DEVELOPMENTAL CHANGE IN SENSITIVITY TO AUDIOVISUAL SPEECH CONGRUENCY AND ITS RELATION TO LANGUAGE IN INFANTS

Masahiro IMAFUKU1,2,3 and Masako MYOWA1

1Graduate School of Education, Kyoto University, Japan
2Graduate School of Arts and Sciences, The University of Tokyo, Japan
3Japan Society for the Promotion of Science, Japan

In this study, we used eye-tracking to investigate selective visual attention paid to congruent and incongruent audiovisual speech of Japanese infants (6 to 12 months old) and adults. Infants’ receptive and expressive language abilities at 12 months were measured through a questionnaire completed by their caregivers. We found that 6-month-olds looked at the mouth longer in the audiovisual congruent condition than in the incongruent condition, whereas 12-month-olds did not show any significant differences in time spent in looking at the mouth. Furthermore, the time spent looking at the mouth among 6-month-olds, in both the audiovisual congruent and incongruent conditions, was positively correlated with receptive language abilities with 12 month olds. These findings suggest that sensitivity to congruent and incongruent audiovisual speech changes during the first year, and attention to the speaker’s mouth at 6 months are good predictors of later receptive language abilities.

Key words: human infants, audiovisual speech perception, language acquisition, eye-tracking

INTRODUCTION

Infants are exposed to multisensory cues in their daily social interactions. In particular, during verbal communication, the mouth of the talking face provides important speech cues (i.e., both speech sounds and lip movements) to the listener. These redundant audiovisual cues are thought to contribute to language acquisition in infancy (Bahrick, Hernandez-Reif, & Flom, 2005; Gogate & Bahrick, 1998).

Previous studies have reported that infants prefer audiovisual stimuli characterized by temporal synchrony rather than asynchrony. For example, Dodd (1979) found that 10-to 16-week-old infants preferred looking at facial movements that were synchronized with speech rather than asynchronized. Moreover, several studies using the preferential looking paradigm have reported that human infants prefer looking at faces whose lip movements
correspond to vowels heard simultaneously, such as /a/ or /i/ (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003). Such tendencies have been observed even during the neonatal period (Aldridge, Braga, Walton, & Bower, 1999). These findings suggest that the ability to detect audiovisual correspondence is present very early in life, although the processes underlying such multisensory integration become more sophisticated with age (Lewkowicz & Hansen-Tift, 2012; Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009).

Recently, many studies have shown that typically developing infants pay increased attention to the mouth of a talking face as they become older (Frank, Vul, & Saxe, 2012; Lewkowicz & Hansen-Tift, 2012; Tenenbaum, Shah, Sobel, Malle, & Morgan, 2013; von Hofsten, Uhlig, Adell, & Kochukhova, 2009). For example, Lewkowicz and Hansen-Tift (2012) found that the amount of visual attention to a native speaker’s mouth, relative to the speaker’s eyes, changed throughout the first year of life. They showed that 4-month-old infants looked at the eyes of a talking face longer, whereas 6-month-olds, who were in the first stage of canonical babbling (Oller, 2000), began to pay attention to the mouth, and 12-month-olds gradually began shifting their attention away from the mouth as they became attuned to their native language. These results indicate that infants show increased interest in the speaker’s mouth at approximately 6 months.

