Journal of Experimental Psychology:
Human Perception and Performance
1991, Vol. 17, No. 3, 829-840

0096-1523/91/$3.00

Cross-Modal Speech Perception in Adults and Infants Using Nonspeech Auditory Stimuli

Patricia K. Kuhl and Karen A. Williams
Department of Speech and Hearing Sciences
University of Washington

Andrew N. Meltzoff
University of Washington

Adults and infants were tested for the capacity to detect correspondences between nonspeech sounds and real vowels. The /i/ and /a/ vowels were presented in 3 different ways: auditory speech, silent visual faces articulating the vowels, or mentally imagined vowels. The nonspeech sounds were either pure tones or 3-tone complexes that isolated a single feature of the vowel without allowing the vowel to be identified. Adults perceived an orderly relation between the nonspeech sounds and vowels. They matched high-pitched nonspeech sounds to /i/ vowels and low-pitched nonspeech sounds to /a/ vowels. In contrast, infants could not match nonspeech sounds to the visually presented vowels. Infants' detection of correspondence between auditory and visual speech appears to require the whole speech signal; with development, an isolated feature of the vowel is sufficient for detection of the cross-modal correspondence.

There is a long tradition in speech research of comparing the perception of speech and nonspeech events. A classic case was the attempt to determine whether nonspeech auditory signals were subject to the phenomenon of "categorical perception" (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Studies showed that categorical perception could be replicated with nonspeech sounds that mimicked the critical properties in speech. For example, in early studies speech phonemes such as /b/ and /p/ were demonstrated to be categorically perceived (Abramson & Lisker, 1970). In addition, later research revealed that nonspeech sounds which preserved the critical temporal feature that distinguished these phonemes, voice-onset time (VOT), were perceived categorically. Thus, categorical perception was demonstrated for temporally offset tones (Pisoni, 1977) and noise-buzz sequences (Miller, Wier, Pastore, Kelly, & Dooling, 1976). Infants demonstrated the categorical perception effect with both speech and nonspeech signals as well (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Jusczyk, Pisoni, Walley, & Murray, 1980; Pisoni, Carrell, & Gans, 1983).

These data on nonspeech contributed in two ways to the development of theory. First, the data were influential in debates concerning the mechanisms underlying categorical perception. Tests that used nonspeech stimuli examined whether a stimulus that had no linguistic relevance but that nonetheless captured a prominent feature of the speech stimulus was sufficient to produce the phenomenon; thus nonspeech data were used to argue that the phenomenon of categorical perception was not special to the processing of speech sounds (Cutting & Rosner, 1974; Diehl & Walsh, 1989; Kuhl, 1987a, 1987b; Pisoni, 1977; Pisoni et al., 1983; Repp, 1984; see Harnad, 1987, for review). Second, and of equal importance, nonspeech findings helped isolate the essential properties of speech signals (in the voicing case, a critical timing difference), and this aided the identification of the acoustic events underlying the perception of speech.

Speech versus nonspeech comparisons typically involve purely auditory events; they are unimodal tests. The problem posed in the current experiments was whether nonspeech stimuli are sufficient in a cross-modal test of speech perception, one in which speech was both auditorially and visually perceived. The novel aspects of these studies are that we used nonspeech stimuli, rather than real speech sounds, to examine whether both adults and infants could relate them to visual speech events and whether there might be developmental changes in this capacity.

A growing body of literature shows that adult perceivers take more than auditory information into account during speech perception. The movements of the articulators are visible concomitants of speech events, and research shows that this visual information strongly influences adults' perception of speech (Dodd & Campbell, 1987; Green & Kuhl, 1989, 1991; Massaro, 1987; Massaro & Cohen, 1983; McGurk & MacDonald, 1976; Summerfield, 1979, 1987).

It has also been demonstrated that the ability to detect auditory–visual correspondence for speech (as we do when we lipread) exists very early in life. Kuhl and Meltzoff (1982, 1984) tested 4-month-old infants in a lipreading task. Infants were shown two facial images side by side of a woman producing two vowels, /a/ as in *pop,* and /i/ as in *peep.* At the same time, they were auditorially presented with a single vowel (either /a/ or /i/). The hypothesis was that infants would look longer at the face that matched the sound they heard. The results supported this hypothesis. Infants who heard /a/ looked longer at the face that produced /a/, and infants who heard /i/ looked longer at the face that produced /i/ (Kuhl & Meltzoff, 1982, 1984). Thus, by 4 months of age, infants recognize correspondences between particular mouth movements and particular speech sounds. They recognize

This research was supported by National Institutes of Health Grant HD 18286 to Patricia K. Kuhl. We thank Erica Stevens and Craig Harris for help with the statistical analyses.

Correspondence concerning this article should be sent to Patricia K. Kuhl, Department of Speech and Hearing Sciences, University of Washington (WJ-10), Seattle, Washington 98195.

that a certain lip posture, such as an open mouth, goes with a certain sound, such as the vowel sound /a/. This auditory–visual matching effect has been replicated and extended to other vowel pairs such as /i/ and /u/ (Kuhl & Meltzoff, 1988) and to disyllables such as *mama* versus *lulu* and *baby* versus *zuzi* (MacKain, Studdert-Kennedy, Spieker, & Stern, 1983).

Thus, a number of studies demonstrate adults' and infants' abilities to relate auditory and visual speech information. The current studies were designed to examine the underlying basis of these abilities. The goal was to take apart the auditory stimulus to identify the critical stimulus features that are necessary and sufficient for the detection of the cross-modal match between a visually perceived phonetic gesture and its concomitant sound. In these studies we presented both adults and infants with tasks that assessed their abilities to relate nonspeech stimuli (pure tones of various frequencies and three-tone complexes derived from the vowels) to real /i/ and /a/ vowels presented auditorially or visually. Just as the use of nonspeech stimuli has helped to define the auditory features that were necessary to reproduce effects such as categorical perception, we thought that these manipulations would help identify the necessary and sufficient signals for cross-modal matching between audible and visible instantiations of speech.

Relating Nonspeech Signals to Vowels: The Predominant Pitch of Vowels

The notion that vowels can be modeled with nonspeech sounds such as simple tones was recognized very early. Isaac Newton's notebooks described how one could create the impression of a series of vowels beginning with low-pitched vowels like /u/ and /a/ up to high ones such as /i/ by filling a "very deepe flaggon with a constant streame of beere or water . . ." (Newton, 1665, cited in Ladefoged, 1967, p. 65; see also Helmholtz, 1885/1954). A more experimental approach has verified that vowels have a "predominant pitch" (Chiba & Kajiyama, 1958; Fant, 1973; Farnsworth, 1937). For example, Farnsworth (1937) presented subjects with 22 random pure tones ranging from 375–2400 Hz. Subjects were instructed to label each of the tones as 1 of 12 vowels. The results showed that tones in the high-frequency range were labeled as high-front vowels such as /i/, whereas tones in the low to middle frequencies were labeled as low vowels such as /a/ and /u/. Fant (1973) instructed subjects to label 12 pure tones ranging from 250 to 4000 Hz as 1 of 9 vowels. Results of this study again showed that particular frequency regions were associated with particular vowels. Identification of the vowel /i/ predominated at 4000 Hz, and identification of /a/ predominated at 1000 Hz. Taken together, these experiments demonstrate that vowel quality is associated with a predominant pitch that can be represented by a single pure tone and that high-to-low ordering of vowels, with /i/ being high and /a/ being low, is very consistent.

