Reprint Series
10 December 1982, Volume 218, pp. 1138–1141

SCIENCE

The Bimodal Perception of Speech in Infancy

Patricia K. Kuhl and Andrew N. Meltzoff

The Bimodal Perception of Speech in Infancy

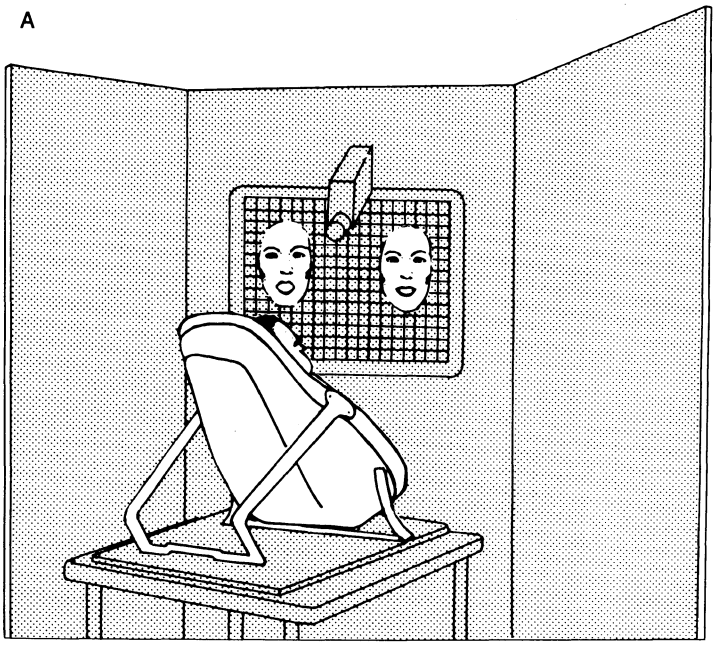
Abstract. *Infants 18 to 20 weeks old recognize the correspondence between auditorially and visually presented speech sounds, and the spectral information contained in the sounds is critical to the detection of these correspondences. Some infants imitated the sounds presented during the experiment. Both the ability to detect auditory-visual correspondences and the tendency to imitate may reflect the infant's knowledge of the relationship between audition and articulation.*

In conversation, speech is often produced by talkers we can both see and hear. We see talkers' mouths move in synchrony with the sounds that emanate from their lips and recognize that the sequence of lip, tongue, and jaw movements correspond to the sounds we hear. Our recognition of these correspondences underlies our ability to lip-read. Recent experiments have demonstrated the impact of vision on speech perception and suggest that in adults speech is represented, at some level, bimodally (*1*).

The experiments reported here show that 18- to 20-week-old infants can detect the correspondence between auditorially and visually perceived speech; in other words, they too manifest some of the components related to lip-reading phenomena in adults. This demonstration of the bimodal perception of speech in infancy has important implications for social, cognitive, and linguistic development.

The infants were shown two side-by-side filmed images of a talker articulating, in synchrony, two different vowel sounds (Fig. 1A). The sound track corresponding to one of the two faces was

0036-8075/82/1210-1138$01.00/0



B	Familiarization			Test
Visual stimuli	Face 1	Face 2	Midline gaze	Both faces
Auditory stimuli	...	...		/a/.../a/.../a/.../a/
Time	10 seconds	10 seconds		2 minutes

Fig. 1. (A) Experimental arrangement of an infant placed in an infant seat within a three-sided cubicle, 46 cm from the two facial displays. (B) Experimental procedure.

presented through a loudspeaker directly behind the screen and midway between the visual images. The visual stimuli consisted of two 16-mm film loops, each containing a face repeating a sequence of ten /a/ vowels (as in *pop*) and ten /i/ vowels (as in *peep*). The articulations were produced once every 3 seconds by the same female talker (*2*). One film loop displayed the /a/ face on the left and the /i/ face on the right; the other loop displayed them in the reverse orientation. The faces were 21 cm long and 15 cm wide; their centers were separated by 38 cm. The auditory stimuli were 16-mm sound tracks containing sequences of /a/'s and /i/'s presented at an average intensity of 60-dB sound pressure level (range, 55 to 64 dB). Either sound track could be played with either film loop. Stimulus durations fell within a narrowly constrained range (*2*), assuring, together with the precise alignment of the sound and film tracks, that each sound track was temporally synchronized to both faces.

