Yeung, H.H., & Werker, J.F., Lip movements affect infant audiovisual speech perception. Psychological Science, 24(5), pp. 603-612.

 DOI: https://doi.org/10.1177/0956797612458802

Lip Movements Affect Infant Audiovisual Speech Perception

H. Henny Yeung & Janet F. Werker

University of British Columbia

Author Note

H. Henny Yeung & Janet F. Werker, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, B.C., Canada V6T 1Z4

Henny Yeung is now at the Laboratoire Psychologie de la Perception (CNRS UMR 8158) and the Université Paris Descartes, Paris Sorbonne Cité

Correspondence concerning this article should be addressed to Henny Yeung, Laboratoire Psychologie de la Perception, Université Paris Descartes, 45 rue des Saints-Pères, 75006 Paris, France. E-mail: henny.yeung@parisdescartes.fr

Abstract

Speech is robustly audiovisual from early in infancy. Here we show that audiovisual speech perception in 4.5-month-old infants is further influenced by sensorimotor information from lip movements made while chewing or sucking. Experiment 1 consisted of a classic audiovisual matching procedure, where two simultaneously displayed talking faces (visual [i] and [u]) were presented with a synchronous vowel (audio /i/ or /u/). Compared to a baseline condition with nothing in the mouth, looking was selectively biased *away* from the audiovisual matching face when infants produced lip movements similar to the heard vowel, but returned to baseline when infants produced lip movements similar to the competing vowel. Experiment 2 confirmed that these sensorimotor effects interacted with the heard vowel, as looking patterns differed when infants produced identical lip movements while hearing an unrelated vowel (audio /a/). These findings suggest that the development of speech perception and speech production may be mutually informative.

Keywords: infant, multisensory, sensorimotor, audiovisual, speech perception

Lip Movements Affect Infant Audiovisual Speech Perception

Speech perception is robustly multisensory from early in development. For example, infants recognize corresponding input from audio and visual (AV) modalities from at least 2 months of age, spontaneously looking more at the face in a side-by-side display of two talking faces that matches a synchronously presented audio track (Kuhl & Meltzoff, 1982; 1984; Kuhl, Williams, & Meltzoff, 1991; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983; Patterson & Werker, 1999; 2003). Subsequent work has shown that this AV matching exhibits unique EEG signatures in the brain (Bristow et al., 2009), and shapes the learning of phonetic categories (Teinonen, Aslin, Alku, & Csibra, 2008). As has been demonstrated in adults (McGurk & MacDonald, 1976), visual speech information can even alter how auditory speech is perceived in infants (Burnham & Dodd, 2004; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Rosenblum, Schmuckler, & Johnson, 1997).

What mechanisms support this link between audio and visual modalities in infancy? AV speech linkages are unlikely to be learned associations between talking faces and speech sounds, as matching is seen between non-native speech and faces (Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009; Walton & Bower, 1993). Another possibility is that infants detect shared temporal or amodal properties across modalities (Bahrick, Lickliter, & Flom, 2004), which is the basis on which many AV correspondences are detected from birth (Lewkowicz, 2010). However, amodal properties like temporal synchrony are experimentally controlled in many studies where infants nevertheless accomplish

AV matching (Kuhl & Meltzoff, 1982; 1984; MacKain et al., 1983; Patterson & Werker, 1999; 2003). An alternative hypothesis is that infants map speech information in auditory and visual modalities onto a common articulatory representation (Kent & Vorperian, 2007; Kuhl & Meltzoff, 1984; 1988). First, AV matching is facilitated for speech relative to non-speech sounds (Kuhl & Meltzoff, 1984; Kuhl et al., 1991), suggesting that the speech signal has unique, domain-specific properties that map onto human faces. Second, infants sometimes produce congruent oral gestures when in AV speech contexts (Kuhl & Meltzoff, 1982; 1984; 1996; Legerstee, 1990; Patterson & Werker, 1999).

