MATCHING PHONETIC INFORMATION IN LIPS AND VOICE
IS ROBUST IN 4.5-MONTH-OLD INFANTS

by

Michelle Louise Patterson

B.A. (Honours), Queen's University, 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

Department of Psychology

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

May 1998

© Michelle Louise Patterson, 1998

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of PSYCHOLOGY

The University of British Columbia
Vancouver, Canada

Date July 28, 1998.

Abstract

Past research (Kuhl & Meltzoff, 1982; 1984) claims that 4.5-month-old infants can match phonetic information in the lips and voice. These studies used female faces poking through cloth to occlude possible distractions. Attempts to replicate these findings have not produced convincing results. The present studies were conducted to replicate and extend past research by examining how robust the ability to match phonetic information in lips and voice is at 4.5-months of age. If speech is represented intermodally in young infants then they should show evidence of matching with more *ecologically valid* visual stimuli and also with *male* faces and voices. Also, more infants might be expected to imitate the vowels when the model's lips and voice match than when they do not match. Sixty-four infants were seated in front of two side-by-side video monitors displaying filmed images of a female or a male face, each articulating a different vowel sound (/i/ or /a/) in synchrony. The sound track was played through a central speaker and corresponded to one of the two vowels but was synchronous with both. Infants spent approximately equal amounts of time looking and smiling at both the female and the male faces ($p > .05$). However, infants looked longer at the face that matched the heard vowel for both female and male stimuli ($p < .01$). Also, infants showed articulatory imitation in response to the matching face/voice stimuli ($p < .05$). The finding that bimodal phonetic matching holds with a more ecologically valid face and with male stimuli supports the hypothesis that infants are able to link phonetic information presented in the lips and voice. This suggests an integrated, multi-modal representation of articulatory and acoustic phonetic information at 4.5-months of age.

TABLE OF CONTENTS

ABSTRACT	ii
TABLE OF CONTENTS	iii
LIST OF FIGURES	iv
ACKNOWLEDGMENTS	v
INTRODUCTION	1
Matching speech sounds and articulation	2
STUDY 1	7
Method	7
Participants	7
Stimuli	7
Filming and selecting stimuli	8
Synchronizing the stimuli	9
Equipment and Test Apparatus	10
Procedure	10
Scoring	11
Results and Discussion	12
STUDY 2	13
Method	15
Participants	15
Apparatus and Procedure	15
Results	15
GENERAL DISCUSSION	17
REFERENCES	24
FIGURES	29
APPENDIX	35

LIST OF FIGURES

<u>Figure 1.</u> Diagram of testing and control rooms.....	30
<u>Figure 2.</u> Flow diagram of sequence of events experienced by one infant in one condition.....	31
<u>Figure 3.</u> Percentage of total looking time (PTLT) to the face that matched the heard vowel for female and male stimuli.....	32
<u>Figure 4.</u> Number of infants who looked longer at the face that matched the heard vowel for female and male stimuli.....	33
<u>Figure 5.</u> Mean duration (sec) of infant articulatory imitation when adult lips and voice presented /a/ or /i/.....	34

ACKNOWLEDGMENTS

I was very lucky to be able to work with an excellent team of people who provided the guidance, support, feedback, and encouragement that enabled me to complete this thesis. First of all, I thank the parents who took the time to bring their infants into the lab; without their generosity this work would not have been possible. I also thank Sharon, Monika, and Jennifer for all their help baby-sitting and recruiting infants. I appreciate the support and helpful suggestions from other members of my lab (thank you Chris, Judi, Renee).

Creating the stimuli and setting up the equipment for this work was very frustrating at times. My deepest thanks go out to Ray Hall, Ian Hinkle, and Steve Ratcliffe for their technical knowledge and assistance. Thank you to Brenda, Christine, Erina, Sean, Matt, and Joe for posing as faces and voices.

My deepest thanks goes to my thesis supervisor, Dr. Janet Werker, whose enthusiasm, support, and insightful feedback made this thesis a truly enriching experience for me. She always found the time to counsel me through a set back or go over my revisions. Thank you for your encouragement and patience. Thank you to my committee members, Charlotte Johnston and Richard Tees, who provided thoughtful comments and suggestions.

Last, but certainly not least, thank you to my friends and family who listened and encouraged me at every step.

Thank you all, I am sincerely grateful.

This research was supported by an NSERC PGS-A graduate fellowship.

INTRODUCTION

It is now well-established fact that there is more to speech than meets the ear (e.g., McGurk & MacDonald, 1976; Summerfield, 1979, 1987). The movements of the articulators are visible concomitants of speech events and have strong influence over adult speech perception (Dodd & Campbell, 1984; Green & Kuhl, 1989; Massaro & Cohen, 1990). Visible articulatory information also influences speech perception in children with the extent of influence changing as a function of the child's age (McGurk & MacDondald, 1976) and articulatory ability (Desjardins, Rogers & Werker, 1997). Clearly, speech is a multi-modal phenomenon. It has been shown that the ability to recognize audio-visual temporal correspondence for speech exists very early in life (e.g. Dodd, 1979) and there is increasing evidence that infants may also be able to recognize matches between the phonetic information in the lips and voice (Kuhl & Meltzoff, 1982; 1984; 1988; Walton & Bower, 1992). The present studies will examine how robust the ability to match phonetic information in the lips and voice is at 4.5-months of age, as well as the extent to which this ability is affected by stimulus complexity and speaker gender. Although speech may be a special case (Lieberman & Mattingly, 1985), understanding how audio and visual information are organized to form a unified speech percept and whether there is a common metric that recognizes the equivalence between information entering different sensory channels is a central issue in theories of intermodal perception (see Lewkowicz & Lickliter, 1994).

In general, infants' ability to pick up "amodal" cues specifying intermodal relations between objects and sounds has been well documented (Bahrick, 1983; 1988; Humphrey, Tees, & Werker, 1979; Spelke, 1976; 1979). Amodal information is tied to the structural properties of an action or event but is not specific to a particular sensory modality. That is, the same information can be detected by several modalities. Examples include temporal information, such as rhythm or tempo, and

object properties, such as size, shape, texture, and substance (Walker-Andrews, 1994). Of particular interest to my thesis is infants' sensitivity to the intermodal relations in phonetic information which are embedded within the larger event of speaking.

