Developmental Psychology

Audiovisual Speech Perception in Infancy: The Influence of Vowel Identity and Infants' Productive Abilities on Sensitivity to (Mis)Matches Between Auditory and Visual Speech Cues

Nicole Altvater-Mackensen, Nivedita Mani, and Tobias Grossmann

Online First Publication, November 23, 2015. http://dx.doi.org/10.1037/a0039964

Developmental Psychology

0012-1649/15/$12.00 http://dx.doi.org/10.1037/a0039964

Audiovisual Speech Perception in Infancy: The Influence of Vowel Identity and Infants' Productive Abilities on Sensitivity to (Mis)Matches Between Auditory and Visual Speech Cues

Nicole Altvater-Mackensen
Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Nivedita Mani
Georg-August-University Göttingen

Tobias Grossmann
University of Virginia

Recent studies suggest that infants' audiovisual speech perception is influenced by articulatory experience (Mugitani et al., 2008; Yeung & Werker, 2013). The current study extends these findings by testing if infants' emerging ability to produce native sounds in babbling impacts their audiovisual speech perception. We tested 44 6-month-olds on their ability to detect mismatches between concurrently presented auditory and visual vowels and related their performance to their productive abilities and later vocabulary size. Results show that infants' ability to detect mismatches between auditory and visually presented vowels differs depending on the vowels involved. Furthermore, infants' sensitivity to mismatches is modulated by their current articulatory knowledge and correlates with their vocabulary size at 12 months of age. This suggests that—aside from infants' ability to match nonnative audiovisual cues (Pons et al., 2009)—their ability to match *native* auditory and visual cues continues to develop during the first year of life. Our findings point to a potential role of salient vowel cues and productive abilities in the development of audiovisual speech perception, and further indicate a relation between infants' early sensitivity to audiovisual speech cues and their later language development.

Keywords: audio-visual speech perception, phoneme learning, babbling

Speech addressed to infants mostly involves face-to-face interaction and typically provides the infant with bimodal speech input, that is, both auditory and visual speech information. Infants seem to be sensitive to the congruency between these speech cues from early on as they prefer to look at silent speaking faces whose articulations are congruent with a heard sound relative to incongruent articulations (Kuhl & Meltzoff, 1982, 1988; Patterson & Werker, 1999, 2003). Furthermore, infants detect correspondences between sequentially presented auditory sounds as well as between variant sequentially presented auditory sounds and visual mouth gestures, suggesting that they are similarly sensitive to speech cues within and across modalities (Bristow et al., 2009). These findings suggest that infants' audiovisual speech perception relies on abstract categories linking auditory and visual speech information (Bristow et al., 2009; Patterson & Werker, 1999), and that a basic ability to match auditory and visual speech cues might be innate or develop with little experience early in infancy (e.g., Kuhl & Meltzoff, 1982).

However, more recent work suggests that audiovisual speech perception undergoes significant development in the first year of life, paralleling the development in auditory speech perception (for a review of native language attunement in the auditory domain see Saffran, Werker, & Werner, 2006). In particular, infants lose sensitivity to visual differences between languages (Weikum et al., 2007) while their sensitivity to the correspondence of auditory and visual speech cues that are not relevant to their native language similarly declines (Pons, Lewkowicz, Soto-Faraco, & Sebastian-Galles, 2009). At the same time they seem to refine their perception of native audiovisual speech contrasts (Lewkowicz, 2000; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). These findings raise questions as to the factors that influence this development.

Several researchers have argued that changes in speech perception are linked to changes in infants' ability to produce native speech sounds (see, e.g., Vihman, 1996; Kuhl, 2000; Westermann & Reck Miranda, 2004). Such a link between speech perception and production seems especially plausible (and might be particularly fruitful to investigate) for audiovisual speech perception because visual speech cues are inherently linked to the articulatory gestures leading to its production. Thus, infants' experience in producing a sound might modulate their ability to match a sound to the (visual) gesture

Nicole Altvater-Mackensen, Research Group Early Social Development, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany; Nivedita Mani, Georg-Elias-Müller Institute for Psychology, Georg-August-University Göttingen; Tobias Grossmann, Department of Psychology, University of Virginia.

Correspondence concerning this article should be addressed to Nicole Altvater-Mackensen, Research Group Early Social Development, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, D-04103 Leipzig, Germany. E-mail: altvater@cbs.mpg.de

producing it. For instance, it might become easier for an infant to detect audiovisual (in)congruencies in native sounds if she is able to produce the sounds herself because she can recruit her own knowledge about the link between an articulatory gesture and the resulting auditory sound pattern.

Indeed, there is some evidence that productive knowledge influences audiovisual speech perception: Eight-month-olds are more sensitive to the match between (nonnative) auditory and visual speech cues of bilabial trills, which often occur in early babbling, than of whistles, which are unlikely to be part of an infant's productive repertoire (Mugitani, Kobayashi & Hiraki, 2008); 4.5-month-olds' ability to match auditory and visual speech cues of native vowels is influenced by concurrently performed (nonlinguistic) lip movements (Yeung & Werker, 2013; see also Legerstee, 1990); and the likelihood of integrating auditory and visual speech cues is modulated by preschoolers' productive language skills (Desjardins, Rogers, & Werker, 1997). Taken together these studies suggest a potential influence of productive abilities on audiovisual speech perception. Yet, given the age of the participants (Desjardins et al., 1997), the nonnative status of the assessed sound contrast (Mugitani et al., 2008) and the nonlinguistic nature of the performed motoric actions (Yeung & Werker, 2013), they provide only limited information about the influence of emerging productive abilities on native speech perception in the first year of life. Thus, it remains unclear if there is a functional link between speech perception and production that modulates learning of (audiovisual) native sound categories. The current study therefore investigated the link between an infant's ability to match native auditory and visual speech cues and her ability to produce different sounds in babbling.

To examine the link between productive and perceptive development, the choice of an appropriate sound contrast is crucial. In particular, the perception of the chosen sounds should improve during development and their production in babbling should follow different time lines. Investigating infants' audiovisual perception of vowels might reveal such developmental differences: Infants perceive vowels less categorically than consonants (e.g., Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) and are also biased in their auditory vowel perception in that they perceive some vowel contrasts better than others (for a review see Polka & Bohn, 2003, 2011). This might render it difficult to associate vowels with distinct visual gestures and lead to substantial improvement in perception during development. Indeed, even adults often confuse vowels that have similar visual and acoustic properties, such as /y/-/u/ or /Ø/-/o/, but are rather accurate in their perception of open vowels that do not have close counterparts, such as /a/ (van Son, Huiskamp, Bosman, & Smoorenburg, 1994; see also Richie & Kewley-Port, 2008). Because vowels occur early in infants' vocalizations (Stark, 1980), their perception might further be modulated by productive knowledge. Two particularly suitable contrasts are /a/-/e/ and /a/-/o/. These vowels are articulatory and acoustically less distinct than the corner vowels /a/, /i/, and /u/ that have been examined in earlier studies, rendering them potentially more difficult. Moreover, low and front vowels, such as /a/ and /e/, tend to dominate over back vowels, such as /o/, in infants' early productions (see review in de Boysson-Bardies et al., 1989). Thus, infants are likely to show differences in which sounds are part of their productive repertoire, making these sounds suitable to investigate the interplay between perception and production.

The current study tested German 5.5- to 6-month-olds' ability to detect mismatches between the auditory and visually presented native vowel contrasts /a/-/e/ and /a/-/o/. Infants were presented with videos of a female mouthing a vowel that was either congruent with the vowel they concurrently heard or not, while their looking times to the videos were measured. If infants were sensitive to the congruency between auditory and visual speech cues, we expected them to look longer at matching than mismatching videos (see Mugitani et al., 2008). To ensure that infants were able to acoustically discriminate the vowels, we also presented them with an auditory discrimination task. Again, a preference paradigm was used. We measured infants' attention to trials in which two vowels were presented in alternation (alternating trials) and trials in which one vowel was repeated (nonalternating trials). If infants discriminated the vowels, we expected them to listen longer to alternating trials than to nonalternating trials (see Best & Jones, 1998). In addition, we collected parental reports on infants' babbling, that is, which sounds consistently occurred in infants' productions (see Procedure section for more detail). We tested infants aged 5.5 to 6 months because they are old enough to have a sufficiently large attention span to complete the experiment. Importantly, they are also old enough to start producing well-formed native syllables in their babbling (Oller, 1978; Stark, 1980), but are still young enough to be in the midst of native language attunement in perception (see Polka & Werker, 1994 for the finding that perceptual attunement for vowels occurs between 4 and 10 months of age). We hypothesized that, if infants' ability to detect audiovisual mismatches is related to their productive knowledge, they would be more likely to detect audiovisual (in)congruencies if they produce similar sounds in babbling. This would provide evidence that perceptive and productive development is interrelated in language acquisition. It would furthermore show that infants do not match auditory and visual cues for all sounds equally well, but that they improve their sensitivity to audiovisual congruencies in the course of development.