This articulatory hypothesis is bolstered by other research, which suggests that AV speech perception is linked with articulatory movements in adults (Hickok & Poeppel, 2007; Pulvermüller & Fadiga, 2010). For example, *perceiving* either auditory or visual speech affects speech production (Galantucci, Fowler, & Goldstein, 2009; Kerzel & Bekkering, 2000), while *producing* articulatory movements affects auditory speech perception in similar ways as seeing visual speech (Sams, Möttönen, & Sihvonen, 2005). In development, however, previous research has shown only that *perceiving* auditory speech affects early vocalizations: native language sound patterns affect newborn cries (Mampe, Friederici, Christophe, & Wermke, 2009), vowel-like utterances (Ruzza, Rocca, Boero, & Lenti, 2006) and babbling in infants (de Boysson-Bardies, Sagart, & Durand, 1984; Whalen, Levitt, & Goldstein, 2007), as well as the earliest word productions by toddlers (McCune & Vihman, 2001). Conversely, no studies have

ever manipulated infants' *production* of oral gestures to ask whether this affects either auditory, visual, or AV speech perception.

With few exceptions (Best, 1995; Kuhl & Meltzoff, 1988; McCune & Vihman, 2001; Vihman, 1993; Werker, 1993), there has been correspondingly little theoretical discussion about sensorimotor influences on the development of speech perception. This likely stems from the fact that infants begin perceiving sophisticated phonetic patterns long before oral gestures in young infants can be classified as articulatory (i.e., speech-related). For example, auditory language input may affect the production of early oral gestures, like babbling, which implies that such gestures are continuous with later word production, and are thus speech-related (McCune & Vihman, 2001). On the contrary, others argue that many aspects of babbling reflect universal constraints on the development of the motor system not specific to speech (Davis & MacNeilage, 1995). Moreover, the types of muscle movements made when infants and toddlers babble, suck, or chew do not appear to be continuous with mature speech motor control (Steeve, Moore, Green, Reilly, & McMurtrey, 2008).

In summary, the basis of infants' rich AV speech sensitivities remains unclear, although some suggest that articulatory information plays an important role. It is well established that articulatory information is linked to speech perception in adults, but it is not known whether a similar relation exists in young infants, and if such a relation exists, how it could be related to infants' relatively immature motor development. In two experiments, we tested whether very simple

sensorimotor features of non-speech oral gestures are related to AV speech processing.

Experiment 1

Experiment 1 relied on the similarity between the lip movements produced when adults articulate /i/ and /u/, and when 4.5-month-old infants engage in chewing and sucking. This is an age before infants start babbling or otherwise begin producing clear speech, and an age commonly tested in the developmental AV speech literature. A matching procedure was used that replicated many previous reports: a video of two talking faces (either visual [i] or visual [u]) was displayed side-by-side while a synchronized audio track matching one of the faces (either audio /i/ or audio /u/) was presented for a 2-minute period (Baier, Idsardi, & Lidz, 2007; Kuhl & Meltzoff, 1982; 1984; Kuhl et al., 1991; Patterson & Werker, 1999; 2003).

For a *baseline* group, the AV matching procedure was simply replicated with nothing in the mouth. For two other experimental groups, infants produced lip movements during the task that could be described as either "/i/-like" lip spreading, or "/u/-like" lip rounding. Both are illustrated in Figure 1. The *lip-sound match* group produced lip movements that matched the heard vowel (and mismatched the competing vowel), while the *lip-sound mismatch* group produced lip movements that mismatched the heard vowel (and matched the competing vowel; see Table 1).

If these lip movements were indeed related to AV matching, then results should differ across the three groups. One possible pattern was an articulatory

assimilation effect, echoing previous work in the adult speech literature (e.g., Sams et al., 2005). On this account, the lip-sound match group should activate motor features linked to corresponding audio and visual representations of the heard vowel, perhaps also suppressing competing representations. For example, lip-spreading should activate motor features shared with audio /i/ and visual [i], facilitating /i/-[i] matching, while lip-rounding should lead to the converse, facilitating /u/-[u] matching. The opposite prediction holds for the lip-sound mismatch group, where lip-spreading should suppress features of audio /u/ and visual [u], impairing /u/-[u] matching, while lip-rounding should impair /i/-[i] matching. In summary, an assimilation effect predicts that the lip-sound match group should be similar to the baseline group in showing a bias *towards* the AV matching face, and that the lip-sound mismatch group should show the converse pattern: a bias *away* from the AV match.

