7 Speech as an Intermodal Object of Perception

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: Erlbaum.

Patricia K. Kuhl
Andrew N. Meltzoff
University of Washington

Theoretical treatments of perceptual development typically focus on the visual world. Here, surfaces have to be broken up and perceived as entities that are bounded and separate from one another. This is essential for object perception. An equally challenging but different domain in which to pose questions about perceptual development involves infants' perception of "auditory objects," in the case discussed here, speech.

It is illuminating to think of speech as an "object of perception." Consider the consonants and vowels, the sounds that form the building blocks of speech. These phonetic segments are the entities used when we execute a sequence of articulatory gestures needed to pronounce a word. Vocalizing the word *split* very slowly highlights this point. The junctures between the units of *split* are smooth, but there are five distinct gestures or targets that must be sequenced in order to produce the word. The gesture for /s/ can be separated from that of /p/ and /l/, and so on. Eliminating one of the gestures, for instance the /p/ or the /l/, results in the new words *slit* and *spit*. Moreover, the correct sequencing of segments is invariant whether the word is spoken or written. In the former case the sequence of articulatory gestures has to be mouthed in a given order; and in the latter a series of orthographic symbols is sequenced using motions of the hand with a writing instrument. The same is true in the perception of the word. Whether in reading the printed word or in listening to a speaker, the ordered sequence of units has to be deciphered for the word to be perceived. The consonants and vowels of speech, although they cannot be held like a ball or a cup, are independently manipulable entities with phenomenal reality. They are the "units" of speech and, as such, are objects of perception.

It is of interest to theory building to understand how infants develop a sense of the critical properties defining objects and their separateness from other objects. Visual objects have to be recognized even when they sit on top of other things and might appear to be continuous with them. They have to be recognized when they are rotated in space, appear at different distances, or are partially occluded. According to some theories, an infant's ability to pick up and handle an object enriches his tendency to define it as a separate entity distinct from all others. Visual objects are tangible, manipulable things.

Speech sounds are not "things" that can be reached for, touched, and held. In fact, defining speech sounds in any physical sense is difficult. The characteristics defining a speech sound's identity as well as the characteristics defining its unity or separateness are notoriously complicated. Speech segments are described as exhibiting a lack of "invariance" (the identity problem in speech) and "linearity" (the unity problem) (Chomsky & Miller, 1963).

Regarding the lack of invariant acoustic properties in speech segments, research has repeatedly demonstrated that the acoustic cues underlying the perception of individual phonetic segments are context dependent (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). One context cue that exemplifies this is the age and gender of the talker. Because talkers' mouths differ dramatically in size and shape, even a simple vowel such as /a/ results in very different physical signals when it is produced by talkers of different age and sex (Peterson & Barney, 1952). Moreover, when a given sound is produced by the same talker but in different phonetic environments (e.g., the /d/ in /da/, /di/, and /du/) it is "co-articulated," that is, it is strongly influenced by its neighboring consonants and vowels. This causes rather large differences in the acoustic events that signal /d/ in the three cases. A comparable situation in vision would be if an object, such as a cup, actually changed its physical shape when it sat adjacent to different objects, or rested on different surfaces. This would make defining the cup's "true" features very difficult indeed. Presumably, however, there would be some invariant set of features that remain constant across all contexts in which the cup appears, and these features could be described as the criterial ones defining the object. In speech, it is just such a search for the invariant cues to speech objects that has proven so frustrating and difficult. (But see Blumstein and Stevens, 1981, for recent progress.) The mapping between acoustic information and phonetic perception is enormously complex; so much so that to date no computer can be programmed to recognize the phonetic structure of ongoing speech across a variety of talkers.

No less problematic is the issue of unity, or as it is more classically referred to in speech, the linearity or segmentation problem. The speech stream as it appears acoustically is continuous. It cannot be temporally segmented so that the surface layout of the acoustic stream relates in a one-to-one fashion to the ordered sequence of speech sounds that are perceived. Speech sounds do not lie out in order like beads on a string even though they are perceived that way. Thus, although speech sounds have phenomenal reality—they are produced as a planned sequence of motor gestures and perceived as an ordered sequence of segments—delineating their boundaries and their defining characteristics physically, either in exact acoustic or motor terms, has so far proven impossible.

Finally, to make things more complex, we know that there are multiple sources of information that contribute to the perception of speech segments. For example, no less than six distinct auditory events, spread out in time, control the perception of the voicing feature (Klatt, 1975). Not only do acoustic cues that are relatively close to the unit influence it, but events that are quite remote, say three to four words away from it, influence its identity. This is the rule, not the exception, in defining speech units. How these multiple sources of information are integrated and weighed is a central problem in speech perception.

Most recently, investigations on the multiple-sources-of-information problem in speech have raised a new issue—the multimodal delivery of that information. The question is whether information delivered through a modality other than audition can influence the perception of speech segments. In object perception, the answer to this is quite clear. Information about objects is not restricted to that delivered by a single sensory modality. Objects in the world—cups, coins, keys—can be seen as well as touched, and information from both modalities can contribute to the identity and unity of the object. Thus, objects in the world are intermodally specified, and research has shown that observers use both kinds of information to identify objects.

But what of the objects of speech? Speech is normally considered the sole province of audition. Yet speech can be seen. During typical conversations we see the talker's face, and watch the movements of lip, tongue, and jaw that are concomitant byproducts of the speech event. So in some sense speech is both auditory and visual. But is speech an intermodal event for the listener/observer? Are the visual events that accompany the auditory signal taken into account in determining the identity of the unit, or are they simply ignored? And if adults take speech to be an intermodal event, how is knowledge of its intermodal nature acquired by infants?

This is a new and complex issue in speech. Here the mapping between physical cues and phonetic percepts goes beyond the realm of the single modality typically associated with it. As such, it becomes an intermodal mapping problem. How such a complex array of information—including that delivered by eye and by ear—is organized in development is the subject of this particular chapter. We show that speech is already an intermodal object of perception for young infants, and that this has some interesting theoretical implications.

THERE IS MORE TO SPEECH THAN MEETS THE EAR: EFFECTS OF VISION ON SPEECH PERCEPTION

In adults the sight of a person producing speech contributes to its perception by hearing-impaired people. Lip-reading was used to teach deaf people to speak at least as early as the mid 1500's; Pablo Bonet in 1620 credits Ponce de Leon with having "taught the dumb to speak" at that time (Deland, 1920). There is ample modern-day research to indicate that normal-hearing adults also benefit from watching a talker's mouth movements, especially in noise (Sumby & Pollack, 1954). The "cocktail party effect," watching the face of the talker at a noisy party, is a common example. We do it without being aware of it, presumably because it feels as if vision helps us to hear the talker. Those of us who wear glasses are also familiar with the impression that it is harder to hear without our glasses on.

These commonplace examples notwithstanding, the full theoretical impact of the role of vision in speech perception was not recognized until relatively recently. One factor bringing the issue to the center of attention was demonstrations that a normal-hearing observer is strongly influenced by the sight of the talker's articulatory movements, even when the auditory signal is perfectly clear, and not degraded by noise. A powerful demonstration of this occurs when the auditory and visual information in speech, which is normally redundant, is put in conflict. When auditory information specifying the disyllable /baba/ is combined with visual information specifying the disyllable /gaga/, the illusory percept /dada/ is perceived (McGurk & MacDonald, 1976). This effect is robust at least for that particular pair of sounds and has been replicated in several labs (Green & Kuhl, 1986; Kuhl, Green, & Meltzoff, in preparation; Massaro & Cohen, 1983; Summerfield, 1979). We are just beginning to understand some of the factors governing the integration of discrepant auditory and visual speech information by adults, and several competing theoretical explanations have emerged (Kuhl et al., in preparation).

We do know that the role of vision is not restricted to situations involving a specific mouth movement that relates to a particular phonetic unit in a syllable. Here we cite two quite different examples. The first example involves the perception of the phonetic distinction /b/ versus /w/. This phonetic distinction is influenced by many different sources of information. One of them is the overall rate at which a speaker is talking (Miller & Liberman, 1979). Systematic increases or decreases in the rate of speech necessitate systematic changes in the specific acoustic cues required to perceive /b/ versus /w/. This means that the listener takes rate of articulation information into account when evaluating the acoustic information. Both the acoustic cues and the overall rate information are used to decide whether the speaker said /b/ or /w/. What is surprising is that this source of information, rate of speech, can be presented visually rather than auditorially, with no diminution in the effect. Green and Miller (1985) had observers watch a speaker who used either a fast or a slow rate of speech, while listening to the same acoustic information. The question was whether or not watching the fast versus slow speech would influence the /b-w/ judgments. The results showed that picking up the rate information visually resulted in a replication of the same effect observed when the rate information was delivered auditorially. Nearly identical changes in the acoustic information were shown to be needed to maintain the perception of /b/ as opposed to /w/. Apparently, even information as global as "rate of speech" can be provided through the optic channel.