Despite these growing attractions to infants’ visual attention toward the speaker’s mouth, how the sensitivity to audiovisual congruency change during infancy remains poorly understood. Some studies have shown that infants could detect congruency of audiovisual speech (Altvater-Mackensen & Grossmann, 2015; Lewkowicz, 2010; Pons & Lewkowicz, 2014). To detect the discrepancy between speech sounds and lip movements in incongruent audiovisual speech, infants need to gain an understanding of the relationship between audiovisual speech cues. Recent neuroimaging studies have demonstrated that at approximately 6 months, infants’ left inferior frontal brain regions are activated, which has a crucial role in the coupling of auditory and articulatory speech information, more strongly in response to congruent audiovisual speech than to incongruent audiovisual speech (Altvater-Mackensen & Grossmann, 2016). In contrast, adults show greater responses to incongruent audiovisual speech (Ojanen et al., 2005). Altvater-Mackensen and Grossmann (2016) explained the contradiction: infants, who start to attune to the characteristics of their native language (Lewkowicz & Hansen-Tift, 2012), are more sensitive to familiar audiovisual speech as shown in previous behavioral studies with infants younger than 6 months (Altvater-Mackensen & Grossmann, 2015; Dodd, 1979), while adults have increased sensitivity to unfamiliar audiovisual speech because incongruent conditions require higher demands of processing for conflicting audiovisual information (Ojanen et al., 2005). These findings suggest that the sensitivity to audiovisual congruency changes with development. Furthermore, as infants acquire knowledge of the native language audiovisual relationship between 6 and 11 months of age (Pons et al., 2009), it is possible that the sensitivity to audiovisual congruency develops through the first year of life. However, few studies have examined how infants of various ages view a speaker’s face when normal audiovisual speech congruency is violated.

Furthermore, in language acquisition by infants, visual attention to the mouth of a
talking face is crucial. In adults, redundant audiovisual speech cues enhance language comprehension when compared to audio-only speech (Sumby & Pollack, 1954). A similar pattern also appears to occur in infants. When infants around six months of age receive redundant audiovisual information, they can directly access tightly and coupled multimodal information, and these redundant stimuli facilitate their learning (e.g., Bahrick, Flom, & Lickliter, 2002; Bahrick et al., 2005; Bahrick & Lickliter, 2000; Gogate & Bahrick, 1998). Previous studies have shown that visual attention toward the speaker’s audiovisual redundant cues (i.e., the mouth) contributes to language development in infants. Specifically, 6-month-olds, who looked for longer periods at redundant audiovisual speech cues, subsequently demonstrated higher levels of expressive language at 24 months (Young, Merin, Rogers, & Ozonoff, 2009). Attention to the mouth at 12 months was also associated with language outcomes at 18 and 24 months (Tenenbaum, Sobel, Sheinkopf, Malle, & Morgan, 2015). This demonstrates that individual differences of visual attention toward the mouth in redundant audiovisual speech contexts are indicators of language acquisition patterns during infancy. In contrast, the relationship between visual attention toward the mouth and language has not been tested when audiovisual speech cues are incongruent.

In the present study, we used eye-tracking to examine the visual attention of 6- and 12-month-old infants, as well as that of adults, to congruent and incongruent audiovisual speech. Furthermore, we examined the developmental link between visual attention to the mouth at six and 12 months, and receptive or expressive language abilities at 12 months. We investigated whether there were developmental changes in the selective visual attention of infants to congruent and incongruent audiovisual speech, and the relationship between infants’ selective visual attention and their language abilities. We assumed first that, because congruent audiovisual speech information is highly salient for infants around six months of age (Hyde, Jones, Flom, & Porter, 2011; Kubicek et al., 2014), these infants would attend more to the mouth when audiovisual speech cues are congruent. Second, infants acquire a certain level of expertise in their native language by 12 months of age (Lewkowicz & Hansen-Tift, 2012; Pons et al., 2009). As such, we assumed that time spent looking at the mouth in the audiovisual incongruent condition may increase through the first year. Third, as adults have already acquired the knowledge of audiovisual speech, they would show selective visual interest during incongruent audiovisual speech by violating their prior knowledge. Finally, assuming that infants who attended more to the mouth of a talking face acquired redundant audiovisual speech cues, we predicted a relationship between the visual attention infants give to a speaker’s mouth and their language abilities in audiovisual congruent speech. Furthermore, if the sensitivity to audiovisual congruency reflects language acquisition patterns, this relationship would be observed even in audiovisual incongruent speech, as well as congruent speech.