A second possibility was an articulatory contrast effect, echoing theories of action-perception from outside the speech domain. Here it is thought that common representations or processes are shared between perceptual and motor systems, and thus engaging motor processes can withhold related information from perceptual analysis, sometimes even biasing perceptual judgements in the opposite direction (Hamilton, Wolpert, & Frith, 2004; Schütz-Bosbach & Prinz, 2007). A contrast effect in the lip-sound match group predicts that lip-spreading should impair audiovisual /i/-[i] matching, while lip-rounding should impair /u/-[u] matching. The opposite pattern is predicted in the lip-sound mismatch group: Lip-spreading should facilitate /u/-[u] matching, while lip-rounding should facilitate /i/-

[i] matching. In summary, a contrast effect predicts that the lip-sound match group should be dissimilar to the baseline group in showing a looking bias *away* from AV matching face, and that the lip-sound mismatch group should show the converse pattern: the same bias as the baseline group *towards* the AV match.

Methods

Stimuli[1]. Two videos containing [i]- and [u]-faces were used, identical except the side on which each vowel appeared. Each was constructed from 10 clips of [i]- or [u]-articulations, the onsets of which were synchronized and occurred every 2 s. The duration of mouth opening and the onset of blinking were also synchronized, and these ten clips were looped until each video played continuously for approximately 2 min.

Stimuli videos were presented with an audio track, which was recorded in a separate session by the same woman watching the videos of herself articulating [i] and [u]. She produced vowels as closely as possible with the original audio tracks, and ten tokens were used to create new tracks, where the vowel onsets were edited to synchronize with the onsets in the original tracks. Durations of mouth-opening in the video were longer than those of the vowels ($M_{[i]}$ = 1.36 s; $M_{[u]}$ = 1.32 s; $M_{/i/}$ = .44 s; $M_{/u/}$ = .52 s), resembling the temporal dynamics between face and voice in the original recordings.

Procedure. Infants were seated in a caregiver's lap while eye-gaze was recorded with a Tobii 1750 eye-tracker positioned 60 cm from the infant at an angle of 30 degrees. Each face display covered a 9.8 cm x 9.8 cm square, symmetrically oriented around the center and separated horizontally by 2.7 cm.

During the test video, sound pressure levels ranged between 60 - 64 dB, and sound emanated from two speakers behind a cardboard barrier surrounding the eye-tracker. An experimenter monitored infants through a video feed.

The procedure began with gaze calibration, where a looming blue ball appeared in the center and four corners of the screen, accompanied by beeping sounds. Calibration points were marked when infants appeared to fixate on the relevant location. The test procedure began immediately afterwards, closely following previous paradigms (Patterson & Werker, 1999; 2003). One face was silently displayed for 9 s, followed by the other face for 9 s. Both faces were displayed together in silence for another 9 s, and finally the screen went blank for 3 s before the 2 min test movie was played (Figure 2). Which side appeared first, as well as the side of each face were counterbalanced across infants.

Lip-spreaders chewed or mouthed part of a larger object (i.e., too large to be a choking-hazard), and spread their lips to accommodate the object's width. Most infants were given a wooden teething ring[2] provided by the experimenter (1.2 cm in thickness and 6.8 cm in diameter; see Figure 1) (N = 22), but a few infants preferred another commercially available teething toy (N = 6), or their caregiver's horizontally oriented finger (N = 2), or a combination of any of these objects at different points during the test period (N = 2). Lip-rounders sucked on part of an object, usually a pacifier (N = 28). For a few infants, however, either the tip of their caregiver's finger (N = 3) or a combination of the finger and the pacifier (N = 1) was used at different points during the test.

Caregivers were instructed to attend to their baby, and to avoid fixating the visual display. In the baseline group they were also instructed to prevent their infants from chewing on any hands or clothing, while parents in the other groups were instructed to prevent the finger, object, or pacifier from being spit out. In the event that an object was not in their infant's mouth, caregivers were asked to adjust or replace it immediately. Clean teething rings or pacifiers were available under caregiver's chair for this purpose.

Participants. The analysis included ninety-six infants (48 female) with an average age of 4 months 18 days (*R* = 4;0 - 5;3), who heard English at least 30% of the time by parental report (*M* = 89%; *R* = 30% - 100%). They were randomly assigned to one of the three experimental groups, except that infants were occasionally re-assigned to a different group if they refused to chew and/or suck, or to balance gender across experimental orders. Ten additional infants were tested, but excluded due to experimenter error or equipment failure. Thirty-four additional infants were also tested, but excluded based on three *a priori* criteria derived from preliminary gaze analysis: if <4 calibration points could be recorded (*N* = 2); if recorded gaze was <40 s in the 2 min test period (i.e., a third of the total), which happened when infants were excessively fussy or disinterested, or if their position shifted so that the eye-tracker was unsuccessful at calculating gaze (*N* = 28); finally, if infants looked <1 s at one of the faces, demonstrating a side-bias (*N* = 4). This latter criterion assumed that these infants had trouble disengaging from one face, and followed previous reports (Kuhl & Meltzoff, 1984; Patterson & Werker, 1999; 2003).

