded like speech. Our question was to what extent these nonspeech signals could be related to vowel stimuli, especially in a cross-modal matching experiment.

The adult tests involved a number of conditions. For the auditory task the vowel was presented as an auditory stimulus; for the cross-modal task the vowel was presented as a face pronouncing the vowel; for the amodal task the vowel was simply "imagined" by having subjects think about the vowel (an amodal task, because the stimulus was not in any sensory modality). Infants were tested only in the cross-modal format. The question here was whether infants would detect a cross-modal match between the visually presented faces and a nonspeech stimulus that captured a prominent feature of speech. Everything was the same as in Kuhl and Meltzoff (1982), but rather than hearing one of the real vowels presented auditorially, they heard either a pure-tone stimulus or a three-tone analog. The amplitudes (loudness) of the nonspeech signals were varied to match the amplitudes of the original vowels. As the mouths opened the nonspeech signal grew louder; as the mouths closed the nonspeech signal became softer. The fact that the auditory amplitude envelope was appropriate for the stimulus being seen created a situation in which it was not trivially obvious that the mouth could not be producing the sound that was being presented. In fact, the data showed that infants fixated the faces just as long in the nonspeech conditions in this experiment as they did in the speech conditions tested in Kuhl and Meltzoff (1982). The question was whether variations in frequency of the pitch resulted in differential looking at the faces, as was the case when the real /a/ and /i/ vowels were presented.

The results revealed clear developmental differences. Adults successfully related pure tone signals to the vowels /a/ and /i/. Adults matched pure tones to vowels unimodally (when the vowels were auditorially presented), cross-modally (when the vowels were visually presented), and amodally (when the vowels were imagined). In all cases adults matched low-frequency pure tones to the vowel /a/ and high-frequency pure tones to the vowel /i/, in line with the predominant pitch idea. The results on the adult tests with the three-tone analogs of /a/ and /i/ were a bit more complicated but strongly supported the same conclusion (Kuhl et al., 1991). Once again, adults had no difficulty relating three-tone analogs to vowels presented auditorially and visually.

The infant results were quite different. Infants in both experiments—whether listening to pure tones of various frequencies or three-tone nonspeech analogs derived from the vowels—showed no ability to match nonspeech auditory signals to visually presented vowels. They smiled at the faces and looked at them just as long as they had in previous experiments;

however, there was no cross-modal effect. The nonspeech analogues of /a/ versus /i/ did not differentially affect which face infants fixated.

Implications For Intermodal Theory

The results provided support for two inferences: Adults can relate speech stimuli such as the vowels /a/ and /i/ to nonspeech stimuli on the basis of a simple isolated feature, and infants do not rely on the same simple feature for intermodal speech perception under similar test conditions. These speech findings have implications for both auditory–visual and general theories of intermodal processing.

First consider the connection adults perceive between vowels and pitch, and the possible basis for this perception. Two alternatives can be put forward. The perceptual/linguistic alternative argues for a fairly direct mapping between vowels and tones of a particular frequency. On this view, spectral features that are responsible for the perceived predominant pitch of the vowel are derived during the perceptual analysis of the sound. If featural properties such as grave-acute are automatically derived in perceptual processing, then the link between vowels and pitch is based on the psychological reality of decomposing speech into featural elements. A second alternative is that the link between /a/ vowels and low sounds and between /i/ vowels and high sounds may be mediated by more metaphorical thinking (Gentner & Grudin, 1985; Ortony, 1979) as part of a larger cognitive network. Some work in our laboratory on "phonetic symbolism" and the semantic qualities associated with vowels shows that adults think of /a/ as a "strong" sound, whereas /i/ is "weaker." The attribute "strong" is typically associated with maleness, and male voices are predominantly low in pitch, which could be the network through which the association is made between /a/ vowels and low tones. If true, this would be more "cognitively mediated" than the first alternative. The current findings show that features embodied in nonspeech stimuli can be used in cross-modal and amodal speech perception, but do not decisively sort between these alternatives.

These results with adults become more interesting when considered in relation to the findings from infants. Infants did not display the same link between vowel and pitch exhibited by adults. Their visual fixations were not differentially affected by the nonspeech stimuli; however, when real speech stimuli were used, infants did make differential visual choices. Thus, 4-month-old infants detect face–voice matches when speech stimuli are presented auditorially while failing to do so when the auditory stimulus is stripped down to its simplest featural component, as in a pure tone, or when three-tone nonspeech analogs of the vowels are presented.

It appears that infants' detection of cross-modal correspondence for speech requires the whole speech stimulus. (A whole stimulus is a signal that is

sufficient to allow the *identification* of the speech signal. Synthetic speech signals qualify as whole stimuli by this definition; they do not include all of the speech information present in a natural utterance, but still allow the identification of the speech stimulus.) From a developmental viewpoint, the findings suggest that the intermodal perception of speech does not progress through a developmental sequence that goes from "parts" to "wholes." Infants do not begin relating faces and voices on some simple feature, and then gradually build up a connection between the two that involves, on the auditory side, an identifiably whole speech stimulus. Perceptual developmental theorists have suggested that infants may at first be maximally responsive to wholes, especially in the form of complex natural stimuli that are later differentiated into component aspects (e.g., Bower, 1982; E. J. Gibson, 1969; J. J. Gibson, 1966). The case of intermodal speech perception provides data compatible with such a developmental model, inasmuch as older but not younger subjects detect an intermodal match when provided only a "part" of the stimulus—one that cannot be independently identified as speech.

The hypothesis just stated raises a point about the boundary conditions of intermodal perception for infants. It begins to tell us when intermodal processing breaks down. Of course, this depends on the assumption that nonspeech signals can be processed by infants in the first place. They can be. Previous unimodal tests showed that nonspeech stimuli supported many of the same phenomena as the full speech signal. The phenomenon of categorical perception has been shown in adults with nonspeech signals (e.g., Diehl & Walsh, 1989; Pisoni, Carrell, & Gans, 1983). Categorical perception of nonspeech signals has also been shown by infants for both two-tone and three-tone analogs derived from real speech syllables (Jusczyk, Pisoni, Walley, & Murray, 1980; Jusczyk, Pisoni, Reed, Fernald, & Myers, 1983). There seems to be a striking dissociation between unimodal and intermodal tasks.

This dissociation is further corroborated by other work from our laboratory in which 4- to 5-month-old infants were tested with nonspeech signals in another cross-modal task, one that involves vocal imitation. In the case of vocal imitation, infants have to relate the auditory perception of the vowels /a/ and /i/ to their own motor productions of speech. In our work on vocal imitation, infants listened either to speech stimuli (the vowels /a/ and /i/) or the nonspeech pure-tone signals used in the present studies. The results again showed that nonspeech signals were not effectively related to articulation. In response to speech signals, infants produced speechlike utterances. However, in response to the nonspeech signals infants did not produce speechlike vocalizations; they listened intently but did not produce speech (Kuhl & Meltzoff, 1988).