Soon after birth, infants are able to discriminate and categorize vowel sounds. Trehub (1973) showed that infants between 1- and 4-months of age can discriminate changes between /a/ and /i/ and between /u/ and /i/ when these vowels are isolated or follow a common consonant. Marean and Werner (1992) successfully trained 2-month-olds to respond when the vowel changed from /a/ to /i/ and to refrain from responding when the vowel did not change category despite variation in the spectral cues associated with pitch and talker. Similar results have been shown for infants aged 4-months (Kuhl, 1983). Therefore, infants can discriminate vowel categories that are quite similar acoustically at an early age. Although it is clear that young infants can discriminate features within different modalities, there is much less known about when and how knowledge of the intermodal nature of speech is acquired by infants.

Matching speech sounds and articulation

There is some evidence that infants are able to match equivalent information in facial and vocal speech. Dodd (1979) found that 2- to 4-month-olds looked longer at a woman's face when the speech sounds and lip movements were in synchrony than when they were asynchronous. However, detection of synchrony alone does not reveal knowledge of the match between phonemes and articulatory movements.

Kuhl and Meltzoff (1982) conducted the first study that specifically examined infants' ability to detect a match between articulatory movements and vowel sounds. Using the preferential looking technique, Kuhl and Meltzoff presented 4.5-month-olds with side-by-side, filmed images of a woman articulating the vowels /i/ and /a/ in synchrony. The woman's face was framed by black cloth to occlude neck,

hair, and ears. On average, infants spent 73% of the total looking time on the matching face, and 24 of the 32 infants looked longer at the matched face. These results were replicated with a new set of 4.5-month-olds (Kuhl & Meltzoff, 1984) and with a new pair of vowels (/i/, /u/) (Kuhl & Meltzoff, 1988); however, in both cases the effect was slightly weaker (64% of total looking time spent on matching face). Although, it appears that young infants can relate speech information presented acoustically and visually, it is not clear how robust this ability is at 4.5-months of age.

In order to identify the stimulus features that are necessary and sufficient for detecting cross-modal equivalence, Kuhl and colleagues (Kuhl & Meltzoff, 1984; Kuhl, Williams, & Meltzoff, 1991) selected pure tones of various frequencies that isolated a single feature of a vowel without allowing it to be identified. Adults and 4.5-month-old infants were presented with tasks that assessed their abilities to relate the pure tones to the articulatory movements for /i/ and /a/. Although adults could still detect the match, infants showed no preference for the match when the auditory component was reduced to simple tones. These results suggest that the gaze preferences observed with intact vowel stimuli were not based on simple temporal or amplitude commonalities between audio and visual streams, but rather were likely based on matching of spectral information contained in the auditory component with articulatory information in the lips. According to Kuhl et al., because spectral information, unlike temporal and amplitude information, depends largely on articulatory changes, sensitivity to the relationship between spectral information and visual speech is based on linking phonetic features picked up by both ear and eye. Therefore, 4-month-olds may be sensitive to specific natural structural correspondences between the acoustic and visual properties of articulation.

Infants' ability to detect the match between mouth movements and speech sounds has been extended by two independent studies, but the results were neither strong nor convincing. MacKain, Studdert-Kennedy, Spieker, and Stern (1983) presented infants with simultaneous displays of two women articulating three pairs of CV disyllables in a repeated-measures design. Infants between 5- and 6-months of age looked longer at the sound-specified display, but only for three of the six disyllables and only when the matched face appeared on the right-hand side. The authors suggest that looking to the right side facilitates intermodal speech perception and thus indicates that, in infancy, the left hemisphere is predisposed to process cross-modal speech-speaker correspondences. However, asymmetries of lateral gaze have not been validated as an index of cerebral lateralization (see Rose & Ruff, 1987). Clearly, independent replication of the original Kuhl and Meltzoff (1982) study is needed.

Walton and Bower (1992) replicated Kuhl and Meltzoff's (1988) bimodal matching effect with /i/ and /u/ using an operant-choice sucking procedure. Four-month-olds sucked more to receive the appropriate face-voice pairings. In a second study, the /u/ face was paired with the /u/ sound (match), the /i/ sound (impossible), or a French /y/ sound (possible but unfamiliar). Six- to 8-month-olds sucked significantly more to receive both possible face-voice pairs than to receive the impossible pair. In fact, novelty seemed to enhance the appeal of articulatory possibility, with greatest mean number of sucks to the French /y/. Thus familiarity does not appear to be the cause of infants' preference for matched face-voice pairings. Walton and Bower also reported that results from "first look" data suggest that infants did not have to work out that a mismatch was impossible, but that it was perceived rapidly. However, infants' ability to discriminate a French /y/ from an English /u/ was not tested, thus the basis for infant's performance in this task is unclear.

Indirect evidence for infants' sensitivity to a speaker's mouth movements has been obtained from studies of infants' imitation of vocalizations. Matching of pitch prosody and phonetic structure has been previously documented in young infants (Kuhl & Meltzoff, 1988; 1991; Papousek, Papousek, & Bornstein, 1985). Two-month-old infants are capable of pre-speech mouth movements and imitate other mouth movements, such as tongue protrusion, lip pursing, and mouth opening (Meltzoff & Moore, 1983). Legerstee (1990) examined the role of audition and vision in eliciting early imitation of speech sounds. Audio and visual components of the vowels /a/ and /u/ were presented separately to 3- to 4-month-old infants. In one condition, vowel sounds were presented to infants through headphones while an adult mouthed the matching visual component of the sound in synchrony. In a second condition, a different group of infants heard either /a/ or /u/, but the adult mouthed the mismatching visual component in synchrony. Only infants who were exposed to matching audio-visual information were observed to imitate the vowels.