To follow-up on infants' language development, we asked parents to fill out a vocabulary questionnaire 6 months after they participated in the experiment, that is, around the infants' first birthday. Because earlier native language attunement in auditory speech perception is related to more advanced vocabulary development (Tsao, Liu, & Kuhl, 2004), we hypothesized that a similar relation exists for the development of audiovisual speech perception and later language development, that is, that increased sensitivity to the congruency between auditory and visual speech cues at 6 months of age would predict more advanced vocabulary development at 12 months of age. Because earlier studies further found a relation between infants' early babbling and their later vocabulary size (Majorano, Vihman, & DePaolis, 2014; McCune & Vihman, 2001), we also hypothesized that—given the reliability of our parental report measure—infants' productive abilities at 6 months of age would be related to their vocabulary size 6 months later.

Method

The data reported in the current article was collected as part of a larger study that also addressed social influences on the acqui-

sition of phonological categories. The experimental data as well as the vocabulary data presented here is drawn from the same sample as the data presented in Altvater-Mackensen and Grossmann (2015). The current article and Altvater-Mackensen and Grossmann (2015) complement each other in that they address different questions, rely on different variables and present different analyses.

Participants

Forty-four German monolingual 6-month-olds participated in the experiment (16 female; age range: 5;11 (months; days) to 6;04, mean age: 5;23). All infants were born full-term with normal birth weight and had no reported hearing or vision impairment. Seven additional infants started to cry and could not be tested, one infant was excluded from analysis because his looking times were more than two standard deviations away from the mean. Infants were recruited via the participant pool of the first author's institute. Parents gave informed consent to participate in the study and received 7.50 Euro and a toy for their infant for participation.

Stimuli

Visual stimuli for the auditory discrimination task consisted of a video of a moving colored toy water wheel against a black background (Stager & Werker, 1997). Video frames were 1,200 pixels wide and 880 pixels high, resulting in a width of 32 cm and a height of 24 cm on screen. The video was accompanied by successive repetition of six different tokens of /a/ in nonalternating trials, and three tokens of each /a/ and /e/ or /a/ and /o/ in alternating trials (see Best & Jones, 1998 for the use of alternating and nonalternating trials to assess sound discrimination in infants). Tokens were separated by approximately 1.5 s of silence, leading to a trial length of 15 s. All vowels were spoken by a female native speaker of German using infant-directed speech (see below for further details on the acoustic characteristics of the vowels).

Visual stimuli for the audio-visual matching task consisted of videos showing a woman mouthing /a/, /e/, and /o/. Each video entailed six consecutive utterances of the respective vowel. Each utterance started and ended with the mouth completely shut in neutral position. Visual stimuli were hyperarticulated to mimic infant-directed speech (see below for further details on the visual characteristics of the vowels). Each vowel articulation was separated by approximately 3 s in which the woman kept a friendly open face and smiled at the infant, leading to a trial length of 30 s. Her eye-gaze was always directed toward the infant. All videos were zoomed and cropped so that they only showed the woman's head against a light-gray wall. Video frames were 1,024 pixels wide and 1,000 pixels high, resulting in a width of 27 cm and a height of 26 cm on screen. Figure 1 shows an example frame of the mouth position for each of the fully articulated vowels.

Audio stimuli for the audio-visual matching task consisted of six different tokens each of /a/, /e/, and /o/, spoken in infant-directed speech. Stimuli were recorded by the same woman who produced the visual stimuli and the auditory stimuli of the discrimination task. The length of the vowels was timed to match the length of the mouthing in the videos. The final stimuli were created by dubbing the audio recordings of the vowels onto the videos of the woman mouthing the vowels. For matching trials, visual and auditory vowels were the same. For mismatching trials, visual and auditory vowels did not fit, that is, seen /a/ was accompanied by heard /e/ or /o/, and vice versa (see Mugitani et al., 2008 for the use of matching and mismatching trials to assess sensitivity to audiovisual congruency). Note that visual and auditory stimuli used in matching and mismatching trials were identical, only their pairing changed across trial types.

Three additional familiarization trials were created using different recordings of the same woman uttering each vowel twice in a block of three repetitions followed by an engaging smile and raise of her eyebrows. All stimuli were digitally recorded in a quiet room with a sampling rate of 24 frames per second and 44.100 Hz. Vowels were matched in volume (mean intensity: /a/ = 76.0 db, /e/ = 77.3 db, /o/ = 77.1 db), fundamental frequency (mean pitch: /a/ = 191.5 Hz, /e/ = 196.8 Hz, /o/ = 199.8 Hz), and length (mean duration: /a/ = 1.83 s, /e/ = 1.79 s, /o/ = 1.84 s). Furthermore, /e/ and /o/ matched in vowel height (mean F1: /a/ = 986.3 Hz, /e/ = 433.2 Hz, /o/ = 462.8 Hz) and differed similarly from /a/ in vowel backness (mean F2: /a/ = 1597.8 Hz, /e/ = 2754.3 Hz, /o/ = 956.2 Hz). To estimate the visual distance between vowels, we measured the horizontal and vertical opening of the mouth, that is, we assessed the distance in pixels between left and right lip corner as well as the distance between upper and lower lip based on a still image of each fully articulated vowel (see Green et al., 2010 for a similar method). Vowels differed in horizontal lip opening (mean distance between left and right lip corner in pixels: /a/ = 106.33,

Figure 1. Example of the mouth positions for fully articulated /a/, /e/, and /o/ (adapted with permission from "Learning to Match Auditory and Visual Speech Cues: Social Influences on the Acquisition of Phonological Categories" by N. Altvater-Mackensen & T. Grossmann, *Child Development, 86,* p. 366. Copyright 2014 by Society for Research in Child Development, Inc.). Note that the original videos were colored. The individual whose face appears here gave signed consent for her likeness to be published in this article.

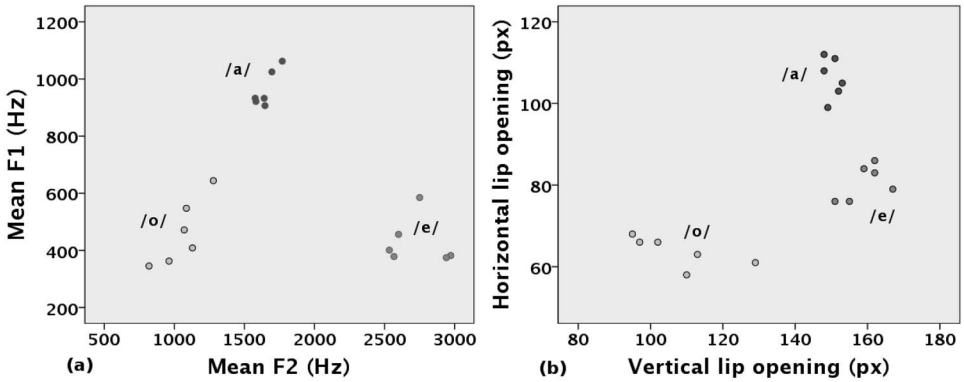

Figure 2. Vowel tokens used in the audiovisual matching task plotted by their F1 and F2 frequencies in hertz (a), and their horizontal and vertical lip opening in pixels (b). Each circle represents one auditory and visual token, respectively.

/e/ = 80.67, /o/ = 63.67) as well as vertical lip opening (mean distance between upper and lower lip in pixels: /a/ = 150.17, /e/ = 159.33, /o/ = 107.67). Although the difference in lip opening on both axes was smaller for the contrast /a/-/e/ than for the contrast /a/-/o/, the three vowels formed distinct visual clusters. Figure 2 plots each vowel token used in the audiovisual matching task based on first and second formant frequency (a) and based on horizontal and vertical lip opening (b) to visualize the acoustic and visual differences between vowel categories in our stimulus set.