Taken together it can be inferred that, in cross-modal speech tasks, young infants need the whole signal, one that is identifiable as a speech sound.

Although infants need this more complete specification in order to link the perception and production of speech, they do not need it in unimodal tasks; moreover, for adults, a "part" of the stimulus is sufficient in cross-modal (and amodal) situations. The ontogenesis of this ability to use parts in a cross-modal setting and the reason for its absence in early infancy are currently being investigated in our laboratory in longitudinal studies.[1]

ADULT AUDITORY–VISUAL ILLUSIONS AND INTERMODAL SPEECH PERCEPTION

The next series of experiments utilized adult subjects to examine in more detail the nature of the code or metric that is used to combine information about speech from two modalities. To investigate this we moved from studying cross-modal matches to cross-modal illusions, in which the information from the two modalities is clearly discrepant. In this case the percept does not derive from the detection of "invariant" information; there is no invariant that can be recognized in the two modalities. Rather, the percept results from the unification of discrepant information. By systematically varying the signals one can uncover the nature and form of the information at the time that the input from the two modalities mix.

The auditory–visual "illusion" reported by McGurk and MacDonald (1976) is a robust phenomenon (Green & Kuhl, 1989, 1991; Green et al., 1991; Massaro, 1987a, 1987b; Summerfield, 1987). The illusion results when auditory information for /b/ is combined with visual information for /g/. Perceivers report the phenomenal impression of /d/ despite the fact that this information was not delivered to either sense modality. Speech scientists are now beginning to find out how the phenomenon works. Of primary

[1]It might be useful to clarify the "wholes" versus "parts" argument. Cross-modal tasks can be accomplished on the basis of what some might call simple attributes, such as the synchrony between simple tones and flashes of light (e.g., Lewkowicz, chap. 8, this volume). However, in cases such as tone–light synchrony it is not clear that our whole–part distinction comes into play. Synchrony is the most prominent aspect of these stimuli, the gestalt. It makes little sense to say the stimulus is "broken down" into synchrony. In the cases we are addressing, the speech signal is a whole, but it can be broken down into parts. Thus, our point about the limitations of tones in the cross-modal situation is not one about tones per se, but "parts versus wholes." In such cases the developmental question becomes whether infants first operate on the whole and then differentiate it into parts, or conversely. Our data indicate that for infants perceiving speech, the cross-modal system (but not the unimodal system) requires the whole signal to map audition to articulation, as measured by both the lip-reading and the vocal imitation studies. We do not hold that tones and other such stimuli can never support cross-modal relations in infancy, but suggest that cross-modal relations for speech (and perhaps more generally) may at first require whole stimuli; later, separate parts are sufficient. We have shown this for speech, and it would be interesting to see if a similar pattern would obtain with the wholes versus parts of visual objects (faces, geometric solids) in a cross-modal task. Although some unimodal tests have been conducted, there are few developmental cross-modal tests on this point.

interest is the nature of the interaction between the two modalities and the manner in which optic and acoustic information is mixed to yield the unified percept of /d/ when there was no /d/ presented in the stimulus.

Early accounts of the process suggested that the information in each modality was featurally categorized and then combined in some sort of additive process. The hypothesis was that the visual modality provided "place" information whereas the auditory modality provided "manner" information. The "place of articulation" feature describes the location in the mouth where the primary constriction of the airflow takes place. When producing a /b/, /p/, or /m/, for example, the primary constriction takes place at the lips, and the sound is said to have a bilabial place of articulation. In contrast, the sounds /t/, /d/, and /n/ result from a primary constriction created when the tongue tip touches the alveolar ridge behind the teeth and the sounds are said to have an alveolar place of articulation. Place of articulation features are visible on the face of the talker: You can see whether a person makes a bilabial speech sound by looking to see if the two lips close. In contrast the manner of articulation feature is nearly impossible to see on the face of the talker. The manner feature refers to the way in which a sound was produced. For example, sounds that are produced by lowering the velum and allowing air to escape from the nose are said to have a nasal manner of articulation (see Kuhl & Meltzoff, 1988, for further details). It is the manner feature that distinguishes two speech sounds with the same place of articulation, such as /b/ and /m/, which are visually identical.

Our recent data, as well as those of others, provide convincing evidence that at the point of integration the information is in a precategorical state (Grant et al., 1985; Green & Kuhl, 1989, 1991; Green et al., 1991; Green & Miller, 1985; Massaro, 1987a, 1987b). That is, the speech stream has not yet been rigidly coded as having a defined and specific place or manner of articulation before the intermodal integration takes place. The principal question now is the form of this precategorical information that makes such illusory auditory–visual blends possible.

Visually Caused Shifts in the Phonetic Boundaries Underlying Manner Features

Two studies conducted by Green and Kuhl (1989, 1991) showed that the "vision provides place information and audition provides manner information" hypothesis cannot be sustained. The data demonstrate that vision affects even the assignment of the manner feature, and this underscores the depth of communication between the visual and auditory pick up of speech information.

Green and Kuhl (1989) utilized a well-established phenomenon in auditory speech perception to study whether features were assigned prior

to the integration of information in the cross-modal perception of speech. The well-established phenomenon is a change in the location of the category boundary on a voiced–voiceless continuum (a manner feature) that occurs with changes in place of articulation (Abramson & Lisker, 1970; Miller, 1977). Thus, within the auditory modality, it is known that auditory place influences decisions about auditory manner. The question posed by Green and Kuhl (1989) was whether visually specified place could also be influence (auditory) manner information; that is, whether the location of the phonetic boundary for the manner feature would shift when the place of articulation was specified by eye instead of by ear.

Observers were presented with an auditory /ibi/ and a visual /igi/. As expected, subjects perceived an illusory syllable, the syllable /idi/. The question was: Given that the perception of place information was a blend of both auditory and visual information, was the voicing information solely determined by the auditory signal, because no voicing information was available visually, or was even the perception of voicing (a classic "auditory" feature) affected by the visual information?

The results showed that the location of the voicing category boundary shifted in the auditory–visual condition. That is, when the /ibi/ and /ipi/ stimuli were presented in an auditory-alone condition, a voiced-voiceless category boundary typical for bilabial stimuli was obtained. However, when observers heard these same auditory stimuli while watching the visual /igi/, the voiced-voiceless category boundary was shifted to one that was appropriate for an alveolar place of articulation (the /d/–/t/ continuum) because through the illusion subjects now perceived a continuum ranging from /idi/ to /iti/. This was true even though the auditory information remained the same in the two conditions.