This difference in the predominant pitch of vowels was captured in distinctive feature theory (Jakobson, Fant, & Halle, 1969) by the feature *grave–acute.* The grave–acute contrast referred to the location of the main concentration of energy in the vowel. When the mean of the vowel's first two formants is low in frequency, its "center of gravity" is lower (grave) than if the mean of the vowel's first two formants is higher in frequency (acute). Thus, the vowel /a/ is considered grave, whereas /i/ is acute.

The experiments reported here further investigated the relationship between nonspeech stimuli and vowels. We examined both simple pure tones that capture a single feature of the vowel and three-tone complexes whose frequencies matched the formant frequencies of the natural vowels. In Experiments 1 and 2 we used pure tones and extended previous research by using nonspeech stimuli in two ways: by testing infants as well as adults and by extending the tests to cross-modal, amodal, and unimodal settings. Two questions were the following: Can nonspeech stimuli with the acoustic simplicity of a pure tone be reliably "matched" to vowel sounds presented as either an auditory or visual or "imagined" stimulus? If so, are such matches in the direction predicted by the predominant-pitch hypothesis, that is, a pattern in which the vowel /i/ (whether heard, seen, or imagined) is matched to pure-tone stimuli in the high-frequency region, and the vowel /a/ (whether heard, seen, or imagined) is matched to auditory pure-tone stimuli in the low-to-midfrequency region? In Experiments 3 and 4 we used three-tone complexes to test the hypothesis that more complex nonspeech stimuli can be reliably matched to speech stimuli by adults and infants. Multitone complexes have been used in previous unimodal studies in both adults and infants (Jusczyk et al., 1980; Miller et al., 1976; Pisoni, 1977; Pisoni et al., 1983). In the present studies we extended the use of multitone complexes to a cross-modal setting.

Experiment 1

In Experiment 1 adults were asked to match nonspeech stimuli (pure tones) to speech events (vowels). Adults were tested in three conditions. In the first, they were asked to adjust an auditorially presented pure tone until it provided the best match to an auditorially presented /i/ or /a/ vowel. The vowel category presented (either /i/ or /a/) did not consist of a single token but instead many instances of the category. Multiple tokens were used to ensure that listeners' matches were not based on a single acoustic component physically present in a single /i/ or /a/ vowel such as a particular harmonic of the vowel. Listeners had to pick a tone that matched the overall sound quality of the entire set of vowels they heard. In the second condition, adults adjusted the pure tone until it matched an imagined /i/ or /a/ vowel. In this condition, subjects relied on their internal representations of the vowels /i/ and /a/ rather than on any physical instantiation of them; this ensured that listeners could not base their matches on a physical component present in the stimulus. In the third condition, subjects were presented with pure tones of various frequencies and asked to identify which of two visually presented articulatory acts (/i/ and /a/) was the best match to the tone.

Method

Subjects

The subjects were 72 adults from 23 to 47 years of age. They had no known history of hearing problems. Most of the subjects were

students in the Department of Speech and Hearing Sciences at the University of Washington and were paid $10 for their participation in the study. Each had had at least one class in phonetic transcription, but few had any course work in experimental phonetics.

Conditions

In the auditory (A) condition, subjects listened to an audio recording of a female talker who produced a series of /i/ vowels or a series of /a/ vowels. Subjects then manipulated the frequency dial on a pure-tone audiometer to produce a tone that best matched the vowels. In the imagined (I) condition, subjects were instructed to imagine the vowels contained in words printed on a card. Then, subjects were told to manipulate the pure-tone dial to select the tone that best matched the imagined vowel. In the visual (V) condition, subjects viewed a film of two faces, side by side, of a female talker silently articulating the vowels /i/ and /a/. Subjects then listened to a pure tone of low or high frequency that was synchronized to the movements of the two mouths. They were asked to identify the face (either the /i/ face or the /a/ face) that best matched the tone.

Twenty-four subjects were tested in both the A and the I conditions. The order of the conditions (A and I) and the order of presentation of the two vowels (/i/ and /a/) were counterbalanced across subjects. An additional 48 subjects were tested in the V condition: 24 in a low-tone condition and 24 in a high-tone condition.

Stimuli

Vowels. The vowels presented in the A condition consisted of 10 different tokens each of the isolated vowels /i/ and /a/. The vowel tokens used in this study were the same ones used in Kuhl and Meltzoff (1982, 1984). The vowels were spoken by a female talker. The average of the first three formant frequencies for the vowel /i/ were 416 Hz, 2338 Hz, and 2718 Hz; comparable values for the vowel /a/ were 741 Hz, 1065 Hz, and 3060 Hz. Average durations of the vowels were 1.12 s (/i/) and 1.15 s (/a/).

The /i/ and /a/ vowels were produced with a rise–fall fundamental frequency contour characterized by an initial rapid rise in frequency over the first 200 ms, followed by a longer, more gradual decline in the fundamental frequency contour over the remainder of the vowel. The average starting frequency of the contours was 204 Hz, the average peak was 276 Hz, and the average final frequency was 160 Hz. In the I condition, the printed word *heat* was used to represent the vowel /i/, and the printed word *hot* was used to represent the vowel /a/.

In the V condition, the vowels were filmed faces. The faces were arranged on two 16-mm film loops; one contained the left face pronouncing /i/ and the right face pronouncing /a/, and the other loop had the reverse spatial arrangement. Each film loop consisted of 10 /i/ and 10 /a/ articulations paired side by side and aligned so that the opening and closing of the two mouths was synchronized for each oral gesture. When projected, the faces were about life size, 21 cm long and 15 cm wide; their centers were separated by 38 cm. The faces were filmed so that the woman's ears and hair could not be seen. Vowel articulations were exaggerated in both duration and degree of mouth opening. The mean lip-opening-to-closing duration was 1.97 s for the /i/ stimuli and 1.92 s for the /a/ stimuli; one articulation occurred every 3 s.

Tones. The tones presented to the adults in the A and I conditions were produced by a clinical audiometer (Madsen Electronics, OB 77). The audiometer had a discrete frequency dial that could be adjusted to present pure-tone stimuli of various frequencies. The dial could be adjusted to present frequencies between 125 Hz and 8000 Hz in discrete steps (125, 250, 500, 750, 1000, 1500, 2000, 3000, 4000, 6000, and 8000 Hz). In the A and I conditions, subjects could select any one of these tones as the best match for the vowels /i/ and /a/.

In the V condition, subjects were presented with one of six pure tones (750, 1000, 1500, 2000, 3000, or 4000 Hz). Pilot studies had suggested that subjects related auditorially presented /i/ vowels to frequencies of 2000 Hz and above and related /a/ vowels to pure-tone frequencies located below 2000 Hz. Thus we categorized the first three tones (750, 1000, and 1500 Hz) as 'low' and the remaining ones (2000, 3000, and 4000 Hz) as 'high.' The tones were generated by a sine-wave synthesis program on a DEC PDP 11–34 computer. For each frequency, there were 10 pure-tone stimuli, one for each of the original /i/ and /a/ vowels. The pure-tone stimuli matched the original vowels in duration and in their amplitude envelopes over time. The stimuli ranged in duration from 1,125 to 1,425 ms. The tones were generated at sound pressure level (SPL) values that resulted in adult judgments of equal loudness for the original vowels when adults were seated in the experimental position. The pure-tone stimuli were stored on the computer and triggered on-line by the original vowels. (The trigger delay was 40 μs and hence imperceptible.) The subjects' task was to identify the articulatory act (either /i/ or /a/) that best matched the tone. Eight subjects were tested at each frequency, 24 in each of the low- and high-frequency conditions.