To ensure that matches and mismatches in the audiovisual stimuli are detectable and do not elicit any McGurk-effect (McGurk & MacDonald, 1976), we validated the stimuli in a prestudy with adults, testing eight German native speakers (one male; mean age: 25 years, age range: 20–38 years) in a reaction time (RT) task. Stimulus videos from the infant study were cut in half so that each video contained three repetitions of a vowel and had a trial length of 15 s. The stimulus set for the adult prestudy thus included six matching videos, two videos each for /a/, /e/, and /o/, and eight mismatching videos—two videos each for seen /a/ accompanied by heard /e/ and /o/ and for heard /a/ accompanied by seen /e/ and /o/. Each participant was presented with all 14 videos in random order on a 44 cm wide notebook screen. For each video, the participant had to indicate as fast as possible whether the seen and heard vowels matched or mismatched by pressing the appropriate key on the keyboard. Results showed that participants categorized the stimuli correctly in almost all cases, only two trials were wrongly classified (one participant classified one matching /e/-video as mismatching, one participant classified one mismatching /e/-video as matching) This suggests that adult native speakers accurately perceived the videos as matching and mismatching. To further investigate potential delays in classifying mismatching stimuli, we also analyzed the RT data. Two participants initially did not understand that they had to answer as fast as possible and their responses to the first trials had to be excluded from analysis (six trials with RTs more than two standard deviations above the mean were excluded). The remaining data was aggregated by subject and submitted to a repeated-measures ANOVA with trial type (matching/mismatching) and vowel (e/o) as within-subjects factor. Results show a main effect of vowel, $F(1, 6) = 7.413$, $p = .035$, but no effect of trial type and no interaction between factors ($ps > .17$). A post hoc paired sample t test comparing the RTs to /e/- and /o/-videos aggregated across trial type revealed that although participants answered marginally faster for /o/-videos ($M = 3.08$ s, $SD = 0.75$) than for /e/-videos ($M = 3.34$ s, $SD = 0.67$), this difference was not significant, $t(7) = 1.602$, $p = .15$. Taken together, the results of the prestudy confirm that adult speakers of German rated the auditory and visual stimuli as compatible with one another in matching videos, and incompatible with one another in mismatching videos.

Procedure

When invited to participate in the experiment (about a week before the experiment took place), parents were informed that they would be asked about their infant's vocal productions to give them some time to observe their infant and to remember which sounds the infant produced. Before taking part in the experiment, parents were interviewed about their infant's babbling behavior. We asked them to describe their infant's babbling to assess whether infants showed canonical babbling, and to name the sounds that the infant produced. After noting every sound that the parent spontaneously named, we once again specifically asked which vowels and consonants the infant produced. Questions were open to avoid any bias in parents' answers. Because parents had no phonetic training, we did not assume that their reports would be phonetically accurate. Parents rather relied on their knowledge about the German sound inventory and named the sounds that most closely resembled their infants' productions. If they were unsure how to classify a sound, they mimicked their infants' babbling, leaving it to the experimenter to decide which would be the appropriate sound category.[1] Thus, this parental report measure does not provide any phonetically accurate description of an individual infant's sound inventory. Such information can only reliably be obtained by recording and carefully transcribing infants' productions. However, because parents most likely only remember and report those sounds that are

[1] Note that German has a very close phoneme-grapheme correspondence. Thus, reports should be similar regardless of whether mothers—being literate—relied on graphemic representations of sounds or on their auditory representation.

reliably and consistently produced by their infant, this measure provides an estimate (a) of whether an infant produces a variant of a certain phoneme, and (b) of how many different phonemes the infant targets in his or her babbling.

For the perception test, infants were seated on their parent's lap facing a 52 cm wide and 32.5 cm high TV screen at a distance of 40 cm from the screen. Parents wore headphones playing music intermixed with speech during the experiment. The intermixed speech consisted of sentences taken from another experiment as well as of the vowels used in the current experiment. So the music effectively blinded parents' perception of the auditory stimuli and of the match between auditory and visual stimuli in the audiovisual task. Parents were instructed to interact as little as possible with their infant. A camera mounted below the screen recorded infants' eye-movements during the experiment. Auditory stimuli were presented via loudspeakers that were located behind the screen. Based on the video image, the experimenter started a trial when the infant was looking at the screen and continued to indicate throughout the trial whether the infant was looking at the screen or not by pressing the corresponding button on a keyboard. Each trial lasted until the infant was looking away for more than 2 consecutive seconds or until completion. In between trials a flashing light was displayed in silence to reorient infants toward the screen.

Each infant was first presented with the three familiarization trials showing the woman uttering /a/, /e/, and /o/, to familiarize infants with the speaker and her characteristics. This was immediately followed by the discrimination task. Infants were presented with a total of eight nonalternating /a/ trials, four alternating /a/-/e/ trials, and four alternating /a/-/o/ trials. Half of the alternating trials started with /a/, the other half started with /e/ and /o/, respectively. Trial order was pseudorandomized so that each nonalternating trial was followed by an alternating trial and so that /a/-/e/ and /a/-/o/ trials were evenly distributed across the 16 discrimination trials.

The discrimination task was immediately followed by the audiovisual matching task. Infants were presented with nine matching trials, three trials each for /a/, /e/, and /o/, and eight mismatching trials. Two mismatching trials each paired visual /a/ with auditory /e/ and /o/, and auditory /a/ with visual /e/ and /o/, respectively. Trial order was pseudorandomized so that no more than two consecutive trials contained the same auditory or visual stimulus and so that the different vowels as well as matching and mismatching trials were evenly distributed across the 17 audio-visual trials. On average, the experiment took approximately 10 min.

Around the infants' first birthday, that is, 6 months after the experimental session, we sent out letters to the parents and asked them to fill out a subpart of a standardized German questionnaire on language development for 12-month-olds (ELFRA 1; Grimm & Doil, 2000), assessing infants' receptive and productive vocabulary.

Data Analysis

We included the data from all 44 infants in the analysis. Infants provided data for 16 trials in the discrimination task and on average 16 trials (range: 6–17) in the audiovisual matching task. Note that the pattern of results did not change if infants that provided data for less than 17 trials in the audiovisual matching task ($n = 7$) were excluded. Based on the online codings of the experiment, we calculated the mean looking time to alternating and nonalternating trials in the discrimination task and to matching and mismatching trials in the audio-visual matching task, aggregated by vowel contrast and infant.

To assess reliability of the online codings, the data from 15% of the infants was reassessed offline using a digital video scoring system. A trained coder indicated for each 40 ms frame of the video whether the infant was looking at the screen or away. The coder was blind to experiment phase and trial type. The coding output was aligned with information about the phase of the experiment and the auditory stimulus presented. Mean looking times to the different trial types were calculated as described above and compared to the online codings. Reliability between online and offline codings was 99% ($r = .994$, $p < .001$).

Based on parental reports, we further examined whether an infant produced /a/-like, /e/-like, and /o/-like sounds in his or her babbling and calculated the overall size of the reported sound inventory, that is, how many different vowels and consonants the parents identified in their infants' babbling.

Based on the language questionnaire that we sent out 6 months after the experiment, we further assessed infants' vocabulary development at 12 months of age. The questionnaire contains a list of 164 words from 13 semantic classes, such as animals, body parts, and activities. To estimate infants' vocabulary size, we calculated how many words mothers marked as being understood by the infant. Eleven infants (four female) had to be excluded from the vocabulary analysis because parents did not send back the questionnaire.

Results

A paired-sample t test showed that infants preferred matching over mismatching trials in the audiovisual matching task, $t(43) = 3.097$, $p = .003$, $d = .23$; 32 out of 44 infants looking longer at matching trials, indicating that they were sensitive to the congruency between auditorily and visually presented vowels. Similarly, infants preferred alternating over nonalternating trials in the discrimination task, $t(43) = 6.368$, $p < .001$, $d = .45$; 37 out of 44 infants looking longer at alternating trials, confirming that they also acoustically discriminated the vowels. Figure 3 displays the mean difference in looking time between alternating and nonalternating trials in the discrimination task and between matching and mismatching trials in the audiovisual matching task (see also Altvater-Mackensen & Grossmann, 2015).

Differences Between Vowels

To investigate if performance in the audio-visual matching task differed for the two vowel contrasts, we ran a repeated-measures ANOVA with trial type (matching/mismatching) and vowel (e/o) as within-subjects factor. Results showed a main effect of trial type, $F(1, 43) = 8.931$, $p = .005$, $\eta_p^2 = .172$; and an interaction of trial type and vowel, $F(1, 43) = 13.372$, $p = .001$, $\eta_p^2 = .237$. Post hoc comparisons indicated that infants were sensitive to the congruency between auditorily and visually presented vowels for the contrast /a/-/o/, $t(43) = 5.509$, $p < .001$, $d = .60$; 33 out of 44 infants looking longer at matching trials, but not for the contrast /a/-/e/ ($p > .90$). To ensure that any failure to detect mismatches in audiovisual perception cannot be attributed to difficulties in the acoustic discrimination of the vowels, we assessed performance in

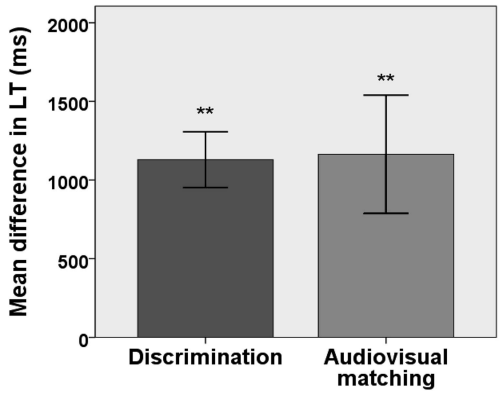

Figure 3. Mean difference in looking time (LT) to alternating minus nonalternating trials in the discrimination task and matching minus mismatching trials in the audiovisual matching task. Error bars indicate ±1 *SE*, asterisks indicate a significance level of ** $p < .01$ (adapted with permission from "Learning to Match Auditory and Visual Speech Cues: Social Influences on the Acquisition of Phonological Categories" by N. Altvater-Mackensen & T. Grossmann, *Child Development, 86,* p. 368. Copyright 2014 by Society for Research in Child Development, Inc.).

the discrimination task separately for the contrasts /a/-/e/ and /a/-/o/. Results showed that infants acoustically discriminated both /o/, $t(43) = 4.803$, $p < .001$, $d = .38$; 34 out of 44 infants looking longer at alternating trials, and /e/ from /a/, $t(43) = 5.108$, $p < .001$, $d = .49$; 35 out of 44 infants looking longer at alternating trials. Figure 4 displays the mean difference in looking time between trial types in the audiovisual matching and the discrimination task separately for the contrasts /a/-/o/ and /a/-/e/.