The result is interesting because it indicates that although a single modality (in this case, the auditory modality) provided the sensory input about voicing, the visual stimulus still influenced the perception of voicing. This suggests that the integration of information from the two modalities takes place prior to the time that it is categorized into phonetic features. Data and theory advanced by Massaro (1987a, 1987b) and Summerfield (1987) also support this inference.

Integral Processing of Visual and Auditory Speech Information

A second study provided converging evidence that the integration of information from the two modalities takes place prior to the time that features are assigned. A speeded classification design created by Garner (1974) was used. Garner showed that when two dimensions of a stimulus are processed

"integrally," the reaction time to classifying syllables on one dimension is significantly increased by variations in information in the other dimension. When the dimensions are processed independently this increase in reaction time does not occur. Auditory experiments have shown that when classifying information along the voicing dimension, variation in the place feature results in increased reaction times (Eimas, Tartter, Miller, & Keuthen, 1978), which indicates that the two features are processed integrally rather than separately. It is well established, in both speech studies and other studies using visual objects, that when two features are processed separately, variation in an irrelevant second dimension does not cause an appreciable increase in reaction time (Eimas et al., 1978; Garner, 1974).

Green and Kuhl (1991) examined the reaction time to classify four auditory–visual syllables (/b/, /p/, /d/, /t/) that varied along two dimensions—place and voicing. In the study, subjects were asked to classify the speech syllables along the voicing dimension (classifying them as either voiced or voiceless), or along the place dimension (classifying them as either bilabial or alveolar). The voicing information in the four syllables varied only in the auditory domain, whereas the place information varied only in the visual domain. Green and Kuhl reasoned that if speech was not deeply bimodal then subjects ought to be able to selectively attend to separate modalities (auditory for voicing classification and visual for place classification), and thus process the featural information separately. The results showed that when classifying the auditory stimuli according to the voicing feature, variation in the (visual) place feature produced an increase in classification times, even though vision does not overtly contribute information regarding the voicing feature; similarly, when classifying the visual stimuli according to the place feature, irrelevant variation in the (auditory) voicing feature resulted in significant increases in classification times. This "interference effect" indicates that even when the featural information is typically carried by a specific modality (auditory-voicing and visual-place), the information from the two modalities is integrated prior to phonetic feature assignment and not treated as separate. We believe that this is due to the fact that the auditory and visual information are not initially classified featurally and then combined, but that precategorical auditory and visual information are mapped onto a stored representation at the same time (see also Massaro, 1987a, 1987b).

These studies show that at the time that the auditory–visual information is mixed, speech is not featurally classified, suggesting that it has maintained some of the detail of an analog form. If the information at the point of conflux is extremely detailed then one might be able to disrupt integration by making the information in the two modalities so noticeably different that the two streams could not mix. The next experiment addressed this point.

The Cross-Gender Speech Illusion Experiment

In cognitive psychology, cross-modal inputs that could not have derived from a common biological or physical source are said to violate the "unity" assumption (Welch & Warren, 1980). Results show that violations of the unity assumption impede intermodal perception. For example, in the ventriloquism effect, a large spatial separation between auditory and visual input (which suggests that information could not have derived from a common source) profoundly dampens the effect (Warren, Welch, & McCarthy, 1981).

The goal of the Green, Kuhl, Meltzoff, and Stevens (1991) study was to violate the unity of source for auditory–visual speech information and test whether listeners perceived a unified phonetic percept despite the fact that such a percept would have to be derived from two obviously different talkers. We created a situation in which there was an obvious discrepancy between the gender of the talker presenting the information in the two modalities. A visual male face was combined with the voice of a female talker, and vice versa. We took pains to choose our speakers such that the gender incompatibility was highly salient. A very male-looking football player's face was paired with a high and feminine-sounding female voice, and vice versa. There was no mistaking the gender mismatch for these auditory–visual stimuli.

The rationale of the study was twofold. The first was to test whether violating the unity assumption would prevent multimodal combination at the phonetic level of speech. We thought that intermodal relations for phonetic units might involve a mandatory process wherein visual phonetic information cannot be ignored by the listener, and that the two inputs might be combined despite the clear violation of the unity assumption. Second, we were interested in the detail versus abstractness of the information at the point of integration. The specific talker producing a speech sound greatly alters the acoustic detail (absolute frequencies) of the phonetic unit. If the information at the point of integration preserves that detail then it might be difficult to integrate phonetic information across talkers of different gender. However, if the information about the identity of the phonetic unit is more abstract ("talker neutral"), then the fusion of a male face and a female voice into one unified phonetic percept might still occur.

The results showed that the gender discrepancy was readily apparent. Subjects readily reported the mismatch and judged it "peculiar" and "funny" to hear a high-pitched voice come out of a whiskered, large male face. Nonetheless, the data revealed that the integration of auditory and visual information was as pervasive in the gender-discrepant situation as in the gender-congruent case. The number of auditory–visual illusions was not significantly different in the two situations. An interesting finding was the strong dissociation between the judgments of gender identity versus those of

phonetic identity. There was no blending in the gender judgments: The stimuli simply looked like males and sounded like females (or vice versa) with no blending. Conversely, there was blending at the level of the phonetics: The perceivers were not able to report what they saw or what they heard, because they perceived something else, something that was not presented to either modality, a phonetic unit that was a blending of the two modality streams.

Evidently, violations in the unity assumption indicating that the input could not have derived from a common biological source do not disrupt phonetic perception. They suggest that at the time of auditory–visual integration, the phonetic information from the two modalities is in a somewhat abstract form that neutralizes differences across talkers. When available, both auditory and visual speech information, even though noticeably discrepant, are used to derive a unified phonetic interpretation of the speech signal. It is as if integrating the auditory and visual information is mandatory.

Intermodal Visual and Tactual Speech Studies: The Whole Is Greater Than the Sum of Its Parts

Further studies have shown that the information fed into the two separate modalities is not simply additive. The speech that is perceived is more than the sum of its unimodal parts. This suggests that multimodal input maps onto stored representations of linguistic information that go beyond the raw input from either modality alone.

