Exploring early developmental changes in face scanning patterns during the perception of audio-visual mismatch of speech cues

Tomalski, P.*[1,2], Ribeiro, H.[1], Ballieux, H.[1], Axelsson, E.[1], Murphy, E.[1], Moore, D.G.[1] and Kushnerenko, E.*[1,2]

[1] Institute for Research in Child Development, School of Psychology, University of East London, London, UK.

[2] Faculty of Psychology, University of Warsaw, Warsaw, Poland

Running title: Infant face scanning of audio-visual speech cues

* Corresponding authors: Przemyslaw Tomalski and Elena Kushnerenko, Institute for Research in Child Development, School of Psychology, University of East London, Water Lane, London E15 4LZ, UK; e-mail: tomalski@mac.com or e.kushnerenko@gmail.com, tel. +44 208 223 4513.

No. words: 4600 (main text and references)

Keywords

Infancy, eye-tracking, face scanning, audio-visual (AV) speech integration, audio-visual mismatch

Abstract

Young infants are capable of integrating auditory and visual information and their speech perception can be influenced by visual cues, while 5-month-olds are able to detect a mismatch between the mouth articulation and the speech sound. From 6 months of age infants gradually shift their attention away from eyes and towards mouth in articulating faces, potentially to benefit from intersensory redundancy of audio-visual (AV) cues. Using eye-tracking we investigated whether 6-9 month-olds show similar age-related increase of looking to the mouth, while observing congruent and/or redundant vs. mismatch and non-redundant speech cues. Participants distinguished between congruent and incongruent AV cues as reflected by amount of looking to the mouth. They showed age-related increase in attention to the mouth, but only for non-redundant, mismatched AV speech cues. Our results highlight the role of intersensory redundancy and audio-visual mismatch mechanisms in facilitating the development of speech processing of infants under 12 months of age.

Acknowledgments

We would like to thank all participating infants and parents for their contribution. We acknowledge the financial support of Eranda Foundation and the University of East London. We would like to thank Robin Panneton and Mark H. Johnson for helpful comments and Glorianne Spiteri and Caroline Frostick for assistance with data collection.

Introduction

Human infants show rapid development of speech processing capabilities in the first year of life. One crucial aspect of early phonological development is the ability to integrate auditory and visual speech cues. Several studies have demonstrated that infants attend to visual cues during audio-visual (AV) speech perception tasks. Very young infants can learn arbitrary face-voice associations (Brookes, et al., 2001), and by 4 months they detect AV asynchrony when observing speech production (Lewkowicz, 2010). Infants aged 2 and 4 months prefer watching faces with mouth articulations matching auditory vowels (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003). Thus, in the first months of life infants are already able to detect corresponding patterns of mouthing and auditory speech in vowel production.

These early capacities are not reflected in increased attention to the mouth in the first months of life. Only after the age of 6 months infants gradually increase their looking at the mouth when scanning dynamic faces (Hunnius & Geuze, 2004). When viewing faces speaking their native language, infants begin to shift attention from the eyes to the mouth between 4 and 8 months, yet attention to the mouth declines again after 12 months of age (Lewkowicz & Hansen-Tift, 2012). Such a pattern of developmental change, especially between 6 and 12 months, suggests that attention to visual speech cues plays a vital role in the development of speech perception. Several mechanisms have been proposed to explain the role of visual cues in facilitating speech perception development (see: Lewkowicz & Hansen-Tift, 2012). Firstly, visual cues may enhance auditory speech perception by increasing the saliency of ambiguous or under-specified parts of the speech stream and by providing redundant AV information (Bahrick, Lickliter, & Flom, 2004; Campbell, 2008). Secondly, increased attention to articulation at a time when infants engage in canonical babbling may facilitate their

own speech production either through imitation, or motor learning reinforced by caregiver's feedback (Howard & Messum, 2011). In summary, attention to the mouth when observing visual speech cues provides infants with vital information that can facilitate their phonological development. Existing evidence suggests that infants pay attention to the mouth especially between 6 and 12 months of age when learning about their native language.

