INFANT BEHAVIOR AND DEVELOPMENT 7, 361–381 (1984)

The Intermodal Representation of Speech in Infants*

PATRICIA K. KUHL AND ANDREW N. MELTZOFF
University of Washington

Infants' abilities to detect auditory-visual correspondences for speech were tested in two experiments. Infants were shown two visual images side-by-side of a talker articulating, in synchrony, two different vowel sounds, while a sound matching one of the two vowels was auditorially presented. Infants' visual fixations to the two faces were video-recorded and scored by an independent observer who could neither see the faces nor hear the sounds. The results of Experiment 1 showed that the auditory stimulus systematically influenced infants' visual fixations. Infants looked longer at the face that matched the sound. In Experiment 2, the same visual stimuli were presented, but the auditory stimuli were altered so as to remove the spectral information contained in the vowels while preserving their temporal characteristics. Performance fell to chance. Taken together, the experiments suggest that infants recognize the correspondences between speech information presented auditorially and visually, and moreover, that this correspondence is based on the spectral information contained in the speech sounds. This suggests that infants represent speech information intermodally.

speech perception intermodal representation faces auditory stimuli
visual fixation cross-modal matching

Speech perception has traditionally been studied almost exclusively as an auditory phenomenon. Yet conversational speech is often produced by a talker we can both see and hear. What effect does visual information have on the perception of speech?

Research on adults suggests that the effect of the visual modality is considerable (Erber, 1975). Sighted people, both hearing-impaired and normal, demonstrate the ability to "lipread"—to derive linguistic information by watching a talker's mouth movements. While the clinical impact of this ability has long been recognized (Johnson, 1775), the importance of lipreading phenomena for theories of speech perception has often been overlooked.

Normal adults use visual information when they listen to speech in a noisy environment (Erber, 1969; Ewertsen & Birk-Nielsen, 1971; O'Neill, 1954; Sumby & Pollack, 1954), or when the auditory information in speech is degraded by filtering (Binnie, Montgomery, & Jackson, 1974; Sanders & Good-

*Portions of this research were presented at the 91st meeting of the Acoustical Society of America (Chicago, 1981) and a brief report appeared in Kuhl and Meltzoff (1982). The research was supported by a grant from the National Science Foundation (BNS 8103581) to P.K.K.; preparation of the manuscript was supported by this grant and by grants from the National Science Foundation to A.N.M. (BNS 8309224) and the Spencer Foundation to A.N.M. and P.K.K. Requests for reprints should be sent to Patricia K. Kuhl, Department of Speech and Hearing Sciences, University of Washington, Seattle, WA 98195.

rich, 1971). Listeners also rely on visual information when the speech of one talker is presented against a background of other talkers (Summerfield, 1979), as in a "cocktail party" situation.

The data show that there is a substantial contribution of visual information in each of these circumstances. For example, the Sumby and Pollack (1954) study found that seeing the face of a talker whose speech was presented in noise was equivalent to increasing the signal by 15–20 decibels. More recently, Grant, Ardell, Kuhl, & Sparks (in press) presented an isolated pure tone whose amplitude and pitch followed those of the fundamental frequency of a talker who was reading from text. The pure tone preserved the rhythm and stress pattern of speech and was perceived as voice-like. When the tone was presented in the absence of seeing the talker's face, no syllables, words, or phrases could be identified. However, when subjects saw the talker while listening to the tone, the spoken text became 80% intelligible. Thus, while visual information is not essential to the perception of speech (blind adults perceive speech normally), numerous studies demonstrate that vision can be and is used to derive information about speech under certain circumstances.

What kind of linguistic information does vision provide? Watching a talker's mouth movements provides two kinds of information: prosodic and phonetic. While careful work on the prosodic information available through lipreading has not been done, it appears that prosodic information in the form of syllable timing and rhythm is provided by the temporal sequences of mouth openings and closings. Since consonants are produced with a relatively constricted vocal tract and vowels with a relatively open vocal tract, the open-close cycle that results when consonant-vowel syllables are combined provides a rough visual marking of the boundaries of syllables (Erber, 1977). Syllabification is essential to the perception of stress and rhythm in speech.