A second example demonstrates that the effects of visual information on speech perception are not restricted to single syllables. This experiment involved the perception of ongoing speech. It demonstrated that speech can be perceived quite readily under extremely impoverished listening conditions if the face of the talker is in full view. Grant, Ardell, Kuhl, and Sparks (1985) presented listeners with a pure-tone signal that followed the fundamental frequency (pitch) of a talker who was reading prose. In the first test condition, the listener did not face the reader, so no visual information was available. Only the tone was presented. By itself, the tone provided no information about speech. Not a single word, syllable, or phoneme could be identified. It was simply a tone that changed in frequency. In the second condition, the listener turned and faced the talker, watching the talker speak while listening to the tone. The listener was to repeat, word for word, everything the talker said, so that the degree of speech reception could be precisely assessed. Results showed that the listener/observer could successfully repeat about 70% of the material, as opposed to about 40% when vision alone was provided. The value of visual information, particularly in conjunction with auditory information, was thus firmly demonstrated.

Taken together, these studies provide powerful evidence that speech perception is not the sole province of audition. It can be; we can hear perfectly well with our eyes closed. But when provided, information from the visual channel is taken into account. In fact, our work (Kuhl et al., in preparation) and that of others (Massaro & Cohen, 1983) suggest that the perceiver is compelled to take visual information into account when it is present. It cannot be ignored. What mechanism relates speech information from two such disparate sources, the eye and the ear? What is the ontogenesis of the ability to equate optic and acoustic information for speech? We return to this question after examining the nature of the information that vision supplies.

THERE IS MORE TO SPEECH THAN MEETS THE EYE: LIMITATIONS ON THE VISUAL CHANNEL

Thus far we have discussed the surprising extent to which visual information of various sorts contributes to speech perception. It is now appropriate to dispel any notion that the reader may have formed that all or even most of speech can be perceived by eye. A relatively small proportion of the information in speech is visually available.

We can divide the information in speech into two broad classes. One involves the phonetic structure of speech – the consonant and vowel segments we have been describing. The second involves the prosodic aspects of speech, the intonation (pitch) pattern of the utterance, its stress and rhythm.

Consider first the contribution of vision to phonetic perception. Can all of the consonants and vowels be identified using visual information alone? The answer to the question is definitely no. The simplest way to explain this answer is to describe the component features, the "distinctive features" (Jakobson, Fant, & Halle, 1969), that make up consonant and vowel segments. Some of these features are observable visually, but others are not.

The "place of articulation" feature is the most visually available speech feature. The place feature describes the location in the mouth where the primary constriction of the airflow in the vocal tract takes place. When producing a /b/, /p/, or /m/, for example, the location of primary airflow constriction is the lips, so we refer to the sound as having a *bilabial* place of articulation. Obviously, a bilabial articulation is highly visible. But other levels of the place feature are not as prominent. An *alveolar* articulation, where the tongue tip touches the ridge in back of the teeth, involved in the production of sounds such as /t, d, n/, can be seen under good lighting conditions, but it is more difficult. Sounds produced even further back in the mouth, the *velars* such as /k, g, ŋ/, are nearly impossible to see. Fricatives with intermediate places of articulation, such as the alveolars, /s/ and /z/; the labiodentals /f/ and /v/; the linguadentals, /θ/ and /ɚ/; and the palatals, /ʃ/ and /ʒ/, depend on the amount of training of the observers and the lighting conditions under which the talker was filmed. Under the best conditions observers who have had training with a given speaker can successfully identify at 75% accuracy nine different place categories for the consonants of English (Walden, Prosek, Montgomery, Scherr, & Jones, 1977). These categories are the *visemes*, that is, visibly distinguishable phonemes.

The second major class of distinctive feature, the "manner of articulation" features, are usually not distinguishable by vision alone. The manner features distinguish phonetic segments produced at the same place of articulation. For example, the segments /p/, /b/, and /m/ are distinguished by two manner features, voicing and nasality. But these manner features cannot be seen on the lips of the talker. Producing the sounds /p/, /b/, and /m/ in front of a mirror will illustrate how similar they look. Virtually no information regarding these manner features can be seen.

We turn now to prosodic information, that which identifies the intonation, stress, and rhythm of an utterance. It too is largely unavailable visually. The intonation (pitch) of an utterance is controlled by the glottis, located deep in the larynx. There is no visible correlate to glottal vibration and thus no visible manifestation of intonation. Linguistic stress, which is cued by changes in intonation, loudness, and duration, is also not directly perceivable by eye. Although there has been little experimental work to verify it, a likely feature being used in studies where prosodic cues are picked up visually is information about the durations and junctures of syllables provided by the opening and closing of the mouth. (Of course, there may be other associated body changes that often go along with linguistic stress, such as arm movements, head nods, and/or eyebrow movements, but here we are addressing ourselves to the necessary concomitants of prosodic information that are manifest in the articulatory movements themselves.)

To summarize, much of speech misses the eye. Of the phonetic features that make up consonant and vowel segments, only the place feature can be seen. The manner feature is not visible. Some prosodic information, in the form of syllabification, is probably available. and this is a powerful cue in on-going speech, but no direct manifestation of intonation or linguistic stress can be gleaned through the visual channel. The contribution of vision to speech perception is therefore carried by a very small number of speech features. An interesting fact is that the highly visible place features are more disruptable by auditory perturbations than their invisible cousins. Miller and Nicely (1955) showed that the introduction of noise and/or filtering affects the perception of place features dramatically. Just the opposite is true for manner features. They are virtually invisible, but are auditorially robust. Thus, there is an interesting complementarity between the various speech features when delivered through the auditory versus the visual modality.

SPEECH THROUGH THE TACTUAL SENSE

As discussed, place information can be perceived by the visual channel. But before we propose a "place by eye, manner by ear" hypothesis (MacDonald & McGurk, 1978), a rather more startling finding needs to be dealt with. Information about some speech features—notably manner features, such as voicing, nasality, and frication—can be successfully delivered through the skin.

Much research has been directed toward tactile aids for deaf listeners (Kirman, 1973). The important point for this discussion is that there is ample evidence that certain speech features delivered through the skin can be integrated with information delivered through another modality such as vision or audition. For example, work in our labs has shown that manner information delivered through the skin can be combined with place information delivered visually, resulting in the correct perception of syllables differing in place, voicing, and nasality (Sparks, Ardell, Bourgeois, Wiedmer, & Kuhl, 1979; Sparks, Kuhl, Edmonds, & Gray, 1978). These studies used electrocutaneous stimulation via a matrix of 144 electrodes. The matrix displayed the spectrum of speech in a frequency × amplitude display that contained 36 different frequency channels with eight amplitude levels at each frequency. The device displayed the information spatially and was worn as a belt circling the abdomen. Tests showed that although the place features were not reliably detected by subjects wearing the belt, the voicing and nasality features were. When observers combined the information obtained through the tactile channel with place information perceived by eye, perception of syllables was near perfect (Sparks et al., 1978).

More recently, Grant, Ardell, Kuhl, and Sparks (1986) conducted a tactile study that closely paralleled their earlier one exploring the role of vision in speech perception. Recall that in Grant et al. (1985) subjects viewed a talker reading prose while they were auditorially presented with a pure tone that followed the fundamental frequency of the talker's voice. In Grant et al. (1986) the same test conditions were replicated using the tactile channel as an aid to lipreading. The test conditions again involved the perception of ongoing speech when information obtained by watching a talker speak was combined with fundamental-frequency information (*fo*). But this time, rather than delivering the *fo* information auditorially, they devised an electrocutaneous device that could be worn on the forearm. It consisted of eight electrodes that were arranged spatially in a line from wrist to elbow. Each electrode covered a limited range of frequencies and vibrated when the *fo* was in its specified range. As in the previous study, the receiver was exposed to the electrocutaneous information, both observing and without observing the talker. The results demonstrated that the tactually delivered *fo* information significantly increased speech reception over that obtained by lipreading alone.

Although there have been many studies on the tactile reception of speech, controlled studies are still few in number. What is clear is that some speech information can be delivered to the skin and integrated with that perceived by eye or by ear. The speech information obtained through tactile aids is not as dramatic as that obtained through vision, and training is necessary, but there is no doubt that if the processing limitations of the skin are taken into account (Sparks et al., 1978), speech information can be delivered tactually. The important theoretical point is that the limiting factor is not the impenetrability of the speech-processing mechanism. 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 AMODAL REPRESENTATION OF SPEECH?

On the basis of the work discussed here, one can suggest that the speech-processing mechanism may be amodal in nature. Information about place features can enter via the visual or auditory modality and information about manner features can enter through the skin or ear. To date, there are no models of speech perception that explain or predict that this should be the case. The pick-up of speech information from such varied input systems poses a profound problem for models of the speech-recognition mechanism.