Despite these advances, relatively little is known about how infants process AV speech when offered conflicting cues during this vital period of development. Lewkowicz and Hansen-Tift (2012) have proposed that infants beyond 6 months of age increase their attention to the speaking mouth in order to access redundant AV speech cues and to facilitate learning about native speech. We have tested this hypothesis by investigating face-scanning patterns while 6-to-9-month-olds observed congruent (and/or redundant) vs. conflicting and non-redundant AV speech cues. According to the intersensory redundancy hypothesis the age-related increase in attention to the articulating mouth would be expected while observing congruent and redundant AV speech cues, but not necessarily while observing incongruent and conflicting cues. We used the 'McGurk illusion' stimuli from Kushnerenko et al. (2008) to test this hypothesis.

McGurk and MacDonald (1976) were first to demonstrate that relevant visual information (lip articulation) influences the perception of speech sounds. When an auditory /ba/ syllable is dubbed onto a video of a face articulating /ga/, adults commonly perceive the resulting combination as /da/ (fusion effect), while the reverse combination leads to a perception of a non-fusible /bga/ in half of adult participants. Although some studies have argued that audio-visual intergration during early and middle childhood (3-5 and 7-8 years) appears to be less robust than in adults

(MacDonald & McGurk, 1978; Massaro, Thompson, Barron, & Laren, 1986), recent studies have shown that even young infants are prone to the McGurk effect (Burnham & Dodd, 2004; see also: Rosenblum, Schmuckler, & Johnson, 1997).

Apart from behavioural results, the detection of mismatch between auditory and visual speech cues has been documented in electrophysiological studies, with a mismatch-related ERP response found over fronto-lateral sites (Bristow, et al., 2009; Kushnerenko, et al., 2008; Mottonen, Krause, Tiippana, & Sams, 2002; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007). Kushnerenko and colleagues (2008) demonstrated that 5-month-olds already show the audiovisual event-related mismatch response (AVMMR) to conflicting combination of cues (VbaAga) and that this response is different in trials where both cues can be fused into a single percept. Importantly, the AVMMR is distinct from potentials evoked by either the auditory or visual components of each bimodal stimulus. It is likely that this early capacity for detecting mismatch between auditory and visual speech information is a signature of an important neural mechanism that may assist infants' learning about native speech sounds. Furthermore, in a group of 6-to-9-month-olds, Kushnerenko et al. (under review) have found a strong correlation between the duration of looking to the mouth while watching the incongruent VbaAga cues and the size of the AVMMR right-central positivity in ERPs.

Given these findings of ERP markers of audio-visual mismatch, we have investigated whether infants beyond 6 months of age show increased attention to the mouth while watching congruent (and redundant) vs. incongruent AV speech cues. In order to examine the age-related change reported by (Lewkowicz & Hansen-Tift, 2012) we measured attention to the eyes and the mouth regions of infants aged 6-7 months and 8-9 months while they observed two canonical, congruent speech stimuli: visual /ba/ –

auditory /ba/ (VbaAba) and visual /ga/ – auditory /ga/ (VgaAga); as well as two incongruent, crossed stimuli: visual /ga/ – auditory /ba/ (VgaAba) cues which adults may fuse to an illusory percept /da/ and the opposite, non-fusible (mismatch) combination (VbaAga), which adults may perceive as /bga/. An additional silent and still face stimulus was presented to test whether infants' looking times to the mouth and to the eyes would vary as a function of the presence or absence of speech and lip movements.

According to the intersensory redundancy hypothesis we predicted that infants' attention will be captured more by the audiovisually redundant stimuli, that is congruent /ba/ and /ga/ and possibly fusible VgaAba combination. On the contrary, attention to the mouth for the salient mismatch between modalities (non-fusible VbaAga, /bga/) will be decreased. Based on the Lewkowicz and Hansen-Tift (2012) data, we expected developmental changes in attention to the mouth versus eyes area in infants aged 6 to 9 months with an age-related increase in their attention to articulating mouth. We further predicted that this increased attention to visual cues would be present for congruent and/or redundant AV speech information (canonical /ba/ and /ga/ and fusible VgaAba).

