Speaker Matters: Natural inter-speaker variation affects 4-month-olds' perception of audio-visual speech

Jovana Pejovic[1,2], Eiling Yee[3], and Monika Molnar[4]

[1]BCBL. Basque Center on Cognition, Brain, and Language

[2]Center of Linguistics, University of Lisbon, Portugal

[3]Department of Psychological Sciences, University of Connecticut

[4]Department of Speech-Language Pathology, University of Toronto

Address for correspondence:

Jovana Pejovic

Laboratorio de Fonetica & Lisbon Baby Lab

Centro de Linguistica, Faculdade de Letras,

Universidade de Lisboa,

Alameda da Universidade, 1600-214 Lisboa, Portugal.

e-mail: jpejovic@edu.ulisboa.pt

Abstract

In the language development literature, studies often make inferences about infants' speech perception abilities based on their responses to a single speaker. However, there can be significant natural variability across speakers in how speech is produced (i.e., inter-speaker differences). The current study examined whether inter-speaker differences can affect infants' ability to detect a mismatch between the auditory and visual components of vowels. Using an eye-tracker, 4.5-month-old infants were tested on auditory-visual (AV) matching for two vowels (/i/ and /u/). Critically, infants were tested with two speakers who naturally differed in how distinctively they articulated the two vowels within and across the categories. Only infants who watched and listened to the speaker whose visual articulation of the two vowels were most distinct from one another were sensitive to AV mismatch. This speaker also produced a visually more distinct /i/ as compared to the other speaker. This finding suggests that infants are sensitive to the distinctiveness of AV information across speakers, and that when making inferences about infants' perceptual abilities, characteristics of the speaker should be taken into account.

Keywords: speech perception development; audio-visual matching; infant; visual and auditory perceptual salience; eye-tracking

Introduction

Infants use both visual-articulatory as well as auditory information when processing spoken language. At just two months of age they are able to match auditory speech sounds to visual-articulatory features produced by speakers in videos (Baier, Idsardi, & Lidz, 2007; Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003; Yeung & Werker, 2013). However, little is known about what factors might affect infants' ability to match auditory and visual speech. Recent findings suggest that the visual distinctiveness of speech sounds interacts with infants' auditory-visual (AV) speech matching ability: German-learning 5.5-6-month-old infants were able to detect AV mismatch when they were presented with visual and auditory instances of the vowel pair /a-o/, but not when they were presented with /a-e/ (Altvater-Mackensen, Mani and Grossmann, 2015). The authors suggested that this difference is due to the fact that visually, the vowels /o/ and /a/ are more distinct than are /a/ and /e/. That is, the lips are rounded for /o/, but are spread horizontally for both /a/ and /e/. Thus, the difference between the lip-rounding associated with the vowel /o/ and the lip-spreading associated with /a/ may have facilitated the detection of the mismatch (Altvater-Mackensen et al., 2015).

Yet, when it comes to the auditory and visual-articulatory features of speech, there are differences not only across speech sound categories, but also across speakers. For instance, individual differences have been observed for jaw height within the production of several American-English vowel categories (e.g., Johnson, Ladefoged, & Lindau, 1993), suggesting that the same vowel category is produced by somewhat different articulatory features across speakers. Inter-speaker differences in speech sound production have also been observed in caregivers, who vary considerably in their visual-articulatory characteristics when producing infant-directed speech (e.g., Green, Nip, Wilson, Mefferd, & Yunusova, 2010). Moreover, it is possible that there is a relationship between the acoustic characteristics of speech produced

by caregivers (which correlates strongly with the visual-articulatory properties of speech) and speech perception development in infants. One study suggests that the acoustic distinctiveness of caregivers' speech (i.e., exaggerated vowels) correlates positively with infants' performance on native consonant discrimination (Liu, Kuhl, & Tsao, 2003).

In the current study we asked whether naturally occurring differences across speakers in how visually distinctively they produce different vowels (across and within categories) can modulate infants' AV processing. We first selected two female speakers who appeared to naturally exhibit differences in how visually distinctive their productions of the vowel /i/ were from their productions of the vowel /u/. The speakers also seemed to exhibit differences when they were compared on their productions of the same vowel category (e.g., /i/). Then, following the procedures described in Altvater-Mackensen et al. (2015), we quantified these differences to verify that the speakers indeed differed on the visual articulatory distinctiveness of the vowels[1]. We predicted that infants' AV matching ability would interact with inter-speaker differences in the visual distinctiveness of the vowels. Specifically, we suspected that the natural differences in lip-spreading across speakers (i.e., how wide they open their mouth during the production of the selected vowels) would modulate infants' AV matching ability, as measured by their amount of attention to AV match and mismatch videos[2].

[1] We, of course, also observe accompanying differences in acoustic distinctiveness; we return to these differences in the General Discussion.