**Method**

*Participants*

The participants included 21 Japanese infants; six months of age ($M = 192.5$ days; $SD = 8.81$ days; 13 male), 21 infants 12 months of age ($M = 376.6$ days; $SD = 8.82$ days; 11 male) and 14 adults ($M = 22.4$ years;
Four infants (two each from the 6-month and 12-month groups) and two adults were excluded due to eye-tracking issues. None of the participants had a history of serious illness or auditory deficits, and all the infants had been born at full-term and were healthy at the time of the experiment. Language assessment for all infants (including those who were six months old when given their eye-tracking tests) was conducted at 12 months of age by their primary caregiver. Four members of the original 6-month-old group and one from the 12-month-old group were excluded from this portion of the study because their carers did not complete the language assessment. Language assessment data were successfully collected from 17 infants in the original 6-month-old group (M = 380.9 days; SD = 8.92 days; 12 male) and from 20 infants in the 12-month-old group (M = 377.9 days; SD = 6.76 days; 10 male). Informed consent was obtained from parents of the infants or from the adults themselves prior to the commencement of the study. Approval was granted by the Ethics Committee of Kyoto University (24-p-9), and the study was conducted in accordance with the standards specified in the 1964 Declaration of Helsinki.

**Materials and Procedure**

Infant participants were seated on a parent’s lap, and adult participants were seated on a chair, with their eyes approximately 60 cm from a monitor in a soundproof room. A five-point calibration procedure was performed prior to data collection. A video camera was aimed at the participants’ face to ensure that they were attending to the display.

Eye movements were recorded using a Tobii X60 (Tobii Technology) near-infrared gaze tracking system, which was integrated with a 22-inch display monitor that presented the movies. The minimum fixation duration was set to 60 ms. Stimulus presentation and data recording were controlled by a computer (Dell Precision T7500) with Tobii Studio software (Tobii Studio 2.1.12, Tobii Technology).

Participants viewed video clips (subtending 20° × 16° of the visual angle) of a woman telling two different stories in Japanese, as if speaking to an infant. There were two conditions for each story: (1) the audiovisual congruent condition, in which the temporal timing of auditory and visual stimuli was synchronized, and (2) the audiovisual incongruent condition, whereby the correspondent relationship between the auditory and visual stimuli was disrupted by way of playing the visual stimuli backwards. Incidentally, the language could be identified even when the visual speech stimuli was temporarily reversed (Ronquest, Levi, & Pisoni, 2010). Each condition lasted 14 seconds and was presented twice; making a total of four trials for each participant. The video clips were created and edited using Adobe Premiere Pro CS3 (Adobe System Inc.).

We used the MacArthur Communicative Development Inventory, adapted for Japanese (JCDI; Words and Gestures; Ogura & Watamaki, 1998), to evaluate the infants’ receptive and expressive language abilities. Primary caregivers completed the JCDI (Ogura & Watamaki, 1998) when the infants were 12 months old. The following four items were used to capture receptive and expressive language: (1) signs of understanding (e.g., responding to one’s name), (2) phrase understanding (e.g., “Are you hungry?”), (3) starting to talk (e.g., vocal imitation), and (4) vocabulary (baby talk, such as “woof woof”).

**Data analyses**

**Eye-tracking data.** Eye movement data were analyzed using Tobii’s standard statistics package. We defined three areas of interest (AOIs; see Fig. 1): the actor’s face, eyes, and mouth. The time period of fixation on each AOI was summed across two trials in each condition. We used the sum of fixations for each AOI across trials because infants’ fixations across videos are thought to reflect individual differences in infants’ speech perception (Altvater-Mackensen & Grossmann, 2016; Lewkowicz & Hansen-Tift, 2012). Firstly, we analyzed the time period of fixation on the face (6-month-olds: congruent, M = 18.43 sec, SD = 5.30, incongruent, M = 18.93 sec, SD = 6.18; 12-month-olds: congruent, M = 20.86 sec, SD = 4.76, incongruent, M = 19.23 sec, SD = 5.82; adults: congruent, M = 23.30 sec, SD = 4.15, incongruent, M = 22.47 sec, SD = 4.22). Then, we calculated the proportion of time spent looking at the mouth rather than the eyes (time looking at mouth/time looking at eyes + time looking at mouth) in accordance with Lewkowicz and Hansen-Tift (2012).