Infant imitation was a serendipitous finding in the original Kuhl and Meltzoff (1984) study. It was noted that 10 out of 32 infants showed evidence of vocal and/or facial imitation of the adult female, however, these findings were not formally analyzed. In a later study, Kuhl et al. (1991) observed that infants differentially imitated speech versus nonspeech sounds; moreover, when infants were listening to speech they were most likely to imitate the auditory signals they saw (Kuhl & Meltzoff, 1988).

Together, results from studies of infants' detection of audio-visual structural correspondences and vocal imitation provide converging evidence for the intermodal organization of speech in early infancy and have important implications for language development. However, independent replication of Kuhl and Meltzoff's (1982) bimodal matching effect is required. If prelinguistic infants are sensitive to the temporal and structural congruence of audio-visual speech

information then a stronger argument can be made for the existence of invariant phonetic information across both facial and heard speech and/or a specialized module which facilitates the integration of audio-visual speech information (Liberman & Mattingly, 1985). It seems likely that intermodal perception may play a critical role in the development of speech perception and production. Of particular interest, if invariant phonetic information is available in both face and voice, one would expect infants to match visual and acoustic information about speech with more ecologically valid stimuli (full heads, including hair, neck, and shoulders). Furthermore, if infants are able to detect invariant phonetic information then they should match phonetic information in the lips and voice of both female and male talkers.

The present studies were designed to replicate and extend Kuhl and Meltzoff's bimodal matching effect. Specifically, we examined how robust the ability to match audio and visual vowels is in 4.5-month-old infants. If infants do perceive unified percepts based on detection of amodal attributes and/or correspondences, it is essential to determine to what information infants are sensitive and to describe the limits and mechanisms underlying such sensitivity. In this endeavor, we examined whether infants' ability to match phonetic information in lips and voice is affected by stimulus complexity and speaker gender. To address the question of stimulus complexity we tested infants in a procedure very similar to that used by Kuhl and Meltzoff except the full head, including shoulders, neck, and hair, was visible. To test the effect of speaker gender we tested infants using both female and male stimuli. Acts of articulatory imitation were analyzed to see if infants imitate differentially to matching versus mismatching audiovisual stimuli and/or to female versus male faces and voices.

Study 1: Matching a Female Face and Voice

The first study was designed to replicate Kuhl and Meltzoff's (1982; 1984) finding that 4.5-month-old infants can match seen and heard information specifying vowels. The procedure was, therefore, very similar to the preferential looking paradigm used by Kuhl and Meltzoff, with a few exceptions. Stimuli were presented on two different video monitors instead of on one large projection screen, however, the distance between faces was approximately the same. In order to make the stimuli more ecologically valid, the woman's entire head, including hair, neck, and shoulders was visible to the infant, instead of having the face framed by black cloth.

Method

Participants

Mothers were recruited from a local maternity hospital shortly after giving birth or responded to an advertisement in the local media. Informed consent was obtained from all parents who participated with their infants (see Appendix). The final sample consisted of 32 infants, 16 male and 16 female, ranging in age from 17.7 to 20.6-weeks ($M = 19.2$ weeks, $SD = 2.4$ weeks). An additional 20 infants were tested and excluded from analyses for the following reasons: cried/fussy (9), did not look at both stimuli during familiarization period (4), looked at same screen for entire test phase (2), equipment failure (5). Infants had no known visual or auditory abnormalities, including recent ear infections. Infants who were preterm (< 36 weeks), low birth weight (< 2500 grams), or otherwise at-risk for normal development were not tested.

Stimuli

A computer-based multi-media authoring package (mTropolis) was used to combine, control, and present the digitized audio and visual stimuli. Using this technique, infants were shown two filmed images displayed on separate, side-by-side computer monitors of a female face articulating two different vowels in synchrony.

The sound track corresponding to one of the vowels was presented through a speaker midway between the two images. In order to examine infants' knowledge that particular types of speech sounds are produced by particular mouth movements, the possibility that infants might detect face-sound correspondences based purely on temporal cues had to be ruled out. The two visual images were presented in synchrony and the sound was aligned with the images so that it was temporally synchronous, and equally so, with both the "matched" and the "mismatched" mouth movements.

Filming and selecting stimuli

Stimuli were made simultaneously for both Study 1 and 2. In a recording studio, we used a Sony Beta camera (SP-637) to record the images on colour film of a female and a male producing the vowels /a/ and /i/ against a black background. Male and female faces were selected for similar colouring and attractiveness (both were Caucasian and fair-haired). The female had shoulder-length hair and the male's hair came to the bottom of his ears. Both male and female wore a white turtleneck and neither wore any jewelry or make-up. First, the male was filmed, producing the vowel /a/ to the beat of a metronome set at 1 beat per 3 s, for a total of 2 min of recording. He was instructed to articulate each vowel with equal intensity. This recording was then played back over a TelePrompter, and all other vowels (male /i/ and female /a, i/) were produced in synchrony with the male's /a/.

As in Kuhl and Meltzoff (1984), the audio stimuli used in the experiment were not those emanating from the visual articulations. This helped ensure that no idiosyncratic characteristics of a particular production would link the audio and visual displays. A different male and female were selected to record the audio stimuli. Audio recordings were made in a sound-proof recording booth using a studio-quality microphone and were recorded onto audio tape (Sony Betacam SP-

637). Speakers were asked to articulate the vowels /i/ and /a/ in infant-directed speech with equal intensity and duration.

One visual /a/, one visual /i/, and one instance of each vowel sound for both the female and the male stimuli were chosen by three judges who rated what they deemed to be the five best visual and audio stimuli¹. The facial images were chosen such that the duration of the individual articulations fell within a narrowly defined range that overlapped for the two vowel categories, the head did not move, and one eye blink occurred after each articulation. The duration of each visual stimulus was measured to specify the length of time that the lips were parted. For the female, this duration was .94 s for /a/ and .65 s for /i/. For the male, this duration was 1.27 s for /a/ and 1.28 s for /i/.