In one study, a speaker sat at a window in a soundproof booth and read aloud from a novel (Grant et al., 1985). An observer sat outside the booth and could not see or hear the speaker. The observer listened to a pure-tone signal that followed the fundamental frequency and the amplitude of the speaker's voice. When presented by itself, the pure tone was unintelligible. No words, syllables, or phrases of the novel could be heard. Then the observer was instructed to turn and face the window, thereby bringing the speaker into view. Under these conditions the observer could obtain visual information about the speech that was produced. The results were quite dramatic. Speech perception jumped from the 37% that could be perceived with visual cues alone (pure lipreading) to nearly 80% intelligibility. Subjects reported that turning toward the speaker and allowing them to see the speaker's mouth movements while listening to the pure tone produced an astounding change in what they "heard." It is striking that the intelligibility of the information from each of the two modalities was more than additive. Considered separately, the two signals provided "37% intelligibility" (0% from audition alone + 37% from lipreading). However, when the two were combined, intelligibility was 80%.

An even more startling finding concerns speech information perceived through the skin. Research tested whether the pure-tone information follow-

ing the fundamental frequency of the voice of the speaker could be presented tactually rather than auditorially (Grant, Kuhl, Ardell, & Sparks, 1986). Subjects viewed the talker reading a novel but there was no auditory information delivered. Instead, the pure-tone information was delivered through an electrocutaneous device that could be worn on the forearm. It consisted of eight electrodes, each of which covered a limited range of frequencies, that were arranged spatially in a line from wrist to elbow. The receiver was exposed to the face of the talker and tested both with and without the electrocutaneous information. (The electrocutaneous information by itself was 0% intelligible.) The results demonstrated that the tactually delivered information significantly increased speech perception over that obtained by lipreading alone. When "feeling" and seeing the speech, intelligibility increased 20% over lipreading alone.

These two experiments demonstrate that the information fed into the two separate modalities is not simply added. This is similar to the auditory–visual findings reported at the neurophysiological level by Stein and Meredith (1990, 1993) using nonspeech stimuli.

Further experiments investigated what kinds of speech information could be delivered through the skin. In these studies, electrocutaneous stimulation occurred via a matrix of 144 electrodes that presented the entire spectrum of speech in a frequency × amplitude array. The device displayed the information spatially and was worn as a belt circling the abdomen. The studies demonstrated that manner information delivered through the skin could be combined with place information delivered visually, resulting in the correct perception of individual syllables (Sparks, Kuhl, Edmonds, & Gray, 1978) and excellent perception of connected discourse (Sparks, Ardell, Bourgeois, Wiedmer, & Kuhl, 1979).

Speech information can be delivered to the skin and integrated with that perceived by eye or by ear. The important theoretical point is the penetrability of the speech-processing mechanism by the information picked up by touch (or vision). Speech is not solely the province of audition, nor even of audition plus vision. Information delivered tactually also appears to have access to the speech-processing mechanism. The speech code is thoroughly intermodal in nature.

IMPLICATIONS FOR THEORY: VIEWING FACES AND SPEECH THROUGH AN "INTERMODAL LENS"

A variety of phenomena concerning intermodal functioning have been discussed both in infants and adults. We here intend to draw out the theoretical implications of these findings. Discussed are the developmental aspects of intermodal perception, its bases, and functional utility.

Infant Babbling as Viewed Through an Intermodal Lens

It has been discussed how speech is perceived through different sensory modalities—auditory, visual, and even the tactile sense. We now turn to speech production for clues about the developmental history of the intermodal organization of speech.

There is ample evidence to suggest that by adulthood we have a set of rules specifying the relation between sounds and speech movements. The mapping is not restricted to the production of articulatory acts that are overlearned. We can reach an auditory target if we hold a foreign object in our mouths (a pencil, food, or even a novel object). This has been demonstrated experimentally in studies in which an adult, instructed to produce a vowel such as /i/, is suddenly prevented from doing so by the introduction of a weight or load imposed on the lip or jaw. Under these conditions, the speaker produces a perfectly adequate /i/ vowel, but uses a different set of muscles than those typically used in the production of that sound. Detailed measurements of the muscle movements show that in this situation compensation is virtually immediate, prior to the time that auditory feedback could have led to the compensation (Perkell, Matthies, Svirsky, & Jordan, 1993). Such rapid motor adjustments suggest a flexible set of rules; we call it an auditory-articulatory "map," relating articulatory movements to sound.

How does this auditory-articulatory map develop? It can be suggested that one important developmental contributor is the practice infants gain from their own self-produced sound. Normal infants the world over produce speech milestones on a fairly predictable schedule. At 3 months infants will "coo," producing vowel-like utterances; by 7 months, infants will "babble," producing reduplicative consonant-vowel syllables like *babababa* (e.g., Ferguson, Menn, & Stoel-Gammon, 1992; Locke, 1993). The classical, now outdated view, was that these speech-production milestones were maturationally driven, perhaps due to the unfolding of an internal motor program (Lenneberg, 1967). However, data derived from a comparison of speech production in normal children, deaf children, and blind children shows how deeply early speech production is affected by environmental input.

It has been shown, for example, that deaf infants do not babble in the way that is universal among hearing infants. They do not babble on the same time schedule that hearing infants do, and the durations of their babbled utterances do not match those of normal infants (Oller & Eilers, 1988; Oller & Lynch, 1992). Moreover, the phonetic content of the babbled utterances of hearing-impaired infants is different than that of normal infants (Stoel-Gammon, 1988; Stoel-Gammon & Otomo, 1986). Hearing-impaired infants babble using a disproportionally high level of bilabial sounds—/b/, /m/, and /d/—sounds that are easily seen. Normally hearing infants include a higher proportion of sounds such as /g/ that cannot be readily seen and

require audition to perceive in detail. Thus, speech production does not mature independent of experience, but is modified by the auditory environment. Perhaps more surprising, it appears that the lack of visual information during development can have an effect on infant vocal productions. Blind children learn sounds that have visible articulation more slowly than sighted children and manifest a different pattern of articulatory errors (Mills, 1987). These alterations in the typical pattern of speech development may occur because blind infants cannot see how other speakers move their articulators to achieve particular targets. This interpretation is in line with results from facial imitation, which directly demonstrated that seeing others' mouth movements influences the oral movements of infants. The natural experiment of blindness shows that the absence of such visual input alters early speech production.

Infant speech production is thus influenced by experience and is a thoroughly intermodal affair—what babies produce with their own articulators is profoundly influenced both by what they see and hear. Infants who are engaged in cooing and babbling in their bassinets are engaged in serious business: They are mastering quite general rules about the auditory consequences of their own vocal tract manipulations. They are solidifying an auditory-articulatory intermodal map of speech. In developing this map they use auditory and proprioceptive information from the self and visual information from others to learn what to do with their own vocal tracts when producing speech. They learn, for example, that raising versus lowering the tongue blade has a particular kind of impact on the sound that is emitted.