Methods

Participants

The final sample of 32 infants recruited from the East London area, UK, was equally divided into two age groups (6- to 7-month-olds and 8- to 9- month-olds) of 16 participants (11 girls and 5 boys in each). Eight additional infants were excluded due to fussiness (*n*=2) or low quality of eye-tracking data (<40% samples recorded, *n*=6). The mean age was 200.4 days (*SD*=10.97) for the younger group, and 269.0 days (*SD*=28.2) for the older group. The sample had a mixed ethnic composition (16 Caucasian, 5 Afro-Caribbean, 4 Asian, and 7 mixed ethnicity). In all but three cases (two in the older and one in the younger group) English was regularly spoken at infants' homes. The study received approval from the local university ethics committee.

Stimuli

Video recordings of a female, native English speaker articulating /ba/ and /ga/ sounds were edited to create single clips containing one instance of each speech sound articulation. The crossed speech sound stimuli were created by mixing the audio track with the incongruent articulation, thus producing visual /ba/+auditory /ga/ and visual /ga/+auditory /ba/ stimuli. The sound onset was adjusted in each clip at 360 ms from the stimulus onset and the auditory syllable lasted for the following 280-320 ms (see: Kushnerenko, et al., 2008). Each single clip was 760 ms long, each trial contained 10 repetitions of a single clip (10 instances of articulation) and was 7600 ms long. The video stimuli were rendered with a digitization rate of 25 fps. Stereo soundtracks were digitized at 44.1 kHz (16-bit resolution). The silent face trial consisted of a still image of the same speaker's face (single frame taken from the video recording with mouth

closed) displayed for an equivalent amount of time (7600 ms per trial). For stimuli sizes see Figure 1.

Experimental Procedure

In order to minimise the effects of lack of familiarity with incongruent AV stimuli, infants were previously familiarized with both canonical and crossed speech stimuli (three different speakers, total of 100 articulations per condition presented in random order). Participants were seated on a parent's lap in a dimly lit room in front of a Tobii T120 eye-tracker monitor (17"), at a distance of 60 cm. The parents' view of the stimuli was obscured, so that it would not interfere with the infants' eye-tracking recording. Eye movements were monitored online and recorded with a 120 Hz sampling rate. Following a successful calibration routine (5 points), each participant observed a total of ten experimental trials (2 trials per condition x 5 conditions). Before each trial, the participants' attention was attracted to the screen using colourful animations with sound, and terminated as soon as the infant fixated them. The first two and the last two trials were the canonical VbaAba and VgaAga trials, with their order counterbalanced between subjects. In between them, two trials of each incongruent condition and of silent face still images were displayed, in a random order. No effects of trial order on the total looking times were found (all *p*s>.22). The entire test lasted not more than 5 minutes.

_____Figure_1_here_____

The eye-movement data were analysed by specific Areas-Of-Interest (AOIs): mouth, eyes and the entire face oval (see Figure 1). The sum of total fixation lengths from

both trials in each condition was calculated for each participant and each AOI using Tobii Studio package and Tobii fixation filter. The data were ln-transformed to normalise their distribution. Data for each AOIs were analysed in 5 x 2 mixed-model ANOVAs with condition as a within-subject factor (silent face, VbaAba, VgaAga, VbaAga and VgaAba) and age group as a between-subject factor. Additional 4 x 2 ANOVAs (condition x age group) and planned pair-wise contrasts were carried out for the speech conditions (VbaAba, VgaAga, VbaAga and VgaAba), along with Pearson's correlations (reported two-tailed). Greenhouse-Geisser corrected and Bonferroni-corrected *p*-values were used where necessary.