A comparable process was used to choose the sequences of /a/'s and /i/'s from the audio recording. The duration of the vowel sounds was important since it is possible for the duration of the mouth opening to be longer than the sound duration but not vice versa. Therefore, we ensured that the vowel sounds were of the same or shorter duration than the mouth opening. For the female, the duration of the sound was .61 s for /a/ and .63 s for /i/. For the male, the duration of the vowel sound was .62 s for /a/ and .73 s for /i/.

Synchronizing the stimuli

Once the audio and visual stimuli were chosen, the sound tracks had to be synchronized with each of the faces. Using a studio-quality editing bench (Sony Betacam Recorder/Player UVW-1600; Mackie sound board 1202-VLZ), the sound was dubbed onto each visual stimulus at the point of maximal mouth opening. The films were then digitized and entered into a multi-media computer programme

¹ Kuhl and Meltzoff (1982; 1984) chose 10 audio and visual /a/s and 10 audio and visual /i/s to make two film loops. Since we digitized the audio and visual stimuli onto an I-Omega CD, the file size was limited such that we could only pick three instances of each audio and visual stimulus. When transferring these files to the multimedia authoring programme, we chose one instance of each audio and visual stimulus to speed up the running of the programme and to reduce the likelihood of crashing.

which locked the appropriate faces in phase. Each articulation was repeated to form a continuous series of articulations occurring once every 3 s. When displayed on the monitors, the faces were approximately life-size, 17 cm long and 12 cm wide. Their centres were separated by 41 cm. The sounds were presented at an average intensity of 60 ± 5 dB SPL.

Equipment and Test Apparatus

The study was conducted in the Infant Studies Centre at the University of British Columbia. As illustrated in Figure 1, the control room housed a Power Mac 7300/180 computer, a video recorder, and a television monitor. The stimuli were projected on two 17" colour video monitors in the testing room. Black curtains covered the entire wall so that only the monitor screens and the camera lens were visible. The infant was seated in an Evenflo infant seat secured to a table and the caregiver was seated to the infant's right. When seated, infants were 46 cm from the facial displays. The speaker (Sony SRS-A60) was located behind the black curtain and centred midway between the two facial images. During testing, a 60-Watt light in a lamp shade (to yield soft, diffuse light) was suspended 1 m 10 cm above the infant. A camera (JVC GS-CD1U) was positioned behind a small hole located between and above the two monitors. A mirror was positioned on the back wall of the testing room so the camera (JVC-BR1600U) would record the baby as well as what was appearing on the monitors on the video recorder.

Procedure

The experimental procedure, very similar to that described by Kuhl and Meltzoff (1984), involved two phases: Familiarization and Test. During Familiarization, each visual stimulus (the /a/ and the /i/ face) was presented individually without sound for 9 s. After this 18 s period, both faces were presented

simultaneously without sound for 9 s². Both stimuli were then occluded for 3 s before the Test phase began. During the 2 min Test phase, both faces were presented (one on each monitor) and one sound (either /a/ or /i/) was played. The sound presented to the infant, the left-right positioning of the two faces, the order of familiarization, and infant sex were counterbalanced. Therefore, infants who were presented with the female stimuli were randomly and equally assigned to one of eight different conditions (see Figure 2).

Scoring

Coding was performed using a Panasonic S-VHS AG-1960 recorder which allowed frame-by-frame analysis. Coders were blind to the stimuli presented to the infant. Interobserver reliability was assessed by rescoring 25% of the subjects. The percentage agreement for each second in the sampled periods ranged from 96.8% to 99.2%.

For the Familiarization and Test periods, duration of gaze was scored for each second when the infant appeared, on the video recording, to be looking either to the right or to the left of midline. Individual gaze-on seconds were summed for each display and divided by the total time spent looking at the displays to obtain the percentage of total looking time (PTLT) that the infant looked at each display during the test period.

Finally, video tapes were scored for evidence of infant articulatory and/or vocal imitation of the vowels presented, both acoustically and visually. For articulatory imitation, coders recorded the duration of the mouth movement if the mouth was wide open (/a/), if the lips were spread (/i/), and if the cheeks were lifted and the mouth upturned (smile). All other mouth movements (e.g. sucking,

² Kuhl and Meltzoff (1982;1984) did not include the simultaneous presentation of both faces in their familiarization phase. This phase is typically included in studies of infant word comprehension (Hirsh-Pasek & Golinkoff, 1992). The logic behind including this phase is to teach infants that both displays can be on simultaneously. Also, it is sometimes used as a check for infant side bias.

spitting, tonguing, etc.) were not recorded. For vocal imitation, coders recorded the duration of vocalization if it sounded like "aw" or "ee". All other infant sounds (e.g. crying, lip smacking, etc.) were not recorded.

Results and Discussion

Infants looked longer at a particular face when the appropriate vowel sound was heard. Overall, infants spent 79.4% of the testing time looking at the faces. On average, infants spent 64.8% of the total looking time on the matching face, which was significantly greater than a chance value of .50, $t(31)=3.27$, $p<.01$, as shown in Figure 3. Of the 32 infants tested, 25 looked longer to the sound specified display than to the incongruent display, as shown in Figure 4; this is significantly different from chance ($p<.001$), according to a binomial test. Other factors counterbalanced in the design were entered into five separate one-way ANOVAs; none of these factors were significant. Infants did not look significantly more to the right versus the left screen, $F(31)=.707$, $p>.05$, they did not prefer to look at the /a/ face versus the /i/ face, $F(31)=.187$, $p>.05$, they did not prefer the face shown last during familiarization, $F(31)=1.89$, $p>.05$, nor were there any significant differences between male and female infants, $F(31)=.671$, $p>.05$. Furthermore, contrary to McKain et al.'s (1983) results, infants did not look significantly longer at the matching face when it was on the right versus the left side, $F(31)=.732$, $p>.05$. Out of 32 infants tested, 17 showed evidence of articulatory imitation and three showed evidence of vocal imitation. Because the numbers of infants imitating is small, analyses will be conducted after combining data from both Studies 1 and 2.