Procedure

In the A condition, adults were seated at the control panel of the audiometer. They listened to the vowel stimuli, presented once every 3 s over earphones (TDH-39, with AMX cushions). The subjects presented the pure tones to themselves by depressing a button and had complete control over the knob that adjusted the frequency of the pure tone. They were instructed to manipulate the dial until they had identified a pure tone that provided the 'best match' to the vowel. Subjects were instructed to present themselves with a tone in the interstimulus interval (approximately 1.5 s) between the taped vowel stimuli. The order of presentation of the vowels (/i/ first or /a/ first) was counterbalanced. In the I condition, the adults sat at the audiometer control panel and were shown a card with a word printed on it. They were told to imagine the vowel in the word, produced in isolation. Then they were asked to manipulate the dial until they identified the pure tone that was the best match to the imagined vowel.

The order of the A and I conditions was counterbalanced across subjects. Prior to the start of each condition, subjects were given a practice period during which they manipulated the dial to sweep through the entire frequency range from which they could choose. In both conditions, subjects had 5 min to complete the task. No one reported any difficulty with the task, and most completed it well before the 5-min time limit.

The laboratory set-up used to test subjects in the V condition consisted of an experimental suite and a control room; both were sound-treated. The control room contained a DEC PDP 11–34 computer, a Siemens Model 2000 16-mm dual system projector, a timer-controlled shutter system, and a Sony video recorder and monitor. The adult subjects were seated in a chair in front of a three-sided cubicle, on the front panel of which was projected the two visual faces. Behind the front panel, in midline position with respect to the two faces, was an Electrovoice SP-12 loudspeaker through which the auditory stimuli were presented.

The adults were familiarized with the two faces for 10 s each before both faces were turned on. The order of familiarization of the faces was counterbalanced across subjects. Once both faces were projected, the pure tone was auditorially presented from the loudspeaker. The pure tones emanating from the loudspeaker had amplitude envelopes that matched the original vowels so that the tones were soft when the mouth first opened, grew louder as the mouth grew wider, and became

softer again as the mouth closed. The visual stimuli were arranged on a continuous loop of film so that the 10 paired articulations were repeated until a subject completed the experiment. Subjects watched the faces and listened to the tones for 2 min, after which they were asked to specify the face that best matched the pure tone. In essence, the procedure with adults was nearly identical to that used to test infants in the Kuhl and Meltzoff (1982, Experiment 2) study, except that verbal instructions were used.

Results

Pure-Tone Matching to Auditorially Presented Vowels

The subjects' choices of the best match between the pure tones and the vowels are displayed in the first two columns of Table 1. As shown, 22 of 24 subjects matched pure tones to the vowels in the predicted direction, specifying a higher tone as the best match to the vowel /i/ than that chosen as the best match to the vowel /a/ ($p < .0001$, binomial test). The mean value designated as the best match for the /i/ vowel was 1479.17 Hz (range = 250 Hz–4000 Hz), whereas the mean value designated as the best match for the vowel /a/ was 520.83 Hz (range = 125 Hz–1500 Hz), $t(23) = 5.13$, $p < .0001$. The modal choice of a pure-tone match for the vowel /i/ was 1500 Hz, whereas the modal choice for the vowel /a/ was 250 Hz.

Table 1
Adult Subjects' Matches of Pure Tone (Hz) to Auditorially Presented (A), Imagined (I), and Three-Tone Nonspeech Analogs of the Vowels /i/ and /a/

	Condition					
	A		I		Three-tone	
Subject	/i/	/a/	/i/	/a/	/i/	/a/
1	1500	250	2000	250	500	1000
2	4000	1500	1000	125	1000	2000
3	1000	250	750	125	500	1000
4	750	250	2000	1000	500	1000
5	1000	750	1000	500	500	1000
6	1500	1000	500	250	500	1000
7	750	250	2000	500	750	1000
8	500	250	1000	125	500	1000
9	250	250	2000	500	500	750
10	2000	1500	1500	500	500	3000
11	1500	500	500	250	500	750
12	250	250	3000	250	3000	1000
13	750	250	2000	500	500	1000
14	2000	750	750	250	500	1000
15	1500	500	2000	1000	3000	1000
16	3000	750	750	250	500	1000
17	1500	500	1000	500	500	1500
18	500	250	2000	250	2000	3000
19	750	250	1000	250	500	1000
20	3000	250	2000	250	1000	2000
21	250	125	4000	1500	250	1000
22	3000	750	750	125	500	2000
23	250	125	4000	1500	500	1500
24	4000	1000	2000	750	250	1000

Note. Conditions A and I used the same subjects; a new group of subjects was tested in the three-tone analog condition.

Pure-Tone Matching to Imagined Vowels

The subjects' choices of the best match between the pure tones and the imagined /i/ and /a/ are displayed in the third and fourth columns of Table 1. Each of the 24 subjects matched the vowel /i/ to a higher tone than the vowel /a/. The mean value designated as the best match to the vowel /i/ was 1645.83 Hz (range = 500 Hz–4000 Hz), whereas the mean value designated as the best match to the vowel /a/ was 479.17 Hz (range = 125 Hz–1500 Hz), $t(23) = 7.95$, $p < .0001$. The modal choice of a pure tone to match the imagined /i/ was 2000 Hz, whereas the modal choice for the imagined /a/ was 250 Hz.

A 2 × 2 analysis of variance examining the values assigned across conditions (A and I) and vowels (/i/ and /a/) revealed a nonsignificant effect for condition, $F(1, 23) = 0.51$, $p > .50$, and a highly significant effect of vowel, $F(1, 23) = 46.8$, $p < .001$. Thus, listeners' perception of the predominant pitch of vowels is a very robust phenomenon. Regardless of whether the listeners are provided with concrete instances of the vowels or whether they are asked to imagine the vowels, listeners systematically match the vowel /i/ to a tone that is high in pitch and match the vowel /a/ to a tone that is low in pitch.

Pure-Tone Matching to Visually Presented Vowels

The results of the V condition, in which subjects chose which of two visually presented faces best matched a specific pure tone, are shown in the first two columns of Table 2. Recall that 24 subjects were presented with a low tone (either 750, 1000, or 1500 Hz) and that another 24 subjects were presented with a high tone (either 2000, 3000, or 4000 Hz). A chi-square test of the relation between tone frequency and visual articulation revealed a highly significant effect, $\chi^2(1, N = 48) = 18.9$, $p < .0001$. Of the 24 adults who were auditorially presented with a high tone, 21 chose the /i/ face as the best match to the tone ($p < .0001$, binomial test). Of the 24 adults who were auditorially presented with a low tone, 19 chose the /a/ face as the best match to the tone ($p < .005$, binomial test).

In summary, Experiment 1 demonstrated three things about adults' perception of the relation between pure tones and vowels. First, adults matched pure tones in the high-frequency range to an auditorially presented /i/ vowel and matched pure tones in the low-frequency range to an auditorially presented /a/ vowel. Second, this association between the vowel /i/ and high tones and the vowel /a/ and low tones occurred even in an amodal test, when the subject simply imagined the vowel to which the pure tone was matched. Finally, adults could match a nonspeech pure tone (a purely auditory stimulus) to faces pronouncing /a/s and /i/s (visual stimuli), and for this cross-modal case we obtained the same correspondence seen in the auditory condition and the imagined condition: High tones correspond to visual /i/ vowels, and low tones correspond to the visual /a/ vowels.