The experimental procedure was one of familiarization and testing (Fig. 1B). During familiarization, an infant was shown each face separately for 10 seconds without sound. Following this 20-second period, the faces were briefly covered until the infant's gaze returned to midline. Then the sound (either /a/'s or /i/'s) was turned on and both faces were presented for the 2-minute test phase. The sound presented to the infants, the left-right positioning of the two faces, the order of familiarization, and the sex of the infant were counterbalanced. The subjects were 32 normal infants ranging in age from 18 weeks and 0 days to 20 weeks and 1 day ($\bar{X}$ = 19.3). The only source of visible light in the room was that provided by the films themselves. An infrared light was suspended above the test cubicle. An infrared camera and microphone provided audiovisual recordings of the infants. The infant's visual fixations were scored from videotape by an independent observer who could neither hear the sound nor see the faces presented to the infant (*3*).

We hypothesized that the auditorially presented vowel would systematically influence the infants' visual fixations. Specifically, we predicted that if infants detected the correspondence between the auditorially and visually perceived speech information, they would look significantly longer at the face that matched the sound. The results were in accord with this prediction. The percentage of total fixation time devoted to the matched versus mismatched face was calculated for each infant. The mean percentage devoted to the matching face was 73.6 percent, which is significantly different from the 50 percent chance level [t (31) = 4.67, $P < .001$]. Twenty-four infants looked longer at the face that matched the sound presented than at the mismatched face ($P < .01$, binomial test). There were no significant left-right preferences, face preferences, or familiarization order effects.

Experiment 1 demonstrated that 18- to 20-week-old infants detect a cross-modal relationship between the auditory and visual products of articulation. It did not isolate the auditory information necessary for the detection of these correspondences. Experiment 2 constituted an initial attempt to do so. The original auditory stimuli were altered to remove the spectral information necessary to identify the vowels (formant frequencies) while preserving their temporal characteristics (amplitude and duration). These computer-generated signals were pure-tone stimuli centered at the average frequency of the female talker's fundamental (200 Hz). Their onset-offset characteristics and their amplitude envelopes were synthesized to duplicate those of the original vowels. If infants in experiment 1 were relying on temporal information to link particular face-voice pairs, then they should still look longer at the "matched" face, even though it was represented only by its sine-wave analog (*4*). Alternatively, if the spectral information contained in the vowels was necessary for the detection of these auditory-visual correspondences, performance should now drop to chance.

Thirty-two infants ranging in age from 18 weeks and 1 day to 20 weeks and 0 days ($\bar{X}$ = 19.4) were tested using the same procedure as experiment 1 and the new stimuli. The mean percentage of fixations to the matched face dropped to chance (54.6 percent, $P > .40$); only 17 infants looked longer at it. Experiment 2 thus suggested that some aspect of the spectral information was necessary. It will now be important to determine if perception of the vowel's identity is required to produce the effect, or whether an auditory signal approximating the spectral pattern of the vowel without identifying it is sufficient.

An infant's ability to detect equivalences between auditorially and visually perceived speech has implications for theories of social, cognitive, and linguistic development. From a social perspective, the recognition that a given audi-

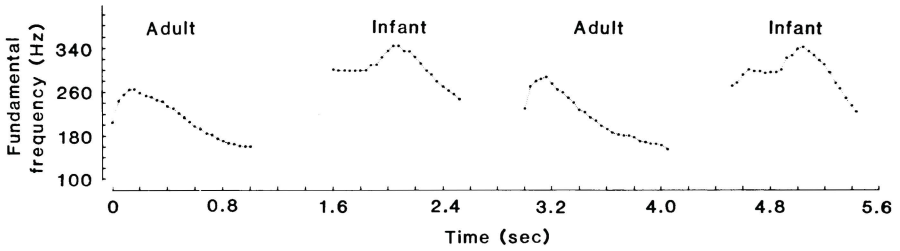


Fig. 2. The fundamental frequency (pitch) of the adult's and infant's vowels are shown as a function of time. Both the adult's and infant's contours are characterized by an initial rise and then a gradual fall in the fundamental frequency. Their durations are also similar. This display illustrates the infant's tendency to take turns. The infant's first vocalization was produced 1/2 second after the adult's, the second, 1/4 second after.

tory signal emanates from a mouth moving in a particular way may help direct the infant's attention toward a specific speaker. This in turn may play a role in coordinating joint actions between infants and caretakers (*5*). These results are also relevant to an emerging view of infant cognitive development. This view holds that young infants are predisposed to recognize intermodal equivalences in the information picked up by different perceptual modalities, and this ability underlies their success on a variety of cross-modal tasks (*6, 7*).