Results

Attention to the mouth and eyes in the canonical and crossed speech conditions

Looking times to the mouth. A 4 x 2 ANOVA for the four AV speech sound conditions showed significant differences in attention to the mouth between two crossed (VbaAga & VgaAba) and two canonical (VbaAba & VgaAga) stimuli (condition x age group ANOVA, main effect of condition, $F(3,90)=3.52$, $p=.025$, $\eta_p^2=.11$). The effect was explained by longer looking to the mouth in the fusible condition (VgaAba) than either in the non-fusible (VbaAga) or canonical stimuli (pair-wise contrasts, all ps<.02). Importantly, a significant interaction of condition and age group was found ($F(3,90)=4.37$, $p=.006$, $\eta_p^2=.13$), which was explained by a significant contrast between the fusible and the non-fusible conditions ($p=.006$) and between the non-fusible and canonical /ga/ ($p=.033$). While 6- to 7-month-olds looked longer at the mouth in the fusible than in the non-fusible trials, 8- to 9-month-olds looked at the mouth equally long during both crossed stimuli. No main effect of age group was found ($F(1,30)=0.13$, $p=.73$).

A one-way ANOVA for the younger group (6- to 7-month-olds) showed a main effect of condition ($F(3,45)=5.56$, $p=.002$, $\eta_p^2=.27$). This effect was driven by a significant difference between the fusible and non-fusible conditions (pair-wise comparison, $p<.004$) as well as between fusible and canonical /ga/ ($p<.022$). Thus, infants in the younger age group looked significantly longer at the mouth in the fusible than the non-fusible and /ga/ conditions. They also looked for less time at the mouth in the non-fusible condition than the canonical /ba/ ($p= .015$) and canonical /ga/ conditions (approaching significance, $p=.062$). No difference was found in 6- to 7-month-olds' looking times between the two canonical conditions ($p=.72$).

In the same analysis conducted for infants aged 8-9 months, the main effect of condition approached significance ($F(3,45)=2.82$, $p=.07$, $\eta_p^2=.16$), with participants looking significantly longer at the mouth in the fusible and non-fusible conditions in comparison with the canonical /ba/ (both *p*s<.05). No looking time difference between the fusible and the non-fusible condition was found (*p*>.9), while the difference between the two crossed stimuli and the canonical /ga/ did not reach significance (both *p*s>.17).

Looking times to the eyes. Longer looking times to the mouth in the fusible condition were not related to a differential decrease of looking to the eyes. The looking time data showed no significant effects or interactions (all *p*s>.30). Thus, greater attention to the mouth in the crossed conditions did not result in a systematic decrease in the duration of looking to the eyes.

_____ Table 1 here_____

Correlations of looking times to the mouth area with age

Previous analyses have indicated the presence of age-related changes in infants' attention to the mouth while they were observing the non-fusible (VbaAga) video. This was confirmed by a significant positive correlation of participant age with the total time spent looking at the mouth in the non-fusible condition (Pearson's *r*=.401, *p*=.023, two-tailed), with older infants watching the mouth longer during this condition (Figure 2). Looking times to the mouth in no other condition significantly correlated with age (all *r*s<.18, all *p*s>.32).

_____Figure_2_here_____

Looking to the mouth as a proportion of looking to the entire face

We further investigated whether participants spent more time fixating on the mouth in the crossed conditions at a cost of decreased looking to other parts of the face. Firstly, a two-way ANOVA on total looking times to the entire face yielded no main effect of condition ($F(3,90) = 1.42$, $p = .24$), but only a significant interaction of condition and age group ($F(3,90) = 2.92$, $p = .039$, $\eta_p^2 = .089$). The interaction was explained by a nearly-significant difference between the two crossed conditions across the age groups ($F(1,30) = 3.85$, $p = .059$, $\eta_p^2 = .114$). Secondly, looking times to the mouth were calculated as a percentage of the total time spent looking at the face oval and submitted to a two-way ANOVA, which resulted in a significant main effect of condition ($F(3,90) = 5.02$, $p = .003$, $\eta_p^2 = .14$). The condition x age group interaction approached significance ($F(3,90) = 2.19$, $p = .1$, $\eta_p^2 = .07$). However, planned pair-wise contrasts revealed a significant interaction of age group and looking duration to the crossed stimuli (VbaAga vs. VgaAba, $F(1,30) = 6.07$, $p = .02$, $\eta_p^2 = .17$). This difference was driven primarily by much shorter looking to the mouth in the mismatch (VbaAga) than the fusible (VgaAba) conditions by 6- to 7-month-olds (71.42% vs. 52.34%, respectively) compared with a similar proportion in 8- to 9-month-olds (64.86% vs. 60.03%). Thus, the age-related difference in looking times to the mouth between the crossed conditions confirmed the previous raw looking times data.