Experiment 2

The subjects in Experiment 2 were infants rather than adults. In this study, we posed the visual condition to infants

to determine whether they also matched high tones to the visual representation of the vowel /i/ and low tones to the visual representation of the vowel /a/. Save for minor age-appropriate modifications (the use of an infant seat, a camera to record infant visual fixations, and a light), the apparatus and stimuli were the same as in Experiment 1, and the procedure was altered only to delete the verbal instructions. In fact, the procedures used were identical to those found to be successful in previous tests of lipreading in infants (Kuhl & Meltzoff, 1982, 1984).

Method

Stimuli

The visual stimuli for this experiment were the /i/ and /a/ faces used by Kuhl and Meltzoff (1982) and were identical to those presented to the adults in the V condition of Experiment 1.

Apparatus

A camera (Panasonic, Model WV-1354A) was positioned through a hole in the middle panel midway between the two projected faces for the infant tests. It filmed a close-up image of the infant's face. Behind the hole was a small light that could be flashed at the infants at particular points during the experiment to bring the infant's attention to midline.

Subjects

The subjects were 96 infants, 48 female and 48 male. Infants ranged in age from 18 to 20 weeks (M = 19.2 weeks). Selection criteria for the study were that infants had to be older than 36 weeks gestational age at birth and have no history of middle-ear infection. The parents of prospective subjects were called from newspaper birth announcements, and their infants were screened with the aforementioned criteria. There were 16 infants tested for each of six pure-tone stimulus frequencies, thus yielding 96 infants. Within a group of 16 subjects, infants were counterbalanced with respect to left–right facial orientation, order of facial familiarization, and sex.

Procedure

The procedure fully duplicated that described in our earlier cross-modal work, using real vowels instead of tones (Kuhl & Meltzoff, 1982, 1984). It also replicated that used with adults in the visual condition of Experiment 1 with the exception that infants sat in an infant seat, and instead of by verbal instruction, their attention was brought to midline by a light that flashed through a hole in the front panel of the cubicle. Once the infant's attention was focused on the screen, each face was shown separately for 10 s without sound. After this initial 20-s period (while the faces were still off) the presentation of the pure-tone stimuli began, and a midline gaze was again obtained from the infant by flashing the light. Once midline was achieved the light was turned off and both faces were presented for the 2-min test period.

Table 2
Number of Adults and Infants Selecting /i/ and /a/ Faces as the Best Match to Pure Tones of High and Low Frequency and to Three-Tone Nonspeech Analogs of /i/ and /a/

Auditory stimulus	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	/i/	/a/	/i/	/a/	/i/	/a/	/i/	/a/
High tone	21	3	17	31				
Low tone	5	19	16	31[a]				
Three-tone /i/					2	14	6	10
Three-tone /a/					12	4	5	11

[a] There was one tie in this group of infants, giving a row total of 47 rather than 48.

Scoring and Data Analysis

Scoring of the infants' visual fixations was done by an observer who viewed the infant's face from the videotape record. The observer scored the direction of visual gaze (right, left, or no facial fixation) during the 2-min test interval. Inter- and intraobserver reliabilities were determined by rescoring 24 subjects who were randomly drawn from each frequency group so that 4 infants per group were rescored. Both reliability measures were .99, as calculated by a Pearson r.

Results

The results of this study are most interesting when put in the context of related work on the detection of auditory–visual correspondences for speech in infancy. In the original Kuhl and Meltzoff (1982) study and in a replication of that experiment (Kuhl & Meltzoff, 1988), infants were tested with speech stimuli (vowels). The results of this past work showed that infants' visual preferences were governed by what they heard. Infants who heard the /i/ vowel looked longer at the face that produced /i/, whereas infants who heard the /a/ vowel looked longer at the face that produced /a/. In contrast, the data from this experiment showed that when infants listened to nonspeech (pure-tone) stimuli while looking at the same /i/ and /a/ faces, they did not produce any cross-modal effect. There was no evidence that the auditory signal played a role in infants' looking preferences. Instead, infant visual fixation was governed by a preference for a particular oral gesture (the /a/ face) and did not depend on the frequency of the pure-tone auditory stimulus. Thus, unlike adults tested in the same condition in Experiment 1, infants did not detect a match between pure tones and faces.

The results of the current experiment show that across all conditions infants devoted 59.8% of their total fixation time to the /a/ as opposed to the /i/ face, which is significantly greater than the 50% chance value, $t(95) = 3.12$, $p < .005$. At the level of individual subjects, 62 of the 96 infants looked longer at the /a/ face than at the /i/ face ($p < .01$, by binomial test). Table 2 provides the data broken down by frequency. As shown, of the 48 infants presented with a high tone (2000, 3000, and 4000 Hz), 31 infants looked longer at the /a/ face, and 17 infants looked longer at the /i/ face. Of the 48 infants who listened to a low tone (750, 1500, and 2000 Hz), 31 looked longer at the /a/ face, and 16 looked longer at the /i/ face; there was one tie.

We also performed an analysis on the other factors counterbalanced in the experiment. The mean percentage of total fixation time spent looking at the first familiarization face

fixation time spent looking at the first familiarization face was 56.0%, $t(95) = 2.02$, $p < .05$, with 58 of the 96 infants looking longer at the first familiarization face ($p = .05$, by binomial test). This first-face preference was significant only when the first face was /a/, however; when the first face was /i/, the first-face preference did not differ significantly from the 50% chance level. Infants did not show a preference for the right as opposed to the left face.

Discussion of Experiments 1 and 2

In Experiment 1, adults were tested in unimodal, cross-modal, and amodal (imagined) conditions. They compared (a) auditorially presented vowels to a pure-tone stimulus, (b) imagined vowels to a pure-tone stimulus, and (c) a visually presented vowel sound to a pure-tone stimulus. In Experiment 2, infants were tested in the same cross-modal task as the adults, which involved the detection of a match between the nonspeech pure-tone stimuli and the visually presented faces.

The results of Experiment 1 showed that adult subjects related nonspeech stimuli (pure tones) to speech in all three different modes. The results conformed to the predominant-pitch hypothesis, in which the vowel /i/ was matched to a tone that was high in frequency and the vowel /a/ was matched to a tone that was low in frequency. This is in accordance with distinctive feature theory, which describes /i/ as an "acute" vowel and /a/ as a "grave" vowel (Jakobson et al., 1969). In the auditory condition, adults made a match even though the vowel category presented (either /i/ or /a/) did not consist of a single token but many instances of the category. The use of multiple tokens ensured that listener's matches were not based on a single acoustic component physically present in the set of /i/ or /a/ vowels, such as a particular harmonic of the vowel. Rather, listeners were asked to pick a tone that matched the overall 'sound quality' of the set of vowels they heard. In fact, neither the subjects' modal choice of a pure-tone match for the two vowels nor the mean of their choices corresponded to either the actual first or the second formant frequency of the two vowels. The fact that adults' matches in the auditory condition were not governed by a specific physical component of the vowels was confirmed by the results of the imagined (amodal) vowel condition. In this condition, subjects relied on their internal representations of the vowels /i/ and /a/ rather than on any physical instantiation of it. Yet they uniformly chose high tones to represent /i/ and low tones to represent /a/. Across the two conditions (auditory and imagined), the high–low relationship between the two vowels was manifested in 46 of 48 instances. Thus, adults display a robust ability to match complex speech signals like vowels to stimuli with the simplicity of a pure tone.