**Language scores.** We defined two categories of language scores: receptive (three sub-items: first signs of understanding, phrase understanding, and vocabulary comprehension) and expressive (two sub-items: starting to talk and speech production). Mean receptive and expressive scores were 18.53 (SD = 7.60) and 1.88 (SD = 2.80), respectively, in the 6-month-old group, and 21.05 (SD = 8.94) and 1.81 (SD = 2.20), respectively, in the 12-month-old group. We calculated Pearson’s correlations between the proportions of time spent looking at the mouth and the language scores.
Results

Infants' and Adults' Visual Attention to Audiovisual Speech

We analyzed the time of fixation on the face with a two-way analysis of variance (ANOVA), using age (6 months, 12 months, and adults) as the between-subjects factor, and condition (congruent and incongruent) as the within-subjects factor. We found no significant main effects for audiovisual speech congruency, $F(1, 53) = 1.18$, $p = 0.28$, $\eta^2_p = 0.02$. In addition, no significant interaction between age and condition emerged, $F(2, 53) = 1.26$, $p = 0.29$, $\eta^2_p = 0.05$. Thus, the visual attention to faces of neither infants nor adults differed between the two conditions. Furthermore, there was a statistically significant main effect of age, $F(2, 53) = 3.32$, $p = 0.04$, $\eta^2_p = 0.11$. Post-hoc analyses (Bonferroni) have revealed that total fixation time on the face in 6-month-olds was significantly lower than in adults ($p = 0.04$).

We then analyzed the proportion of time looking at the mouth with a two-way mixed ANOVA, with age (six months, 12 months, adults) as the between-subjects factor, and audiovisual speech congruency condition (congruent, incongruent) as the within-subjects factor (Fig. 2). We found there was a statistically significant interaction between age and condition, $F(2, 53) = 5.13$, $p = 0.01$, $\eta^2_p = 0.16$. Simple main effect tests revealed that for 6-month-old infants, the proportion of time spent looking at the mouth was higher in the audiovisual congruent condition than in the incongruent condition ($p = 0.04$). For adults, the proportion of time spent looking at the mouth was higher in the audiovisual incongruent condition than in the congruent condition ($p = 0.03$). For 12-month-olds, there was no significant difference between the congruent and incongruent conditions ($p = 0.25$). In the congruent condition, the proportion of time spent looking at the mouth in both 6-month-olds ($p = 0.03$) and 12-month-olds ($p < 0.01$) was significantly higher than in adults, but not in the incongruent condition. There was a statistically significant main effect of age, $F(2, 53) = 4.57$, $p = 0.02$, $\eta^2_p = 0.15$. Post-hoc analyses (Bonferroni) revealed that the
proportion of time looking at the mouth in 12-month-olds was significantly higher than in adults ($p = 0.01$), but there was no significant difference between 6-month-olds and adults ($p = 0.41$). There was no statistically significant main effect of condition, $F(2, 53) = 0.04$, $p = 0.84$, $\eta_p^2 = 0.00$.

**Relationship between Visual Attention and Language Scores in Infants**

The correlations between the proportion of time spent looking at the mouth and language scores are shown in Table 1. The proportion of time looking at the mouth in both the audiovisual congruent and incongruent conditions at six months was positively correlated with receptive language scores at 12 months ($n = 17$, $r = 0.72$, $p < 0.01$; $n = 17$, $r = 0.73$, $p < 0.01$). The other combinations between the ratio of time spent looking at the mouth and language scores did not yield any significant differences ($p > 0.05$ in all cases).