Results

Gaze analysis was conducted in the two regions of interest over the displayed faces (see Figure 2) without applying any fixation filters or interpolative calculations. Total gaze was entered into an omnibus ANOVA with between-subjects factors of experimental group (baseline, lip-sound match, lip-sound mismatch), gender (male, female), vowel (heard /i/, heard /u/), and side of the AV match (left, right). No significant effects were found (*alpha* = .05), and looking at the faces was captured for an average of 81.03 s, *SD* = 21.94 s, during the 2 min test period. In the remaining time, gaze could not be localized, or infants looked at other regions of the screen.

Proportion looking at each face was then calculated (see Table 1). The proportion spent on the AV matching face was entered as a dependent variable into an omnibus ANOVA with the same between-subjects factors as above. As shown in Figure 3, results showed only a main effect of experimental group, $F(2, 72) = 3.46$, $p = .037$, $\eta^2 = .088$ (*alpha* = .05). Corrected post-hoc comparisons (two-tailed, Fisher's LSD) showed that infants looked more at the AV matching face in the baseline group, $M = .58$, $SD = .23$, compared to the lip-sound match group, $M = .43$, $SD = .27$, $t(72) = 2.38$, $p = .020$, *95% CI* [.03, .28]. Infants also looked more at the AV match in the lip-sound mismatch group, $M = .57$, $SD = .24$, compared to the lip-sound match group, $t(72) = 2.17$, $p = .033$, *95% CI* [.01, .27]. However, infants looked equivalently at the AV match in the baseline and lip-sound mismatch groups, $t(72) = .21$, $p = .84$, *95% CI* [-.12, .14].

Discussion

Results indicate that producing simple lip movements while chewing or sucking affects performance in an infant AV speech matching procedure. Specifically, looking differed from baseline when lip movements matched the heard vowel, but was unchanged from baseline when lip movements mismatched the heard vowel. This clearly indicates an articulatory contrast effect, such that infants *suppressed* speech representations similar to produced lip movements.

What specific processes explain this pattern of results? One possibility is that sensorimotor input biases visual preferences *away* from just the face most similar to the produced lip shape. On this account, the lip-sound match and mismatch groups looked more at the dissimilar facial expression, irrespective of what vowel was presented. While suggesting that facial expression matching is powerful enough to override AV matching, this account does not necessarily suggest that sensorimotor and auditory information interact.

A second possibility is that the observed effect reflects an interaction with audio and visual modalities. On this account, motor information in the lip-sound match group suppressed infants' ability to match audio and visual representations of the heard vowel, perhaps increasing activation of the competing representation. This resulted in a bias away from the AV matching face, compared to baseline. For the lip-sound mismatch group, motor information selectively suppressed information about the competing vowel, which resulted in performance similar to the baseline condition. This suggests that motor information selectively interacts with AV speech processing when it is aligned with both auditory and visual modalities.

Experiment 2

Experiment 2 distinguishes between the two possible interpretations of the effect from Experiment 1. Another group of infants lip-spread and lip-rounded in the same test procedure, except that the presented vowel (audio /a/) was *neutral* with respect to either the achieved lip shapes or the presented faces (see Table 1). On the first account (i.e., facial expression matching), this group of infants should continue to prefer only the face that produced a *mismatching* expression. On the second account (i.e., AV-motor interactions), this group should show a different pattern from infants who produced lip movements in Experiment 1, as auditory information in the presented vowel is unrelated to both the lip movements and the displayed faces.

Methods

Stimuli. The same stimuli from Experiment 1 were used, except the audio track contained ten tokens of the vowel /a/. These tokens were recorded by the same speaker in the same manner as the original /i/ and /u/ vowels (see Baier et al., 2007), and had correspondingly similar durations, $M_{/a/}$ = .44 s. Tokens were again placed at the onsets of the original audio track to create the videos.

Procedure. The stimuli and procedure were identical to Experiment 1. Lip-spreaders were given the same teething ring as before (N = 14), except one preferred a commercially available teething toy, and one a combination of the ring and a horizontally oriented finger. Lip-rounders sucked on a pacifier (N = 14), except for two infants, who preferred the tip of their caregiver's finger.