Adults and infants were also tested in a cross-modal matching task that required them to detect a correspondence between the pure-tone nonspeech stimuli and visual instantiations of the vowels /i/ and /a/. Subjects listened to pure-tone stimuli ranging from 750 Hz to 4000 Hz that were paired with the visual presentation of two faces mouthing the vowels /i/ and /a/. The results showed that adults did so with ease. Of the 48 adults tested, 41 detected a match in the direction predicted by the predominant-pitch hypothesis. That is, adults listening to high tones selected the /i/ face as the best match, whereas adults listening to the low tones selected the /a/ face as the best match. Infants behaved quite differently. Infants' facial fixations appeared to be influenced only by the visual stimulus; there was no effect of auditory pure-tone frequency and thus no cross-modal effect. Across all stimulus frequencies, infants showed significant preferences for a single face (/a/), particularly when it appeared first during the familiarization phase.

From a theoretical perspective, the principal findings of interest are two dissociations that need to be explained: (a) between the adults and infants (adults detected a match between nonspeech stimuli and speech presented visually, whereas infants did not) and (b) between infants tested in previous unimodal experiments with nonspeech stimuli and infants tested here with nonspeech stimuli in a cross-modal setting.

If we focus on the second dissociation, the current findings contrast sharply with previous results on infants' perception of nonspeech auditory signals. Naturally, the present results would have been of only limited interest if infants had not already been shown to respond to nonspeech stimuli in other circumstances. As reviewed, it has been shown that infants reproduce categorical perception effects when nonspeech sounds are used. The dissociation between these established results and the present ones is provocative. The critical difference could reside either in the nature of the stimulus used here (pure tones) or in the fact that the task involved a cross-modal rather than a unimodal ability. Considering the nature of the stimulus, perhaps infants could not relate pure-tone stimuli to the faces because the pure-tone stimuli were insufficiently like speech. In the previous experiments on infants, in which nonspeech stimuli were used, the signals were two- or three-tone analogs of the speech events (Jusczyk et al., 1980; Pisoni et al., 1983). They contained two or three peaks in spectral energy that matched the formant frequencies of the speech stimuli, whereas the pure tones used here contained only one. The multitone analogs are more complex and thus more like speech. If more complex nonspeech analogs were used, would adults and infants match the nonspeech signals to speech? This question was posed in Experiments 3 and 4.

Experiment 3

In Experiment 3, adults were presented with three-tone nonspeech analogs of the vowels /i/ and /a/. Adults were tested in two conditions. In the A condition they were asked to adjust an auditorially presented pure tone until it provided the best match to the auditorially presented three-tone analogs of /i/ and /a/ vowels. This condition was included because the hypothesis resulting from Experiment 1 was that adults used the predominant pitch of a nonspeech stimulus to make a match to the vowels. We thus wanted to identify the predominant pitch of the three-tone analogs. Just as in Experiment 1, the three-tone analogs presented to adults consisted of a set of stimuli rather than a single one. Each /i/ and

/a/ vowel in Experiment 1 was used to generate a three-tone stimulus. Multiple tokens were used to ensure that listeners' matches were not based on a single acoustic component physically present in a single three-tone analog. Listeners had to pick a tone that matched the overall sound quality of the entire set of three-tone analogs they heard. In the V condition, adults were presented with the three-tone analogs (derived from either /i/ or /a/) and were asked to identify which of two visually presented vowels (/i/ and /a/) was the best match to the three-tone analogs they heard.

Method

Subjects

The subjects were 56 adults from 21 to 38 years of age who had no known history of hearing problems. Most of the subjects were students in the Department of Speech and Hearing Sciences at the University of Washington and were paid $10 for their participation in the study. Each had had at least one class in phonetic transcription, but few had any course work in experimental phonetics.

Conditions

In the A condition, subjects listened to an audio recording of the series of three-tone analogs. They then manipulated the frequency dial on a pure-tone audiometer to produce a tone that best matched the three-tone analogs. In the V condition, subjects viewed a film of two faces, side by side, of a female talker silently articulating the vowels /i/ and /a/ while they listened to a set of three-tone analogs (derived from either /i/ or /a/). The multitone complexes were synchronized to the mouth movements just as had been the case with the pure tones in Experiment 1. Subjects were asked to identify the face (either the /i/ face or the /a/ face) that best matched the tonal complexes. Twenty-four subjects were tested in the A condition; a new group of 32 subjects were tested in the V condition, 16 with the /i/ three-tone analogs and 16 with the /a/ three-tone analogs.

Stimuli

The three-tone analogs used in the A and V conditions were derived from the formant frequency values of the 10 tokens of /i/ and the 10 tokens of /a/ used in Experiment 1. The formant frequencies of the original vowels were analyzed by computer (DEC PDP 11–34), and these values were used to generate the three-tone complexes. Each three-tone complex was composed by combining the center frequencies of the first three formants of one of the original /i/ and /a/ vowels. We rounded each formant frequency value to the nearest 10-Hz (F1) or 50-Hz (F2 and F3) equivalent when generating the three-tone analogs; the amplitudes of the three tones were equal. The values used are shown in Table 3. The durations and amplitude contours of the three-tone analogs matched those of the vowels from which they were derived. The three-tone complexes sounded like piano chords, and when queried, subjects reported that they did not perceive the three-tone complexes as speech events.

In the A condition, the three-tone analogs were matched to a tone generated by the clinical audiometer used in Experiment 1. The audiometer had a discrete frequency dial that could be adjusted to present pure-tone stimuli between 125 Hz and 8000 Hz in discrete steps (125, 250, 500, 750, 1000, 1500, 2000, 3000, 4000, 6000, and 8000 Hz). In the V condition, subjects were presented with one set of the three-tone analogs (either those derived from /i/ or from /a/) while they watched the same filmed faces used in Experiment 1. As in Experiment 1, the three-tone analogs were stored on the computer and triggered on-line by the original vowels.

Table 3
Frequencies of the Tones Used to Generate the 10 Three-Tone Analogs of the /i/ and the /a/ Vowels

Vowel /i/			Vowel /a/		
F1	F2	F3	F1	F2	F3
420	2500	2800	650	1000	3100
400	2800	3400	750	1200	3200
400	2600	3000	700	1100	3000
400	2700	3300	720	1100	3100
420	2800	3350	650	1100	3100
400	2600	2900	800	1000	3100
420	2400	2800	800	1100	2900
450	2400	2800	780	1000	3000
420	2600	2900	780	1050	3000
430	2500	2900	780	1000	3100

Procedure

The same procedure used in Experiment 1 was used here. Instead of real vowels, subjects listened to the three-tone analogs presented once every 3 s over earphones (TDH-39, with AMX cushions). Subjects presented the pure tones to themselves by depressing a button on the audiometer, and they had complete control over the knob that adjusted the frequency of the pure tone. They were instructed to manipulate the dial until they had identified a pure tone that provided the 'best match' to the set of three-tone analogs. Subjects were practiced users of the audiometer and were instructed to present themselves with a tone in the interstimulus interval between the taped three-tone stimuli (about 1.5 s). The order of presentation of the tonal complexes (/i/ analogs first or /a/ analogs first) was counterbalanced.