The profound problem is captured in the following examples. A bilabial articulation, such as /b/, can be signaled either by the sight of lips coming together or by the sound of formant transitions that rise in frequency. To address another case, the timing difference that separates the release of a stop consonant and the onset of voicing (which distinguishes /b/ from /p/ and is called the voice-onset time or VOT), can be presented either auditorially, in the form of two acoustic events, or tactually, in the form of differentially timed vibratory pulses. How is information delivered across different input modalities equated by the speech-processing mechanism? And when place information is delivered by eye and manner by ear or skin, how is the information organized to form a phenomenally unified speech percept? Is there a common metric, one that is modality-neutral, *amodal*, that recognizes the equivalence between information entering different channels?

One way to approach these problems is to ask how the developing system comes to be organized. When, for example, do infants recognize that visual information about speech, particular mouth movements and postures, correspond to particular speech sounds?

In 1980, we embarked on a research program to find this out. One of us had already explored the cross-modal perception of objects (Meltzoff & Borton, 1979) and found that 4-week-olds could relate objects presented visually to those previously explored tactually. He had also done research indicating that young infants could perceive subtle differences in mouth movements, inasmuch as they could differentially imitate mouth movements on the basis of vision alone (Meltzoff, 1985; Meltzoff & Moore, 1977, 1983a). The other had done extensive work on infants' auditory perception of speech, primarily the categorization and representation of speech sounds (Kuhl, 1985, 1986a, 1986b, 1987). She had also examined adults' abilities to perceive speech information delivered through visual and tactile channels (Grant et al., 1985, 1986; Sparks et al., 1978, 1979). The time was ripe to examine infants' perception of intermodal relations for speech. What followed was a series of collaborative experiments designed to pose "lipreading" problems for infants (Kuhl & Meltzoff, 1982, 1984a). The studies produced three important findings: (a) young infants can indeed related speech information presented auditorially and visually, (b) reducing the speech signal to a simple acoustic feature that distinguishes the vowels is not a sufficient stimulus to produce the matching effect in infants, and (c) infants provide evidence of vocal imitation, thus manifesting another aspect of the intermodal organization of speech.

STUDIES ON INFANTS' INTERMODAL PERCEPTION OF SPEECH

Our research program was designed to discover whether young infants could relate the sight of a person producing a particular speech sound to its auditory concomitant (Kuhl & Meltzoff, 1982). In order to pose the problem to infants we devised a situation in which infants were shown two filmed faces of the same woman articulating two different vowel sounds. One face articulated the vowel /a/ as in *pop* and the other /i/ as in *peep*. While viewing the two faces infants were auditorially presented with one of the vowel sounds, either /a/ or /i/. The experimental set-up is shown in Fig. 7.1.

We reasoned that if infants could relate specific articulatory postures to their auditory equivalents, they would look longer at the face that "matched" the sound. Two problems had to be solved in devising an experimental situation that allowed us to test this hypothesis. The more obvious of the two is that the auditory signal had to be delivered from a neutral location to avoid any spatial localization cues as to the correct location (right or left) of the matched face. If the sound emanated from a location nearer one or the other of the two faces, infants might be influenced by this and look at the face nearest the sound source. Such a test would show that infants could localize

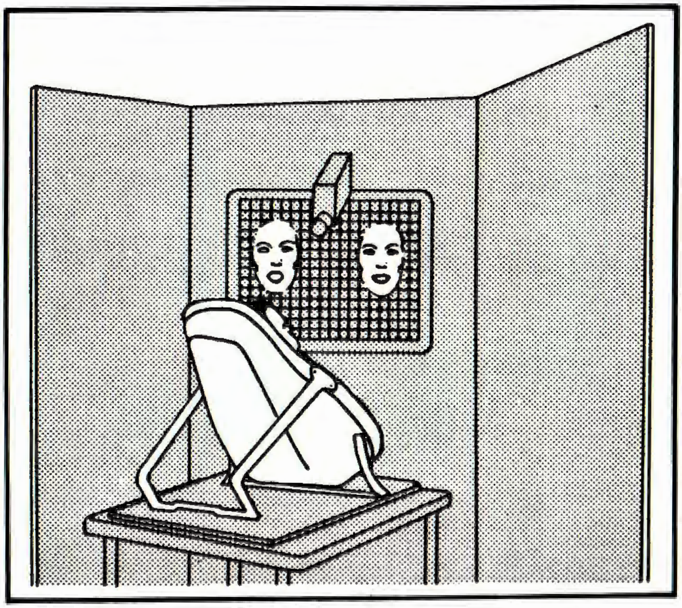

FIG. 7.1 Technique used to test infants' cross-modal perception of speech (from Kuhl & Meltzoff, 1982).

sound but not that they could solve the cross-modal matching problem. To guard against this we placed the loudspeaker midway between the two faces.

A second problem was less obvious, and more difficult to solve. Because speech is dynamic and unfolds in time we had to make sure that the match between the auditory and visual stimuli was not based on their temporal characteristics. Consider the following situation. If the visual /a/ mouth remained open for a longer period of time than the visual /i/ mouth, and the /a/ sound was likewise longer, then infants' detection of a match between visual /a/ and auditory /a/ could be based on purely temporal characteristics. Similarly, the time course of the mouth openings (opening, maximum, and closing) and the amplitude envelope of the sounds (gradually louder to a maximum and then gradually softer) had to be well matched. We took a variety of steps to control these temporal envelope cues (Kuhl & Meltzoff, 1984a).

First, rather than using a single face to represent /a/ visually and a single face to represent /i/ visually, we used a series of productions (20) of each. This was done so that no idiosyncratic feature of a single articulation could influence the detection of the match. The same was true for the auditory stimuli. Twenty auditory /a/'s and 20 auditory /i/'s were chosen for use, and importantly, they were not the /a/'s and /i/'s originally produced by the chosen visual stimuli. Thus there were no idiosyncratic features to link the two domains. All of the auditory and visual stimuli fell within a narrow and overlapping range. We felt that slight variation in duration and loudness was good because it tended to focus infants' attention on the category /a/ and the category /i/ rather than on single features of any one stimulus. It also made the stimuli appear to be natural productions of a long series of vowels.

The stimuli were used to create two film loops each 3 min in duration (one with /a/ on the right and /i/ on the left, and the other opposite this), and two auditory loops (one for each of the two vowels). A special projector was used to replay the stimuli, and it allowed a mechanical link between the audio and visual 16-mm tracks so that once started they could not get out of sync. Each of the visual loops could be combined with each of the audio loops to create four conditions. The alignment of the audio and visual loops was done in a professional studio and was made easy by the careful control over selection of the stimuli for use. When complete, watching the /a/ or the /i/ face while listening either to the /a/ or /i/ auditory soundtrack provided no clue to which face the sound was better aligned to temporally. They were both equally good. This, of course, was a critical point. If babies could detect a match based on temporal features, then we could say nothing about infants' detection of correspondence for the *phonetic information* in auditory and visual speech, and the question of the intermodal representation of speech per se could not be raised. Infants could simply have detected a temporal match between two stimuli. Our procedures ensured that this could not be the case (Kuhl & Meltzoff, 1984a). Infants could solve the cross-modal matching test only by detecting a phonetically relevant correspondence between the /a/ articulation and its auditory concomitant and between the /i/ articulation and its auditory concomitant. There were no spatial cues or temporal cues that would allow them to do so any other way.

The experiment was conducted in two phases, a familiarization phase and a test phase. During familiarization, infants were shown each visual stimulus, in the absence of sound, for 10 sec. Then, once the infant's gaze returned to midline, the 2-min test phase began during which both faces were presented side by side and the sound (either /a/'s or /i/'s) was turned on. All of the obvious features were counterbalanced: right-left facial orientation, sound presentation, side of first-face presentation during familiarization, and sex of the infant.

Thirty-two normal infants ranging in age from 18 to 20 weeks were tested. They were placed in an infant seat facing a three-sided cubicle, as shown in Fig. 7.1. The room was darkened so that the only light provided was that resulting from the films. Infants were videorecorded using an infrared camera, and audiorecorded from a microphone suspended above the infant's head. An observer, who was uninformed about the infant's test condition and could neither see the visual stimuli nor hear the sound presented to the infant, scored the videotaped visual fixations to the right or left stimulus.

We predicted that if infants detected correspondence between auditorially and visually presented vowels, their visual fixations would be systematically influenced by the sound they heard. The results supported this prediction. Infants who heard /a/ looked longer at the face producing /a/, and infants who heard /i/ looked longer at the face producing /i/. The total amount of visual fixation time devoted to the matched face was 73.6%, significantly greater than the 50% chance value ($p < .001$). Twenty-four of the 32 infants looked longer at the matched face ($p < .01$ by the binomial test). There were no other significant effects. The matching effect was equally strong for the /a/ and the /i/ stimuli, and equally strong when the matched stimulus appeared on the infant's right side as opposed to the left side.

The results suggested that by about 4 months of age infants can relate /a/ sounds to /a/ faces and /i/ sounds to /i/ faces. They can in some sense equate the auditory form of a vowel sound and its visual equivalent. Such a result was important for theory, and our first step was to replicate and extend our findings.