Participants. The analysis included thirty-six infants (18 female) with an average age of 4 months 20 days (*R* = 4;4 - 5;12), who heard English at least 30% of the time according to parental report (*M* = 87%; *R* = 50% - 100%). Five additional infants were tested, but excluded due to experimenter error (*N* = 2) or for hearing English less than 30% of the time (*N* = 3). Twenty-six others were also excluded based on the three *a priori* criteria from Experiment 1: if <4 calibration points were recorded (*N* = 0); if recorded gaze was <40 s (*N* = 25); or if a side-bias was observed (*N* = 1).

Results

Results were analysed together with those infants from Experiment 1 who also produced lip movements. Total gaze was entered into an omnibus ANOVA with between-subjects factors of experiment (Experiment 1, 2), gender (male, female), lip shape (lip-spreading, lip-rounding), and side of the visual face that matched the achieved lip shape (left, right). No significant effects were found (*alpha* = .05), and overall looking at the faces averaged 82.34 s, *SD* = 20.30 s, during the 2 min test period.

To explicitly test the facial expression matching hypothesis, the proportion of looking at the visual face that matched the achieved lip shape (i.e., the lip-matching face) was entered as a dependent variable into an omnibus ANOVA with the same between-subjects factors as above. As shown in Figure 4, results showed only a main effect of experiment, $F(1, 80) = 4.77$, $p = .032$, $\eta^2 = .056$ (*alpha* = .05), as infants in Experiment 1 looked significantly less at the lip-

matching face, $M = .43$, $SD = .25$, compared to infants in Experiment 2, $M = .54$, $SD = .21$.

Discussion

Infants in Experiment 2 achieved lip shapes identical to those achieved in Experiment 1, but heard a neutral vowel (audio /a/) rather than a vowel matching one of the displayed faces (audio /i/ or audio /u/). Unlike the results from Experiment 1, infants in Experiment 2 did not show a bias towards the visual face mismatching the achieved lip shape. This indicates that the effects observed in Experiment 1 cannot be explained by facial expression matching without any interaction with the heard vowel.

General Discussion

Our findings reveal that sensorimotor information is directly implicated in AV speech processing from early in infancy. Specifically, looking patterns observed in Experiment 1 reflect two distinct trends: a selective bias away from the AV matching face in the lip-sound match group, and a return to baseline looking (back towards the AV matching face) in the lip-sound mismatch group. Experiment 2 confirms that this effect reflects an AV-motor interaction and does not reflect simple facial expression matching.

The contrast effect we observed is opposite to what is typically obtained in adult speech research, where ambiguous speech information is often biased *towards* adults' articulations (i.e., an assimilation effect). Nevertheless, our results are compatible with the rich literature showing both assimilation and

contrast effects between visual perception and bodily action (e.g., Hamilton et al., 2004; Schütz-Bosbach & Prinz, 2007). For example, one reported contrast effect shows that executing certain arm movements (i.e., drawing rising arcs) while viewing related visual displays (i.e., dots move in arc-shaped trajectories) biases perceptual identification away from features shared with the performed actions (i.e., dots appear to move in flatter arcs) (Grosjean, Zwickel, & Prinz, 2009). When such effects are observed in adults, it is hypothesized that shared or overlapping information between perception and action is withheld (or inhibited) in perceptual analysis (Grosjean et al., 2009), as judgments are biased towards perceptual hypotheses that do not recruit the same features as the performed action (Hamilton et al., 2004). Lip-spreading or lip-rounding could have a similar effect on AV speech, biasing infants' perceptual preferences towards the contrasting vowel.

The action-perception literature suggests principles by which one alternately elicits assimilation versus contrast effects. In some cases, the speed with which movements are executed can affect whether assimilation or contrast effects are detected (Grosjean et al., 2009), and the presently reported patterns might similarly vary as a function of task, or as a function of the fluency with which articulatory configurations in infants are achieved. In other words, as infants develop mastery over more complex articulatory schemas, and achieve closer approximations to what adults are doing, then contrast effects may prove to be unstable, perhaps disappearing at one age and later reappearing as an assimilation effect at another.