**DISCUSSION**

The present study had two aims. First, we examined whether selective visual attention to congruent and incongruent audiovisual speech differed between 6-month-olds, 12-month-olds, and adults. Second, we investigated whether visual attention to the mouth of a talking face was associated with the language scores of infants.

As predicted, we found that visual attention to the mouth differed as a function of age and audiovisual condition. Six-month-olds spent more time looking at the mouth in the congruent condition than in the incongruent condition. This finding is consistent with previous studies which have shown that infants younger than 6 months of age prefer consistent over inconsistent audiovisual presentations (Altvater-Mackensen & Grossmann, 2015; Dodd, 1979; Kubicek et al., 2014; Lewkowicz, 1986; Patterson & Werker, 1999, 2003). Furthermore, using event-related potential (ERP), for infants aged 5 months, the early auditory component was greater for synchronous audiovisual speech than
asynchronous audiovisual speech (Hyde et al., 2011). This suggests that 6-month-old infants have sensitivity to congruent multimodal information. We extended on previous findings by measuring the amount of time the participants spent looking at the speaker’s mouth. Our results indicate that congruent audiovisual speech cues are crucial for selective attention of infants at 6 months of age.

Interestingly, we found that 12-month-olds showed no significant difference in time spent looking at the mouth between the audiovisual congruent and incongruent conditions. Our results are in line with a previous study, which demonstrated that 12-month-olds did not show a visual preference either for congruent or incongruent audio-visual speech in the preferential looking task, although 6-month-olds showed a preference for congruent audiovisual speech (Kubicek et al., 2014). A possible reason for 12-month-olds showing no difference between the two audiovisual conditions is their growing understanding of the typical congruent audiovisual speech cues associated with their native language (Lewkowicz & Hansen-Tift, 2012). Some researchers have indicated that native language phonetic perception ability develops between 6 and 12 months of age (Kuhl et al., 2006) and that infants acquire knowledge of the relationship between audiovisual speech cues at approximately 12 months of age (Pons et al., 2009). In addition, although infants at approximately 6 months of age are preference to congruent audiovisual speech (Altvater-Mackensen & Grossmann, 2015; Dodd, 1979), adults show preference to incongruent audiovisual speech (Ojanen et al., 2005). As discussed above, it is thought that incongruent audiovisual speech places higher demands on processing of conflicting audiovisual information in participants who have more mature patterns of audiovisual speech processing. For this reason, 12-month-olds are thought to show increased interest in the speakers when audiovisual speech cues are incongruent. However, they do not yet show adult patterns of visual attention. This may be because the audiovisual speech processing system is still immature in 12-month-olds and even develops during childhood (Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014) and from childhood to adulthood (Lalonde & Holt, 2016).

In the present study, adults spent more time looking at the mouth when audiovisual speech cues were incongruent rather than congruent. This is in agreement with a previous ERP study in adults which demonstrated that, compared to synchronous speech,

| Eye-tracking Condition | Receptive language | Expressive language |
|------------------------|--------------------|--------------------|
| 6 months \((n = 17)\)  |                    |                    |
| Congruent              | .72**              | .11                |
| Incongruent            | .73**              | .21                |
| 12 months \((n = 20)\) |                    |                    |
| Congruent              | -.02               | .29                |
| Incongruent            | -.13               | .16                |

**\(p < 0.01\).
asynchronous speech elicited greater amplitudes for early components believed to reflect auditory processing (Pilling, 2009). This previous study has suggested that adults are sensitive to the inconsistency between speech sounds and lip movements. Conversely, adults demonstrated less time spent looking at the speaker’s mouth when they were exposed to congruent audiovisual speech of their native language (Lewkowicz & Hansen-Tift, 2012). This would be due to the reduction adults have in their need to access redundant audiovisual speech cues. In our study, it is possible that the increased visual attention to the mouth among adults in the audiovisual incongruent condition reflected entirely mature patterns of audiovisual speech processing.