Influence of Productive Abilities

According to parental report the majority of infants produced /a/-like (33 infants) and /e/-like sounds (27 infants) in their babbling, while only 15 infants produced /o/-like sounds.

To determine a possible influence of infants' production of a vowel on their audiovisual matching abilities, we ran separate repeated-measures ANOVAs for each vowel (e/o) with trial type (matching/mismatching) as within-subjects factor and reported production (yes/no) as between-subjects factor. Results showed a main effect of trial type for /o/, $F(1, 43) = 24.544$, $p < .001$, $\eta_p^2 = .369$. No other main effects or interactions reached significance ($ps > .14$). (Note that categorizing infants by their production of spread and rounded vowels more generally provides similar results.) Post hoc comparisons confirmed that infants detected mismatches for /a/-/o/, regardless of whether they were reported to produce, $t(14) = 2.485$, $p = .026$, $d = .49$; 10 out of 15 infants looking longer at matching trials, or not produce an /o/-like sound in their babbling, $t(28) = 4.996$, $p < .001$, $d = .64$; 23 out of 29 infants looking longer at matching trials, while infants showed no sensitivity to audiovisual mismatches for /a/-/e/, regardless of whether they were reported to produce an /e/-like sound or not ($ps > .20$).

Again, we conducted paired-samples *t* test on the mean looking time to alternating and nonalternating trials in the discrimination task separately for those infants that, according to parental report, did and did not produce vowels similar to /e/ and /o/ in their babbling. Results confirmed that infants discriminated /o/ and /a/, regardless of whether they were reported to produce, $t(14) = 2.935$, $p = .011$, $d = .34$; 11 out of 15 infants looking longer at alternating trials, or not produce /o/-like sounds, $t(28) = 3.740$, $p = .001$, $d = .41$; 23 out of 29 infants looking longer at alternating trials. Similarly, infants discriminated /e/ from /a/, regardless of whether they were reported to produce, $t(26) = 2.876$, $p = .008$, $d = .36$; 19 out of 27 infants looking longer at alternating trials, or not produce /e/-like sounds, $t(16) = 4.950$, $p < .001$, $d = .67$; 15 out of 17 infants looking longer at alternating trials. Figure 5 displays the mean difference in looking time between trial types for the audiovisual matching and the discrimination task for /a/-/o/ and /a/-/e/ depending on infants' reported production of /o/- and /e/-like vowels in babbling.

Success in the audiovisual matching task depends on the recognition not only of /o/ and /e/, but also of /a/. Eighteen infants were reported to produce both /a/ and /e/ and 13 infants were reported to produce both /a/ and /o/. We further investigated whether the conjoint presence of /a/-/e/ and /a/-/o/ in infants' babbling inventory modulated performance in the audiovisual matching task, running separate repeated-measures ANOVAs for each vowel

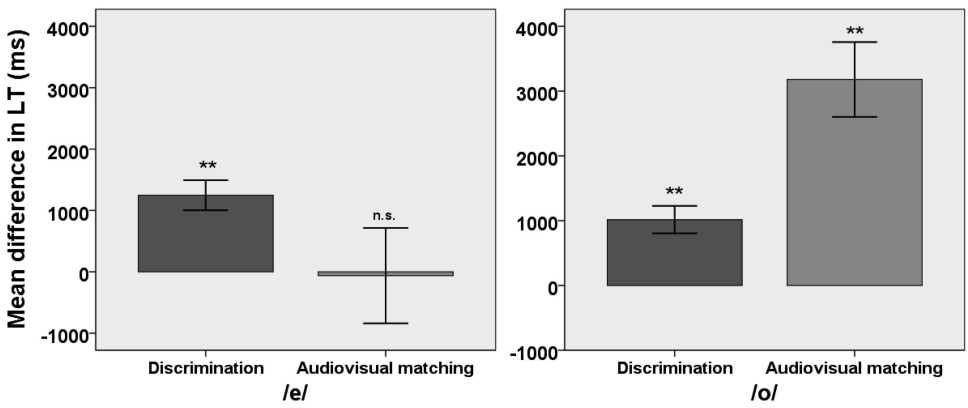

Figure 4. Mean difference in looking time (LT) to alternating minus nonalternating trials in the discrimination task and to matching minus mismatching trials in the audiovisual matching task plotted by vowel. Error bars indicate ±1 *SE*, asterisks indicate a significance level of ** $p < .01$, *ns* indicates nonsignificant effect.

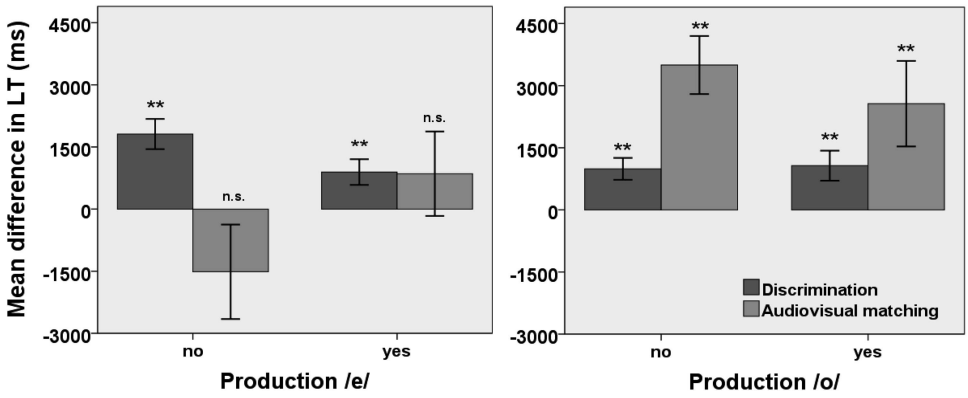

Figure 5. Mean difference in looking time (LT) to alternating minus nonalternating trials in the discrimination task and to matching minus mismatching trials in the audiovisual matching task plotted by vowel and its production. Error bars indicate ±1 *SE*, asterisks indicate a significance level of ** $p < .01$, *ns* indicates nonsignificant effect.

(e/o) with trial type (matching/mismatching) as within-subjects factor and conjoint production (yes/no) as between-subjects factor. Results for /o/ showed a main effect of trial type, $F(1, 42) = 23.611$, $p < .001$, $\eta_p^2 = .360$, while results for /e/ revealed a main effect of conjoint production, $F(1, 42) = 4.402$, $p = .042$, $\eta_p^2 = .095$, and an interaction between trial type and conjoint production, $F(1, 42) = 7.474$, $p = .009$, $\eta_p^2 = .151$. No other main effects or interactions reached significance ($ps > .17$).

Post hoc comparisons confirmed that infants detected audiovisual mismatches involving /a/-/o/, regardless of whether they were reported to produce both /a/ and /o/, $t(12) = 2.546$, $p = .026$, $d = .59$; nine out of 13 infants looking longer at matching trials, or not, $t(30) = 4.872$, $p < .001$, $d = .61$; 24 out of 31 infants looking longer at matching trials. Similarly, auditory discrimination for the contrast /a/-/o/ was successful in producers, $t(12) = 2.464$, $p = .03$, $d = .37$; nine out of 13 infants looking longer at alternating trials and nonproducers, $t(30) = 4.066$, $p < .001$, $d = .40$; 25 out of 31 infants looking longer at alternating trials. In contrast, infants' sensitivity for audiovisual mismatches involving /a/-/e/ was modulated by the conjoint production of /a/ and /e/, with producers showing a (nonsignificant) preference for matching trials, $t(17) = 1.809$, $p = .088$, $d = .32$; nine out of 18 infants looking longer at matching trials, and nonproducers showing a (nearly significant) preference for mismatching trials, $t(25) = -2.017$, $p = .055$, $d = -.31$; 10 out of 26 infants looking longer at matching trials. Again, infants discriminated between /a/ and /e/ regardless of whether they were reported to produce both sounds in their babbling, $t(17) = 2.646$, $p = .017$, $d = .38$; 13 out of 18 infants looking longer at alternating trials, or not, $t(25) = 4.411$, $p < .001$, $d = .56$; 22 out of 26 infants looking longer at alternating trials.