Provocatively, our results indicate that speech is not just a multisensory system in early infancy, but also a sensorimotor one. We show that coarse-grained sensorimotor information about the articulators (e.g., lip-rounding, lip-spreading, jaw-opening, etc.) is available to perceptual systems processing AV speech. Further work is needed to determine the precise format of this information: Is it amodal, based on gestural events (Best, 1995), or is it somatosensory, based on feedback from skin receptors about oral gestures (Ito, Tiede, & Ostry, 2009)? No matter the precise format, such sensorimotor features likely become more specific in development, becoming embedded in richer and more coordinated gestural movements. As such, sensorimotor oral influences may become increasingly restricted to speech-like gestures at older ages (see also Best, 1995). We further hypothesize that this development may similarly track the development of speech motor control, where patterns of muscle activity begin to distinguish babbling from chewing and sucking from 9 months of age, and then speech from other oral gestures from around 15 months of age (Steeve et al., 2008).

The established view in developmental research is that sensorimotor information has little influence on the perceptual development of speech in the first-year. Audio feedback from babbling and other infant vocalizations may constitute a kind of indirect sensorimotor influence (McCune & Vihman, 2001; Vihman, 1993), but previous work has stopped short of demonstrating direct influences on perception. Our results show that sensorimotor information from explicitly non-speech oral gestures, like chewing and sucking, are indeed linked

to AV speech perception by at least 4.5 months of age, before any clearly speech gestures are produced. This not only demonstrates a direct sensorimotor link to speech perception earlier in development than previously thought, but also suggests that the development of sensorimotor systems could be pivotal in explaining why developmental changes happen when they do in both AV speech perception (Pons et al., 2009) and auditory speech perception at large.

Acknowledgements

This work was supported by NSERC and James S. McDonnell Foundation grants to J.F.W. and a fellowship from the Fondation Fyssen to H.H.Y. We thank Rebecca Baier, Bill Idsardi, and Jeff Lidz for sharing stimuli as well as Eric Bateson, Jim Enns, Bryan Gick, Ali Greuel, David Lefkowicz, and Athena Vouloumanos for comments on the manuscript.

Notes

[1] The authors of Baier et al., 2007 generously offered these stimuli.

[2] Item #1004 manufactured by Camden Rose ®, Ann Arbor, MI, USA.

References

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, *13*(3), 99-102. doi:10.1111/j.0963-7214.2004.00283.x

Baier, R., Idsardi, W. J., & Lidz, J. (2007). Two-month-olds are sensitive to lip rounding in dynamic and static speech events. In J. Vroomen, M. Swerts, & E. Krahmer (Eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing*. Retrieved from http://spitswww.uvt.nl/Fsw/Psychologie/AVSP2007/papers/BaierIL_AVSP2007.pdf

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Eds.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-204). Timonium, MD: York Press.

de Boysson-Bardies, B., Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language*, *11*(1), 1-15.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (2009). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, *21*(5), 905-921. doi:10.1162/jocn.2009.21076

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk

effect. *Developmental Psychobiology*, *45*(4), 204-220.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech & Hearing Research*, *38*(6), 1199-1211.

Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception & Psychophysics*, *71*(5), 1138-1149. doi:10.3758/APP.71.5.1138

Grosjean, M., Zwickel, J., & Prinz, W. (2009). Acting while perceiving: assimilation precedes contrast. *Psychological Research*, *73*(1), 3–13.

Hamilton, A., Wolpert, D., & Frith, U. (2004). Your own action influences how you perceive another person's action. *Current Biology*, *14*(6), 493-498. doi:10.1016/j.cub.2004.03.007

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393-402. doi:10.1038/nrn2113

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1245-1248. doi:10.1073/pnas.0810063106

Kent, R. D., & Vorperian, H. K. (2007). In the mouths of babes: Anatomic, motor, and sensory foundations of speech development in children. In R. Paul (Ed.), *Language disorders from a developmental perspective: Essays in honor of Robin S. Chapman.* (pp. 55-81). Mahwah, NJ: Lawrence Erlbaum.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 634-647. doi:10.1037/0096-1523.26.2.634

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 Behavior & Development*, *7*(3), 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 Symposium on Child Development. (Vol. 20, pp. 235-266). Hilldale, NJ: Lawrence Erlbaum.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, *100*(4), 2425-2438.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 829-840.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11442-11445. doi:10.1073/pnas.0804275105

Legerstee, M. (1990). Infant use of multimodal information to imitate speech sounds. *Infant Behavior and Development*, *13*(3), 343–354.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony.