The replication experiment was undertaken with 32 additional infants and a new research team (Kuhl & Meltzoff, 1984b). All other details of the experiment were identical. The results again showed that infants looked longer at the face that matched the sound they heard. Of the total fixation time, infants spent 62.8% fixating the matched face ($p < .05$), and 23 of the 32 infants demonstrated the effect ($p < .01$). Recently another team of investigators has also replicated this cross-modal matching effect for speech using disyllables such as *mama* versus *lulu* and *baby* versus *zuzi* in a design similar to ours (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983).

Our next question was whether we could extend the cross-modal speech effect to another vowel pair. If the effect was of importance it would have to be upheld in more than a single vowel pair. We chose to test the /i-u/ pair, thus including the third member of the "point" vowels, /i, a, u/, in the set of vowels tested. The point vowels are maximally distinct, both acoustically and articulatorily, and define the three endpoints of the triangle in vowel space (Peterson & Barney, 1952).

The test was conducted just as it had been previously, only this time infants watched faces producing the vowels /i/ and /u/, and listened to either /i/ or /u/ vowels. Thirty-two new infants were tested. The results showed that the effect could be extended to a new vowel pair. The mean percentage of fixation time to the matched face was 63.8% ($p < .05$), and 21 of the 32 infants looked longer at the matched face ($p < .05$) (Kuhl & Meltzoff, 1984b).

Having demonstrated in three studies infants' abilities to detect equivalences in speech information presented visually and auditorially, we devised a series of experiments to uncover the basis of the effect. We not only wanted to know that 18-week-olds could perform a neat trick akin to lipreading, we wanted to know *how* they did it. Our approach was to search for the effective stimulus. We wanted to know what aspect of the auditory signal was necessary and sufficient to evoke the matching response. Was it necessary that the auditory signal contain enough information to identify the vowel? Or would it be sufficient to use a single isolated property of the auditory stimulus, one that preserved a distinctive feature of the vowel, but did not allow the vowel to be identified? We designed studies to take apart the auditory stimulus in order to identify the critical stimulus features that governed the effect, believing that such studies would contribute to theories concerning the nature and basis of infants' cross-modal speech perception abilities.

EXAMINING THE BASIS OF CROSS-MODAL MATCHING FOR SPEECH

Studies on the cross-modal perception of visual objects have not yet systematically taken apart the stimulus to determine the basis of the cross-modal effect. One difficulty in doing so with objects is that there is no well-developed theory that fully specifies the "distinctive features" of objects. Here speech has an advantage. "Distinctive Feature Theory" (Jakobson et al., 1969) isolates the component features of speech objects, and many diverse experiments attest to the psychological reality of these features.

The distinctive features of vowels are defined primarily in terms of spectral (frequency) information rather than in terms of temporal or amplitude (loudness) information. As such, they are related to the locations of the formant frequencies. Our first step in identifying the effective stimulus was to verify a fact that we inferred from our previous work but had not directly tested. The fact we inferred was as follows: Because we had matched the auditory and visual vowel stimuli on all temporal and amplitude parameters, we inferred that infants' matches must be based on the *spectral properties* of the auditory signal, that is, the pattern of frequency differences that signaled the /a/ versus /i/ vowels. We thus hypothesized that if we altered this spectral information somehow, by taking the formant frequencies out of the sounds for example, infants could no longer succeed on the cross-modal task. In our first study exploring the effective stimulus governing the effect (Kuhl & Meltzoff, 1982, 1984a) we set out to test this hypothesis directly.

The /a/ and /i/ vowels used in the first study were altered to remove the spectral information that distinguished the sets of vowels (their formant frequencies) while leaving whatever temporal and amplitude information that remained. Using computer analysis techniques, we extracted the time-intensity curves (the amplitude envelopes) of the vowels and their precise durations. Then we computer synthesized pure-tone stimuli with a frequency of 200 Hz (the average value of the female talker's fundamental frequency), one for each of the original 20 /a/ and 20 /i/ vowels. Each pure-tone stimulus exactly followed the amplitude envelope of its speech-stimulus original. Thus, we created 40 new auditory stimuli, devoid of spectral information but matched in every detail to the temporal and amplitude cues that remained in the original stimuli.

These pure-tone stimuli could not be identified as /a/ or /i/, yet when they were played while looking at the faces, the resulting display was quite engaging. Because the temporal properties of the tones matched the original vowels, the tones became louder as the mouths grew wider and softer as the mouths drew to a close. Thus, if infants in our task could discover a match between auditory and visual stimuli on time-intensity cues alone, they should succeed. If, however, the spectral properties of the vowels were necessary, the results should drop to chance. Arguing that the temporal-envelope properties of the stimuli were insufficient for success in our original experiment, we favored the spectral hypothesis.

The results were in support of the spectral hypothesis. In the absence of spectral information, infants' cross-modal performance dropped to chance. The mean percentage of fixation time to the matched stimulus was 54.6% ($p > .50$), with only 17 of the 32 infants demonstrating the effect. Inspection of the overall visual fixation data revealed that infants spent just as long looking at the faces in this experiment as they had in the previous three experiments in which they heard speech sounds rather than tones, so it was not as though they found these stimuli uninteresting. However, a match between the tones and the faces could not be detected. We had shown, then, that the temporal envelope of the vowel stimuli used in our experiment was not sufficient to produce the cross-modal effect. Some aspect of the spectral information was necessary, as we had hypothesized.

But what aspect of the spectral information was needed? Did the information in the auditory stimulus have to be sufficient to identify it as an /a/ or an /i/ in order for the match to be detected? Or would a simpler spectral property be sufficient?

Recall that distinctive features are the components that make up speech segments. One feature that distinguishes /a/ and /i/ vowels is the *acute-grave* distinction (Jakobson et al., 1969). This refers to the location of the main concentration of energy in the vowel. When the vowel's formants are low in frequency, its center of gravity is lower than if the vowel's formants are high in frequency. A low center of gravity produces a *grave* sound, but a high center of gravity produces an *acute* sound. The vowel /a/ is grave whereas /i/ is acute.

A second spectral feature that distinguishes /a/ and /i/ is the *compact-diffuse* distinction. This refers to the relative spread of energy across the formant frequencies. The energy can be described either as *diffuse*, in which case the formant frequencies are fairly well separated on the spectrum, or *compact*, in which the energy in the formants is more concentrated. The vowel /a/ is compact with its formants spaced closely together, and the vowel /i/ is diffuse.

Of the two features distinguishing /a/ and /i/, the grave-acute feature is very prominent perceptually. It is easy to judge that the vowel /i/ is high in pitch, whereas the vowel /a/ is low. Experiments in the adult speech perception literature have verified the fact that listeners can identify a predominant pitch in the vowels (Chiba & Kajiyama, 1941; Fant, 1973; Farnsworth, 1937).

Our own recent work verifies the fact that adults can match auditorially presented vowel sounds to auditorially presented pure tones. More important, work in our lab has now extended these adult experiments to tests involving the pitch feature in a cross-modal speech perception task. That is, we examined whether adults could match auditorially presented pure tones to visually presented vowels. We will briefly describe the results of both our auditory–auditory (A–A) tests and our auditory–visual (A–V) tests with adults (Kuhl, Merrick, & Meltzoff, in preparation).

Two A–A matching studies were conducted with adults. In both, listeners were asked to adjust the frequency of a pure tone until it matched, as closely as possible, a reference vowel. In one case the reference vowels were the actual /a/ and /i/ vowels used in our original infant experiment. In the second, the reference vowels were not played for the subjects. We simply told them to "imagine" them. Regardless of the test condition, real or imagined, adults adjusted the tone to a high frequency for the vowel /i/, usually a frequency above 2000 Hz. For the vowel /i/ in both test conditions they adjusted the tone to a mid-frequency, usually between 750 Hz and 1200 Hz. In other words, our studies on adults' perception of the pitch of vowels showed that they can match auditorially presented vowels to auditorially presented pure tones, thus replicating work done previously on the pitch of vowels. It also extended these findings to imagined stimuli.

Our next question was whether the ability to relate pure tones to vowels by adults could be replicated cross-modally. This had never been tested before. This study involved auditory–visual matches between pure tones presented auditorially and the visually presented articulatory movements. The study was a replication of our cross-modal test using infants. The adults watched the same /a/ and /i/ faces, but instead of listening to vowel sounds, they listened to one of nine pure tones ranging from a very low to a very high frequency: 125 Hz, 250 Hz, 500 Hz, 750 Hz, 1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, and 4000 Hz.

As in our infant tests, the adults sat facing the three-sided cubicle. They were first familiarized with the two visual stimuli in counterbalanced order. Then both faces were shown, and one of the nine pure tones was presented in synchrony with the faces. After a 2-min test period the adult was asked which of the two faces was a better match to the tone. Eight adults were tested in each of the nine frequency conditions, for a total of 72 subjects. These adults were not the same ones who had been tested in our A–A pure-tone vowel tests.