To further investigate whether there is a general influence of infants' productive abilities on their sensitivity to audiovisual mismatches, we correlated the number of different sounds produced in infants' babbling with their preference for matching trials. Indeed, higher productive abilities, as expressed by larger babbling inventories, led to a stronger preference for matching trials, $r(44) = .261$, $p_{(\text{one-tailed})} = .044$.[2] Infants' preference for alternating trials in the discrimination task, however, did not correlate with the size of their reported babbling inventory ($p_{(\text{one-tailed})} > .16$). A repeated-measures ANOVA with trial type (matching/mismatching) as within-subjects factor and babbling inventory (small/large based on the median split) as between-subjects factor showed a main effect of trial type, $F(1, 42) = 10.519$, $p = .002$, $\eta_p^2 = .200$, and a nonsignificant interaction between trial type and babbling inventory, $F(1, 42) = 2.997$, $p = .091$, $\eta_p^2 = .067$. Post hoc comparisons confirmed that infants with larger babbling inventories detected audiovisual mismatches, $t(20) = 3.210$, $p = .004$, $d = .35$; 17 out of 21 infants looking longer at matching trials, while infants with smaller babbling inventories showed no sensitivity for the matching between auditory and visual speech cues ($p > .25$). Again, paired-sample t tests confirmed that infants successfully discriminated the vowels regardless of whether they had small, $t(22) = 6.353$, $p < .001$, $d = .48$; 21 out of 23 infants looking longer at alternating trials, or large babbling inventories, $t(20) = 3.323$, $p = .003$, $d = .43$; 16 out of 21 infants looking longer at alternating trials. Figure 6 plots infants' preference for matching trials in the audiovisual matching task as a function of the number of different sounds being reported as produced in babbling (a), and displays the mean difference in looking time between trial types in the audiovisual matching and the discrimination task depending on the size of infants' babbling inventory (b).

[2] Given previous studies showing that advanced development in speech perception and more mature babbling is positively related to later language development (McCune & Vihman, 2001; Tsao et al., 2004), we expected that the size of infants' babbling inventories would positively correlate with their later vocabulary size, and that better matching abilities would positively correlate with both the size of infants' babbling inventories and their later vocabulary size. We therefore report one-tailed p values for these correlations. Note that the more conventional two-tailed p values would only approach significance for the correlation between size of babbling inventory and matching preference ($p = .088$) and the correlation between size of babbling inventory and vocabulary size ($p = .062$), even though post hoc comparisons show consistent performance differences between groups.

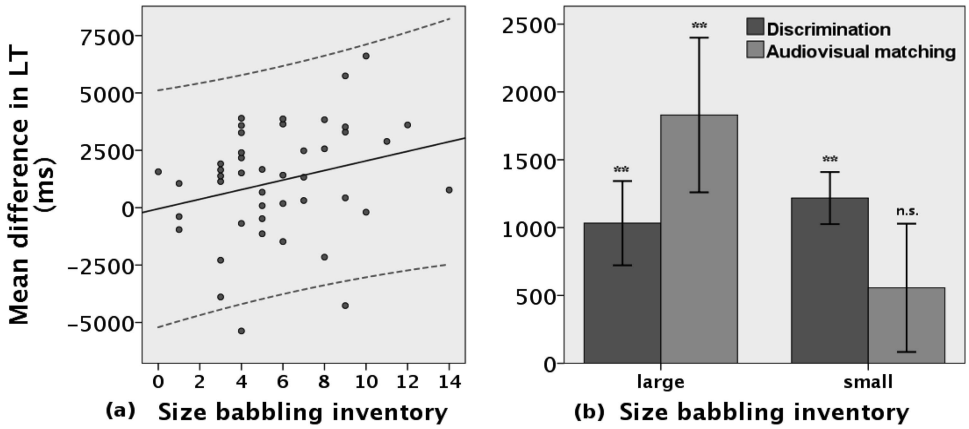

Figure 6. (a) Mean difference in looking time to matching minus mismatching trials plotted against the size of infants' babbling inventory, that is, the number of different sounds parents reported as produced in babbling. The gray line depicts the linear regression line ($R^2 = .068$), dotted lines indicate the 95% confidence interval. (b) Mean difference in looking time (LT) to alternating minus nonalternating trials in the discrimination task and to matching minus mismatching trials in the audiovisual matching task plotted by inventory group (based on the median split). Error bars indicate ±1 *SE*, asterisks indicate a significance level of ** $p < .01$, *ns* indicates nonsignificant effect.

Relation to Later Vocabulary Size

According to parental reports collected when the infants were 12 months of age, infants understood on average 35.2 words (range: 4 to 91). Infants' vocabulary size at 12 months of age positively correlated with their reported babbling inventory at 6 months of age, $r(33) = .328$, $p_{(\text{one-tailed})} = .031$, (see Footnote 2) suggesting a consistent development from early babbling to early word learning. Even though infants with larger babbling inventories at 6 months of age knew on average 11 words more at 12 months of age than infants with smaller inventories, this difference did not reach significance in an independent sample *t* test ($p > .16$). Figure 7 plots the number of words reported as understood at 12 months as a function of the number of different sound categories parents reported as being produced in babbling at 6 months.

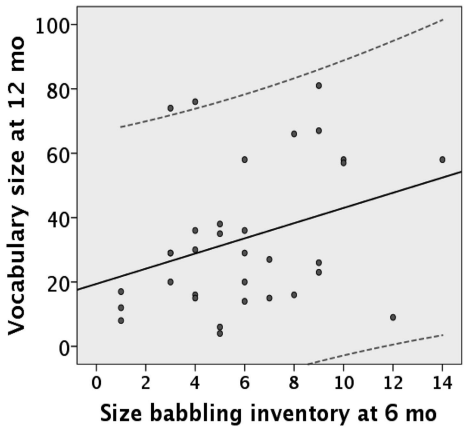

Figure 7. Mean size of infants' babbling inventory, that is, the number of different sounds parents reported as produced in babbling, plotted against infants' vocabulary size, that is, the number of words parents reported as understood at 12 months of age. The gray line depicts the linear regression line ($R^2 = .108$), dotted lines indicate the 95% confidence interval.

To investigate the potential relation between infants' early audiovisual speech perception and their later language development, we correlated infants' preference for matching trials in the audiovisual matching task with their vocabulary size. Infants' sensitivity to mismatches between auditory and visual speech cues was indeed associated with more advanced vocabulary development 6 months later, $r(33) = .360$, $p_{(\text{one-tailed})} = .020$ (see Footnote 2; see also Altvater-Mackensen & Grossmann, 2015). To further investigate the relation between early audiovisual speech perception and later vocabulary size, we ran a repeated-measures ANOVA with trial type (matching/mismatching) as within-subjects factor and vocabulary group (large/small based on the median split) as between-subjects factor. Results showed a main effect of trial type, $F(1, 31) = 8.353$, $p = .007$, $\eta_p^2 = .212$, and an interaction between trial type and vocabulary group, $F(1, 31) = 4.505$, $p = .042$, $\eta_p^2 = .127$.

Post hoc comparisons on the interaction between vocabulary size and trial type showed that infants with larger vocabularies detected audiovisual mismatches, $t(16) = 3.970$, $p = .001$, $d = .96$; 15 out of 17 infants looking longer at matching trials, while infants with smaller vocabularies were not sensitive to the matching between auditory and visual speech cues ($p > .63$). Again, we conducted paired-sample *t* tests on the mean looking time to alternating and nonalternating trials in the discrimination task separately for those infants who had high and low vocabulary scores to exclude difficulties in the acoustic discrimination of the vowels. Results confirmed that infants discriminated the vowels, regardless of whether they had large, $t(16) = 4.443$, $p < .001$, $d = 1.08$; 15 out of 17 infants looking longer at alternating trials, or small vocabularies, $t(15) = 3.338$, $p = .004$, $d = .83$; 14 out of 16 infants looking longer at alternating trials. Figure 8 displays infants' preference for matching trials in the audiovisual matching

task as a function of the number of words reported as understood in the vocabulary questionnaire (a), and plots the mean difference in looking time between trial types in the audiovisual matching and the discrimination task depending on infants' vocabulary size (b).

Visual Control Study

Results indicate that infants are more sensitive to matches between auditory and visual speech cues for the contrast /a/-/o/ than for the contrast /a/-/e/. The difference between vowels is not likely to arise from auditory difficulties as infants showed equal sensitivity to both sound contrasts in the auditory discrimination task. Yet, although we relied on similar visual differences as previous studies reporting successful audiovisual matching (e.g., Kuhl & Meltzoff, 1982), infants might have had difficulties to visually distinguish the vowels. To exclude this possibility, we ran a control study testing visual discrimination of the contrast /a/-/e/ and /a/-/o/ in 5.5- to 6-month-olds.