The laboratory setup used to test subjects in the V condition was identical to that used in Experiment 1. Rather than listen to tones while watching the facial stimuli, adults listened to the three-tone analogs. All other aspects were identical to those used in Experiment 1.

Results

Pure-Tone Matching to Three-Tone Analogs

Subjects' choices of a pure-tone frequency to match the three-tone analogs are displayed in Columns 6 and 7 of Table 2. The subjects in Experiment 3, who were presented with nonspeech analogs of vowels rather than real vowels, reversed the relationship between vowel and tone shown in Experiment 1. In Experiment 1, subjects presented with real /i/ vowels chose high-frequency tones as the best match, whereas those presented with real /a/ vowels chose low-frequency tones as the best match (Table 1, Columns 2–5). In the present experiment, subjects presented with three-tone analogs of /i/ chose

low-frequency tones as the best match and those presented with three-tone analogs of /a/ chose high-frequency tones as the best match. Twenty-two of the 24 subjects chose a lower frequency as the best match to the three-tone /i/ analog than the frequency chosen as the best match to the three-tone /a/ analog ($p < .0001$, binomial test). The mean value designated as the best match for the three-tone /i/ analog was 802.08 Hz (range = 250 Hz–3000 Hz), and the mean value designated as the best match for the three-tone /a/ analog was 1312.50 Hz (range = 750 Hz–3000 Hz), $t(23) = -2.75$, $p < .05$. The modal choice of a pure-tone match for the three-tone /i/ analog was 500 Hz, and the modal choice for the three-tone /a/ analog was 1000 Hz.

Three-Tone Analog Matching to Visually Presented Vowels

The results of the V condition, in which subjects chose which of two visually presented faces best matched the three-tone vowel analogs, are shown in Table 2. Recall that 16 subjects were presented with the three-tone /i/ analogs and that another group of 16 subjects were presented with the three-tone /a/ analogs. As shown, subjects three-tone analog and visual face matches were dictated by the perceived pitch of the three-tone analogs. That is, adults matched the three-tone /i/ analogs (perceived as having a predominantly low pitch according to the A condition in this experiment) to the /a/ face and matched the three-tone /a/ analogs (perceived as having a predominantly high pitch) to the /i/ face. A chi-square test of the relation between three-tone analogs and visual vowels revealed a highly significant effect, $\chi^2(1, N = 32) = 10.3$, $p < .01$. Of the 16 subjects who were presented with the three-tone /i/ analogs, 14 chose the /a/ face as the best match ($p < .005$, binomial test). Of the 16 subjects who were presented with the three-tone /a/ analogs, 12 chose the /i/ face as the best match ($p < .05$, binomial test).

In summary, Experiment 3 demonstrated two things about adults' perception of the relation between nonspeech stimuli and vowels. First, adults associated auditorially presented three-tone /i/ analogs to pure tones in the low-frequency range and auditorially presented three-tone /a/ analogs to pure tones in the high-frequency range. This is a reversal of the results of Experiment 1, when real vowels were matched to tones. In that experiment, /i/ vowels were associated with high tones and /a/ vowels with low tones. The vowels' tonal analogs reverse the relation between vowel and tone. Second, this reversal is maintained when the three-tone analogs are matched to vowels presented visually. Adults matched the auditorially presented three-tone /i/ analogs to the visual face that pronounced /a/ and the auditorially presented three-tone /a/ analogs to the visual face that pronounced /i/.

Experiment 4

In Experiment 4, we presented the three-tone analogs to infants in the V condition to determine whether they would detect a match between the more complex three-tone analogs and the visual representation of the vowels /i/ and /a/. The procedures used were identical to those used with infants in Experiment 2.

Method

Stimuli

The visual stimuli for this experiment were identical to those used in Experiment 2. The three-tone vowel analogs were identical to those used in Experiment 3 with adults.

Apparatus

The apparatus was identical to that used in Experiment 2.

Subjects

The subjects were 32 infants, 16 female and 16 male, from 18 to 20 weeks of age ($M = 19.4$ weeks). Selection criteria for the study were that infants be older than 36 weeks gestational age and have no history of middle-ear infection. The parents of prospective subjects were called from newspaper birth announcements, and their infants were screened with the aforementioned criteria. There were 16 infants tested for each set of three-tone vowel analogs. Within a group of 16 subjects, infants were counterbalanced with respect to left–right facial orientation, order of facial familiarization, and sex.

Procedure

The procedure replicated that used in Experiment 2, with the exception that infants heard three-tone vowel analogs rather than pure tones while they looked at the two faces.

Scoring and Data Analysis

Scoring of the infants' visual fixations was done by an observer who viewed the infants' faces from the videotape record. The observer scored the direction of visual gaze (right, left, or no facial fixation) during the 2-min test interval. Inter- and intraobserver reliabilities were determined by rescoring a randomly selected 25% of the subjects. Both reliability measures were .99 as calculated by a Pearson r.

Results

The results of Experiment 4 confirm again that infants did not detect a match between nonspeech stimuli and visually presented vowels (Table 2) even though in this experiment the nonspeech stimuli were three-tone analogs of the vowels and thus more complex than the pure-tone stimuli in Experiment 2. There was no cross-modal effect; a chi-square test of the relation between three-tone analogs and visual vowels

showed that the effect did not approach significance, $\chi^2(1, N = 32) = 0.139$, $p > .40$.

As in Experiment 2, infants in both conditions tended to prefer the /a/ face: they devoted 59.1% of their total fixation time to the /a/ as opposed to the /i/ face, which was nearly identical to the preference seen overall for the /a/ face in Experiment 2 (59.8%). In this experiment, however, with fewer conditions (two vs. six) and thus fewer infants tested (32 vs. 96), this was not significantly greater than the 50% chance value, $t(31) = 1.49$, $p > .10$. Twenty-one of the 32 infants looked longer at the /a/ face than at the /i/ face ($p = .055$, binomial test).

An analysis of the other factors counterbalanced in the experiment showed no other significant factors. Infants did not show an overall looking preference that was based on the order of presentation during the familiarization phase; the mean percentage of total fixation time spent fixating the first face was 44.2%, $t(31) = -.932$, $p > .30$. Infants did not show a preference for the right face as opposed to the left face.

Discussion of Experiments 3 and 4

The results of Experiments 3 and 4 are surprising in a number of respects. Consider first the adult data, which showed a reversal in the perception of the predominant pitch for the three-tone analogs of vowels as opposed to the real vowels. Experiment 3 revealed that for adults, the three-tone analogs of the vowels /i/ and /a/ were not perceived to have a predominant pitch which matched that of the vowel from which it was derived. In particular, adults matched the nonspeech analog of /i/ to a low pure tone and the nonspeech analog of /a/ to a high pure tone, which is the reverse of the vowel–tone relation seen in Experiment 1. The predominant pitch of the three-tone complexes appeared to rely on the lowest tones, the ones derived from either the first or the second formant. The lowest tone is considerably lower for /i/ than for /a/. There is precedent for domination by the lowest tone in determining the pitch of multitone complexes (Remez & Rubin, 1984; Remez, Rubin, Nygaard, & Howell, 1987; see also the work on the "dominance region" by Plomp, 1967; Ritsma, 1967).