These results support prior findings that the heard vowel influences infants' visual preferences (Kuhl & Meltzoff, 1982; 1988; Walton & Bower, 1992). Infants between 4.5- and 5-months of age looked longer at the synchronized display of a female face producing articulatory patterns that specified the speech they heard than

at an alternative, synchronized display of the same female face displaying a different, mismatching articulatory pattern. Therefore, even when the visual stimulus is more complex (i.e., includes the full head, including, hair, ears, neck, and shoulders), infants can match the lip movements that distinguish the vowel sounds /a/ and /i/ with the appropriate vowel sound. The fact that infants are able to perform such audio-visual matches at such a young age (i.e., with fairly limited experience with faces and voices) supports claims that infants may be born with some form of neural circuitry that facilitates learning of seen and heard speech. However, these results do not address the possibility that matching phonetic information in the lips and voice may be largely influenced by experience with faces and voices. Infants may have learned to associate and integrate modality-specific face-voice information in their own environments and generalize such arbitrary relations to the woman in our experiment. If this is the case, infants need not be predisposed to detect audio and visual concomitants of speech.

One way to clarify the roles of nature versus nurture in this case may be to compare young infants' responses to male versus female stimuli. In general, infants have had more exposure to female compared to male faces and voices; therefore, if the bimodal matching effect is based on an arbitrary relationship that is learned with experience, one would expect a substantially weaker effect with male stimuli than with female stimuli.

Study 2: Matching Phonetic Information in a Male Face and Voice

Male and female voices differ in terms of fundamental frequency and formant structure (Ladefoged, 1993), therefore, it is important to test whether cross-modal matching of articulatory gestures and speech sounds generalizes beyond vocal characteristics specific to the female voice. In utero, infants are exposed to the maternal voice and this exposure has been shown to influence post-natal

preferences. For example, newborns will suck to hear their mother's voice (DeCasper & Fifer, 1980) but not their father's voice (DeCasper & Prescott, 1984). Even after birth, most infants receive more exposure to female faces and voices. Of the 52 infants tested in Study 1, in only one case was the father the primary caregiver and three babies spent approximately equal amounts of time with both parents.

There is reason to expect that experience might play a role in matching phonetic information in the lips and voice, at least in the perception of consonants. DeJardins and Werker (submitted) found that some 4-month-old infants showed evidence of perceiving an integrated percept when presented with mismatching audio and visual CV syllables, however, integration was neither strong nor consistent. Similarly, Desjardins et al. (1997) found that, compared to preschoolers who did not make production errors, children who made production errors on consonants performed poorly when lip-reading consonants in a visual-only condition and were less influenced by the visual component in an audiovisual speech perception task. Finally, Siva, Stevens, Kuhl, and Meltzoff (1995) tested adults with cerebral palsy who could not produce speech in an audiovisual speech perception task. These individuals were influenced only by the auditory stimulus; thus auditory /aga/ - visual /aba/ was perceived as /aga/. Adult controls reported the expected 'combination' percept of /abga/. These findings suggest that experience producing speech influences the ability to perceive some audiovisual stimuli.

If infants truly perceive structural correspondences between visual and auditory speech, they should be able to perform cross-modal matches with both female and male stimuli. However, if experience plays a significant role in the ability to match phonetic information in the lips and voice, the matching effect may not be as strong with male faces and voices.

Method

Participants

Infants were recruited in a manner identical to Study 1. The final sample consisted of 32 infants, 16 male and 16 female, ranging from 16.5 to 20.5 weeks ($M=19.5$ weeks, $SD=3$ weeks). An additional 15 infants were tested but excluded from analyses for the following reasons: cried/fussy (4), did not look at both stimuli during familiarization (5), locked onto one screen during test phase (1), equipment failure (4), mother interfered (1).

Apparatus and Procedure

Stimuli were presented in the same manner as in Study 1. The only difference was that the male face and voice was used instead of the female face and voice. All other aspects of the equipment, testing, and scoring procedures were identical to Study 1. Inter-observer reliability was again calculated after rescoring 25% of the subjects. The percent agreement for infant looking during each second in the sampled periods ranged from 95% to 100%.

Results

The primary dependent variable was the percentage of the total looking time (PTLT) spent looking at the matching face. As in Study 1, several "nuisance" variables were examined to see if they contributed to differences in looking times to the matching face. Side of vowel presentation, side of match, order of familiarization, and infant sex were each entered into separate one-way ANOVAs. The results of these analyses were all nonsignificant ($p>.05$), thus these variables did not contribute to differences in looking time and were thus dropped from the main analysis.

Overall, infants spent 73.1% of the testing time looking at the faces. As in Study 1, infants spent significantly more time looking at the matching face than would be expected from chance responding. As shown in Figure 3, on average,

infants spent 62.7% of the total looking time on the matching face, which is significantly different from chance responding, according to a single-sample t test, $t(31)=3.02$, $p<.01$. At the individual subject level, 24 of the 32 infants looked longer at the sound specified face (see Figure 4), which is significant ($p<.05$) according to the binomial test.

To examine any difference in PTLT spent on the "match" with male versus female stimuli, data from both Study 1 and 2 were entered into a 2-between, fixed-effects ANOVA with Sound (/a/, /i/) and Speaker Gender as the main effects. Combining the data from both studies afforded a considerable increase in power while also allowing analysis of the effects of Speaker Gender separately. The main effects of both Sound, $F(63)=.061$, $p>.05$, and Speaker Gender, $F(63)=.109$, $p>.05$, as well as the interaction between these variables, $F(63)=.779$, $p>.05$, were nonsignificant. Therefore, across both studies, infants did not look significantly longer at the /a/ versus the /i/ face and did not look significantly longer at the male match versus the female match. Thus, young infants who have had more experience with female faces and voices are able to determine which facial articulation matches a heard vowel sound even when the face and voice are male.