Two kinds of studies serve as metrics for infants' acquisition of auditory-articulatory rules, studies of vocal imitation and studies of auditory–visual speech perception. Both assess infants' connections between audition and articulation, and both can be used to chart developmental progress (Kuhl & Meltzoff, 1984, 1988; Studdert-Kennedy, 1986).

Studies of vocal imitation show that as young as 12 weeks of life infants alter their vocalizations to match the vocalizations they hear another produce (e.g., Kuhl & Meltzoff, 1988, 1994; Legerstee, 1991). In our recent study, infants saw a video presentation of a woman articulating one of three vowels, /a/, /i/, or /u/. The infants' vocalizations were analyzed by a phonetically trained listener who transcribed the vocalizations of the infants (while remaining uninformed about the actual stimulus).

The results provided evidence of vocal imitation in 12-week-old infants. The analysis showed that infants produced more /a/-like vocalizations in response to hearing the model say /a/ then in response to hearing the model say either /i/ or /u/. Similarly, they produced more /u/-like vocalizations in response to the model's /u/ than in response to either her /a/ or /i/. Spectrographic analyses of infants' /a/-, /i/-, and /u/-like vowels indicated that the formant patterns of the infants' vocalizations resembled those

characteristic of the adult model's vowels (Kuhl & Meltzoff, 1994). These data indicate that by 12 weeks of age infants have learned something about what to do with their articulators to produce a sound that matches one they hear. They have begun to bring their articulatory and auditory systems into register with one another.

A second measure of infants' relating auditory and articulatory instantiations of speech is infants' ability to detect a match between an auditory presentation of a sound and the sight of a person producing that sound, such as we presented in the lipreading experiments (Kuhl & Meltzoff, 1982, 1984). In this situation the speaker is someone other than themselves. The task demands that infants recognize articulatory movements by eye and relate them to the concomitant auditory information. The fact that 18-week-old infants recognize the correspondence between the sound of the vowel /a/ and the sight of a person with a wide open mouth, and between the sound of the vowel /i/ and the sight of a person with retracted lips, provides converging evidence of infants' acquisition of auditory-articulatory maps.

We are thus emphasizing intermodal connections between the sound, sight, and movements involved in speech, and the role that experience plays in that development. Infants have auditory–motor as well as auditory–visual linkages, and the two may feed on one another during development.

Could real-world cooing and babbling experience contribute to the laboratory effect of infant lipreading? There might be such a developmental effect. It would require that infants relate the articulations they see in our experiment to the auditory-articulatory events they themselves produced when cooing. There is research demonstrating that infants can relate mouth movements they see to their own self-produced mouth movements. Infants can imitate oral gestures. There is thus an underlying ability to map the seen articulatory movements to their own articulations (Meltzoff, 1993). On the auditory side, Kuhl's (1979, 1983, 1985) speech categorization work demonstrates that young infants can recognize the equivalence between the vowels uttered across talkers, including those produced by children and adults. Thus there is also an underlying ability to recognize an equivalence between the heard adult vowels and their own self-produced sounds.

In short, infants have the requisite tools, as manifest by facial imitation and the cross-talker categorization of vowels, to use self-produced speech movements to help solve the intermodal speech task involved in lipreading. The auditory-articulatory mappings experienced during their own cooing and babbling may contribute to infants' ability to recognize cross-modal equivalences for speech when they see those same relations posed on the face of others. The emerging developmental picture is that intermodal abilities support one another, underscoring the web that connects them in ontogenesis.

Stored Targets and Representations and Their Role in Intermodal Theory

The intermodal phenomena we discussed can be organized within three broad classes. The first is a situation in which infants (or adults) detect *perceptual-perceptual matches* between information picked up from two separate perceptual modalities. This encompasses both infant auditory–visual speech perception and the tactual–visual perception of objects. The second involves *perceptual–motor matches.* The visual–motor examples was gestural imitation; the auditory–motor examples were vocal imitation and babbling. Here the infant perceives information in one modality (vision or audition) and this drives a matching event using their own motor systems. The third involves the *blending of discrepant multimodal information into a new and unified percept.* The examples were the auditory–visual blend illusion and the tactual–vision perception of speech. In these cases the information in the two modalities is not equivalent. What unites the two modalities is not "invariant" information picked up by the perceiver. Rather, we suggest that in the case of speech the discrepant information is united by a higher order phonetic representation of speech that acts as a mediator between nonidentical information in the two modalities.

The notion that stored representations link intermodal information provides leverage in discussing several seemingly diverse phenomena. We first apply it to the speech cases and next to the understanding of faces and gestural imitation.

Early exposure to a particular linguistic environment has long-term effects on both speech perception and production. Indeed, the linguistic environment begins to have an effect very early in life. Kuhl and her colleagues showed that exposure to a particular language alters infants' perception of speech by 6 months of age (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Six-month-old infants in two countries, the United States and Sweden, were tested with English and Swedish vowel "prototypes," vowels that were particularly good instances of the category. The results showed that infants in both cultures treated the native-language sounds in a special way: They generalized further around the native-language prototype than around the foreign-language prototype even though psychophysical distance was strictly equated. In explaining these results, Kuhl (1992, 1993a) argued that infants develop stored representations of the native-language sounds that they previously heard and that these affect the processing of current inputs.

We believe that these stored representations for speech in turn serve as targets that guide the motor system. This can be seen in two types of recent studies. First, the results of our studies on vocal imitation in 12-week-old infants show that auditory input is sufficient to drive infant production (Kuhl & Meltzoff, 1982, 1988). Second, it has also been discovered that infants from

different cultures babble in different ways at least by 10 months (de Boysson-Bardies, 1993; de Boysson-Bardies, Halle, Sagart, & Durand, 1989; de Boysson-Bardies, Sagart, & Durand, 1984), suggesting that exposure to native-language speech shapes the particulars of infant production. The inference is that at an early age, speech is represented in a way that unites information from multiple modalities, linking auditory, visual, and motor instantiations of speech. Experience clearly plays a major role in elaborating this representation, inasmuch as infants' representations seem to vary as a function of being reared in different linguistic environments (due to the different auditory and visual input), and also as a function of abnormalities in which the sensory filters or cognitive machinery are atypical (Kuhl, 1993b). Moreover, early exposure seems to have virtually permanent effects on both perception and production, suggesting a "sensitive period" in development. It is difficult to hear speech distinctions that are not used phonemically in one's native language (perception), and it is nearly impossible to lose a foreign accent (production).