Developmental Psychology, *46*(1), 66-77. doi:10.1037/a0015579

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, *219*(4590), 1347-1349. doi:10.1126/science.6828865

Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, *19*(23), 1994-1997. doi:10.1016/j.cub.2009.09.064

McCune, L., & Vihman, M. M. (2001). Early phonetic and lexical development: A productivity approach. *Journal of Speech, Language, and Hearing Research*, *44*(3), 670-684. doi:10.1044/1092-4388(2001/054)

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-8. doi:1012311

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior & Development*, *22*(2), 237-247.

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

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*(26), 10598-10602.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351-

360. doi:10.1038/nrn2811

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347-357. doi:10.3758/BF03211902

Ruzza, B., Rocca, F., Boero, D. L., & Lenti, C. (2006). Investigating the musical qualities of early infant sounds. *Annals of the New York Academy of Sciences*, *999*, 527–529.

Sams, M., Möttönen, R., & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research*, *23*(2-3), 429-435.

Schütz-Bosbach, S., & Prinz, W. (2007). Perceptual resonance: Action-induced modulation of perception. *Trends in Cognitive Sciences*, *11*(8), 349-355. doi:10.1016/j.tics.2007.06.005

Steeve, R. W., Moore, C. A., Green, J. R., Reilly, K. I., & McMurtrey, J. R. (2008). Babbling, chewing, and sucking: Oromandibular coordination at 9 months. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1390-1404. doi:10.1044/1092-4388(2008/07-0046)

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*.

Vihman, M. M. (1993). Vocal motor schemes, variation and the production-perception link. *Journal of Phonetics*, Phonetic development, *21*(1-2), 163-169.

Walton, G. E., & Bower, T. G. R. (1993). Amodal representation of speech in infants. *Infant Behavior and Development*, *16*(2), 233–243.

Werker, J. F. (1993). The contribution of the relation between vocal production and perception to a developing phonological system. *Journal of Phonetics*, Phonetic development, *21*(1-2), 177-180.

Whalen, D. H., Levitt, A. G., & Goldstein, L. M. (2007). VOT in the babbling of French- and English-learning infants. *Journal of Phonetics*, *35*(3), 341-352. doi:10.1016/j.wocn.2006.10.001

Figure Legends

Figure 1. Lip-spreaders chewed on a toy and/or finger, oriented in such a way that infants' lips would be repetitively spread as when articulating the vowel /i/. Lip-rounders sucked on a pacifier or fingertip, and this similarly ensured that infants' lips would be repetitively rounded as when articulating the vowel /u/.

Figure 2. The experimental procedure is illustrated here: (clockwise from upper left) a photo of the eye-tracking apparatus, a schematic video timeline for one version of the procedure, and a still video image that illustrates regions of interest for [u]- and [i]-faces (i.e., dashed and solid gray squares, respectively).

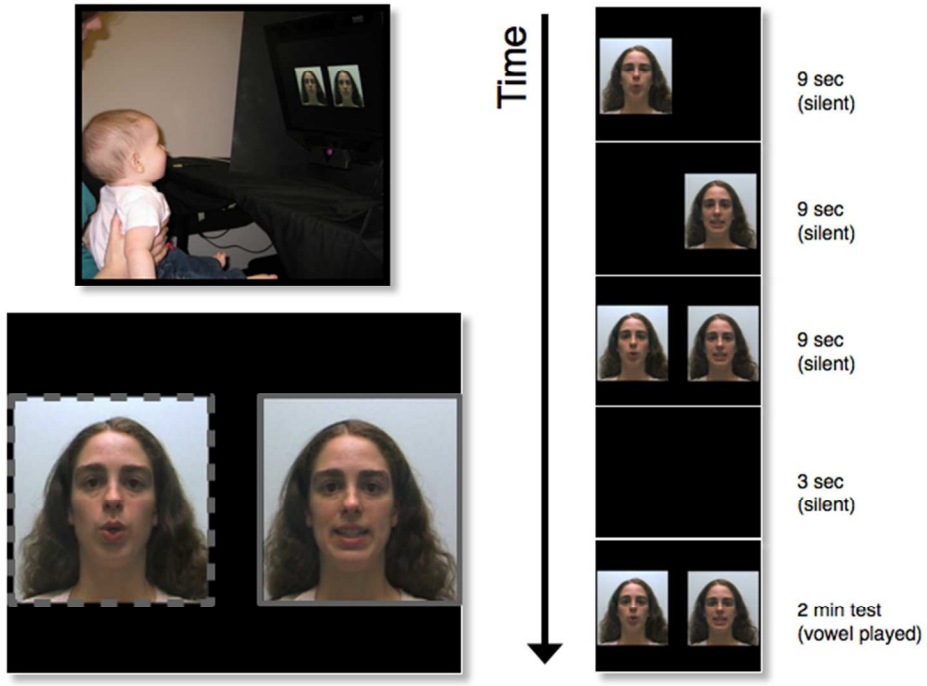
Figure 3. Results for each group in Experiment 1 are plotted as the proportion looking to the face that was an audiovisual match. The lip-sound match group was biased away from the AV matching face compared to both baseline, as well as the lip-sound mistmatch group. Error bars indicate std. errors and asterisks indicate significant differences between groups ($p < .05$).