Attention to face parts in the speaking versus silent face conditions

A 5 x 2 ANOVA with condition and age group for the face oval AOI showed that participants looked equally long at the entire face during the silent face and speech conditions ($F(4,120)=1.84$, $p=.16$, $\eta_p^2=.058$).

Looking times to the mouth. Infants looked longer at the mouth (see Table 1) while watching AV speech stimuli (articulating mouth) than still face images without accompanying sound. A condition x age group ANOVA for this AOI showed a significant effect of condition ($F(4,120)=21.50$, $p<.001$, $\eta_p^2=.42$), with all participants looking significantly longer at the mouth area for all speech conditions than at silent face image of equal duration (all *p*s<.001).

Looking times to the eyes. In contrast, participants looked longer at the eyes in the silent face condition than in the AV speech conditions (significant effect of condition, $F(4,120)=22.95$, $p<.001$, $\eta_p^2=.43$; pair-wise comparisons all *p*s<.001). Altogether, these results demonstrate that participants from both age groups distinguished between speech and non-speech conditions in terms of looking times on the areas of eyes and mouth.

Discussion

Our study provides new data on the development of audiovisual speech integration in pre-linguistic infants. We sought to establish whether 6- to 9-month-old infants show age-related increase in attention to the mouth while freely viewing dynamic audio-visual speech stimuli with congruent (canonical) and incongruent (crossed) patterns of mouth articulation with either /ba/ or /ga/ sound. In particular, we investigated whether this previously demonstrated attentional shift towards articulating mouth from 6 months of age (Lewkowicz & Hansen-Tift, 2012) is present solely for congruent and/or redundant AV speech cues or not.

Our results indicate that infants aged 6 to 9 months increase their attention to the mouth while watching AV speech, but only when auditory and visual cues are in apparent conflict and cannot be fused to a single percept (mismatch VbaAga condition). Inconsistently with the proposition of Lewkowicz and Hansen-Tift (2012) we did not find age-related increase in attention to the mouth for congruent (redundant) speech cues. In effect, it appears that intersensory redundancy hypothesis (Bahrick, Flom, & Lickliter, 2002; Lewkowicz, 2000) explains only looking behaviour of the younger (6-7 month old) but not of the older (8-9 month old) infants. Our study also demonstrates that 6- to 9-month-olds discriminate between congruent and incongruent speech cues in terms of total duration of looking to the mouth, but not to the eyes. Infants from the younger group (6-7 months) discriminated also between the two incongruent (crossed) conditions, with longer looking during the fusible (VgaAba) than the non-fusible (VbaAga) speech cues. In contrast, the older infants (8-9 months) looked equally long at the mouth in both incongruent conditions but longer than in congruent videos.

Fusible condition (VgaAba). Both 6- to 7- and 8- to 9-month-olds looked significantly longer at the mouth while viewing the incongruent fusible than the congruent videos. If infants were able to perceive AV cues in this condition as a unitary percept, as suggested by existing data (Burnham & Dodd, 2004), there would be no differences in looking times compared with canonical /ba/ and /ga/. Apparently, however, the incongruent fusible condition attracted more attention to the mouth area than normal congruent syllables. This cannot be explained solely by mismatch between auditory input and visual cues, as looking behaviour in the other incongruent AV condition (VbaAga) was different. A potential explanation is that while fixating the mouth in the fusible but not in the non-fusible condition, infants experienced a noticeable change in speech percept, which made them attend to mouth for a longer time. In adults the illusory /da/ is perceived only as long as the lip articulation of /ga/ is fixated while hearing the sound /ba/ (McGurk & MacDonald, 1976). Given the evidence for existence of McGurk illusion in infancy it is possible that greater attention to the mouth while watching fusible cues serves to maintain the novel illusory percept.