The important finding of the current study was the determination of significant positive relationships between visual attention to the mouth at six months, in both the audiovisual congruent and incongruent conditions, and receptive language scores at 12 months. In the audiovisual congruent condition, our data supports the view that the visual attention infants paid to the mouth of a talking face contributes to later language acquisition, because mouths offer redundant audiovisual speech cues (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Our results confirm this view, and are consistent with previous behavior studies showing that infants who focused on the speaker’s mouth had a larger vocabulary size at 24 months (Tenenbaum et al., 2015; Young et al., 2009). However, the redundancy effect might not be the only factor explaining why visual attention to the mouth at six months would be associated with receptive language scores at 12 months, because we found this relationship even in the audiovisual incongruent condition. In our study, incongruent audiovisual speech stimuli included two types of information: audiovisual temporal asynchrony and visual stimuli played backwards. This feature might have increased the tendency to perceive the stimuli as unnatural, in comparison to the temporally delayed audiovisual speech stimuli used in previous studies (Dodd, 1979; Lewkowicz, 1986; Hyde et al., 2011). However, the finding that neither infants nor adults seem to pay more attention to the face in the incongruent rather than the congruent condition might indicate that the incongruent audiovisual speech stimuli was not perceived as particularly unusual. In either case, the incongruent audiovisual speech stimuli may offer less useful information than the congruent audiovisual speech stimuli. Some researchers have argued that incongruent audiovisual speech stimuli are not salient for infants, since the stimulus is less redundant (Bahrick et al., 2002; Bahrick et al., 2005; Bahrick & Lickliter, 2000; Gogate & Bahrick, 1998). Instead, the incongruent audiovisual speech stimuli are thought to induce awareness of the violation between audiovisual speech cues. We can therefore assume that infants’ and adults’ levels of visual attention in the audiovisual incongruent condition indicate their sensitivity to the relationship between auditory and visual cues. Accordingly, our findings on the relationship between looking at the mouth under conditions of both audiovisual congruence and incongruence, and receptive language scores, suggest that among infants, attention to the mouth, regardless of what is happening there, can predict later language acquisition.

In summary, this study has revealed a developmental change in the perception of audiovisual speech congruency, as well as a relationship between looking at a speaker’s mouth at age 6 months and subsequent receptive language abilities. In particular, our
results suggest that 12-month-olds show increased interest in incongruent as well as congruent audiovisual speech although it was not the completed adult pattern. From this evidence, it is thought that increased sensitivity to incongruent audiovisual speech may reflect more mature patterns of audiovisual speech processing in the first year of life. Then, the present investigation raises at least one question for future research: what factor is associated with the development of the audiovisual speech processing system in infancy? One possible factor is that visual attention paid to the mouth of a talking face may be associated with speech production abilities, such as babbling, in infants (Lewkowicz & Hansen-Tift, 2012). The theoretical relationship between speech perception and production has been pointed out in previous studies (Liberman & Mattingly, 1985; Westermann & Miranda, 2004). In the study of infants, this relationship has been investigated. For example, infants who could demonstrate an increased amount of vocalization during daily activities were more sensitive to the congruency between auditory and visually presented speech (Altvater-Mackensen & Grossmann, 2015) and infants who could produce more of the consonants during their interactions with a caregiver could discriminate consonants (DePaolis, Vihman, & Nakai, 2013). A recent study showed that infants who prefer to look at the speaker’s mouth exhibited increased response to audiovisual speech in the left inferior frontal region (responsible for speech production; Altvater-Mackensen & Grossmann, 2016). Further investigation into the relationship between speech production abilities and selective visual attention to a talking face will provide more insight into the mechanisms underlying audiovisual speech perception and language acquisition in infants.