Participants

Twenty-three German monolingual 6-month-old infants participated in the experiment (11 female; age range: 5;15 (months; days) to 6;00, mean age: 5;25). Three additional infants did not contribute data because of crying (two) or technical failure (one). Recruitment and reimbursement were identical to the experiment presented above.

Stimuli

Stimuli for the visual discrimination task were created from the matching stimuli of the audiovisual matching task, adopting the design of the stimuli in the auditory discrimination task. Videos consisted of successive repetitions of six different tokens of /a/ in nonalternating trials, and three tokens of each /a/ and /e/ or /o/, respectively, in alternating trials. Each token started and ended with the mouth completely shut in neutral position. Tokens were separated by approximately 1.5 s during which a blank (black) screen was presented, leading to a trial length of 30 s. In addition, the familiarization trials from the experiment presented above were used.

Procedure and Data Analysis

Parents were interviewed about their infants' babbling behavior and infants were subsequently tested in the visual discrimination task. Parents wore blinded glasses instead of headphones playing masking noise during the experiment, to prevent perception of the visual stimuli. To redirect infants' attention more effectively to the screen in between trials, we added a ringing noise to the flashing light. Otherwise the procedure, the set-up and the coding were identical to the interview and the discrimination task of the experiment presented above.

Results

A paired-sample *t* test showed that infants preferred alternating over nonalternating trials, $t(22) = 3.975$, $p = .001$, $d = .43$; 16 out of 23 infants looking longer at alternating trials, indicating that infants were sensitive to the visual difference between the presented vowels. Post hoc comparisons confirmed that infants visually discriminated both /e/, $t(22) = 3.741$, $p = .001$, $d = .45$; 19 out of 23 infants looking longer at alternating trials, and /o/ from /a/, $t(22) = 2.985$, $p = .007$, $d = .38$; 17 out of 23 infants looking longer at alternating trials, making it unlikely that difficulties to

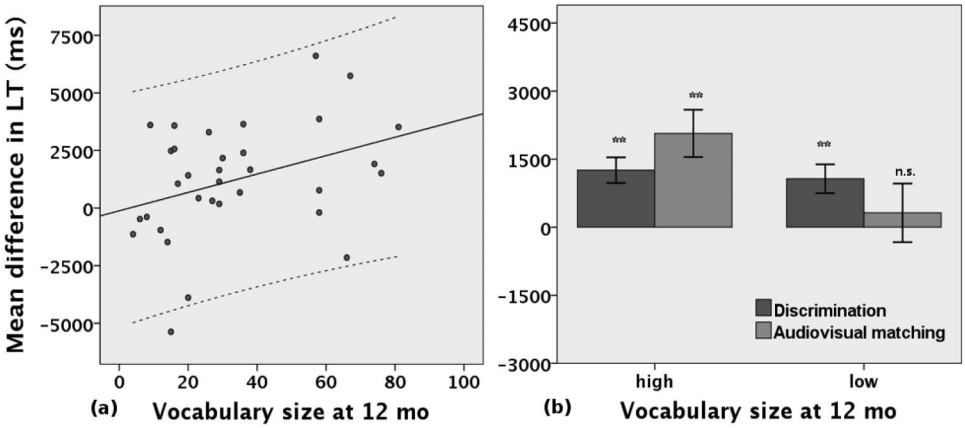

Figure 8. (a) Mean difference in looking time to matching minus mismatching trials in the audiovisual matching task plotted against infants' vocabulary size, that is, the number of words parents reported as understood at 12 months of age (adapted with permission from "Learning to Match Auditory and Visual Speech Cues: Social Influences on the Acquisition of Phonological Categories" by N. Altvater-Mackensen & T. Grossmann, *Child Development, 86,* p. 372. Copyright 2014 by Society for Research in Child Development, Inc.). The gray line depicts the linear regression line ($R^2 = .130$), dotted lines indicate the 95% confidence interval. (b) Mean difference in looking time (LT) to alternating minus nonalternating trials in the discrimination task and to matching minus mismatching trials in the audiovisual matching task plotted by vocabulary group (based on the median split). Error bars indicate ± 1 *SE*, asterisks indicate a significance level of $^{**}p < .01$, *ns* indicates nonsignificant effect.

visually discriminate between the vowels contributed to the observed differences in the audiovisual matching task.[3]

Similar to the experiment presented above, parents reported that infants were more likely to produce /a/-like (19 infants) and /e/-like sounds (15 infants) than /o/-like sounds (four infants). To determine a possible influence of infants' production of a vowel on their visual discrimination abilities, we ran separate repeated-measures ANOVAs for each vowel (e/o) with trial type (alternating/nonalternating) as within-subjects factor and reported production (yes/no) as between-subjects factor. Results showed a main effect of trial type for /e/, $F(1, 21) = 12.729$, $p = .002$, $\eta_p^2 = .377$; and /o/, $F(1, 21) = 8.110$, $p = .01$, $\eta_p^2 = .279$. No other main effects or interactions reached significance ($ps > .18$). (Note that results were similar for the conjoint presence of /a/-/e/ and /a/-/o/ as only two infants who produced /e/ or /o/ did not produce /a/.) A Pearson correlation further confirmed that infants' preference for alternating trials in the discrimination task did not correlate with the size of their reported babbling inventory ($p_{(\text{one-tailed})} > .15$). Figure 9 displays the mean difference in looking time between alternating and nonalternating trials in the visual discrimination task overall and split by vowel contrast.

Discussion

The current study investigated 5.5- to 6-month-olds' ability to match native auditory and visual vowel cues and related it to their ability to discriminate the corresponding vowels, to their reported articulatory abilities and to their later vocabulary size. There are three major findings: First, our results indicate that infants did not perform equally well for all vowel contrasts tested. They showed sensitivity to audiovisual congruencies for the contrast /a/-/o/, but not for the contrast /a/-/e/, suggesting that specific sound characteristics modulate infants' ability to detect audiovisual congruencies in native vowels. Second, infants' sensitivity to audiovisual (mis)matches was related to their productive abilities. Infants who were reported to produce more distinct sound categories in their babbling were more sensitive to the congruency between auditory and visual speech cues. Third, infants' preference for matching audiovisual speech and the size of their babbling inventory at six months of age predicted vocabulary size at 12 months of age, pointing to a relation between early audiovisual speech perception, articulatory abilities, and later language development. In what follows, we will discuss each of these findings, their implications and the questions that they raise in more detail.

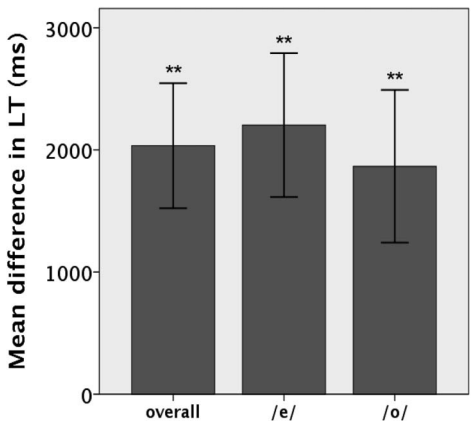

Figure 9. Mean difference in looking time (LT) to alternating minus nonalternating trials in the visual discrimination task, overall and plotted by vowel. Error bars indicate ±1 *SE*, asterisks indicate a significance level of ** $p < .01$.

Vowel Characteristics Modulate AV Speech Perception

As noted above, infants detected mismatches between auditory and visual speech cues for /a/-/o/, but not for /a/-/e/ (see Figure 4). This suggests that even though infants possess basic abilities to detect audiovisual congruencies from early on (Kuhl & Meltzoff, 1982, and subsequent studies), these abilities might not include all speech sounds (see also Mugitani et al., 2008). Alternatively, infants' initial ability to match auditory and visual cues might rely on premature categories that undergo substantial change in the course of development. For auditory sound perception, it has been argued that infants' initial ability to discriminate a broad range of (native and nonnative) sound contrasts relies on general auditory processing mechanisms, and that native language attunement involves the formation of language-specific phonological categories that ultimately alter speech perception (see discussion in Kuhl, 2000). This process involves changes to the perception of both native and nonnative sounds, leading to improved sensitivity to the former and reduced sensitivity to the latter type of sound contrasts (e.g., Kuhl et al., 2006). Audiovisual perception might undergo a similar change, leading to a U-shaped pattern of development where infants lose the ability to perceive a distinction for a certain period in development (as, e.g., reported for the auditory perception of vowel length by Mugitani et al., 2009). Our data does not allow us to draw any conclusions about the developmental path of audiovisual speech perception given that we only tested infants at one time point. Note, however, that adult speakers of German were perfectly able to detect the audiovisual mismatches in our stimuli (see Stimuli section in first study), strongly suggesting that infants eventually also become sensitive to mismatches between /e/ and /a/. Nevertheless, it would be interesting for future studies to test infants not only at the verge of, but also before and after native language attunement. This might reveal how infants' sensitivity to congruencies between specific auditory and visual speech cues changes throughout development.