Because our own and others' data showed that adults' auditory judgments of the mid-frequency tones (primarily 750 Hz to 1500 Hz) were associated with the /a/ vowel and judgments of the high-frequency tones (2000 Hz to 4000 Hz) were associated with /i/, we were most interested in adults' cross-modal judgments of these frequencies. (We had included the very low frequencies 125 Hz–500 Hz because we were also interested in adults' auditory and cross-modal judgments of /u/, a very low vowel. These three frequencies are not considered here.)

The resulting data showed that there was a cross-modal relation between pure tones of certain frequencies and vowels presented visually. For the adults tested with the 750-, 1000-, and 1500-Hz pure-tone stimuli, 19 of the 24 judged them a better match to the /a/ face as opposed to the /i/ face ($p <$ .001 by binomial test). Coversely, of the adults tested with the 2000-, 3000-, and 4000-Hz pure-tone stimuli, 21 of 24 judged them a better match to the /i/ face as opposed to the /a/ face ($p < .001$ by binomial test).

Thus, a clear pattern emerged for adults. Adults matched /a/ vowels to mid-frequency pure tones and /i/ vowels to high-frequency pure tones. This was a robust phenomenon and held true regardless of whether the vowel stimulus was presented visually as silently articulating faces, auditorially as speech signals, or even as "imagined" auditory input. Thus, for adults, a stimulus that instantiates a distinctive feature of vowels, but does not precisely identify the vowel, is a sufficient stimulus for eliciting the cross-modal effect. Remarkably, the stimulus itself need not be perceived as speech, for these pure-tone stimuli are not.

Our next step was to pose the same question to infants. We do not know whether infants relate auditory /i/ vowels to high tones and /a/ vowels to lower ones, but speech categorization tests (e.g., Kuhl, 1983, 1985) would be a good way to examine this, and we are presently doing so. However, our immediate interest was in the cross-modal effect using pure tones with infants. If infants perform as adults do, then a single distinctive feature of the vowels, their predominant pitch, would be a sufficient stimulus for infants in our cross-modal test. This would mean that even for infants the information in the auditory stimulus need not be sufficient to be identified as a vowel, or even as speech.

The infant cross-modal test was run exactly as before in an attempt to see how tones of different frequencies affected infants' visual preferences. Infants viewed the original /a/ and /i/ faces, but instead of listening to /a/ or /i/ vowels, they heard one of the nine pure tones, just as adults had done. Sixteen 18- to 20-week-old infants were tested at each of the nine frequencies, for a total of 144 infants. Left-right facial orientation, familiarization order, and sex were counterbalanced.

Unlike the tests on adults, the results on infants showed no cross-modal effect. Regardless of the frequency of the tone, infants looked longer at the /a/ face than at the /i/ face; at only one frequency (2000 Hz) did /i/ looking exceed /a/ looking, and there only slightly and non-significantly (Kuhl, Merrick, & Meltzoff, in preparation). Apparently, in the absence of a cross-modal effect, infants simply preferred the /a/ face.

The results showed that infants did not match these nonspeech auditory stimuli to visual faces producing speech sounds. This suggests that infants were not basing their success on the original cross-modal experiment on a single distinctive feature of vowels, such as their predominant pitch. Perhaps infants will demonstrate the effect only when the stimulus is speech and the information in the speech sound is sufficient to identify it as one of the visually presented vowels. In other words, infants may require the "whole stimulus" to detect cross-modal matches for speech. They may need to identify/categorize both the auditory and the visual input as the same phonetic unit; a single component feature of the stimulus may not do. We are currently pursuing this hypothesis.

With this series of experiments on the basis of the cross-modal speech effect we had discovered several important facts about infants' intermodal representation of speech. First, our initial experiments demonstrated that very young infants can detect a cross-modal correspondence between speech sounds presented auditorially and the sight of a person producing those same sounds. By 18 weeks of age, infants appear to know that /a/ sounds emanate from lips that are wide open, that /i/ sounds emanate from retracted lips, and /u/ sounds from protruded, pursed lips. That infants equate the auditory and visual concomitants of speech at such an early age is indeed noteworthy for theory.

Second, the detection of a match depends on the spectral rather than the temporal properties of the auditory stimulus. This finding is essential to the argument that it is speech itself (or the component features of speech), rather than some general perceptual property such as timing, that is intermodally represented. Third, our studies aimed at identifying the spectral properties governing detection of the match showed that a distinctive feature such as pitch, when isolated in a nonspeech auditory signal, was not sufficient to reproduce the matching effect in infants. Infants did not match nonspeech sounds to faces producing speech. This finding suggests the possibility that the entire speech sound itself might be necessary for infants to detect a match between auditory and visual speech. Interestingly, our studies showed that the pitch feature was sufficient for adults, and it will be relevant to theory to track the developmental time course of this change in the sufficient stimulus.

We turn now to another related finding emerging from the series of studies. Although seemingly different on the surface, we believe it adds important converging evidence concerning infants' intermodal organization of speech. The evidence concerns infants abilities to imitate the speech signals presented to them. Vocal imitation and cross-modal speech perception are intimately related, as we argue here.

VOCAL IMITATION

Thus far in discussing the intermodal organization of speech we have focused on the perception of speech through different sensory modalities—auditory, visual, and tactile. Now we turn to speech production for further clues about the intermodal organization of speech.

As adults, we can produce a specific auditory target, such as a vowel, on the first try. It is not a trial-and-error process. An auditory signal can be directly related to the motor commands necessary to produce that signal, because adults have rules that dictate the "mapping" between articulation and audition. This mapping is quite sophisticated. Experiments show that if an adult speaker is suddenly thwarted in the act of producing a sound by the introduction of a sudden load imposed on his lip or jaw, compensation is essentially immediate (Abbs & Gracco, 1984). The adjustment can occur on the very first laryngeal vibration, prior to the time the speaker has heard anything. Such rapid motor adjustments suggest a highly sophisticated and flexible set of rules relating articulatory movements to an intended target sound.

How do auditory-articulatory mapping rules develop? Evidence suggests that at least one important mechanism for learning them is vocal imitation.

Among mammals, man is the only animal who gives evidence of "vocal learning," that is, acquiring the culture's vocal repertoire by hearing it and mimicking it. In other words, we learn to speak by listening and imitating. Humans share this ability with a few select avian species, the passerine birds (Marler, 1973), who learn their conspecific song only if they are auditorially exposed to it during a "critical period" early in life (Nottebohm, 1975).

In humans, as in birds, early auditory experience is critical to the development of the vocal repertoire. Deaf infants, even with painstaking instruction, do not learn to speak normally (Davis & Silverman, 1947). The potency of vocal learning is also evidenced in normal-hearing speakers. One needs only to listen to the vocal patterns of foreign speakers to realize that early auditory exposure to a specific language pattern puts an indelible marker on one's speech patterns. Foreigners try to rid themselves of their phonetic errors ("flied lice" for "fried rice") and their foreign accents, and large sums of money are spent in speech classes trying to undo the effects of early auditory experience on speech patterns. It is notoriously difficult to do so.

At some point, then, young infants must become very adept at mimicking the speech patterns they hear others produce. But when are infants capable of imitating the sounds they hear? Some relevant data can be adduced from the earliest age at which infants from different language environments produce phonetic units that are unique to their own native language. The data show that the earliest sounds produced by young infants are characteristic of many different languages (Oller, 1981; Stark, 1980). But by the time first words emerge, infants will produce sounds that are typical of their language, but are rare in other languages. Moreover, these infants will have adapted the accent or "tone" of the language—its cadence, rhythm, and tempo, as well as its characteristic intensity and intonation contours (de Boysson-Bardies, Sagart, & Durand, 1984). Chinese Mandarin toddlers will have begun to sound distinctly Chinese and African Xhosa toddlers will sound distinctly African. They will already have begun to show a mastery of the mother tongue. Some investigators have pushed the time at which one can hear these differences much earlier. Weir (1962) claimed to be able to judge the nationality of 6-month-old infants based on the infants' vocalizations. We can say, then, that at some point prior to the onset of language and perhaps before 6 months of age, infants have acquired enough information about the phonetic units and prosody of their native language to produce it in a way that is characteristic of their culture.

What guides infant imitation of the sound pattern of the language? The infant cannot see his own vocal tract, so his speech productions are not visually guided. Even when the infant can view another talker in front of him, vocal imitation requires more than the reproduction of what he sees the other talker do, because the infant cannot see the critical movements of the tongue, velum, and larynx that are responsible for producing sound. Moreover, although infants may have proprioceptive feedback from movements of their lips and tongues, movement of the velum provides much less proprioceptive feedback. And laryngeal manipulations, such as those required to change the pitch of the voice for intonation contours, provide almost no proprioceptive feedback at all. The critical ingredient for vocal imitation is audition.

Infants vocally imitate by doing two things. They compare the auditory results of their own vocal maneuvers with the auditory results of that produced by someone else. That is, they make an intramodal auditory-auditory match; and second, they develop a set of auditory–articulatory mapping rules that allow them to make adjustments in production to get closer to the auditory target they wish to achieve. Eventually, talkers need not rely on an after-the-fact auditory–auditory comparison, but directly link auditory targets to the motor movements necessary to achieve specific targets. Presumably, this information need not be acquired for each individual sound. Once the auditory–articulatory relation is mapped it can be used to produce a novel sound correctly on the first try. The development of vocal imitation provides a window through which to observe infants' acquisition of these auditory–articulatory mapping rules.