Studies have also demonstrated that the visual channel provides phonetic information. The studies show that vision provides information about the "place of articulation" feature, one that distinguishes sounds like */b/, /d/,* and */g/*. These phonetic units differ in the location of the primary constriction in the mouth (*/b/* = the lips; */d/* = the alveolar ridge; */g/* = the velum), and these differences are detectable by eye when watching mouth movements. Other speech features, however, are not visually distinct. Vision does not provide information about sounds that differ in the "manner of articulation" when they occur at the same place of articulation (such as the sounds */p/, /b/,* and */m/*). These articulations involve the same primary constriction and thus are not visually distinguishable.

The availability of featural information through the auditory and visual modalities forms an interesting complimentary relationship. Research shows that speech information in the two modalities is differentially resistant to the effects of degradation, such that information that is subject to degradation in one modality is available in the other. As we suggested, place information is

available visually, while manner information is generally not. Conversely, while both place and manner information are normally available through the auditory channel, place information is much more subject to the effects of noise and/or filtering. Relatively slight increases in the background noise destroy place information auditorially, while manner information is highly resistant to these effects (Miller & Nicely, 1955). The implication is that the auditory fragility of place information can be effectively counteracted by its availability through the visual modality.

This is best illustrated in research on severely hearing-impaired listeners. Studies show that these individuals are capable of perceiving manner features such as "voicing" and "nasality" by ear, but that they are unable to distinguish the place feature auditorially (Erber, 1972). Specifically, the results demonstrate that severely impaired individuals can still distinguish voiced consonants (*/b, d, g/*) from voiceless consonants (*/p, t, k/*). However, the severely hearing-impaired listener cannot perceive place information auditorially, and thus the three place categories, involving bilabial (*/p, b, m/*), alveolar (*/t, d, n/*), and velar (*/k, g/*) sounds are not distinguished.

In essence, this means that when severely hearing-impaired listeners are auditorially presented with the consonant */p/* (a voiceless, non-nasal, bilabial sound) they correctly perceive the voiceless and non-nasal features, but are unable to perceive its place of articulation. They are therefore unable to identify the sound as */p/* as opposed to */t/* or */k/*. However, when these same hearing-impaired listeners are tested under conditions in which they watch the talker speak, the place feature can be identified by eye and performance on the consonant identification task is near perfect (Erber, 1972). Apparently, information leading to the identification of phonetic features can be independently extracted by the two modalities and then combined.

Current models of the speech perception process cannot account for the integration of auditory and visual information in the perception of speech. Nonetheless, such findings (see also McGurk & MacDonald, 1976; Summerfield, 1979) raise central questions about the representation of speech. At a minimum, they suggest that speech perception is not solely the province of audition. Rather, they suggest that information about speech can be picked up by different modalities and integrated by perceptual mechanisms to form a phenomenally unified phonetic percept. While this much is demonstrated, the manner in which the different modalities interact and the form in which speech sounds are represented to allow this interaction (in articulatory terms, auditory terms, or some more abstract phonetic representation that is not exclusively auditory or articulatory) is still unknown.

From a theoretical standpoint, the fundamental importance of these classic lipreading studies is that they suggest that, for adults, speech information derived from the visual modality can substitute for speech information derived from the auditory modality. They suggest, for example, that the visual

perception of place of articulation (conveyed by the configuration of the lips, tongue, and jaw) can substitute for the auditory perception of place of articulation (conveyed by the configuration of formant frequencies). The question is, how is information that is processed by two separate modalities—audition and vision—equated in speech perception?

One possibility is that adults, through a protracted period during which they both watch and listen to others speak, learn to associate the auditory and visual concomitants of speech. That is, they learn that an auditory /b/ is accompanied by a visible closure of the lips, and so on. If this were the case, then young infants who have not had a long period during which to learn the association between the auditory and visual concomitants of speech would be unable to relate them.