[2] Infants between 3-6 months of age tend to vary with respect to their looking preference in AV matching paradigms. In Altvater-Mackensen et al. (2015), 5.5-6-month-old infants attended longer to AV matching events. However, depending on the vowel they are familiarized with, 3-6-month-old infants may also exhibit a mismatching preference (e.g., when familiarized with the vowel /a/, a matching preference is observed, but when familiarized with the vowel /i/ a mismatching preference is observed; Streri, Coulon, Marie, & Yeung, 2016). Therefore, we are unable to make predictions as to whether match or mismatch trials should elicit longer looking times.

Methods

Quantifying inter-speaker differences

First, we recorded five female speakers uttering two acoustically distinct vowel categories, /u/ and /i/. In line with the demographic characteristics of the region where the study took place (San Sebastian, Spain) all speakers were Spanish-Basque bilinguals. While the two languages differ considerably in terms of their syntax, they rely on virtually identical speech sound repertoires. The target vowels /i/ and /u/ are each part of both the Basque and Spanish vowel inventories, hence Spanish- and Basque-learning infants are regularly exposed to these speech sounds. Also, the Spanish and the Basque versions of the vowels /i/ and /u/ are acoustically identical across the two languages.

All the speakers received the same instructions: They were asked to produce the vowels in an infant-friendly style, as if they were producing these vowels to an infant seated in front of them, while gazing at a camera. The productions were recorded using a Canon LEGRIA HF G10 camera. The speakers were instructed to repeat the same vowel with an approximately 2 second inter-repetition-interval, trying to maintain the same intensity, duration and pitch across tokens. Each speaker was recorded separately, and they received no explicit instructions about how they should produce the vowels (e.g., if they should open their mouth more or less). Once the videos were recorded, the speakers were asked to dub the videos–either saying the vowel that matched the video or saying the vowel that did not match the video (details on video creation are provided in the next section).

First, based on visual inspection of the videos, we selected the videos of two speakers who seemed to produce the vowels in a similar manner (i.e., infant friendly style), but with different visual articulatory cues (i.e., differing on lip-spreading; **Figure 2** presents example frames from the two speakers). Then, to confirm that these cues indeed differed across these

two speakers, we measured the visual articulatory cues via horizontal and vertical lip-opening (i.e., from the left to right lip corner, and from upper to lower lip, respectively) in pixels on a still video frame during a fully visually articulated vowel (see **Figure 1**, left panel; these measures are the same as those used in Altvater-Mackensen et al., 2015, i.e. the measurement occurred on visually maximally opened/spread mouth position during the vowel production). As can be seen in **Figure 1** panel A, for both speakers, the vowel productions clearly differ on horizontal lip-opening (i.e., the vowel /i/ is produced with the lips more spread than the vowel /u/). More relevant to our predictions is that the speakers also differ (in three ways) in how much they open their lips while producing the vowels: First, when producing the vowel /i/, Speaker 2 opens her lips horizontally more than Speaker 1 (the mean distance between the lip corners is 172 pixels in Speaker 2, vs. 132 pixels in Speaker 1). Second, there is greater distinctiveness in horizontal lip opening between the /u/ and /i/ in Speaker 2 than in Speaker 1 (the mean difference between the vowels on horizontal lip-opening is 94 pixels in Speaker 2, vs. 63 pixels in Speaker 1). Third, with respect to vertical lip-opening, Speaker 1 produces the vowels more distinctly than does Speaker 2 (the mean difference between the vowels on vertical lip-opening in pixels is 11 in Speaker 1, vs. 3 in Speaker 2). Thus, measurements of the visual articulatory cues confirmed the existence of potentially relevant inter-speaker variation. These differences are summarized in **Table A2** (Appendix).

Because Speaker 2 produced the two vowels visually more distinctively with respect to horizontal lip-opening than Speaker 1, in line with Altvater-Mackensen et al.'s (2015) findings, we predict that infants watching Speaker 2 will be more likely to succeed on our AV matching task than those watching Speaker 1. However, given that on vertical lip-opening, Speaker 1 produced the vowels slightly more distinctly, if infants are more attuned to differences in vertical than horizontal lip-opening, then their AV matching ability could be better when watching Speaker 1.

We also measured the acoustic characteristics of the two vowels. Unsurprisingly, the visual differences between the vowels described above correspond to acoustic differences. For both speakers, the two vowels form two distinct acoustic categories on F2 (vowel backness; **Figure 1**, panel B) and F3 (vowel roundness, **Figure 1**, panel C). With respect to inter-speaker differences between the vowels, on F1 the mean difference between the vowels is very similar for Speaker 1 and Speaker 2 (25 vs. 35, respectively). On F2, Speaker 1 produced the vowels slightly more distinctively than Speaker 2 (mean difference between the vowels 2160 vs. 1853, respectively). Importantly, in line with Speaker 2's greater visual distinctiveness between the vowels on horizontal lip opening (which can reflect differences in rounding, with more horizontal opening corresponding to less rounding), we observed in F3 that Speaker 1 produced vowels less acoustically distinctively than Speaker 2 (mean difference between the vowels of 101 vs. 327, respectively). Specifically, there is a larger difference between /u/ and /i/ on F3 for Speaker 2 (2962 and 3289, respectively, with larger values reflecting less rounding) than for Speaker 1 (3294 and 3395, respectively). These differences are summarized in **Table A2** in the Appendix.