Thus, in the case of speech, information originally derived from one modality (audition) has long-term effects for later motor behavior, implicating memory. A parallel case can be made for the visual modality. In the case of gestural imitation it is the visually derived information that drives motor production. Facial imitation is doubly interesting because the motor output cannot be monitored using the same sense modality with which the incoming target was perceived. Meltzoff and Moore suggested that early imitation is mediated by a supramodal representation of the adult's act. The newer findings of imitation after a delay (which indicates memory storage) as well as imitation of novel motor patterns and response correction lend support to the theory that a supramodal representation of the adult's behavior serves as the "internal target" that infants use to generate and correct their behavior.

The ability to act on the basis of supramodal representations is postulated to be an aspect of the human perceptual-cognitive system that is present at birth (Meltzoff, 1990); it allows infants to profit from and organize multimodal experience such that the representations of faces and voices become ever more richly specified (as in the arguments regarding babbling). It would be of great value to investigate the neural basis of such a supramodal representational capacity as it develops in the infant. Highly instructive in this regard are the discoveries by Stein and Meredith (1990, 1993; Stein, Meredith, & Wallace, chap. 5, this volume) about multisensory convergence in the brains of nonhuman animals, as well as recent work on brain growth and plasticity by Greenough (Greenough & Alcantara, 1993; Greenough & Black, 1989) and Edelman (1987, 1989). The extrapolation that can be made on the basis of this neural data is that the brain is ready to accept multimodal input and indeed that the input from both self-generated experience (e.g., babbling) and other-generated experience may help

to consolidate the multimodal links, to "grease the skids," as it were, for later perception.

SUMMARY

Faces and speech are among the most important signals that our perceptual systems have evolved to perceive. These biologically important events provide ideal stimuli for exploring the origins, development, and mechanisms of intermodal functioning in humans.

Research shows that newborn infants can relate the facial movements they see to their own unseen facial movements. It was proposed that facial imitation is mediated by active intermodal mapping (the AIM hypothesis). On this view, facial imitation is among the earliest, most complex, and socially significant manifestations of intermodal functioning in the newborn.

Facial imitation involves a mapping from vision to the motor/proprioception domain. Other research indicated that infants can relate objects they feel to those that they see, suggesting a mapping between touch and vision in the first month of life. Further research in our laboratory and others has focused on the generality of these early intermodal connections. We investigated auditory–visual relations by presenting infants with moving faces (vision) and speech sounds (audition). The results showed that 4-month-old infants could recognize what particular facial movements corresponded to what particular speech sound.

The basis and development of such lipreading was investigated by decomposing the auditory speech stimulus into its elementary "features." Nonspeech sounds that were synchronous with the visible movements they saw were presented to infants. The results showed that adults matched these speech "parts" to faces. Infants did not. We concluded that there is a development in the intermodal perception of speech; infants need to hear the whole speech signal, one that is identified as speech, to make the connection to articulation. This may be related to the fact that when they babble they hear and feel whole speech units.

Adult subjects were also used to investigate further the intermodal organization of speech. Here the principal phenomena were an auditory–visual "illusion" and the finding that speech can be perceived by touch as well as by eye and ear. This work underscored the fact that the cross-modal integration of speech information is so powerful that illusory blends are sometimes mandatory, and are obtained even in cases in which the adult "knows" that the visual and auditory sources cannot go together (the cross-gender experiment). This work differs from ordinary cross-modal studies because here there is no match between the information fed into the separate modalities, but rather the formation of a unified percept that

combines discrepant information. Other research shows that the percept is more than the simple sum of the information fed into the two separate modalities. In our view, a stored supramodal representation of speech is the basis for the unitary perception and multiplicative effect.

The role of development, and especially the role of self-produced experience, was highlighted. Infant cooing and babbling was viewed as consolidating an auditory-articulatory map. Such a map could then be used when infants relate a speech sound to a seen articulation, as in lipreading. This would occur because infants have the underlying ability to connect the articulations they feel themselves make to those they see on the face of others (as manifest and practiced in facial imitation). As infants elaborate their knowledge of faces and speech, they are building a network that interconnects a variety of intermodal phenomena, including facial imitation, babbling, and lipreading, using information from one domain to bootstrap their understanding in another.

Faces and speech are intermodal sources of information. Stimulation from the external world coupled with the infant's own self-produced stimulation provides input about the human body, its movement transformations, and auditory concomitants that may affect neural development (Edelman, 1987, 1989; Greenough, Black, & Wallace, 1987; Stein & Meredith, 1993). Our thesis, which is at root an interactive-developmental one, is that many biologically important signals are not only intermodal in the real physical world—things that can be seen, heard, and touched—but also intermodally organized in the infant's mind. Experience both with others' and with their own bodies plays a role in the ontogenesis of this intermodal organization.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by grants from NIH (HD-22514 and HD-18286). We gratefully acknowledge the long-term collaboration of M. Keith Moore and Kerry P. Green and on aspects of the research reported here. We are indebted to Craig Harris and Erica Stevens for their assistance on all aspects of the research.

REFERENCES

Abramson, A. S., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences Prague 1967* (pp. 569–573). Prague: Academia.

Abravanel, E., & DeYong, N. G. (1991). Does object modeling elicit imitative-like gestures from young infants? *Journal of Experimental Child Psychology, 52,* 22–40.

Abravanel, E., & Sigafoos, A. D. (1984). Exploring the presence of imitation during early infancy. *Child Development, 55,* 381–392.

Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior and Development, 6,* 429–451.

Bahrick, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development, 10,* 387–416.

Bahrick, L. E. (1988). Intermodal learning in infancy: Learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development, 59,* 197–209.

Bahrick, L. E., & Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology, 21,* 963–973.

Bower, T. G. R. (1982). *Development in infancy* (2nd ed.). San Francisco: Freeman.

Bower, T. G. R. (1989). *The rational infant.* New York: Freeman.

Bushnell, E. W., & Weinberger, N. (1987). Infants' detection of visual-tactual discrepancies: Asymmetries that indicate a directive role of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 601–608.

Butterworth, G. (1981). The origins of auditory-visual perception and visual proprioception in human development. In R. D. Walk & H. L. Pick (Eds.), *Intersensory perception and sensory integration* (pp. 37–70). New York: Plenum.

Butterworth, G. (1983). Structure of the mind in human infancy. In L. Lipsitt (Ed.), *Advances in infancy research* (Vol. 2, pp. 1–29). Norwood, NJ: Ablex.

Butterworth, G. (1990). On reconceptualizing sensori-motor development in dynamic systems terms. In H. Bloch & B. I. Bertenthal (Eds.), *Sensory-motor organizations and development in infancy and early childhood* (pp. 57–73). Dordrecht, Netherlands: Kluwer.

