Developmental Science 6:2 (2003), pp 191–196

REPORT

Two-month-old infants match phonetic information in lips and voice

Michelle L. Patterson and Janet F. Werker

Department of Psychology, University of British Columbia, Canada

Abstract

Infants aged 4.5 months are able to match phonetic information in the face and voice (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999); however, the ontogeny of this remarkable ability is not understood. In the present study, we address this question by testing substantially younger infants at 2 months of age. Like the 4.5-month-olds in past studies, the 2-month-old infants tested in the current study showed evidence of matching vowel information in face and voice. The effect was observed in overall looking time, number of infants who looked longer at the match, and longest look to the match versus mismatch. Furthermore, there were no differences based on male or female stimuli and no preferences for the match when it was on the right or left side. These results show that there is robust evidence for phonetic matching at a much younger age than previously known and support arguments for either some kind of privileged processing or particularly rapid learning of phonetic information.

Introduction

Though commonly regarded as a purely auditory phenomenon, speech perception is actually an intermodal event. Speech perception is profoundly influenced by visual information in both adults (Dodd & Campbell, 1984; Green & Kuhl, 1989) and children, with the extent of influence changing as a function of the child's age (McGurk & MacDonald, 1976) and articulatory ability (Desjardins, Rogers & Werker, 1997). Indeed, there is strong evidence that infants as young as 4.5 months of age detect equivalent phonetic information in the lips and voice (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999). This is quite remarkable since infants are not able to match other types of information in face and voice such as emotion, age and gender until at least 6 months of age (see Walker, 1982; Bahrick, 1998; Patterson & Werker, 2002). However, no studies have examined bimodal phonetic matching in infants younger than 4 months. The present study examines whether infants at 2 months of age are able to match vowel information in the lips and voice.

Research with young infants suggests that 2-month-olds have the requisite skills to relate phonetic information in the lips and voice. Infants as young as 1 to 4 months are able to discriminate (Trehub, 1973) and categorize (Kuhl, 1979) vowel sounds and even show perceptual constancy for vowel color across variations in pitch and talker (Marean, Kuhl & Werner, 1992; Kuhl & Miller, 1982; Kuhl, Williams & Meltzoff, 1991). Very young infants also notice mouth movements. Indeed, 2-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).

Although it is clear that young infants can discriminate features within different modalities, when and how knowledge of the intermodal nature of speech is acquired by infants is not clear. Infants younger than 4 months of age are able to match temporal information in seen and heard speech. For example, 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.

There is robust evidence that infants aged 4 months and older are able to match equivalent information in seen and heard speech when presented with two side-by-side filmed images of a woman articulating the vowels /i/ and /a/. Infants aged 4.5 months looked longer at the

Address for correspondence: Michelle L. Patterson, Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520, USA; e-mail: michelle.patterson@yale.edu

face that matched the sound (Kuhl & Meltzoff, 1982, 1984). These results were replicated with a new set of 4.5-month-olds and with a new pair of vowels (/i/, /u/) (Kuhl & Meltzoff, 1988). In order to identify the stimulus features that are necessary and sufficient for detecting cross-modal equivalence, Kuhl and colleagues (Kuhl & Meltzoff, 1984; Kuhl *et al.*, 1991) selected pure tones of various frequencies that isolated a single feature of a vowel without allowing it to be clearly identified. Although adults could still relate the pure tones to the articulatory movements for /i/ and /a/, 4.5-month-old 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 seen and heard vowel, but rather were likely based on matching more complex spectral information contained in the voice with articulatory movements. It has been claimed that since 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 primitives' (Kuhl *et al.*, 1991).

Infants' ability to detect equivalence between mouth movements and speech sounds has been extended in three independent studies (MacKain, Studdert-Kennedy, Spieker & Stern, 1983; Walton & Bower, 1993; Patterson & Werker, 1999). MacKain *et al.* presented infants with side-by-side displays of two women articulating three pairs of disyllables (e.g. /mama lulu/). Infants between 5 and 6 months of age looked longer at the sound-specified display, but only when the matching face appeared on the right-hand side. Walton and Bower replicated Kuhl and Meltzoff's (1988) visual preference findings with the vowels /i/ and /u/ using an operant-sucking method with 4-month-olds. Results from infants' first look data suggest that infants did not have to 'work out' that a mismatch was impossible, but that it was perceived rapidly.