Yet, what we can conclude from our data is that at the verge of perceptual attunement, infants are not equally sensitive to all audiovisual (mis)matches, suggesting that infants are yet to learn the association of—at least some—auditory and visual speech cues. The vowel contrasts tested involved similar acoustic differences and infants were well able to tell the vowels apart in an

[3] Note that we tested a different group of infants, here. Certainly, it would be ideal to test the same infants in all three tasks, that is, auditory discrimination, visual discrimination, and audiovisual matching. However, given the resulting length of the experiment and the age of the participants, this would not be feasible.

auditory perception task. Similarly, infants of the same age succeeded in visually discriminating the vowels when no auditory cues were presented. Thus, infants' insensitivity to audiovisual mismatches involving /a/-/e/ are unlikely to result from difficulties to auditorily or visually distinguish between the vowels. Rather, the source of the problem appears to lie in infants' ability to detect (in)congruencies between specific auditory and visual cues. Indeed, matching auditory and visual speech cues is a demanding task as infants have to rely on their previous knowledge to judge if a given acoustic gesture complies with a given visual gesture or not. The apparently more prominent lip rounding in /o/ might contribute to infants' success in detecting mismatches involving /a/-/o/. If so, this would suggest that salient cues help infants to associate auditory and visual speech cues. Consequently audiovisual matching might be more difficult and/or acquired later for those sound contrasts that involve less salient cues. This suggestion is akin to asymmetries in infants' auditory perception of vowel contrasts (see Polka & Bohn, 2003, 2011 for a review). There have been different proposals to explain these perceptual asymmetries, but a prominent hypothesis is that the more peripheral or the corner vowels act as anchors in perception and foster the formation of phonological categories (Kuhl et al., 2008; Polka & Bohn, 2011). A similar process might be at work in the development of audiovisual speech perception with more perceptually salient visual and/or acoustic patterns acting as anchors in perception and learning. Our experiment was not designed to test for potential asymmetries. So it was not possible to establish whether a certain direction of mismatch is easier to detect, for example, if visual /e/ paired with auditory /a/ is more likely to be detected as a mismatch than visual /a/ paired with auditory /e/. However, it would be important to investigate whether infants are indeed more sensitive to acoustically or visually salient cues in audiovisual speech perception.

Another factor that has been found to be important in auditory sound discrimination is the frequency of a particular sound in the ambient language (e.g., Anderson et al., 2003). Sublexical frequency measures for phonological units in the German lexical database CELEX (Baayen, Piepenbrock, & Van Rijn, 1993) reveal that /e/ is more than twice as frequent as /o/ (phoneme token frequency based on word forms: /e/ = 702313, /o/ = 314913; see Hofmann, Stenneken, Conrad, & Jacobs, 2007). We would expect that infants perform better for more frequent sounds if frequency is an important source of the observed differences in audiovisual perception. Yet, infants performed better for mismatches involving less frequent /o/ than for mismatches involving more frequent /e/. It is thus unlikely that frequency can account for the results. Still, it should also be noted that the German phoneme system also includes /ɛ/, which lies acoustically and visually in between /a/ and /e/, while there is no corresponding counterpart in between /a/ and /o/. The additional vowel /ɛ/ might make the difference between /a/ and /e/ less distinct in acoustic as well as visual terms and render the contrast more difficult. This is, however, purely speculative. More research investigating different types of contrasts and specifically manipulating different factors such as frequency, phonological stability, perceptual salience, and so forth is critical to adequately conclude what makes some audiovisual mismatches easier to perceive for infants than others.

Productive Abilities Influence Audiovisual Perception

Our results further suggest that productive abilities influence infants' sensitivity to audiovisual congruencies. We did not find direct evidence that infants' audiovisual speech perception is modulated by whether or not a specific sound category itself forms part of the babbling inventory, but it appears that infants who do not yet produce sounds similar to /a/ and /e/ prefer mismatching trials whereas infants who already produce these sounds tend to prefer matching trials (see Figure 5). This might indicate a change in preference depending on infants' production, paralleling recent findings by DePaolis and colleagues that older infants' listening preference for specific sound patterns is modulated by their productive abilities (DePaolis et al., 2011, 2013). The lack of any significant effect to support this conclusion in our data might be a consequence of the coarse production measure that we used. Given that we relied on parental reports to estimate if infants already produced sounds that are similar to /a/, /e/, and /o/ in their babbling, some of the infants might be misclassified as producers or nonproducers. Future studies that also collect recordings of infants' babbling for phonetic analysis might reveal a more specific relation between the perception and production of individual sound contrasts.

Comparing the standard errors, it is also evident that there is much larger variation in looking times for the audiovisual matching task than for the discrimination task (even though the data comes from the same infants and a comparable number of trials; see Figures 3, 4, and 5). This might indicate that infants' audiovisual speech perception is more variable relative to their auditory speech discrimination at this point in development, speaking against the widespread assumption that infants' sensitivity to audiovisual congruencies is robust from early on. Even though young infants are astonishingly good at matching auditory and visual cues even across speakers and modalities (Bristow et al., 2009; Patterson & Werker, 1999), their abilities have mainly been tested for the vowels /a/, /u/, and /i/ that mark the margins of the vowel space. These maximally disparate cases might be intuitively matchable while sensitivity to less distinct contrasts might emerge over the course of development. This would parallel findings from auditory perception where difficult contrasts are acquired relatively late (Eilers, Wilson, & Moore, 1977; Narayan, Werker, & Beddor, 2010; see also previous section).

Even though our data does not indicate that the production of a specific sound contrast is related to its (audiovisual) perception, this lack of a direct relation between audiovisual matching and productive abilities does not imply that there is no link between speech perception and production. As mentioned above, our parental report measure might not be sensitive enough to reveal such a direct link. Indeed, there is evidence that adults—and potentially infants—recruit motoric knowledge during speech perception (see Imada et al., 2006 for infants and, e.g., Wilson, Saygin, Sereno, & Iacoboni, 2004; Yuen, Davis, Brysbaert, & Rastle, 2010 for adults) and productive experience might well be involved in this association (see, e.g., Westermann & Reck Miranda, 2004, for a computational model). The positive correlation between the size of infants' babbling inventory and their preference for matching trials in the audiovisual matching task (see Figure 6) suggests that infants with more advanced productive skills are better at detecting audiovisual mismatches. This might be the result of a more estab-

lished link between perceptive and productive categories allowing the infant to recruit her own knowledge about specific articulatory gestures and their acoustic results to better predict and match auditory and visual speech cues (see also Desjardins, Rogers, & Werker, 1997). This might even influence further language development as the use of different sound categories in infant babbling and their sensitivity to audiovisual congruencies in speech perception was also related to later vocabulary development (see next section for more detail). Larger babbling inventories might thus indicate that these infants are more advanced in their language development, paralleling previous reports that advanced consonant use in babbling predicts later vocabulary size (e.g., McCune & Vihman, 2001; Majorano et al., 2014).

It should also be noted that infants are very likely to use pacifiers which induce the same mouth shape as the articulatory gesture associated with /o/. Yeung and Werker (2013) show that such nonlinguistic, motoric experience influences speech perception when concurrently performed. The motoric pattern associated with /o/ may be more familiar to the infant from her own experience resulting in a preference for the mouth shape associated with /o/ and better performance for the contrast /a/-/o/ as compared with the contrast /a/-/e/. If so, infants who use pacifiers (or suck on their thumbs) might be more sensitive to audiovisual speech involving /o/ than infants who do not use pacifiers. We did not ask parents if and how often they provided their infants with a pacifier. Therefore, we cannot directly address this assumption. However, we went back to the experimental recordings and checked which infants used a pacifier during the experiment. Eight infants did, but their behavior did not differ from the main pattern of results. Nevertheless, it might be fruitful to test different contrasts more systematically and at different points in development to establish a relation between frequently performed motoric actions and the development of audiovisual speech perception.

Early Audiovisual Speech Perception is Related to Later Language Development

Our last finding addresses the relation between early audiovisual speech perception and later language development. Recent studies have found that earlier native language attunement in the auditory domain predicts more advanced vocabulary development in the second year of life (Rivera-Gaxiola, Klarman, Garcia-Sierra, & Kuhl, 2005; Tsao et al., 2004). Our results parallel these findings and extend it to the audiovisual domain, suggesting that more advanced native speech perception in the first year of life—be it in the auditory or in the audiovisual domain—might be an early marker of more advanced language development. Specifically, infants who show increased sensitivity to native auditory and audiovisual sound categories might have an advantage in using these sound categories to segment speech and to memorize the phonological structure of words, resulting in more efficient word learning and advanced vocabulary development (cf., Kuhl et al., 2008).