In our preferential looking procedure, infants could imitate when looking at the matching face and voice or when looking at the face that did not match the voice. Out of 64 infants, 33 showed evidence of imitation when the face and voice matched and 8 showed evidence of imitation when the face and voice did not match. Because so few infants imitated in the mismatch condition and because imitation in this case is difficult to interpret, only infant imitation that occurred when the face and voice matched was analyzed. Out of the 33 infants who made mouth movements to the matching face and voice, 25 imitated the vowel they saw and heard while 8 infants produced a different vowel. As illustrated in Figure 5, infants spent significantly more time imitating the vowel that matched the lips and

voice than they did producing mouth movements that did not match the lips and voice, $t=2.10$, $p<.05$. Infants spent an equal amount of time imitating both /a/ and /i/ and also spent approximately equal amounts of time smiling to the female and male face ($p>.05$). While duration of smiling to the match versus the mismatch was not significantly different ($p>.05$), more infants smiled at the audio-visual match ($n=12$) versus the mismatch ($n=5$; $p<.05$). Due to the fact that coding of infant vocalization could not be done without being exposed to the sound that the infant was hearing and only 7 out of 64 infants showed evidence of vocalizing, infant vocal imitation was not analyzed.

General Discussion

When given a choice between two identical faces, each articulating a different vowel sound in synchrony, infants between 4.5- and 5-months of age looked longer at both a female and a male face that corresponded with the heard vowel sound. Also infants spent more time producing vowel articulations that matched the lips and voice compared to articulations that did not match. These findings confirm and extend prior reports that infants as young as 4.5-months of age can detect a match between acoustically presented vowel sounds and appropriate facial articulation (Kuhl & Meltzoff, 1982; 1988; MacKain et al., 1983; Walton & Bower, 1992).

As in Kuhl and Meltzoff (1982), the current studies found no preference for the /i/ versus the /a/ face, no right/left side preference, no infant sex differences, and no difference in looking time when the match was on the left versus the right side. This latter finding does not support MacKain et al.'s (1983) finding that matching only occurred when the match was on the right-hand side. It may be the case that the capacity to begin reproducing native language speech sounds in prelinguistic babbling rests on a predisposition of the left hemisphere to recognize the sensorimotor connections between the auditory structure of speech and its

articulatory source, however this is not necessarily reflected in visual preferences at 4-months of age.

It is also possible that the observed effect of symmetrical matching can only be achieved with simple vowel stimuli. Thus the CV disyllables used by MacKain et al. may have been too complex or too rapidly changing for 5-month-old infants. It is possible that only the left hemisphere can pick up this kind of information (see Molfese & Molfese, 1980 for support). Once the infant is experienced with detecting the audio-visual concomitants of vowels, they may then move on to consonants and CV combinations. The finding that vowels are affected by experience earlier than are consonants (Kuhl et al., 1992; Werker & Polka, 1993; Polka & Werker, 1994) may be viewed as support for this hypothesis. With further experience listening to speech, the ability to link audio and visual concomitants of speech might give rise to specific expectations for what is to be perceived through another modality; for example, when we see a gesticulating human face, we expect to hear a human voice. If this is also the case in infancy, one might expect infants to show signs of distress when an articulating human face is paired with synthesized tones played in the same frequency range as the human voice.

Although the PTLT spent on the matching face was slightly less in Study 1 than in Kuhl and Meltzoff's (1982) original work (65% vs. 73%), this was expected since our stimuli were more complex. Kuhl and Meltzoff obscured hair, neck, ears, and shoulders by having a woman poke her face through a hole cut in black cloth, therefore, there was very little to distract infants' attention away from the speaker's mouth movements. Furthermore, the visual articulations in our study were not exaggerated to the extent that they usually are in infant-directed speech, thus making the lip movements less salient than they were in Kuhl and Meltzoff's studies. The present studies show that ecologically valid stimuli of both females and males do not substantially impede 4.5-month-old infants' ability to detect audio-

visual matches based on phonetic information. Finally, due to technological constraints, we used only one instance of each audio and visual stimulus instead of ten variations. If the use of only one auditory/visual exemplar influenced infants' performance on our task, it should have made the task more difficult, if anything, since infants were not able to benefit from the variability present in different manifestations of the same sound/articulation. Therefore, departures from the procedures used by Kuhl and Meltzoff made it more difficult for infants in our study to perform successfully. Together, these methodological differences strengthen our finding that bimodal matching of phonetic cues in the face and voice is robust at 4.5-months of age.

The question that begs asking is how young infants with relatively limited experience with speech are able to link seen and heard speech to achieve a unified percept? It is also interesting to examine how our findings bear on debates regarding the relative contributions of "innate" abilities and experiential factors in speech perception.

A number of explanations have been proposed in the literature. One explanation of our findings is that infants are born with the ability to link appropriate lip shapes and speech sounds. Such an innate bias for processing speech is consistent with the motor theory (Liberman & Mattingly, 1985) and other modular approaches to speech perception (e.g., Fodor, 1983). Proponents of the motor theory claim that phonetic intentions of the speaker are represented in a specific form in the speaker's brain and that there is a perceiving module specialized to lead the listener effortlessly to that representation. The fact that young infants seem to detect audio and visual aspects of vowel articulations relatively easily and also engage in some articulatory imitation provides some support for the motor theory's assumption that the phonetic mode, and the perception-production link it incorporates, is innately specified. Conclusive evidence to test this hypothesis

would require conducting a similar study with newborns; however, our procedure is not suitable for newborns thus support for this explanation awaits further methodological advancements for testing newborns.

Although our findings are consistent with the existence of a specialized module for extracting the phonetic aspects of the speech signal, our findings do not exclude the possibility of a significant role for learning and experience in early speech perception. A number of researchers suggest that there may be an initial response to speech sounds which is rapidly shaped by the language to which the infant is exposed. For example, infants quickly learn to distinguish sound from silence, speech from nonspeech, mother's voice from a female stranger's voice, and syllables from intonation contours (see Jusczyk, 1997 for a review). It seems likely that all intersensory functioning is the product of complex epigenetic processes that begin at conception, such that there is reciprocal feedback between sensory input and both neural and behavioural development (Lewkowicz & Turkewitz, 1994). Thus, we believe it is more likely that any biological predisposition to detect audio-visual concomitants is also open to experience with stimuli in both modalities, and this experience may have a synergistic effect on the development of speech perception and production.