REFERENCES
Altvater-Mackensen, N., & Grossmann, T. 2015. Learning to match auditory and visual speech cues: Social influences on acquisition of phonological categories. *Child Development*, 86, 362–378.
Altvater-Mackensen, N., & Grossmann, T. 2016. The role of left inferior frontal cortex during audiovisual speech perception in infants. *NeuroImage*, 133, 14–20.
Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. 1999. The intermodal representation of speech in newborns. *Developmental Science*, 2, 42–46.
Bahrick, L. E., Flom, R., & Lickliter, R. 2002. Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology*, 41, 352–363.
Bahrick, L. E., Hernandez-Reif, M., & Flom, R. 2005. The development of infant learning about specific face–voice relations. *Developmental Psychology*, 41, 541–552.
Bahrick, L. E., & Lickliter, R. 2000. Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36, 190–201.
Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. 2009. The natural statistics of audiovisual speech. *PLOS Computational Biology*, 5, e1000436.
DePaolis, R. A., Vihman, M. M., & Nakai, S. 2013. The influence of babbling patterns on the processing of speech. *Infant Behavior and Development*, 36, 642–649.
Dodd, B. 1979. Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, 11, 478–484.
Frank, M. C., Vul, E., & Saxe, R. 2012. Measuring the development of social attention using free-viewing. *Infancy*, 17, 355–375.
Gogate, L. J., & Bahrick, L. E. 1998. Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, 69, 133–149.
Hyde, D. C., Jones, B. L., Flom, R., & Porter, C. L. 2011. Neural signatures of face–voice synchrony in
5-month-old human infants. *Developmental Psychobiology, 53*, 359–370.

Knowland, V. C. P., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. C. 2014. Audio-visual speech perception: a developmental ERP investigation. *Developmental Science, 17*, 110–124.

Kubicke, C., Hillairet de Boisferron, A., Dupierrix, E., Pascalis, O., Luxenbruck, H., Gervain, J., & Schwarzer, G. 2014. Cross-modal matching of audio-visual german and french fluent speech in infancy. *PLOS ONE, 9*, e89275. doi:10.1371/journal.pone.0089275.

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

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*, F13–F21.

Lalonde, K., & Holt, R. F. 2016. Audiovisual speech perception development at varying levels of perceptual processing. *Journal of the Acoustical Society of America, 139*, 1713.

Lewkowicz, D. J. 1986. Developmental changes in infants’ bisensory response to synchronous durations. *Infant Behavior and Development, 9*, 335–353.

Lewkowicz, D. J. 2010. Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*, 66–77.

Lewkowicz, D. J., & Hansen-Tift, A. M. 2012. Infants deploy selective attention to the mouth of a talking face when hearing speech. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 1431–1436.

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

Ogura, T., & Watamaki, T. 1998. Research and development of early language development inventory. Research Project, Grant-in-aid for Scientific Research. Kobe University (in Japanese).

Ojane, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., & Sams, M. 2005. Processing of audiovisual speech in Broca’s area. *NeuroImage, 25*, 333–338.

Oller, D. K. 2000. *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum.

Patterson, M. L., & 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. L., & Werker, J. F. 2003. Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*, 191–196.

Pilling, M. 2009. Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research, 52*, 1073–1081.

Pons, F., & Lewkowicz, D. J. 2014. Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica, 149*, 142–147.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. 2009. Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 10598–10602.

Ronquest, R. E., Levi, S. V., & Pisoni, D. B. 2010. Language identification from visual-only speech signals. *Attention, Perception, & Psychophysics, 72*, 1601–1613.

Sumby, W. H., & Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., & Morgan, J. L. 2013. Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study. *Infancy, 18*, 534–553.

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Malle, B. F., & Morgan, J. L. 2015. Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language, 42*, 1173–1190.

von Hofsten, C., Uhlig, H., Adell, M., & Kochukhova, O. 2009. How children with autism look at events. *Research in Autism Spectrum Disorders, 3*, 556–569.

Westermann, G., & Miranda, E. R. 2004. A new model of sensorimotor coupling in the development of speech. *Brain and Language, 89*, 393–400.

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. 2009. Gaze behavior and affect at 6 months: Predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science, 12*, 798–814.

(Manuscript received 17 February, 2016; Revision accepted 1 July, 2017)