Methodological Issues Involved in Studying Vocal Imitation

There are two sets of concerns that need to be addressed in studies of infant vocal imitation. One is interpretive, and relates to the anatomical differences between infants' and adults' vocal tracts. The second is methodological, and relates to the design of experiments attempting to provide evidence of vocal imitation in infants.

Developmental studies of vocal imitation need to take into account the anatomical development of the speech mechanism. A comparison of the infant's and adult's vocal tracts shows that they are quite different (Fig. 7.2). The infant's vocal tract is not simply a miniature version of the adult's. In fact, it resembles that of an adult chimpanzee more than it does an adult human (Lieberman, Crelin, & Klatt, 1972).

The effects of these differences in anatomy are substantial (Kent & Murray, 1982). Three stand out. First, some of the sounds of human speech, like a perfect version of the vowel /i/, simply cannot be produced by the very young infant's vocal tract. Second, the arrangement of the young infants' laryngeal and velopharyngeal structures makes infants obligate nasal breathers

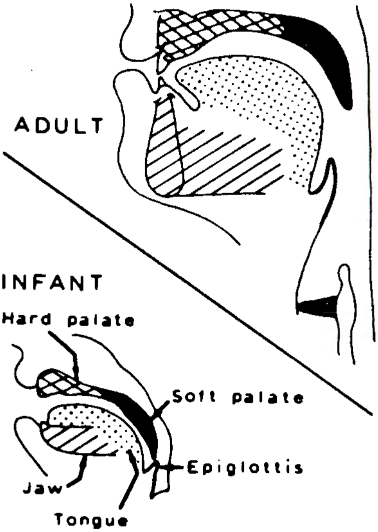

FIG. 7.2 Anatomical differences between infants' and adults' vocal tracts (from Kent & Murray, 1982).

and vocalizers. The infant's vocal tract changes dramatically between birth and 6 months so that, for example, obligate nasalization drops out between 4 and 6 months when the separation of laryngeal and velopharyngeal structures occurs (Sasaki, Levine, Laitman, & Crelin, 1977). Third, the infant's vocal tract is smaller and the vocal folds are shorter than those of the adult, and this makes the infant's fundamental frequency and formant frequencies higher than the adult's.

At least four important consequences follow from these anatomical facts. First, a young infant's failure to match an adult's production of /i/ may not indicate an inability to imitate, but merely the constraints of his vocal apparatus. Second, because of the obligate nasalization, it is important that we neither mistake the nasal resonance for a low formant frequency (and thereby falsely attribute imitation of a sound that has such a component), nor mistakenly ignore imitation of a sound just because it is accompanied by nasalization. Third, because the infant's vocal tract changes rapidly during the first 6 months of life, changes in vocalizations need not be attributed to the infant's adaptation to his linguistic environment, but simply to changes in the anatomy of the oral cavity. Fourth, because the infant's laryngeal mechanism is smaller than the adult's, pitch and formant structure imitation should be sought in the infants' matching of the pattern of pitch contour or of formant structure rather than a match of the absolute frequencies of adult's speech.

Another set of considerations concerning vocal imitation relates to the design of the experiment. As pointed out by Meltzoff and Moore (1977, 1983a, 1983b), there are problems with simply observing adult–infant interaction in a natural setting and assuming that infants' matching of an adult behavior is "imitation." Rather, imitation is demonstrated when two criteria are met (Meltzoff & Moore, 1983b): First, the infant's vocalization has to be elicited by a model's vocalization, and second, the structure or organization of the infant's response has to be shown to be influenced by the structure or organization of the model's vocalization and has to match it in some way.

The first of these stipulates that it is the model's vocalization that elicits a matching response and not vice versa. In laboratory experiments on vocal imitation this can be guaranteed by presenting the stimulus by machine. Natural observations of vocal exchanges between mothers and their infants are usually subject to the question "who is imitating whom?"

The second criterion, ensuring that the structure of the infant's response is influenced by the structure of the model's behavior, is solved by using the "cross-target" design developed for the study of gestural imitation (Meltzoff & Moore, 1977, 1983b). The design helps to control for the spontaneous, nonimitative production of the target response. For example, in order to demonstrate the existence of vocal imitation, it is not sufficient to show that an infant produces an /a/ immediately after an adult produces an /a/. Our work and that of others suggest that when an adult sits in front of 4-month-olds and vocalizes, the infants are likely to vocalize in return. If the vowel typically produced by infants at this age is /a/, then the infant's reaction to an adult's /a/ stimulus is not necessarily an imitated response. It might simply be the infant's typical vocalization. One has to show that the /a/ is not the sound generally produced when the infant becomes aroused, or when an adult sits in front of the infant producing speech.

The cross-target design developed by Meltzoff and Moore addressed this problem by examining *differential* responding. In the vocal imitation case, the design would call for an examination of the number of /a/ vocalizations that occurred in response to an adult's production of /a/, as opposed to the number of /a/ vocalizations that occurred in response to an adult's production of /i/. Imitation can be inferred if the infant produces more /a/'s to the adult's /a/ than to /i/ and conversely produces more /i/'s to the model's /i/ than to /a/. Such differential responding cannot simply be attributed to general arousal in a social response because the infant is presented with auditory signals in both cases and he responds in a differential fashion to each.

A further issue to contend with is distinguishing conditioned responses from imitated ones. If an infant is simply trained to produce a sound in response to an adult sound, it canot be claimed that the infant truly imitated. For example, suppose an adult uses operant conditioning methods to train an infant to produce an /a/ after an adult produces /a/, and to produce /i/ after the adult produces /i/. If such training techniques were used, we could not claim that the infant imitated the adult, because the infant was simply trained to do something. It might have been just as easy to condition the infant to do the reverse, that is to produce /a/ when he heard /i/, and to produce /i/ when he heard /a/. Imitation connotes that the infant produces a spontaneous match, not that he has been trained to produce a response on cue (Meltzoff & Moore, 1983b).

Finally, there are certain measurement problems inherent in vocal imitation. The measurement of sound is particularly complex, whether performed by people or machines. Phonetically trained listeners are needed to categorize young infants' vocalizations accurately. Even then, it is sometimes necessary to use a "forced-choice" technique. This technique involves trained observers who judge which of two target vocalizations the infants have been exposed to on the basis of the sounds they produce. Thus, the observer would be instructed to listen for sounds that were more like one target than the other. If the observers' judgments predict the target significantly more often than chance (50%), then evidence of imitation has been obtained. This technique is needed in cases in which infants are too young to produce an exact match and yet are clearly approximating the target sound that was presented. More sophisticated scoring would, of course, involve instrumental analysis of infants' vocalizations. We believe that studies on vocal imitation should include both perceptual and instrumental measurement. It is important to know both how a phonetically trained person identifies a sound and about that sound's exact acoustic description. In some instances, for example vowels, no instrumentation has yet been invented to provide an absolute categorization of them. For other aspects of speech, such as fundamental frequency contours, duration, and loudness, machine analysis may provide an excellent measurement of the infant's response.

There are several decisions involved in the measurement phase of an experiment on infant vocal imitation, and the approach of choice depends on the nature of the question. Whenever possible, an absolute identification of the infant's production is most desirable. Machines can specify whether an intonation contour was rising or falling, so an absolute identification is possible in those cases. But machines cannot measure whether a particular phonetic unit was produced. An adult scorer can either make the absolute judgment or judge whether the infant's sounds are "more like" one sound than another in a forced-choice task. The major point here is that it takes the proper instruments (computers and fairly elaborate software) and/or phonetic training to do these studies carefully.

From Piaget on, reports have appeared that are highly suggestive of vocal imitation of at least one prosodic aspect of speech, its pitch (Kessen, Levine, & Wendrich, 1979; Lieberman, 1984; Papousek & Papousek, 1981; Piaget, 1962); however, all but one of these studies (Kessen et al., 1979) involved natural interactions between adults and infants, and as such are subject to one or more of the problems previously raised. The Kessen et al. study tested infants in multiple sessions over several months, giving them repeated practice and feedback, so the issue of training is unresolved in the study.

With these issues in mind we sought evidence of vocal imitation in our own experiments on infants' cross-modal perception of speech. The cross-modal studies provided a controlled setting in which to study vocal imitation. Recall our experimental set-up. Infants sit in an infant seat facing a three-sided cubicle. They view a film of a female talker who produces vowel sounds. Half of the infants are presented with one auditory stimulus and the other half are presented with a different auditory stimulus. The stimuli are totally controlled, both visually and auditorially. There are no human interactions with the infant during the test, and thus no chance for the spurious shaping and/or conditioning of a response. The room is a soundproof chamber, and a studio-quality microphone is suspended above the infant to obtain clear recordings that can be perceptually or instrumentally analyzed. Finally, the stimulus on film being presented to the infant occurs once every 3 sec, with an interstimulus interval of about 2 sec. This is ideal for encouraging "turn-taking" on the part of the infant. We have found that infants in this setting are calm and highly engaged by the face-voice stimuli. They often listen for a while, smile at the faces, and then start talking back. Our question was: Do infants' speech vocalizations match those they hear?