According to Meltzoff and Kuhl (1994), the bimodal matching effect suggests that young infants perceive structural correspondences between audio and visual aspects of speech input and that phonetic information is represented intermodally (i.e., amodally). Meltzoff and Kuhl argue that phonetic information available in both face and voice may act as an amodal cue which, in turn, acts as an internal target that infants use to generate and correct their behaviour. However, Kuhl and Meltzoff (1984) did not obtain the bimodal matching effect when the heard sounds were pure tones matched for amplitude and duration to the vowels. This finding suggests that it is spectral properties inherent to the speech signal that enable infants

to link speech sounds with the appropriate articulation. As argued by proponents of motor theory (Liberman & Mattingly, 1985), it may be the case that sounds that conform to possible articulations are perceived as speech and engage a phonetic module while all other nonspeech sounds fail to engage this module.

It is also possible that infants possess an initial sensitivity to audio and visual concomitants, but require minimal exposure to learn such relations. If infants are not predisposed to pick up relations between audio-visual aspects of speech, that is, if the observed matching effect is based primarily on an arbitrary relationship that is learned with experience, we might expect a weaker effect with the male stimuli than with the female stimuli. This was not the case. The average PTLT spent on the matching face was similar for both female and male stimuli (64.8% vs. 62.7%). Despite the fact that most infants in our study had more experience with female faces and voices than with male faces and voices (according to parental report), this asymmetry in experience did not substantially influence their ability to match a heard vowel with the appropriate facial articulation. This finding supports arguments for the existence of some kind of neuronal organization that is functional early in the first year of life and may facilitate the co-ordination of seen and heard speech. This hypothesis is supported by the finding that infants spent approximately the same amount of time producing wide-open and spread-lip articulations when the face and voice were female ($M=3.97$ s, $SD=2.45$) as when the face and voice were male ($M=3.10$ s, $SD=3.81$). If experience played a significant role in developing infants' ability to make auditory-visual matches, one might expect the duration of looking and imitating to be greater with the female versus the male. However, this was not the case in the present studies.

Although the current studies support claims that infants are able to link phonetic information presented in the lips and voice with very limited experience, our findings do not rule out the possibility of trial and error learning as a basis of

the effect (see also Walton & Bower, 1992). Infants at 4-months of age have started producing vowel sounds and may "work out" that a particular lip-voice pairing is possible or impossible.

Despite increasing evidence that there may be inherent constraints operating on infants' acquisition of intermodal knowledge, to date there are few comprehensive models of speech perception that explain or predict that this should be the case. Meltzoff and Kuhl (1994) proposed an interactive-developmental model of speech wherein both cross-modal perception and vocal imitation are linked by a common, higher-level representation of speech. They argue that infants use auditory and proprioceptive information from the self as well as visual information from others to learn what to do with their own vocal tracts when producing speech. Furthermore, these auditory-motor and auditory-visual linkages both may feed off one another during development. Thus, binding of the senses may not be a true mapping of one modality-specific representation onto another, but rather may be a continuous intermeshing of the senses resulting from the complex coaction of organic, organismic, and environmental factors (Werker & Tees, 1992). It is the ability to act on the basis of higher-order representations that is postulated to be an aspect of the human perceptual-cognitive system that is present at birth (Meltzoff, 1990). Such early sensitivity would allow infants to profit from and organize multimodal experience such that representations of faces and voices become increasingly specified. Thus it seems likely that both sides of the integration versus differentiation dichotomy are at work from the start.

In summary, two studies with infants aged 4.5- to 5-months replicated and extended prior findings that infants at this age can match phonetic information presented in the lips and voice. We have shown that this effect holds with a more ecologically valid face and with both female and male stimuli. We also found that infants spent more time producing vowel articulations that matched the lips and

voice compared to articulations that did not match. However, imitation and affect were similar to both female and male stimuli. These looking preferences and imitation findings support an integrated, multimodal representation of articulatory and acoustic phonetic information at 4.5-months of age.

REFERENCES

- Bahrick, L.E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. Infant Behavior and Development, 6, 429-451.
- Bahrick, L.E. (1988). Intermodal learning in infancy: Learning on the basis of two kinds of invariant relations in audible and visible events. Child Development, 59, 197-209.
- DeCasper, A.J. & Fifer, W.P. (1980). Of human bonding: Newborns prefer their mothers' voices. Science, 208, 1174-1176.
- DeCasper, A.J. & Prescott, P.A. (1984). Human newborns' perception of male voices: Preference, discrimination, and reinforcing value. Developmental Psychobiology, 17, 481-491.
- Desjardins, R., Rogers, J., & Werker, J.F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. Journal of Experimental Psychology: Human Perception and Performance, 66, 85-110.
- Desjardins, R. & Werker, J.F. (submitted). The integration of heard and seen speech is not mandatory for young infants.
- Dodd, B. (1979). Lipreading in infants: Attention to speech presented in and out of synchrony. Cognitive Psychology, 11, 478-484.
- Dodd, B. & Campbell, R. (1984). Non-modality specific speech coding: The processing of lip-read information. Australian Journal of Psychology, 36, 171-179.
- Fodor, (1983). The modularity of mind. Boston, MA: MIT Press.
- Green, K. & Kuhl, P.K. (1989). The role of visual information in the processing of place and manner features in speech perception. Perception and Psychophysics, 45, 34-42.