The results are particularly relevant to theories of linguistic development. They suggest that infants relate specific articulatory postures to their concomitant speech sounds. These findings could reflect isolated auditory-visual associations that were learned by watching caretakers speak. Alternatively, they could reflect a more general knowledge, learned or inherent, of the relationship between audition and articulation. Such knowledge might be quite broad, encompassing information about the auditory, visual, and motor concomitants of speech. This latter alternative would imply that in addition to the auditory-visual equivalents demonstrated here, young infants may be cognizant of auditory-motor equivalents, exemplified by vocal imitation, and visual-motor equivalents exemplified by the imitation of visually presented articulatory movements (*8*). Such an intermodal representation of speech would be especially conducive to vocal learning (*9*).

During these experiments we made two observations concerning vocal imitation that support this broader interpretation. (i) Ten infants who heard the vowel stimuli (experiment 1) produced utterances typical of babbling (*10*), whereas only one infant who heard the pure-tone stimuli (experiment 2) did so. The speech stimuli thus seemed more effective in eliciting infant babbling than nonspeech stimuli. (ii) The infants in experiment 1 produced sounds that resembled the adult female's vowels. They seemed to be imitating the female talker, "taking turns" by alternating their vocalizations with hers.

Figure 2 displays a single infant's imitation of the prosodic features of the adult's vowels—their intonation contours and overall durations. The adult produced a "declarative" contour; that is, a rise in the fundamental frequency followed by a longer, more gradual fall in the fundamental frequency. The infant mimics this rise-fall contour producing a pitch pattern that resembles the adult's (although it is higher in frequency because the infant's vocal cords are shorter than the adult's). Rise-fall contours of this type are not common in the babbled utterances of infants at this age (*10*). The overall durations of the vowels, each about 1 second, are also similar. Sustained vowels of this duration, produced without consonant-like elements, are again infrequent in the babbled utterances of infants at this age (*10*). Such vocal productions suggest that infants are directing their articulators to achieve auditory targets that they hear another produce, in other words, that they are capable of vocal imitation (*11*).

We suggest that both the detection of auditory-visual correspondences and vocal imitation reflect a knowledge of the relationship between audition and articulation (*12*). Furthermore, we suggest that these abilities have a common origin—the infant's intermodal representation of speech. Future studies should test the extent to which auditory-visual and auditory-motor equivalence classes are related and the extent to which experience plays a role in their development.

Our findings go beyond these theoretical issues and extend to clinical concerns. On the basis of the data reported here and in other recent infant studies (*6, 7*), we suggest that the bimodal delivery of speech information may facilitate language learning because infants are predisposed to represent information in this way. In particular, infants born deaf might well be aided by the codelivery of visual and tactual information about speech. Such sensory substitution approaches have proven effective in improving speech reception in artificially "deafened" adults, who combine visual information provided by lip-reading with spectral information provided by a tactile aid (*13*).

Infant speech perception has traditionally been studied as an auditory phenomenon (*14*). Here we presented data and arguments showing that it may be profitable to investigate infant speech perception as an intermodal event. Studies of infants' intermodal organization of the auditory, visual, and motor concomitants of speech may bring us closer to understanding the development of the human capacity to speak and comprehend language.

PATRICIA K. KUHL
Department of Speech and Hearing Sciences and *Child Development and Mental Retardation Center, University of Washington, Seattle 98195*
ANDREW N. MELTZOFF
Department of Psychiatry and Behavioral Sciences and *Child Development and Mental Retardation Center, University of Washington*