Chiba, T., & Kajiyama, M. (1958). *The vowel—Its nature and structure.* Tokyo: Phonetic Society of Japan.

Damasio, A. R., Tranel, D., & Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience, 13,* 89–109.

de Boysson-Bardies, B. (1993). Ontogeny of language-specific phonetic and lexical productions. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 353–363). Boston: Kluwer.

de Boysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language, 16,* 1–17.

de Boysson-Bardies, B., Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 11,* 1–15.

de Schonen, S., & Mathivet, E. (1989). First come, first served: A scenario about the development of hemispheric specialization in face recognition during infancy. *Cahiers de Psychologie Cognitive, 9,* 3–44.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America, 85,* 2154–2164.

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

Dodd, B., & Campbell, R. (1987). *Hearing by eye: The psychology of lip-reading.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection.* New York: Basic Books.

Edelman, G. M. (1989). *The remembered present.* New York: Basic Books.

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. (1978). Asymmetric dependencies in processing phonetic features. *Perception & Psychophysics, 23,* 12–20.

Fant, G. (1973). *Speech sounds and features.* Cambridge, MA: MIT Press.

Farnsworth, P. R. (1937). An approach to the study of vocal resonance. *Journal of the Acoustical Society of America, 9,* 152–155.

Ferguson, C. A., Menn, L., & Stoel-Gammon, C. (Eds.). (1992). *Phonological development: Models, research, implications.* Timonium, MD: York Press.

Field, T., Goldstein, S., Vaga-Lahr, N., & Porter, K. (1986). Changes in imitative behavior during early infancy. *Infant Behavior and Development, 9*, 415–421.

Field, T. M., Woodson, R., Cohen, D., Greenberg, R., Garcia, R., & Collins, E. (1983). Discrimination and imitation of facial expressions by term and preterm neonates. *Infant Behavior and Development, 6*, 485–489.

Field, T. M., Woodson, R., Greenberg, R., & Cohen, D. (1982). Discrimination and imitation of facial expressions by neonates. *Science, 218*, 179–181.

Fontaine, R. (1984). Imitative skills between birth and six months. *Infant Behavior and Development, 7*, 323–333.

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

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Lawrence Erlbaum Associates.

Gentner, D., & Grudin, J. (1985). The evolution of mental metaphors in psychology: A 90-year retrospective. *American Psychologist, 40*, 181–192.

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

Gibson, E. J., & Walker, A. S. (1984). Development of knowledge of visual-tactual affordances of substance. *Child Development, 55*, 453–460.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Grant, K. W., Ardell, L. A. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustical Society of America, 77*, 671–677.

Grant, K. W., Ardell, L. A. H., Kuhl, P. K., & Sparks, D. W. (1986). The transmission of prosodic information via an electrotactile speechreading aid. *Ear and Hearing, 7*, 328–335.

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics, 45*, 34–42.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 278–288.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*, 524–536.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics, 38*, 269–276.

Greenough, W. T., & Alcantara, A. A. (1993). The roles of experience in different developmental information storage processes. In B. de Boysson-Bardies, S. de Schonen, P. Juscyck, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 3–16). Boston: Kluwer.

Greenough, W. T., & Black, J. E. (1989). Induction of brain structure by experience: Substrates for cognitive development. In M. Gunnar & C. Nelson (Eds.), *Developmental behavioral neuroscience: Minnesota Symposia on Child Development* (Vol. 24, pp. 155–200). Hillsdale, NJ: Lawrence Erlbaum Associates.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. *Child Development, 58*, 539–559.

Gunderson, V. M. (1983). Development of cross-modal recognition in infant pigtail monkeys (*Macaca nemestrina*). *Developmental Psychology, 19*, 398–404.

Heimann, M., Nelson, K. E., & Schaller, J. (1989). Neonatal imitation of tongue protrusion and mouth opening: Methodological aspects and evidence of early individual differences. *Scandinavian Journal of Psychology, 30*, 90–101.

Heimann, M., & Schaller, J. (1985). Imitative reactions among 14–21 day old infants. *Infant Mental Health Journal, 6,* 31–39.

Helmholtz, H. (1954). *On the sensations of tone as a physiological basis for the theory of music.* New York: Dover. (Original work published 1885)

Jacobson, S. W. (1979). Matching behavior in the young infant. *Child Development, 50,* 425–430.

Jakobson, R., Fant, C. G. M., & Halle, M. (1969). *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: MIT Press.

Jusczyk, P. W., Pisoni, D. B., Reed, M. A., Fernald, A., & Meyers, M. (1983). Infants' discrimination of the duration of a rapid spectrum change in nonspeech signals. *Science, 222,* 175–177.

Jusczyk, P. W., Pisoni, D. B., Walley, A., & Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *Journal of the Acoustical Society of America, 67,* 262–270.

Kaitz, M., Meschulach-Sarfaty, O., Auerbach, J., & Eidelman, A. (1988). A reexamination of newborn's ability to imitate facial expressions. *Developmental Psychology, 24,* 3–7.

Kaye, K. L. (1993, March). *Cross-modal matching in human newborns.* Presented at the meeting of the Society for Research on Child Development, New Orleans, LA.

Kugiumutzakis, J. (1985). *Development of imitation during the first six months of life* (Uppsala Psychological Reports No. 377). Uppsala, Sweden: Uppsala University.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66,* 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development, 6,* 263–285.

Kuhl, P. K. (1985). Categorization of speech by infants. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion* (pp. 231–262). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kuhl, P. K. (1992). Infants' perception and representation of speech: Development of a new theory. In J. Ohala, T. Nearey, B. Derwing, M. Hodge, & G. Wiebe (Eds.), *The Proceedings of the International Conference on Spoken Language Processing* (pp. 449–456). Edmonton, Alberta: University of Alberta Press.

Kuhl, P. K. (1993a). Innate predispositions and the effects of experience in speech perception: The Native Language Magnet theory. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 259–274). Boston: Kluwer.

Kuhl, P. K. (1993b). Developmental speech perception: Implications for models of language impairment. In P. Tallal, A. M. Galaburda, R. R. Llinás, & C. von Euler (Eds.), *Temporal information processing in the nervous system* (Vol. 682, pp. 248–263). New York: New York Academy of Sciences

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

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7,* 361–381.

Kuhl, P. K., & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy: The Minnesota Symposia on Child Psychology* (Vol. 20, pp. 235–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kuhl, P. K., & Meltzoff, A. N. (1994). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America,* under revision.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608.

Kuhl, P. K., Williams, K. M., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 829–840.

Legerstee, M. (1991). The role of person and object in eliciting early imitation. *Journal of Experimental Child Psychology, 51*, 423–433.