Infants generally have had more exposure to female faces and voices compared to male faces and voices. If the matching effect is based on an arbitrary but natural relationship that is learned, one might expect a weaker effect with the male stimuli than with the female stimuli. Patterson and Werker (1999) found that this was not the case with 4.5-month-olds. Infants aged 4.5 months were able to match phonetic information in face and voice equally well with both female and male stimuli (Patterson & Werker, 1999). This finding is particularly noteworthy since infants are unable to reliably match gender information in face and voice with these same stimuli until 8 months of age (Patterson & Werker, 2002). These findings lend support to an innately guided (Jusczyk & Bertoncini, 1988) or prepared system that, in this case, facilitates learning of seen and heard speech. These findings also point to the possibility that infants even younger than 4.5 months may be able to match vowel information in the lips and voice. Yet, no studies to date have examined infants' ability to match phonetic information in the face and voice at ages younger than 4.5 months.

The purpose of the present study was to test infants substantially younger than 4 months of age to determine if the ability to match phonetic information in face and voice is apparent even earlier. The procedure that has been used to date for testing phonetic matching is the preferential looking technique. This procedure is best used with infants older than 8 weeks of age because very young infants tend to 'lock on' to a single display and have difficulty disengaging their attention (Hood, 1995). For these reasons, along with visual angle and acuity considerations, we selected 2 months as the youngest age at which we could reasonably and reliably test infants on this task.

Method

Participants

Mothers were recruited from a local maternity hospital shortly after giving birth or they responded to an advertisement in the local media. The final sample consisted of 32 infants, 16 male and 16 female, ranging in age from 7.8 to 11.1 weeks (M = 9.2 weeks, SD = 1.3 weeks). An additional 19 infants were excluded from analyses due to crying (7), falling asleep (4), not looking at both stimuli during Familiarization (2), total looking time less than 1 min (4), looking at the same screen for the entire Test phase (1) and equipment failure (1). Infants had no known visual or auditory abnormalities, including recent ear infections, nor were infants at-risk for developmental delay or disability (e.g. pre-term, low birth weight).

Stimuli

The same stimuli were used as in Patterson and Werker (1999). Multi-media computer software (mTropolis, version 1.1) on a Macintosh 7300 was used to combine, control and present digitized audio and visual stimuli. Infants were shown two filmed images displayed on separate side-by-side computer monitors of a female or a male face articulating a different vowel (/a/ or /i/) in synchrony. The sound track corresponding to one of the articulated vowels was presented through a speaker

Figure 1 *Examples of the female and male faces articulating /a/ and /i/.*

(Sony SRS-A60) midway between the two images. Since infants can detect face–voice correspondences based on temporal cues, the two visual images were presented in synchrony and the sound was aligned with the images so that it was equally synchronous with the onset of both mouth movements.

The male and female were selected for similar colouring (i.e. Caucasian, fair) and attractiveness. The female had blonde, shoulder-length hair; the male's hair was also blonde and was all-one-length just past his ears. Both the male and female were filmed against a black background, wore white turtlenecks, and were not wearing 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. This 2 min 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/ (see Figure 1).

As in Kuhl and Meltzoff (1984) and Patterson and Werker (1999), a different male and female were selected to record the audio stimuli. Different voices were used to ensure that there were no idiosyncratic cues linking a specific voice to a specific face. Audio recordings were made in a sound-proof recording booth using a studio-quality microphone and were recorded onto audio tape. The speaker was asked to articulate the vowels /i/ and /a/ with equal intensity and duration.

As in Patterson and Werker (1999), one visual /a/, one visual /i/ and one instance of each vowel sound was chosen by three judges who rated what they deemed to be the five best visual and audio stimuli. The facial images were chosen such that duration of individual articulations fell within a narrow range that overlapped for the two vowels, the head did not move and one eye blink occurred after each articulation. For the female, the length of time that the lips were parted was .94 s for /a/ and .95 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 select the audio stimuli. Since duration of mouth opening can be longer than sound duration but not vice versa (for continuants), we ensured that the vowel sounds were of the same or shorter duration than the mouth opening. For the female, duration of the sound was .61 s for /a/ and .63 s for /i/. For the male, this duration was .62 s for /a/ and .73 s for /i/.

The films and audio files were digitized and entered into a customized computer program (using mTropolis, version 1.1) which locked the appropriate faces in phase. Next, the sound was carefully synchronized with each visual stimulus so that it began 1 s (15 frames) after the mouth first started to move. 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, and their centers were separated by 41 cm. The sounds were presented at an average intensity of 60 ± 5 dB SPL.