Our most recent analyses allow us to make two claims about vocal imitation in 4-month-old infants. The first relates to the effective elicitors of vocalizations in young infants. To test this we compared the effectiveness of speech sounds as opposed to nonspeech sounds. The second involves the differential imitation of the phonetic units themselves. To test this we examined infants' matching responses to the /a/ versus /i/ stimuli.

Imitation of Speech Versus Nonspeech Stimuli

Recall that in three of our cross-modal studies infants heard speech sounds. They heard one of the three point vowels, /a/, /i/, or /u/. In the fourth and fifth studies infants heard nonspeech stimuli. These stimuli consisted of 1 of 10 pure tones. In the nonspeech study conducted to test infants' use of temporal envelope cues, the tone was a 200 Hz signal. In the other more elaborate nonspeech study, 9 tones were used, varying from 125 Hz to 4000 Hz. In neither of these nonspeech studies could any of the sounds be identified as speech. They were pure tones, the simplest auditory stimulus than can be produced.

The question we were interested in was this: What happens when infants listen to speech as opposed to nonspeech sounds? Is human speech a more effective elicitor of vocalization than nonspeech?

Our original study suggested that this was the case. Kuhl and Meltzoff (1982) reported that the infants tested in the /a-i/ speech experiment (Experiment 1) versus those tested in the tone experiment (Experiment 2) produced a differential amount of vocalization. Infants who heard speech produced vocalizations typical of speech. That is, they produced vowel sounds of the type referred to as "cooing." The infants who were presented with the nonspeech tone did not produce speech-like vocalizations. They had watched the same faces and heard sounds of the same duration and intensity. They were given just as long to reply. But they did not produce speech. In the 1982 paper we reported that 10 of the 32 infants hearing speech produced speech-like vocalizations whereas only a single infant hearing nonspeech produced speech-like vocalizations ($p < .01$).

We can now extend these results to a much larger sample. To date we have fully analyzed the vocalizations of all of the infants who participated in the two /a-i/ studies for a total of 64 infants. In addition, we have analyzed the vocalizations of the first half (72 infants) of the 144 infants tested in the pure-tone study (Kuhl, Merrick, & Meltzoff, in preparation). Our perceptual scoring system is quite elaborate and includes sufficient categories for both speech and nonspeech so that they can be described fully. The scoring is done by a trained phonetician.

The results strongly show the superiority of human speech in eliciting infant vocalizations. Infants listening to speech produce speech, whereas infants listening to tones do not. Of the 64 infants listening to speech in our sample, 40 produced vocalizations that are typical of speech, whereas only 5 of the 72 infants hearing nonspeech produced sounds of this type ($p < .001$). Infants listening to nonspeech do not tend to produce speech-like vocalizations; instead, they squeal, gurgle, or grunt. The point is, they do not produce speech. Infants talk to faces that are talking to them.

We can say, then, that certain kinds of auditory signals, namely speech sounds, will encourage infants to produce speech sounds of their own, and the presentation of nonspeech sounds induces the infant's own version of nonspeech, squeals, and squeaks.

Imitation of Speech Sounds by 4-Month-Old Infants

We consider here the degree of match between the eliciting speech stimulus and the infant's response. We found evidence for the imitation of two aspects of the speech signal—its prosodic characteristics and its phonetic identity.

In our initial report, we described infants' imitation of the prosodic characteristics of the signal. The pitch contour of the adult model's vowels as well as an infant's response is shown in Fig. 7.3. Instrumental analysis showed that the infant produced an almost perfect match to the adult female's

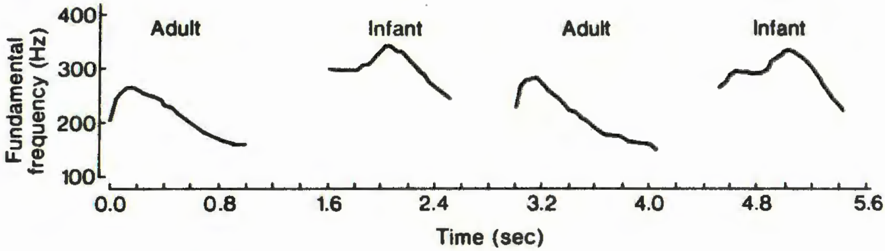

FIG. 7.3. Infant vocal imitation of the adult's production of a rise-fall pitch contour (redrawn from Kuhl and Meltzoff, 1982).

rise–fall pattern of intonation. Although the infant has shorter vocal folds and therefore produces a higher fundamental frequency (note that the infant's pitch is higher in the figure), the pitch "pattern" of a rapid rise in frequency followed by a more gradual fall in frequency duplicates that of the adult. The infant's response also matched the adult's in duration. Both were about 1 sec long. Because vocalizations with this rise–fall pattern and of this long duration are not common in the utterances of 4-month-olds, it was highly suggestive of vocal imitation. We are now conducting experiments in which the pitch pattern and duration of the auditory signal are systematically varied.

A more definitive test of the young infant's ability to imitate relates to the phonetic segments of speech. Half of the infants in our experiments had heard /a/ vowels and the other half had heard /i/ vowels. This allows a good test of the differential imitation of speech sounds.

All of the vowel-like vocalizations produced by the infants in the /a–i/ studies were analyzed. Vowel-like sounds were defined on the basis of the articulatory characteristics typical of vowel sounds. Vowels had to be produced with an open mouth, rather than one that was closed. They had to have a minimum duration of 500 msec. They had to be "voiced," that is, vocalized with normal laryngeal vibration, and could not be aspirated or "voiceless" sounds. They could not be produced on an inhalatory breath. Vocalizations that occurred while the infant's hand was in his mouth could not be reliably scored and were excluded. Consonant-like vocalizations were also scored, but they occurred rarely and were always accompanied by vowel-like sounds.

Once identified, the sounds were scored by a coder who remained uninformed about the eliciting stimulus. The perceptual scoring was done by having a trained phonetician listen to each infant's productions and judge whether they were more "/i/-like" or "/a/-like." Infants at this age cannot produce perfect /i/ vowels, due to anatomical restrictions on their vocal tracts described earlier. They can, however, produce other high front vowels such as /I/ or /ε/. Similarly, a perfect /a/ is rare in the vocalizations of the

4-month-old, but similar central vowels, such as /æ/ and /ʌ/, are producible by infants at this age. Thus, the judgment made by the observer was a forced-choice one concerning whether the infant's vocalizations were more /a/-like or more/i/-like.

Once infants' vocalizations were scored in this way, we could ask if judges could reliably predict whether infants had been exposed to /a/ as opposed to /i/, based on their vocalizations. If judges can do so with greater than chance (50%) accuracy, then there is evidence for vocal imitation. The results confirmed this prediction. Infants produced /a/-like vowels when listening to /a/- and /i/-like vowels when listening to /i/, allowing the judges to predict accurately in 90% of the instances the vowel heard by the infant. These results were highly significant ($p < .01$).

We are now involved in the instrumental analysis of these sounds. This is a long painstaking process, but the computer analyses of infants' vocalizations done to date confirm the fact that infants hearing /a/ produce sounds with acoustic characteristics that are more similar to /a/ than to /i/. Similarly, when listening to /i/ they produce sounds with acoustic characteristics that are more similar to /i/ than they are to /a/. Using distinctive feature theory to guide our instrumental analyses, we measured the acute-grave feature and the compact-diffuse feature in the infants' vowel productions. This required extracting the first and second formant frequencies from each infant vocalization, and then, using the formulas devised by Fant (1973), calculating the degree of "graveness" and of "compactness" for each production. The results demonstrated that infants' vocal responses to /a/ were significantly more grave, that is, they had a lower center of gravity, than their responses to /i/. Similarly, their response to /a/ were significantly more compact, that is, they had formants spaced more closely together, than their responses to /i/. Taken together, the two analyses provide evidence that 4-month-old infants are attempting to imitate the phonetic segments of speech.

To summarize, our studies have revealed two important characteristics of infants' imitation of vocal signals. First, infants differentially duplicate speech versus nonspeech sounds. When listening to speech, infants also produce speech. In contrast, when infants listen to nonspeech sounds they squeal and squeak but tend not to produce speech. Moreover, infants listening to speech imitated the auditory signals they heard. Infants who heard /a/ vowels produced /a/-like sounds, whereas those who heard /i/ vowels produced /i/-like sounds. Given both results, there is good evidence of vocal imitation in infants as young as 4 months of age. There is no possibility of the adult imitating the infant instead of the reverse (a typical concern in more naturalistic vocal-exchange studies) because our "adult" was a machine. To our knowledge this is among the first such evidence of vocal imitation in infants this young under controlled laboratory conditions.