In the case of real speech, listeners are not predominantly influenced by the first formant but by the vowel's "center of gravity." Data on vowel perception suggests that in speech, listeners perceptually average adjacent formants that are close together (within the critical distance of 3 bark from one another); these adjacent formants can be effectively replaced by a single formant that is intermediate in frequency (Chistovich & Lublinskaya, 1979; Syrdal & Gopal, 1986). Formants that are further apart are not averaged. In the vowel /i/, the high upper formants bring the vowel's center of gravity to a higher frequency, whereas the predominantly low formants of /a/ bring its center of gravity down (Carlson, Fant, & Granstrom, 1975). Thus, the center-of-gravity hypothesis is consistent with the results reported here for adults' perception of the relation between vowels and pure tones.

The center-of-gravity hypothesis, however, does not explain the perception of the predominant pitch of the three-tone complexes. The center frequencies of the three-tone complexes were identical to their real-vowel counterparts. It therefore seems reasonable to assume that if real vowels and their three-tone nonspeech analogs were processed the same way, their predominant pitches ought to be the same. What can be made of the dissociation in perception between the speech and nonspeech events used in these experiments?

The differences between the perception of speech and nonspeech could be because multitone nonspeech analogs, though matching speech in some respects, are acoustically quite different from speech. Multitone complexes are created by combining tones whose center frequencies match the center frequencies of the formants contained in the sounds being modeled; thus both the speech and the nonspeech signals contain three major peaks in energy, but the resemblance between the two signals ends there. Nonspeech signals lack a fundamental frequency component and its related harmonics and thus do not exhibit the broad-band formant structure that characterizes speech. Nonspeech complexes are unrelated tones added together, and they frequently resemble a disjunctive chord played on a piano. Data on the perception of nonspeech stimuli by Traunmuller (1981) suggests that the averaging of spectral peaks over the critical three-bark distance may occur only with speech sounds. Thus, the acoustic character of speech may trigger a different kind of analysis of sound, one that is used only when the acoustic signal contains properties that associate it with vocal-tract resonances (Liberman & Mattingly, 1989).[1] It is clear from this case and others that although speech and nonspeech signals produce similar results in some situations, they may yield divergent results as well (cf. Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Diehl & Walsh, 1989; Fowler, 1990; Pisoni et al., 1983; Tomiak, Mullennix, & Sawusch, 1987), which suggests that the issue of the comparability of speech and nonspeech signals needs to be addressed further (Fowler, 1990).

Having considered the implications of the reversal between speech and nonspeech, we now focus on the other principal finding of the experiment, the cross-modal effect that tested the relation of three-tone complexes and faces. Consistent with the original hypothesis of Experiment 1, Experiment 3 demonstrates that adults relate vowels presented visually to three-tone nonspeech signals on the basis of predominant pitch. That is, the perceived predominant pitch of the three-tone complex (as determined in the A condition of this experiment) reliably predicts which of the two articulatory acts (/i/ or /a/) with which it would be associated. Regardless of whether the auditory signals were as simple as pure tones or as complex as multitone nonspeech analogs, those perceived as high in pitch are related to the visual vowel /i/, and those perceived as low in pitch are related to the visual vowel /a/.

---

[1] The caveat here is the work of Remez and his colleagues, which shows that multitone nonspeech complexes that model dynamically varying sentence-length utterances may be identified as real speech by about two thirds of the subjects (Remez, Rubin, Nygaard, & Howell, 1987; Remez, Rubin, Pisoni, & Carrell, 1981). Presumably, the signals used in Remez's experiments are of sufficient length to suggest time-varying resonances in a way that nondynamic isolated vowel analogs cannot.

Consider next the results of Experiment 4, in which the subjects were infants and the stimuli were the three-tone analogs tested in the visual task. The results support the hypothesis that infants do not detect a cross-modal match between nonspeech auditory events and real speech presented visually. Given that we have now seen an /a/ face preference in three experiments involving different nonspeech stimuli (Experiments 2 and 4 of this article and Experiment 2 reported in Kuhl & Meltzoff, 1982, 1984), we conclude that in the absence of the detection of a cross-modal match, infants simply prefer to fixate the face that produces /a/.[2] Apparently, detection of a cross-modal match, as in our original experiment (Kuhl & Meltzoff, 1982, 1984), overrides that preference.

General Discussion

The results of Experiments 1–4 strongly support two findings: Adults can relate speech stimuli such as the vowels /i/ and /a/ to nonspeech stimuli on the basis of perceived pitch, and infants do not demonstrate this under similar test conditions.

Regarding the findings on adults, the results support the strong conclusion that adults associate the vowel /i/ with high frequencies and the vowel /a/ with low frequencies. The association between vowel and pitch was consistently revealed in three conditions: matching a pure tone to a vowel that was auditorially presented, visually presented, or imagined (Experiment 1). The results revealed that regardless of whether the vowels were heard, seen, or imagined, adults associated /i/ vowels with high tones and /a/ vowels with low tones. In Experiment 3, this result was extended to nonspeech stimuli that were more complex: three-tone analogs of vowels. The results with these more complex stimuli confirmed the mapping of vowel and pitch; adults matched three-tone complexes that were perceived to be high in pitch to the visually presented /i/ vowel, whereas they matched three-tone complexes that were perceived to be low in pitch to the visually presented /a/ vowel.

In what sense does an /i/ vowel resemble a high-pitched sound and an /a/ vowel resemble a low-pitched sound? Is similarity detected at a sensory–perceptual level or at a more cognitive (metaphorical) level? The experiments raise but do not answer this question, yet some speculation is in order.

The sensory–perceptual alternative argues for the more direct mapping between vowels and tones of a particular frequency. In this view, spectral features that are responsible for the perceived predominant pitch of the vowel are derived during the perceptual analysis of the sound. If properties such as the grave–acute feature of distinctive feature theory are derived as a function of the perceptual process, then the correspondence between vowels and the pitch of a pure tone is based on the psychological reality of decomposing speech into such featural elements. Theorists such as Fant (1973) have attributed the fact that vowels are perceived to have a predominant pitch to the derivation of such features. There is also evidence that vowel identification involves the calculation of the degree of separation or the ratio relations between formants and between the first formant and the fundamental frequency (for discussion, see Miller, 1989; Syrdal & Gopal, 1986). In one recent model (Syrdal & Gopal, 1986), a formant-distance measure allows the derivation of specific features. Accordingly, it is at least plausible that the mapping between vowel and predominant pitch has a perceptual basis.

On the other hand, the association between /i/ vowels and high sounds and between /a/ vowels and low sounds may be part of a larger network of knowledge and may be mediated by more metaphorical thinking (cf. Gentner, 1988; Ortony, 1979). Pilot work in our laboratory on the semantic qualities associated with vowels is relevant to this alternative. This work suggests that adults think of /a/ as a "strong" sound, whereas /i/ is "weaker." The attribute *strong* may be associated with maleness, and male voices are predominantly low in pitch, in which case the association of /a/ vowels to low tones may be more mediated than the perceptually based alternative previously described. Admittedly, the current data do not allow us to specify the psychological basis for the vowel to pitch mappings; the main contribution of the current work is to call attention to the very robust and pervasive existence of these mappings in adults in three widely different input conditions (unimodally, cross-modally, and amodally) when using two different types of stimuli (pure tones and three-tone analogs).