Equipment and test apparatus

The stimuli were presented on two 17 in. color monitors (Acana 17P) in the testing room. Black curtains covered the wall so that only the monitor screens and the camera lens (JVS GS-CD1U), positioned between and above the two monitors, were visible. The infant was seated 46 cm from the facial displays (visual angle subtended 29 degrees) in an infant seat secured to a table and the caregiver was seated behind the infant. The speaker was centered midway between the two monitors behind the curtain. During testing, a 60-watt light in a lampshade was suspended 1 m 10 cm above the infant.

Procedure

The experimental procedure involved two phases: Familiarization and Test. During the Familiarization phase, the visual stimuli were presented without sound for 27 s. First, each visual stimulus (the /a/ and /i/ face) was presented alone, one on each monitor, for 9 s each. During the final 9 s of the Familiarization phase, both faces were presented simultaneously without sound.[1] Both stimuli

[1] This phase is typically included in studies of infant word comprehension (e.g. Hirsch-Pasek & Golinkoff, 1992). The logic behind including this phase is to teach infants that both displays can be on simultaneously and it can also be used as a check for infant side bias.

were then occluded for 3 s before the Test phase began. During the 2-min Test phase, both faces were presented simultaneously and one of the two sounds (either /a/ or /i/) was played repeatedly. Thus, each infant heard only one sound. Sound presented, left-right positioning of the two faces, order of familiarization and infant sex were counterbalanced across infants.

Scoring

Coding was performed using a Panasonic video recorder which allowed frame-by-frame analysis. Coders were undergraduate students who were blind to the stimuli presented to the infant. Inter-observer reliability was assessed by re-scoring 25% of the participants. Duration of gaze was scored for each second when the infant was looking either at the right or at the left monitor. 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) spent on each display during the Test phase as well as for the 9-sec period of the Familiarization phase where both faces were presented simultaneously. PTLT to the match was also calculated for the first and second minutes of the Test phase separately. Finally, the longest look to the match and the mismatch was recorded for each infant and summed across infants. The percentage agreement for each second in the sampled periods ranged from 95.1% to 99.7% ($M = 98.1\%$) for infant looking.

Results

A paired *t*-test indicated that infant looking during the Familiarization phase was not biased to either the right (6.45 s) or the left (7.73 s) side (paired *t*-test, $p > .05$). Overall, infants spent 79.3% of the Test phase looking at either of the two faces. Infants looked longer at a particular face when the appropriate vowel sound was heard. On average, infants spent 63.95% of the total looking time fixating the matching face, which was significantly greater than chance (50%), $t(31) = 2.92$, $p < .01$. Of the 32 infants tested, 23 looked significantly longer to the sound-specified display than to the incongruent display (binomial test, $p < .05$). When the first and second minutes of the Test phase were analyzed separately, the effect was not present in the first minute (PTLT = 60.73); however, the effect was present in the second minute (PTLT = 64.12, $t(31) = 2.84$, $p < .05$). According to a paired-sample *t*-test, this increase in PTLT from the first to the second minute was not significant ($p > .05$). When PTLT during the first minute was analyzed in 30-sec blocks (59.82 s and 61.64 s), the effect was still not present ($p > .05$). According to a paired-sample *t*-test, infants' longest looks during the Test phase were significantly longer to the match (34.0 s) versus the mismatch (18.7 s, $t(31) = 2.21$, $p < .05$). A 4-way ANOVA (Infant sex, Side of match, Speaker gender, Heard vowel) revealed no significant main effects or interactions ($p > .05$).

Discussion

When given a choice between two identical faces, each articulating a different vowel sound in synchrony, infants aged 2 months looked longer at both a female and a male face that corresponded with the heard vowel sound. This effect was found in the percentage of total looking time spent on the matching face, the longest look to the match versus the mismatch, the number of infants who looked longer at the match versus the mismatch, and in the second minute of the Test phase. To date, this ability has only been reported in infants aged 4 months and older (Kuhl & Meltzoff, 1982, 1984; MacKain *et al.*, 1983; Walton & Bower, 1993; Patterson & Werker, 1999). The mean looking time to the match observed with infants aged 2 months (PTLT = 63.9) was virtually identical to that observed with 4.5-month-olds using the same stimuli (PTLT = 63.7; Patterson & Werker, 1999).

As in Kuhl and Meltzoff (1984), the current study revealed no preference for the /i/ or the /a/ face, no overall side preference, and no infant sex differences. Also, unlike the 4.5-month-olds in Patterson and Werker (1999), infants at 2 months of age did not show a bias for looking longer at the match when it was on the right-hand side. Therefore, the phonetic matching effect appears to be at least as robust at 2 months of age as it is at 4.5 months of age. The fact that matching is evident at 2 months of age, a full six months before infants can match gender information in the same faces and voices (Patterson & Werker, 2002), suggests there is something special about speech information that allows the correspondence to be perceived directly without learning (i.e. an amodal event; see Walker-Andrews, 1994) or that allows for very rapid learning.