Figure 4. Results from those infants in Experiment 1 who achieved lip shapes (and heard /i/ or /u/), are plotted with results from infants in Experiment 2, who achieved the same lip shapes (and heard /a/). Infants in Experiment 2 looked significantly more at the lip-matching face than those infants in Experiment 1, showing that there was not a simply a global tendency to look at the mismatching facial expression. Results from Experiment 1 are thus due to an interaction between the motor and AV speech processes. Error bars indicate std. errors and asterisks indicate significant differences between groups ($p < .05$).

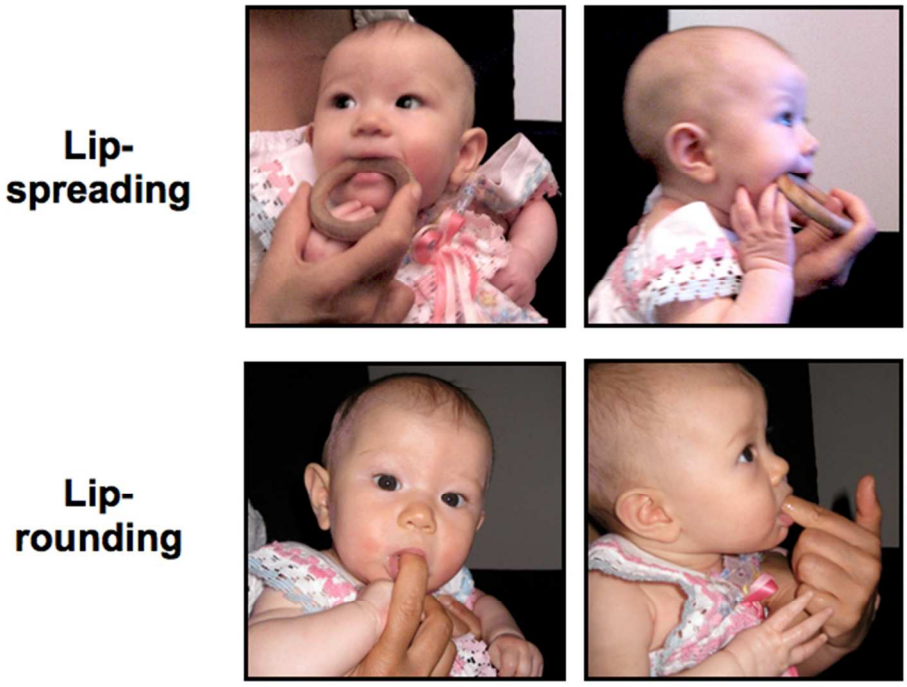
Table 1

Distribution of heard vowels and produced lip movements across conditions. Sub-group means (and std. deviations) indicate proportion looking to the visual [i] face.

Experiment 1	**Heard Vowel = /i/**	**Heard Vowel = /u/**
Baseline (*n* = 32)	*No lip-shape* *M* = .64 (.23)	*No lip-shape* *M* = .47 (.23)
Lip-Sound Match (*n* = 32)	*Lip-spreading* *M* = .47 (.26)	*Lip-rounding* *M* = .62 (.28)
Lip-Sound Mismatch (*n* = 32)	*Lip-rounding* *M* = .55 (.30)	*Lip-spreading* *M* = .42 (.17)
Experiment 2	**Heard Vowel = /a/**	
Lip-Sound Neutral (*n* = 32)	*Lip-spreading* *M* = .61 (.20)	*Lip-rounding* *M* = .52 (.21)



254x190mm (72 x 72 DPI)



254x190mm (72 x 72 DPI)