Non-fusible condition (VbaAga). Our data suggest that between 6 and 9 months of age there may be a developmental transition in infants' visual attention specifically to novel and not experienced before audiovisual speech patterns. While younger infants spent less time fixating the mouth in the non-fusible condition than in other speech conditions, older infants spent more time fixating the mouth in this condition than during congruent ones. This result was confirmed by a significant correlation of total looking to the mouth with age in this condition (non-fusible VbaAga), but not in any other.

For younger infants therefore the congruent audiovisual combinations represent more salient and attention-catching events then mismatched ones. This is in line with the

intersensory redundancy hypothesis and with previous studies showing that 2- and 4-month-old infants prefer watching faces with mouth articulations matching auditory vowels (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003). It should be also noted that the consonant cluster ‘bg’ is illegal in English (and many other languages), since the ‘g’ consonant does not normally follow ‘b’ in a natural speech environment. Thus, in terms of statistical learning, this combination could be regarded as a low probability speech event, and therefore unhelpful and uninformative in learning to categorise speech sounds which is especially important in early stages of language acquisition.

Conversely, older infants observed mouth movements during the presentation of the non-fusible combination longer than during the congruent stimuli. This may suggest that infants’ greater familiarity with native speech sounds and better knowledge of corresponding articulation patterns might result in increased attention to unknown combinations of speech cues. In this case the non-fusible stimulus (VbaAga) would be interpreted by 8- to 9-month-olds as a novel display, not compatible with articulatory and speech sound combinations known from the natural environment.

It is known from familiarization/habituation studies that when relatively little prior exposure to the stimuli of interest is provided, a preference for matching or familiar pairings might be expected, while greater experience with the stimuli should increase a preference for mismatching or novel pairings (see: Houston-Price & Nakai, 2004). While these findings are largely related to the ‘within experiment’ time window, this trend might be applicable to the age-related changes found in our study. For example, it is known that the same amount of familiarization can result in different looking

behaviour by infants of different ages: the older the infant, the more quickly they will develop a novelty preference (Houston-Price & Nakai, 2004).

An alternative account of our results is related to the neural audio-visual mismatch response documented previously in 5-month-olds (Kushnerenko, et al., 2008). In the same age group as our participants Kushnerenko and colleagues have found a strong correlation between the looking time to the mouth and the size of the right-central AVMMR (Kushnerenko, et al., under review). This suggests that the maturation of AV speech processing indexed by neural mismatch response (AVMMR) is also reflected in the pattern of visual attention to incongruent and non-redundant AV speech cues.

To summarise, using eye-tracking measures we have found evidence for infants' ability to discriminate between possible (fusible) and impossible audio-visual speech combinations in terms of total looking duration (6- to 9-month-olds). We have also found evidence for age-related changes in 6- to 9-month-old infants' attention to the mouth during the perception of incongruent, impossible and non-redundant audio-visual speech cues. The age-related shift in attention to non-fusible, mismatched speech cues found here, suggests that an important transition in perceptual learning of speech may occur between 6 and 9 months of age.. Importantly, our data add to the research on intersensory redundancy hypothesis, demonstrating that it is applicable to the early stages of language acquisition, but not to later development (from approximately 8 months onwards). Overall, our study highlights the potential role of audio-visual mismatch detection mechanisms in preverbal language development.

References

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Dev Psychobiol, 41*(4), 352-363.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Curr Dir Psychol Sci, 13*, 99-102.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., et al. (2009). Hearing faces: how the infant brain matches the face it sees with the speech it hears. *J Cogn Neurosci, 21*(5), 905-921.

Brookes, H., Slater, A., Quinn, P. C., Lewkowicz, D. J., Hayes, R., & Brown, E. (2001). Three-Month-Old Infants Learn Arbitrary Auditory-Visual Pairings Between Voices and Faces. *Inf Child Dev, 10*, 75-82.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Dev Psychobiol, 45*(4), 204-220.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci, 363*(1493), 1001-1010.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Dev Psychobiol, 45*(4), 187-203.