- Hirsh-Pasek, K. & Golinkoff, R.M. (1992). Skeletal supports for grammatical learning: What infants bring to the language learning task. In L.P. Lipcott & C. Rovee-Collier (Eds.), Advances in infancy research. Vol. 8. (pp. 299-338). NewJersey: Ablex.
- Humphrey, K., Tees, R.C., & Werker, J.F. (1979). Audio-visual integration of temporal relations in infants. Canadian Journal of Psychology, 33, 347-352.
- Jusczyk, P. (1997). The acquisition of spoken language. Boston, MA: MIT Press.
- Kuhl, P.K. & Meltzoff, A.N. (1982). The bimodal development of speech in infancy. Science, 218, 1138-1141.
- Kuhl, P.K. & Meltzoff, A.N. (1984). The bimodal 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: Earlbaum.
- 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. & Williams, K.A. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255, 606-608.
- Kuhl, P.K., Williams, K.A., & Meltzoff, A.N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. Journal of Experimental Psychology: Human Perception and Performance, 17, 829-840.
- Ladefoged, P. (1993). A course in phonetics. 3rd Edition. NewYork: Harcourt.
- Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. Infant Behavior and Development, 13, 343-354.

- Lewkowicz, D.J. & Lickliter, R. (Eds.), (1994). Development of intersensory perception: Comparative perspectives. Hillsdale, NJ: Earlbaum.
- Lieberman, A.M. & Mattingly, I.G. (1985). Motor theory of speech perception revised. Cognition, 21, 1-36.
- MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left hemisphere function. Science, 219, 1347-1349.
- Marean, G.C. & Werner, L.A. (1992). Vowel categorization by very young infants. Developmental Psychology, 28, 396-405.
- Massaro, D.W. & Cohen, M.M. (1990). Perception of synthesized audible and visible speech. Psychological Science, 1, 55-63.
- McGurk, H. & MacDonald, J.W. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.
- Meltzoff, A.N. (1990). Towards a developmental cognitive science: The implications of cross-modal matching and imitation for the development of representation and memory in infancy. Annals of the New York Academy of Sciences, 608, 1-37.
- Meltzoff, A.N. & Moore, M.K. (1983). Newborn infants imitate adult facial gestures. Child Development, 54, 702-709.
- Meltzoff, A.N. & Kuhl, P.K. (1994). Faces and speech: Intermodal processing of biologically relevant signals in infants and adults. In D.J. Lewkowicz & R. Lickliter (Eds.), The development of intersensory perception: Comparative perspectives (pp.335-369). Hillsdale, NJ: Earlbaum.
- Molfese, D.L. & Molfese, V.J. (1980). Cortical response of pre-term infants to phonetic and nonphonetic speech stimuli. Developmental Psychology, 16, 574-581.

- Papousek, M., Papousek, H., & Bornstein, M. (1985). The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech. In T.M. Field & N. Fox (Eds.), Social perception in infants. Hillsdale, NJ: Earlbaum.
- Polka, L. & Werker, J.F. (1994). Developmental changes in perception of nonnative vowel contrasts. Journal of Experimental Psychology: Human Perception and Performance, 20, 421-435.
- Rose, S.A. & Ruff, H.A. (1987). Cross-modal abilities in human infants. In J.D. Osofsky (Ed.), Handbook of infant development. Hillsdale, NJ: Ablex.
- Siva, N., Stevens, E.B., Kuhl, P.K., & Meltzoff, A.N. (1995). A comparison between cerebral-palsied and normal adults in the perception of auditory-visual illusions. Journal of the Acoustical Society of America, 98, 2983.
- Spelke, E.S. (1976). Infants' intermodal perception of events. Cognitive Psychology, 8, 553-560.
- Spelke, E.S. (1979). Perceiving bimodally specified events in infancy. Developmental Psychology, 15, 626-636.
- Summerfield, A.Q. (1979). Use of visual information in phonetic perception. Phonetica, 36, 314-331.
- Summerfield, A.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. 3-20). Hillsdale, NJ: Earlbaum.
- Trehub, S.F. (1973). Infants' sensitivity to vowel and tonal contrasts. Developmental Psychology, 9, 91-96.
- Turkewitz, G. & Lewkowicz, D.J. (1994). Sources of order for intersensory functioning. In D.J. Lewkowicz & R. Lickliter (Eds.), The development of intersensory perception: Comparative perspectives (pp.3-17). Hillsdale, NJ: Earlbaum.

- Walker-Andrews, A. (1994). Taxonomy for intermodal relations. In D.J. Lewkowicz & R. Lickliter (Eds.), The development of intersensory perception: Comparative perspectives (pp.39-55). Hillsdale, NJ: Earlbaum.
- Walton, G.E. & Bower, T.G.R. (1992). Amodal representation of speech in infants. Infant Behavior and Development, 16, 233-243.
- Werker, J.F. & Polka, L. (1993). Developmental changes in speech perception: New challenges and new directions. Journal of Phonetics, 21, 83-101.
- Werker, J.F. & Tees, R.C. (1992). The organization and reorganization of human speech perception. Annual Review of Neuroscience, 15, 377-402.

Figures

Figure 1. Diagram of testing and control rooms.

Figure 2. Flow diagram of sequence of events experienced by one infant in one condition.

Figure 3. Percentage of total looking time (PTLT) to the face that matched the heard vowel for female (Study 1) and male (Study 2) stimuli.

Figure 4. Number of infants who looked longer at the face that matched the heard vowel for female (Study 1) and male (Study 2) stimuli.

Figure 5. Mean duration (sec) of infant articulatory imitation when adult lips and voice presented /a/ or /i/.

Figure 1. Diagram of testing and control rooms. .

Phase	Familiarization			(Pause)	Test
Time (s)	0-9	10-18	19-27	28-30	31-151
Left TV	/i/	---	/i/	---	/i/
Right TV	---	/a/	/a/	---	/a/
Sound	---	---	---	---	/a/

Note: Heard vowel sound (/i/ or /a/), order of familiarization, side of presentation, and infant sex will be counterbalanced yielding eight conditions.

Figure 2. Flow diagram of sequence of events experienced by one infant in one condition.

Figure 3. Percentage of total looking time (PTLT) to the face that matched the heard vowel for female (Study 1) and male (Study 2) stimuli.

Figure 4. Number of infants who looked longer at the face that matched the heard vowel for female (Study 1) and male (Study 2) stimuli.

Figure 5. Mean duration (sec) of infant articulatory imitation when adult lips and voice presented /a/ or /i/.

Appendix

Consent form for parents