SUMMARY AND CONCLUSIONS

This essay began by considering speech an object of perception, and asking whether the objects of speech, phonetic segments, were intermodally represented in infants. Two lines of evidence were then adduced in favor of this proposition: (a) infants' abilities to recognize correspondences between the auditory and visual products of articulation in cross-modal speech perception experiments, and (b) infants' abilities to imitate speech signals. Both these skills strongly suggest an intermodal organization of speech in early infancy.

The experiments on auditory-visual speech perception provided strong evidence that by 4 months of age they recognize that /a/ sounds go with mouths that are open wide, /i/ sounds with mouths that have retracted lips, and /u/ sounds with mouths whose lips are protruded and pursed. That infants can detect these auditory-visual equivalents at such a young age is surprising and is not predicted or explained by any existing theory of speech perception. The research also showed that a pure tone embodying a feature of the vowels, namely its center of gravity, is not sufficient to produce the cross-modal effect in infants. Infants could not match pure tones to articulatory movements. This is in contrast to adults, who could detect matches between auditorially presented vowel sounds and pure tones, as well as between visually presented vowels and pure tones. Thus, adults can use a component feature of the vowels in these cross-modal tasks, but infants cannot. Infants may require the whole stimulus to detect the equivalence between speech sounds and the faces that produce them. For infants, then, we can say that speech sounds themselves, but perhaps not their component features, are represented intermodally.

A second line of evidence was brought forth to examine the hypothesis that the infant's organization of speech is intermodal in nature. In this case, we examined vocal imitation, the infant's ability to duplicate the sound produced by another. Two facts emerge. First, we discussed data to suggest that speech sounds are more effective elicitors of speech-like vocalizations than nonspeech signals are for infants. Infants hearing speech responded in kind. Infants hearing nonspeech—even though the signals were pure tones that embodied a prominent feature of vowel sounds—did not tend to produce speech. Second, infants who heard speech sounds produced vowel sounds whose spectral characteristics matched those of the vowels they heard. Infants who heard /a/ produced sounds whose perceptual and acoustic characteristics resembled /a/, whereas those who heard /i/ produced sounds whose perceptual and acoustic characteristics resembled /i/. These new findings on infant vocal imitation demonstrate differential imitation of speech in a controlled laboratory environment.

These data on cross-modal speech perception and vocal imitation provide converging evidence for the intermodal organization of speech in young infants. The representation of speech in infants is such that an auditory speech signal can drive two other systems. The data show that an auditory signal drives infants' exploratory looking behavior, causing them to seek out a visual signal that portrays to the eye an event that is phonetically equivalent to the one they hear. The auditory signal also drives infants' motor behavior, prompting them to produce an articulatory maneuver that will result in an event that is phonetically equivalent (to the best of their ability) to the one they hear.

Thus, by 18-20 weeks of age infants relate the speech sounds they hear to ones they see being produced. Moreover, they relate the sounds they hear to the motor movements necessary to produce them, and initiate those movements to create the event themselves. It seems likely that both these phenomena—cross-modal perception and vocal imitation—are linked by some common representation of speech. The notion that the auditory, visual, and motor systems for speech are linked in infants is further reinforced by the finding that for both the auditory-visual (cross-modal) and the auditory-motor (imitation) tasks, a pure-tone stimulus that embodies a feature of the vowel is not a sufficient stimulus to produce matching, whereas the whole stimulus is. This in turn suggests that the infant's representation of speech is specified in units at least as large as the whole phonetic segment.

In summary, it appears that speech is an intermodal object of perception for infants. It can be perceived by eye as well as by ear, and when doing both, infants are prompted to reproduce it themselves. By studying the origins and development of infants' ability for cross-modal perception and vocal imitation we can gain knowledge that contributes not only to theories of speech and language development, but also to more general models of perception and its development.

ACKNOWLEDGMENTS

The preparation of this chapter and the work reported here were supported by grants from NIH (HD18286 and HD22514) and NSF (BNS-8316318). We are indebted to many people for their help. We especially wish to thank Karen Merrick, Kerry Green, and Craig Harris.

REFERENCES

Abbs, J. H., & Gracco, V. L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology, 51,* 705–723.

Blumstein, S. A., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition, 10,* 25–32.

Chiba, T., & Kajiyama, M. (1941). *The vowel—Its nature and structure.* Tokyo: Kaisekan Publishing.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 269–321). New York: Wiley.

Davis, H., & Silverman, S. R. (1947). *Hearing and deafness.* New York: Holt, Rinehart & Winston.

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,* 1–5.

Deland, F. (1920). Ponce de Leon and Bonet. *Volta Review, 22,* 391–421.

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

Farnsworth, P. (1937). An approach to the study of vocal resonance. *Journal of the Acoustical Society of America, 9,* 152–156.

Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to connected discourse perception by speechreaders. *Journal of the Acoustical Society of America, 77,* 671–677.

Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1986). The transmission of prosodic information via an elecrotactile speedreading aid. *Ear and Hearing, 7,* 243–251.

Green, K. P., & Kuhl, P. K. (1986). The role of visual information from a talker's face in the processing of place and manner features in speech. *Journal of the Acoustical Society of America, 80*(Suppl. 1), S63.

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

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

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America, 72,* 353–365.

Kessen, W., Levine, J., & Wendrich, K. A. (1979). The imitation of pitch in infants. *Infant Behavior and Development, 2,* 93–99.

Kirman, J. H. (1973). Tactile communication of speech: A review and analysis. *Psychological Bulletin, 80,* 54–74.

Klatt, D. H. (1975). Voice onset time, frication and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research, 18,* 686–706.

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. (1986a). Reflections on infants' perception and representation of speech. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 19–30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kuhl, P. K. (1986b). Infants' perception of speech: Constraints on characterizations of the initial state. In B. Lindblom & R. Zetterstrom (Eds.), *Percursors of early speech* (pp. 219–244). New York: Stockton Press.

Kuhl, P. K. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. B. Cohen (Eds.), *Handbook of infant perception* (pp. 274–382). New York: Academic Press.

Kuhl, P. K., Green, K. P., & Meltzoff, A. N., (in preparation). *Factors governing the integration of auditory and visual information in speech.*

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

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

Kuhl, P. K., & Meltzoff, A. N. (1984b). *Infants' representation of events: Studies in imitation, cross-modal perception, and categorization.* Paper presented at the Fourth International Conference on Infant Studies, New York.

Kuhl, P. K., & Meltzoff, A. N. (in preparation). *Replication and extension of cross-modal speech perception effects in infants.*

Kuhl, P. K., Merrick, K. M., & Meltzoff, A. N. (in preparation). *Infants' cross-modal perception of speech: Studies on the basis of the effect.*

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Lieberman, P. (1984). *The biology and evolution of language.* Cambridge, MA: Harvard University Press.

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. *American Anthropoligist, 74,* 287–307.

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

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

Marler, P. (1973). Constraints on learning: Development of bird song. In W. F. Norman (Ed.), *The Clarence M. Hicks Memorial Lectures for 1970* (pp. 69–83). Toronto: University of Toronto Press.

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

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

Meltzoff, A. N. (1985). The roots of social and cognitive development: Models of man's original nature. In T. M. Field & N. Fox (Eds.), *Social perception in infants* (pp. 1–30). Norwood, NJ: Ablex.

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

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

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

Meltzoff, A. N., & Moore, M. K. (1983b). The origins of imitation in infancy: Paradigm, phenomena, and theories. In L. P. Lipsitt (Ed.), *Advances in infancy research* (Vol. 2, pp. 265–301). Norwood, NJ: Ablex.

Miller, G. A., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics, 25,* 457–465.

Nottebohm, F. (1975). A zoologist's view of some language phenomena with particular emphasis on vocal learning. In E. H. Lenneberg & H. Lenneberg (Eds.), *Foundations of language development.* New York: Academic Press.

Oller, D. K. (1981). Infant vocalizations: Exploration and reflexivity. In R. E. Stark (Ed.), *Language behavior in infancy and early childhood* (pp. 85–100). New York: Elsevier.

Papousek, H., & Papousek, M. (1981). Musical elements in the infant's vocalization: Their significance for communication, cognition, and creativity. In L. P. Lipsitt & C. K. Rovee-Collie (Eds.), *Advances in infancy research* (Vol. 1, pp. 164–224). Norwood, NJ: Ablex.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Piaget, J. (1962). *Play, dreams, and imitation in childhood.* New York: W. W. Norton.

Sasaki, C. T., Levine, P. A., Laitman, J. T., & Crelin, E. S. (1977). Postnatal descent of the epiglottis in man. *Archives of Otolaryngology, 103,* 169-171.

Sparks, D. W., Kuhl, P. K., Edmonds, A. A., & 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.

Sparks, D. W., Ardell, L., 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.

Stark, R. (1980). Stages of development in the first year of life. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology, Vol. 1, Production* (pp. 73-92). New York: Academic Press.

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.

Walden, B., Prosek, R., Montgomery, A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130-145.

Weir, R. H. (1962). *Language in the crib.* The Hague: Mouton.