References and Notes

1. N. Erber, *J. Speech Hear. Disord.* **40**, 481 (1975); H. McGurk and J. MacDonald, *Nature (London)* **264**, 746 (1976); Q. Summerfield, *Phonetica* **36**, 314 (1979); R. Campbell and B. Dodd, *Q. J. Exp. Psychol.* **32**, 85 (1980).
2. The average durations of the vowels, measured acoustically, were 1120 msec for /i/ (range, 1050 to 1220) and 1150 msec for /a/ (range, 1060 to 1270). The average center frequencies of the first three formants for /i/ were 416, 2338, and 2718 Hz; comparable values for /a/ were 741, 1065, and 3060 Hz.
3. The observer recorded when the infant was looking at the left or right visual display. Both inter- and intraobserver reliability was assessed. The mean difference in the percent-fixation scores for the left (or right) face was 3.3 percent (interobserver) and 1.8 percent (intraobserver).
4. B. Dodd [*Cognit. Psychol.* **11**, 478 (1979)] provided data suggesting that infants detect gross temporal misalignment between mouth movements and sound. While we argued that our alignment procedure effectively ruled out temporal cues as a potential explanation for the effect obtained in experiment 1, experiment 2 addressed this temporal hypothesis directly.
5. J. Bruner, *Cognition* **3**, 255 (1975); D. Stern, J. Jaffe, B. Beebe, S. Bennett, *Ann. N.Y. Acad. Sci.* **263**, 89 (1975).
6. A. N. Meltzoff and K. Moore, *Science* **198**, 75 (1977).
7. T. Bower, *Human Development* (Freeman, San Francisco, 1979); A. N. Meltzoff and R. Borton, *Nature (London)* **282**, 403 (1979). A. N. Meltzoff, in *Infancy and Epistemology*, G. Butterworth, Ed. (Harvester, Brighton, England, 1981), p. 85; ——— and K. Moore, in *Advances in Infancy Research*, L. Lipsitt and C. Rovee-Collier, Eds. (Ablex, Norwood, N.J., 1982), vol. 2.
8. Meltzoff and Moore (*6*) demonstrated imitation of oral gestures in infants less than 1 month old. The gestures were silently produced by an adult model. Since their mouth-opening gesture is similar to the articulatory posture adopted for the production of an /a/ vowel, these data support the hypothesis that infants are capable of imitating some speechlike gestures produced in the absence of vocalization.
9. P. Marler, *J. Comp. Physiol. Psychol.* **71** (1970); F. Nottebohm, *Am. Nat.* **106**, 116 (1972).
10. D. Oller, in *Child Phonology*, Vol. 1, *Production*, G. Yeni-Komshian, J. Kavanagh, C. Fer-

guson, Eds. (Academic Press, New York, 1980), p. 93; R. Netsell, in *Language Behavior in Infancy and Early Childhood*, R. Stark, Ed., (Elsevier, New York, 1981).
11. These observations underscore the need for careful experimental studies on the development of vocal imitation. Previous reports of vocal imitation in young infants have not provided acoustic analyses of either the model's or the infant's vocalizations [J. Piaget, *Play, Dreams and Imitation in Childhood* (Norton, New York, 1962); I. Uzgiris and J. Hunt, *Assessment in Infancy* (Univ. of Illinois Press, Chicago, 1975); W. Kessen, J. Levine, K. Wendrich, *Infant Behav. Devel.* **2**, 93 (1979)].
12. The motor theory of speech perception also outlined an argument in which the auditory and articulatory levels of representation were closely linked [A. Liberman, F. Cooper, D. Shankweiler, M. Studdert-Kennedy, *Psychol. Rev.* **74**, 431 (1967)]. Specifically, the model addressed the issue of speech-sound categorization in adults and argued that it was based on motor mediation. The infant data presented here do not bear on this issue. We posit that at a functional level, 5-month-old infants are cognizant of auditory-articulatory equivalence classes, and we do not suggest that the metric linking the two is defined in motor terms.
13. D. Sparks, P. Kuhl, A. Edmonds, G. Gray, *J. Acoust. Soc. Am.* **63**, 246 (1978).
14. P. Kuhl, in *Handbook of Infant Perception*, P. Salapatek and L. Cohen, Eds. (Academic Press, New York, in press).
15. We thank A. Anderson, D. Grieser, R. Baarslag-Benson, K. Merrick, P. Cameron, and C. Harris for assistance in the experiment and A. Ross, K. Lee, K. Mighell, and M. Sweeney for technical assistance. I. Hirsh, D. Klatt, J. D. Miller, K. Stevens, and Q. Summerfield provided helpful comments on an earlier version of this manuscript. Supported by NSF grant BNS 8103581 to P.K.K.; A.N.M. was supported by grants from the Spencer Foundation and the National Institute of Child Health and Human Development (HD-13024).

9 February 1982; revised 30 July 1982