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

Lewkowicz, D. J. (1985). Bisensory response to temporal frequency in 4-month-old infants. *Developmental Psychology, 21*, 306–317.

Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development, 9*, 335–353.

Lewkowicz, D. J. (1992). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior and Development, 15*, 297–324.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Locke, J. L. (1993). *The child's path to spoken language*. Cambridge, MA: Harvard University Press.

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

Maratos, O. (1982). Trends in the development of imitation in early infancy. In T. G. Bever (Ed.), *Regressions in mental development: Basic phenomena and theories* (pp. 81–101). Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1987a). Speech perception by ear and eye. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53–83). Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1987b). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science, 1*, 55–63.

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

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology, 24*, 470–476.

Meltzoff, A. N. (1990). Towards a developmental cognitive science: The implications of cross-modal matching and imitation for the development of representation and memory in infancy. In A. Diamond (Ed.), *The development and neural bases of higher cognitive functions. Annals of the New York Academy of Sciences, 608*, 1–31.

Meltzoff, A. N. (1993). The centrality of motor coordination and proprioception in social and cognitive development: From shared actions to shared minds. In G. J. P. Savelsbergh (Ed.), *Advances in Psychology Series: The development of coordination in infancy* (pp. 463–496). Amsterdam: Elsevier.

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

Meltzoff, A. N., Kuhl, P. K., & Moore, M. K. (1991). Perception, representation, and the control of action in newborns and young infants: Toward a new synthesis. In M. J. S. Weiss & P. R. Zelazo (Eds.), *Newborn attention: Biological constraints and the influence of experience* (pp. 377–411). Norwood, NJ: Ablex.

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.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology, 25*, 954–962.

Meltzoff, A. N., & Moore, M. K. (1992). Early imitation within a functional framework: The importance of person identity, movement, and development. *Infant Behavior and Development, 15*, 479–505.

Meltzoff, A. N., & Moore, M. K. (1993). Why faces are special to infants—On connecting the attraction of faces and infants' ability for imitation and cross-modal processing. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 211–225). Boston: Kluwer.

Meltzoff, A. N., & Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development, 17*, 83–99.

Miller, J. L. (1977). Properties of feature detectors for VOT: The voiceless channel of analysis. *Journal of the Acoustical Society of America, 62*, 641–648.

Mills, A. E. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 145–161). Hillsdale, NJ: Lawrence Erlbaum Associates.

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development, 59*, 441–449.

Oller, D. K., & Lynch, M. P. (1992). Infant vocalizations and innovations in infraphonology: Toward a broader theory of development and disorders. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 509–536). Timonium, MD: York.

Ortony, A. (1979). *Metaphor and thought*. Cambridge: Cambridge University Press.

Pêcheux, M.-G., Lepecq, J.-C., & Salzarulo, P. (1988). Oral activity and exploration in 1–2-month-old infants. *British Journal of Developmental Psychology, 6*, 245–256.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. L. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /a/: A pilot "motor equivalence" study. *Journal of the Acoustical Society of America, 93*, 2948–2961.

Piaget, J. (1962). *Play, dreams and imitation in childhood*. New York: Norton. (Original work published 1945)

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics, 34*, 314–322.

Reissland, N. (1988). Neonatal imitation in the first hour of life: Observations in rural Nepal. *Developmental Psychology, 24*, 464–469.

Rose, S. A. (1990). Cross-modal transfer in human infants: What is being transferred? *Annals of the New York Academy of Sciences, 608*, 38–50.

Rose, S. A., & Ruff, H. A. (1987). Cross-modal abilities in human infants. In J. D. Osofsky (Ed.), *Handbook of infant development* (pp. 318–362). New York: Wiley.

Sparks, D. W., Ardell, L. A., Bourgeois, M., Wiedmer, B., & Kuhl, P. K. (1979). Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of connected discourse. *Journal of the Acoustical Society of America, 65*, 810–815.

Sparks, D. W., Kuhl, P. K., Edmonds, A. E., & Gray, G. P. (1978). Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental features of speech. *Journal of the Acoustical Society of America, 63*, 246–257.

Spelke, E. S. (1981). The infants' acquisition of knowledge of bimodally specified events. *Journal of Experimental Child Psychology, 31*, 279–299.

Spelke, E. S. (1987). The development of intermodal perception. In P. Salapatek & L. Cohen (Eds.), *Handbook of infant perception: Vol. 2. From perception to cognition* (pp. 233–273). New York: Academic Press.

Stein, B. E., & Meredith, M. A. (1990). Multisensory integration: Neural and behavioral solutions for dealing with stimuli from different sensory modalities. In A. Diamond (Ed.), *The development and neural bases of higher cognitive functions. Annals of the New York Academy of Sciences, 608*, 51–70.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stoel-Gammon, C. (1988). Prelinguistic vocalizations of hearing-impaired and normally hearing subjects: A comparison of consonantal inventories. *Journal of Speech and Hearing Disorders, 53*, 302–315.

Stoel-Gammon, C., & Otomo, K. (1986). Babbling development of hearing-impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders, 51*, 33–41.

Streri, A. (1987). Tactile discrimination of shape and intermodal transfer in 2- to 3-month-old infants. *British Journal of Developmental Psychology, 5*, 213–220.

Streri, A., & Milhet, S. (1988). Équivalences intermodales de la forme des objects enter la vision et le toucher chez les bébés de 2 mois [Intermodal equivalences between vision and touch for the form of objects in 2-month-old infants]. *L'Anneé Psychologique, 88*, 329–341.

Streri, A., & Spelke, E. S. (1988). Haptic perception of objects in infancy. *Cognitive Psychology, 20*, 1–23.

Studdert-Kennedy, M. (1986). Development of the speech perceptuomotor system. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 205–217). New York: Stockton Press.

Studdert-Kennedy, M. (1993). Some theoretical implications of cross-modal research in speech perception. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 461–466). Boston, MA: Kluwer.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36*, 314–331.

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

Vinter, A. (1986). The role of movement in eliciting early imitations. *Child Development, 57*, 66–71.

Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology, 33*, 514–535.

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relationship of eye and voice? *Developmental Psychology, 22*, 373–377.

Walker-Andrews, A. S. (1988). Infants' perception of the affordances of expressive behaviors. In C. K. Rovee-Collier (Ed.), *Advances in infancy research* (Vol. 5, pp. 173–221). Norwood, NJ: Ablex.

Walton, G. E., & Bower, T. G. R. (1993). Amodal representation of speech in infants. *Infant Behavior and Development, 16*, 233–243.

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics, 30*, 557–564.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638–667.