In general, young infants tend to have more exposure to female faces and voices compared to male faces and voices; therefore, if the matching effect is based on an arbitrary but natural relationship that is learned, one might expect a weaker effect with the male stimuli than with the female stimuli. As in the previous work with infants aged 4.5 months (Patterson & Werker, 1999), no significant difference was observed in infant looking

times to the female (62.39 s) versus male (66.10 s) stimuli. This suggests that differential exposure to male and female faces and voices does not influence infants' ability to match phonetic information in face and voice at 2 months of age, thus arguing more for a specialized or amodal explanation than rapid learning.

Of course, evidence of phonetic matching at 2 months of age does not rule out learning as a basis for the effect since these infants have had two months of exposure listening to voices and watching faces. Infants this age likely do have the sensory and learning capabilities to pick up this type of information as infants of only 3 months can learn arbitrary associations between faces and voices (Brookes, Slater, Quinn, Lewkowicz, Hayes & Brown, 2001). Still, the fact that 2-month-olds match seen and heard speech as well as do 4.5-month-olds, that they do so for male as well as female faces, and that they do not match other kinds of face–voice information until a later age all argue for some kind of attentional mechanism which, at the very least, allows for rapid, perhaps innately guided (Jusczyk & Bertoncini, 1988), learning of face–voice phonetic equivalence. The early appearance of this ability may also be evidence of a dedicated mechanism for speech perception in the very young infant (Liberman & Mattingly, 1985). Nevertheless, conclusive results regarding the ontogeny of phonetic matching await methodological advances, which would allow the ability to be tested in newborns.

In summary, these results provide strong evidence for phonetic matching at the youngest age it can reasonably be tested using the standard methodology. Never before have infants been shown to be able to relate such complex, ecologically valid stimuli. Whether or not the ability to match phonetic information in the lips and voice by 4.5-month-old infants is based on amodal specification or learning, innately guided or otherwise, the fact that 2-month-old infants in the current experiment looked longer at the face that matched the vowel sounds shows that matching phonetic information requires relatively little experience with talking faces.

Acknowledgements

The authors would like to thank all of the parents who volunteered to participate in our study with their infants. This research was funded by a grant (#81103) to J. Werker from the Natural Sciences and Engineering Research Council (NSERC), a Social Sciences and Humanities Research Council (SSHRC) Doctoral Fellowship to M. Patterson and support from the Canada Research Chair program (J. Werker).

References

Bahrick, L.E. (1998). Intermodal perception of adult and child faces and voices by infants. *Child Development*, **69**, 1263–1275.

Brookes, H., Slater, A., Quinn, P., Lewkowicz, D.J., Hayes, R., & Brown, E. (2001). Three-month-old infants learn arbitrary auditory–visual pairings between voices and faces. *Infant and Child Development*, **10**, 75–82.

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.

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.

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.

Hirsch-Pasek, K., & Golinkoff, R. (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). Norwood, NJ: Ablex.

Hood, B. (1995). Disengaging visual attention in the infant and adult. *Infant Behavior and Development*, **16**, 405–422.

Jusczyk, P., & Bertoncini, J. (1988). Viewing the development of speech perception as an innately-guided learning process. *Language and Speech*, **31**, 217–238.

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., & 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: Erlbaum.

Kuhl, P.K., & Miller, J. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics*, **31**, 279–292.

Kuhl, P.K., Williams, K.A., & Meltzoff, A.N. (1991). Cross-modal speech perception in adults and infants using non-speech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 829–840.

Liberman, A.M., & Mattingly, I.G. (1985). Motor theory of speech perception revised. *Cognition*, **21**, 1–36.

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

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., Kuhl, P., & Werner, L.A. (1992). Vowel categorization by very young infants. *Developmental Psychology*, **28**, 396–405.

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

Patterson, M., & Werker, J.F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, **22**, 237–247.

Patterson, M., & Werker, J.F. (2002). Infants' ability to match phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology*, **8**, 93–115.

Trehub, S. (1973). Infants' sensitivity to vowel and tonal contrasts. *Developmental Psychology*, **9**, 91–96.

Walker, A.S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology*, **33**, 514–535.

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: Erlbaum.

Walton, G.E., & Bower, T.G.R. (1993). Amodal representation of speech in infants. *Infant Behavior and Development*, **16**, 233–243.

Received: 21 September 2001
Accepted: 11 April 2002