Interestingly, we found a similar association between infants' babbling at 6 months of age and their vocabulary size at 12 months of age. This might indicate that the relation between infants' early sensitivity to audiovisual speech cues and their later vocabulary size is mediated by their productive abilities—or vice versa, that the relation between infants' early productive abilities and their later vocabulary size is mediated by their perceptive abilities. Either way, this would support the notion of a functional link between perceptive and productive abilities in early language development (e.g., Kuhl, 2000; Vihman, 1996; see also previous section).

Conclusion

The current study is the first to investigate the relation between native audiovisual sound perception in the first year of life, emerging productive abilities and later language development. We used two native vowel contrasts, /a/-/e/ and /a/-/o/, that have not been studied so far and employed a modified version of the preferential looking paradigm that allowed us to test both audiovisual sound contrasts as well as their auditory discrimination within the same infant. Interestingly, infants' ability to detect mismatches between auditory and visual speech cues differed depending on the vowels involved, even though there were no differences in their ability to discriminate them. Even though our results provide no direct evidence that the production of similar sounds in babbling modulates the audiovisual perception of these sounds, we found an influence of general productive abilities on infants' sensitivity to audiovisual congruencies. Furthermore, audiovisual speech perception and general productive abilities at 6 months of age were related to vocabulary size at 12 months of age, suggesting a lasting effect on language development.

The current results inform theories on the development of speech perception at three levels. First, the finding that infants do not perform equally well for both vowel contrasts suggest that audiovisual speech perception might be less robust than previously assumed and supports the notion that it undergoes substantial reorganization in the course of development (see also Pons et al., 2009). This process might be modulated by the perceptual salience of auditory and visual speech cues. Second, the finding that infants' productive abilities are—although maybe not specifically—related to their audiovisual speech perception suggests a potential role for emerging productive abilities in perceptual reorganization (see also Kuhl, 2000). Third, the relation between early speech perception and later language development further links phoneme acquisition to early word learning, suggesting a continuous development from early to later stages of language development (see also Tsao et al., 2004).

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. http://dx.doi.org/10.1111/cdev.12320

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech, 46,* 155–182. http://dx.doi.org/10.1177/00238309030460020601

Baayen, H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Philadelphia.

Best, C., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development, 21,* 295. http://dx.doi.org/10.1016/S0163-6383(98)91508-9

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,* 905–921. http://dx.doi.org/10.1162/jocn.2009.21076

de Boysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A cross-linguistic investigation of vowel formants in babbling. *Journal of Child Language, 16,* 1–17. http://dx.doi.org/10.1017/S0305000900013404

DePaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2011). Do production patterns influence the processing of speech in prelinguistic infants? *Infant Behavior and Development, 34,* 590–601. http://dx.doi.org/10.1016/j.infbeh.2011.06.005

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. http://dx.doi.org/10.1016/j.infbeh.2013.06.007

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). *An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks*. http://dx.doi.org/10.1006/jecp.1997.2379

Eilers, R. E., Wilson, W. R., & Moore, J. M. (1977). Developmental changes in speech discrimination in infants. *Journal of Speech and Hearing Research, 20,* 766–780. http://dx.doi.org/10.1044/jshr.2004.766

Green, J. R., Nip, I. S. B., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research, 53,* 1529–1542. http://dx.doi.org/10.1044/1092-4388(2010/09-0005)

Grimm, H., & Doil, H. (2000). *Elternfragebögen für die Früherkennung von Risikokindern* [Parent questionnaires for early diagnosis of children at risk]. Göttingen, Germany: Hogrefe.

Hofmann, M. J., Stenneken, P., Conrad, M., & Jacobs, A. M. (2007). Sublexical frequency measures for orthographic and phonological units in German. *Behavior Research Methods, 39,* 620–629. http://dx.doi.org/10.3758/BF03193034

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. K. (2006). Infant speech perception activates Broca's area: A developmental magnetoencephalography study. *NeuroReport, 17,* 957–962. http://dx.doi.org/10.1097/01.wnr.0000223387.51704.89

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America, 97,* 11850–11857. http://dx.doi.org/10.1073/pnas.97.22.11850

Kuhl, P. K. T., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 363,* 979–1000. http://dx.doi.org/10.1098/rstb.2007.2154

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218,* 1138–1141. http://dx.doi.org/10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy: The Minnesota Symposia on Child Psychology* (Vol. 20, pp. 235–266). Hillsdale, NJ: Erlbaum.

Kuhl, P. K., 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. http://dx.doi.org/10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608. http://dx.doi.org/10.1126/science.1736364

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant Behavior and Development, 13,* 343–354. http://dx.doi.org/10.1016/0163-6383(90)90039-B

Lewkowicz, D. J. (2000). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development, 71,* 1241–1257. http://dx.doi.org/10.1111/1467-8624.00226

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

Majorano, M., Vihman, M. M., & DePaolis, R. A. (2014). The relationship between infants' production experience and their processing of speech. *Language Learning and Development, 10,* 179–204. http://dx.doi.org/10.1080/15475441.2013.829740

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

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748. http://dx.doi.org/10.1038/264746a0

Mugitani, R., Kobayashi, T., & Hiraki, K. (2008). Audiovisual matching of lips and non-canonical sounds in 8-month-old infants. *Infant Behavior and Development, 31,* 307–310. http://dx.doi.org/10.1016/j.infbeh.2007.12.002

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J. F., & Amano, S. (2009). Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology, 45,* 236–247. http://dx.doi.org/10.1037/a0014043

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science, 13,* 407–420. http://dx.doi.org/10.1111/j.1467-7687.2009.00898.x

Oller, D. K. (1978). Infant vocalizations and the development of speech. *Allied Health and Behavioral Sciences, 1,* 523–549.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voices is robust in 4.5-month-old infants. *Infant Behavior and Development, 22,* 237–247. http://dx.doi.org/10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voices. *Developmental Science, 6,* 191–196. http://dx.doi.org/10.1111/1467-7687.00271

Polka, L., & Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication, 41,* 221–231. http://dx.doi.org/10.1016/S0167-6393(02)00105-X

Polka, L., & Bohn, O.-S. (2011). Natural referent vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics, 39,* 467–478. http://dx.doi.org/10.1016/j.wocn.2010.08.007

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 421–435. http://dx.doi.org/10.1037/0096-1523.20.2.421

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. http://dx.doi.org/10.1073/pnas.0904134106

Richie, C., & Kewley-Port, D. (2008). The effects of auditory-visual vowel identification training on speech recognition under difficult listening conditions. *Journal of Speech, Language, and Hearing Research, 51,* 1607–1619. http://dx.doi.org/10.1044/1092-4388(2008/07-0069)

Rivera-Gaxiola, M., Klarman, L., Garcia-Sierra, A., & Kuhl, P. K. (2005). Neural patterns to speech and vocabulary growth in American infants. *NeuroReport, 16,* 495–498. http://dx.doi.org/10.1097/00001756-200504040-00015

Saffran, J. R., Werker, J. F., & Werner, L. A. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In D. Kuhn, R. S. Siegler, W. Damon, & R. M. Lerner (Eds.), *Handbook of child*

psychology: Vol. 2. Cognition, perception, and language (6th ed., pp. 58–108). Hoboken, NJ: Wiley.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388,* 381–382. http://dx.doi.org/10.1038/41102

Stark, R. E. (1980). Stages of speech development in the first year of life. In G. Yeni-Komshian, J. Kavanaugh, & C. Ferguson (Eds.), *Child phonology* (Vol. 1, pp. 73–90). New York, NY: Academic Press.

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 Development, 75,* 1067–1084. http://dx.doi.org/10.1111/j.1467-8624.2004.00726.x

van Son, N., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1994). Viseme classifications of Dutch consonants and vowels. *The Journal of the Acoustical Society of America, 96,* 1341–1355. http://dx.doi.org/10.1121/1.411324

Vihman, M. M. (1996). *Phonological development. The origins of language in the child*. Oxford, UK: Blackwell.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science, 316,* 1159. http://dx.doi.org/10.1126/science.1137686

Westermann, G., & Reck Miranda, E. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language, 89,* 393–400. http://dx.doi.org/10.1016/S0093-934X(03)00345-6

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7,* 701–702. http://dx.doi.org/10.1038/nn1263

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24,* 603–612. http://dx.doi.org/10.1177/0956797612458802

Yuen, I., Davis, M. H., Brysbaert, M., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Science, 107,* 592–597. http://dx.doi.org/10.1073/pnas.0904774107

Received June 20, 2014
Revision received September 4, 2015
Accepted September 25, 2015 ■