These results with adults become even more interesting when considered in relation to the findings from infants. In Experiments 2 and 4, infants did not display the same association of vowel and pitch exhibited by adults. Infants were not differentially affected by the auditory stimuli that were presented. They simply tended to fixate one of the two faces in all conditions. In contrast, Kuhl and Meltzoff (1982, 1984), through the use of the same apparatus and procedure, showed that infants did make differential visual choices when real speech stimuli were presented auditorially; in that case, infants looked at the face that matched the vowel they heard. We may say, then, that 4-month-old infants detect face–voice matches when speech stimuli are presented auditorially and fail to do so when the auditory stimulus is stripped down to its simplest featural component, as in a pure tone or when three-tone nonspeech analogs of the vowels are presented. On the basis of these findings, we put forward the hypothesis that infants' detection of cross-modal correspondence for speech requires the whole speech stimulus. Our operational definition of "whole stimulus" is a signal that is sufficient to allow the identification of the speech signal. (Note that by this definition, synthetic speech signals qualify as whole stimuli, for although they do not include all of the speech information present in a natural utterance, they still allow the identification of the speech stimulus.) To restate the hypothesis within a developmental framework, the findings suggest that infants' cross-modal perception of speech does not originate through

[2] Another possibility is that infants are not simply ignoring the auditory stimulus and demonstrating a basic visual preference but that they somehow relate a nonspeech tonal stimulus—whether simple or complex, or high or low pitch—to a wide-open orifice (/a/) rather than one with retracted lips (/i/). A number of experiments need to be conducted to explore this hypothesis, and it was considered outside the bounds of the current study.

a process that progresses from parts to wholes, in which infants initially relate faces and voices on some simple feature and then gradually build up a connection between the two that involves, on the auditory side, an identifiably whole speech stimulus. Furthermore, on the basis of the comparison of infants and adults, we may infer that at some later age, there develops the ability to detect a cross-modal match given only a part of the stimulus—one that cannot be independently identified as speech.[3]

The result that nonspeech stimuli are not sufficient to generate the cross-modal effect in infants is to be distinguished from the results of previous tests of speech perception that used nonspeech stimuli with infants. In those tests, infants behaved similarly for the speech and the nonspeech sounds. Why is it the case then that infants tested in the present experiments do not map nonspeech parts to the speech wholes presented visually?

The critical difference between the present experiments and those done previously could be the use of a cross-modal paradigm. The requirement of linking two different modalities may require stimulus information in a more complete form. It is plausible that infants are capable of relating pure tones and three-tone analogs to vowels so long as the two stimuli are both presented in a single modality (i.e., auditorially) but that they nonetheless cannot do so in a bimodal task. It would now be of interest to test infants in the auditory task presented to adults in Experiment 1, in which pure tones were matched to vowels presented auditorially. Such tests would be possible with the head-turn conditioning techniques used to study infants' abilities to categorize auditory stimuli (Kuhl, 1979, 1983). The results of those experiments revealed that infants could correctly categorize novel /i/ and /a/ vowels after being trained on a single instance of an /i/ and an /a/ vowel. With that method, infants could be trained on real /i/ and /a/ vowels and then tested with pure tones varying from low to high frequency and with three-tone vowel analogs to see whether they would relate high tones and high three-tone complexes to the /i/ stimulus and low tones and low three-tone complexes to the /a/ stimulus. If infants were successful in relating nonspeech signals to speech signals with this unimodal categorization task, then the critical constraint resides in the cross-modal processing of speech stimuli; in other words, the cross-modal mapping of speech may require the whole stimulus on both sides of the equation. In addition, if infants were successful in relating nonspeech signals to speech in the unimodal categorization experiment, it would provide some support for the perceptual alternative described in the foregoing discussion; young infants would not be expected to possess the kind of metaphorical knowledge inherent in the more cognitive account.

In either event, we have here found that infants do not perceive correspondences between nonspeech auditory signals and the faces of talkers who produce speech, even though they readily did so in the same test situation when the signals were real speech sounds (Kuhl & Meltzoff, 1982, 1984) and even though (as shown here) adults are adept at doing so. The present results suggest a dissociation in how infants treat speech versus nonspeech signals, at least in cross-modal situations. This dissociation corroborates other work from our laboratory on 4- to 5-month-old infants in which infants were tested with nonspeech signals in another cross-modal task. In that case, infants had to relate the perception of the vowels /i/ and /a/ to their own production of those vowels in a vocal-imitation task. Infants listened either to speech stimuli (the vowels /i/ and /a/) or the nonspeech pure-tone signals used in the present studies. The results of the vocal-imitation tests again showed a sharp dissociation between speech and nonspeech signals. In response to speech signals, infants produced speech-like utterances. In response to the nonspeech signals, however, infants did not produce speechlike vocalizations; they listened intently but did not vocalize (for further details, see Kuhl & Meltzoff, 1982, 1988).

Taken together with the results described here, we may infer that in cross-modal tasks in which the perception of speech must be related to the production of speech, young infants need the whole signal, one that is identifiable as a speech sound. This is true in both the auditory–visual situation, in which speech perceived is related to speech produced by another person, and in the imitation situation, in which speech perceived is related to speech produced by oneself. Although infants need a more complete specification of the stimulus to link the perception and production of speech, it is of interest that a part of the stimulus is sufficient to connect the two for adults. The precise time course for the development of this ability and the reason for its absence in early infancy are yet to be determined.

[3] There also exists developmental work and theorizing outside the domain of speech perception which suggests that infants may at first be maximally responsive to wholes (especially in the form of complex natural stimuli) rather than to isolated components (e.g., Bower, 1982; Gibson, 1966; Meltzoff & Kuhl, 1989).

References

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

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception & Psychophysics, 45,* 237–250.

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

Carlson, R., Fant, G., & Granstrom, B. (1975). Two-formant models, pitch and vowel perception. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 55–82). London: Academic Press.

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

Chistovich, L. A., & Lublinskaya, V. V. (1979). The "center of gravity" effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1,* 185–195.

Cutting, J. E., & Rosner, B. S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics, 16,* 564–570.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the

stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America, 85,* 2154–2164.

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

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303–306.

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

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

Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America, 88,* 1236–1249.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59,* 47–59.

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

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

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

Harnad, S. (Ed.) (1987). *Categorical perception: The groundwork of cognition.* New York: Cambridge University Press.

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

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

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

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

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

Kuhl, P. K. (1987a). Perception of speech and sound in early infancy. In P. Salapatek & L. B. Cohen (Eds.), *Handbook of infant perception. Vol. 2: From perception to cognition* (pp. 275–382). New York: Academic Press.

Kuhl, P. K. (1987b). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 355–386). New York: Cambridge University Press.

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

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

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

Ladefoged, P. (1967). *Three areas of experimental phonetics.* London: Oxford University Press.

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

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243,* 489–494.

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

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

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., & Kuhl, P. K. (1989). Infants' perception of faces and speech sounds: Challenges to developmental theory. In P. R. Zelazo & R. G. Barr (Eds.), *Challenges to developmental paradigms* (pp. 67–91). Hillsdale, NJ: Erlbaum.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America, 60,* 410–417.

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

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America, 61,* 1352–1361.

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

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America, 41,* 1526–1533.

Remez, R. E., & Rubin, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception & Psychophysics, 35,* 429–440.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 40–61.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–950.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 243–335). New York: Academic Press.

Ritsma, R. J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America, 42,* 191–198.

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

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

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America, 81,* 755–764.

Traunmuller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America, 69,* 1465–1475.

Received April 6, 1990
Revision received December 19, 1990
Accepted October 22, 1990 ■