Gliga, T., & Csibra, G. (2009). One-year-old infants appreciate the referential nature of deictic gestures and words. *Psychol Sci, 20*(3), 347-353.

Houston-Price, C., & Nakai, S. (2004). Distinguishing Novelty and Familiarity Effects in Infant Preference Procedures. *Infant and Child Development, 13*, 341-348.

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control, 15*(1), 85-117.

Hunnius, S., & Geuze, R. H. (2004). Developmental changes of visual scanning of dynamic faces and abstract stimuli in infants: a longitudinal study. *Infancy, 6*, 231-255.

Jansson-Verkasalo, E., Ruusuvirta, T., Huotilainen, M., Alku, P., Kushnerenko, E., Suominen, K., et al. (2010). Atypical perceptual narrowing in prematurely born infants is associated with compromised language acquisition at 2 years of age. *Bmc Neuroscience, 11*.

Kuhl, P. K., & Meltzoff, A. N. (1982). The Bimodal Perception of Speech in Infancy. *Science, 218*(4577), 1138-1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behav Dev, 7*, 361-381.

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc Natl Acad Sci U S A, 100*(15), 9096-9101.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proc Natl Acad Sci U S A, 105*(32), 11442-11445.

Kushnerenko, E., Tomalski, P., Ribeiro, H., Potton, A., Axelsson, E. L., Murphy, E., et al. (under review). Brain responses to audiovisual speech mismatch in infants are associated with looking behaviour strategies.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychol Bull, 126*(2), 281-308.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Dev Psychol, 46*(1), 66-77.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc Natl Acad Sci U S A, 109*(5), 1431-1436.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Percept Psychophys, 24*(3), 253-257.

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *J Exp Child Psychol, 41*(1), 93-113.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Dev Sci, 11*(1), 122-134.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

Mottonen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Res Cogn Brain Res, 13*(3), 417-425.

Parise, E., Handl, A., Palumbo, L., & Friederici, A. D. (2011). Influence of Eye Gaze on Spoken Word Processing: An ERP Study With Infants. *Child Dev, 82*(3), 842-853.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behav Dev, 1999*(22), 2.

Patterson, M. L., & Werker, J. F. (2003). Two-moth-old infants match phonetic information in lips and voice. *Dev Sci, 6*(2), 191-196.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Percept Psychophys, 59*(3), 347-357.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia, 45*(3), 587-597.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neurosci, 10*, 21.

Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Dev, 75*(4), 1067-1084.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during thefirst year of life. *Infant Behav Dev, 7*, 49-63.

Figure legends

Figure 1. Stimulus position and size in visual angle along with the positioning and size of eyes and mouth Areas of Interest (AOIs).

Figure 2. Curve fit linear regression plot of age for looking times on the mouth AOI (natural log-transformed) for mismatch condition (VbaAga).

Table 1

Condition	**6-7 month-olds**			**8-9 month-olds**		
	Area-Of-Interest			Area-Of-Interest		
	Entire face	*Eyes*	*Mouth*	*Entire face*	*Eyes*	*Mouth*
Silent face	12.70 (0.67)	5.11 (1.05)	2.33 (0.70)	14.17 (0.67)	6.68 (1.05)	2.40 (0.70)
Congruent VbaAba	13.03 (0.96)	2.35 (0.48)	7.86 (1.01)	12.15 (0.96)	2.18 (0.48)	7.10 (1.01)
Congruent VgaAga	12.09 (0.93)	2.06 (0.49)	7.166 (0.81)	11.50 (0.93)	1.90 (0.49)	7.49 (0.81)
Non-fusible VbaAga	11.63 (0.90)	2.44 (0.66)	6.12 (0.99)	14.67 (0.90)	2.89 (0.66)	8.81 (0.99)
Fusible VgaAba	12.06 (0.90)	1.91 (0.50)	8.43 (0.80)	13.29 (0.90)	2.03 (0.50)	8.86 (0.80)

Table 1. Total looking times in seconds for each experiment condition in two participant age groups (both *n*=16) by Area-Of-Interest. Standard error of mean in brackets.