The aim of this experiment was to begin to explore the development of auditory-visual speech perception. We asked whether 4-month-old infants recognized that sounds of a particular type were emitted by mouths moving in a particular way. Specifically, our question was whether infants could detect cross-modal correspondences for speech presented to the auditory and visual modalities.

The problem was posed by showing infants two faces, side-by-side, articulating two different vowel sounds. A sound track matching one of the faces was auditorially presented. We hypothesized that the auditorially presented signal would systematically influence infants' visual preferences. Specifically, we suggested that if infants recognized the correspondence between articulatory gestures and their auditory consequences, they would look longer at the face whose articulatory movements matched the sound presented.

An initial report of the cross-modal speech perception effect was provided by Kuhl and Meltzoff (1982). The purpose of the present paper is to provide full methodological details of the stimulus preparation, experimental procedure, and results, along with a more complete analysis of the theoretical implications of these findings.

EXPERIMENT I

METHODS

Subjects

Thirty-two infants served as subjects. They had no known visual or auditory abnormalities. They ranged in age from 18.0 weeks to 20.1 weeks ($M = 19.3$). Participation in the experiment was solicited by a letter that was sent to the parents of newborns in the Seattle area. Interested parents returned a postcard, which provided details regarding birth and family medical history. Infants who were preterm, low birth weight, or otherwise at-risk for normal development were not tested. An additional 10 infants failed to complete testing due to crying (5), falling asleep (2), or equipment failure (3). Parents were paid $3 for their participation in the study.

Stimuli

An Auditory-Visual Speech Perception technique (AVSP) was developed to present stimuli to the infants. Using this technique, infants are shown two filmed images, side-by-side, of a female talker articulating in synchrony two different vowels. The sound track corresponding to one of these vowels is presented through a loudspeaker directly behind the screen and midway between the facial images. In this study we wanted to examine infants' knowledge that particular types of speech sounds are produced by mouths moving in particular ways. In order to test this point, we had to rule out the possibility that infants might detect face-sound correspondences that were based purely on temporal grounds. Thus, the two visual images had to be presented in perfect temporal synchrony with one another. The two mouths had to open and close at the same time. Moreover, the sound track had to be aligned with the filmed images so that the sound was temporally synchronous, and equally so with both the "matched" and the "mismatched" mouth movements (Kuhl, in press, a).

Filming and selecting temporally matched stimuli. A film was made in a studio of a female talker producing the vowels /a/ (as in "pop") and /i/ (as in "peep"). The talker placed her head through an oval hole in a black velvet cloth so that only her face was filmed. A 16-mm camera (Arriflex BL) recorded the image on color film. Audio recordings were made on a high-quality tape recorder (Nagra, IV-L) and transferred to 16-mm magnetic sound track.

The talker produced the vowels once every 3 s and attempted to articulate them with equal intensity and duration. Rather than using a single production of /a/ and a single production of /i/ and trying to match them on all noncritical dimensions (visual extent of mouth movement, visual rate of mouth movement, auditory intensity, auditory duration, and fundamental frequency), we used a number of different productions to represent the /a/ and /i/ categories. The stimuli selected for use in the experiment were chosen so that the noncritical dimensions fell within a narrowly restricted range that overlapped for the two vowel categories. This procedure helped ensure that recognition of a correspondence between face and sound could not be based on an idiosyncratic property of a single production of the vowels.

Two sequences of 10 /a/'s and two sequences of 10 /i/'s were chosen as stimuli from among the filmed images. They were used to make two film loops. One loop displayed the /a/ face on the right and the /i/ face on the left; the second loop reversed this left-right orientation. The facial images were chosen such that the durations of the individual articulations fell within a narrowly defined range that overlapped for the two vowel categories. The duration of each visual stimulus was measured to specify the length of time that the lips were parted. The average duration of /a/ mouth movements was 1.92 s (range, 1.75–2.08 s). The average duration of /i/ mouth movements was 1.97 s (range, 1.79–2.13 s).