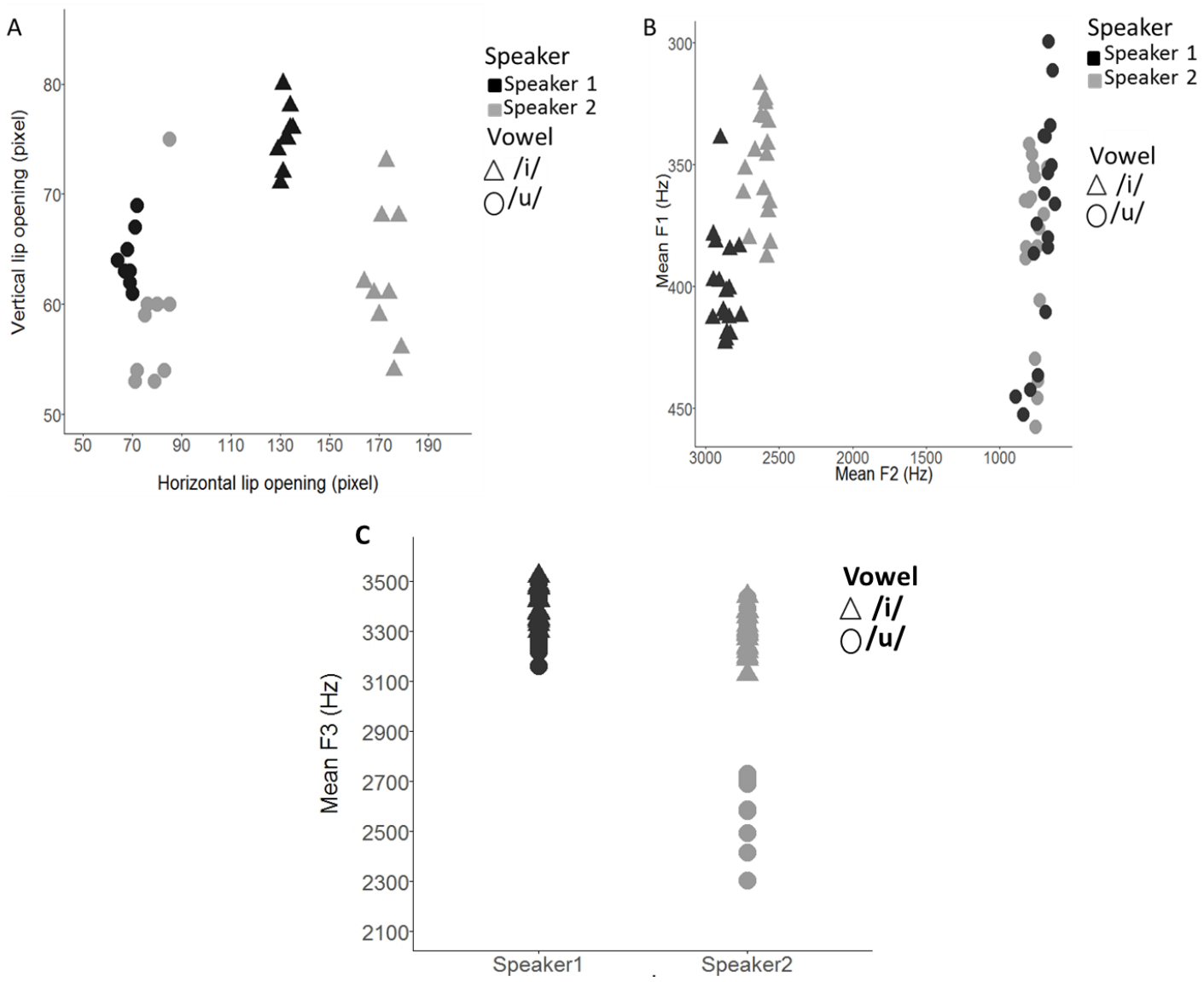

Figure 1.Measures of visual and acoustic vowel distinctiveness across speakers. Panel A depicts vertical and horizontal lip-opening values (in pixels) during full articulation of each token (individual points). Panel B shows the first (F1) and second (F2) formant frequency of each token (individual points). Panel C presents the third formant (F3) frequency of each token. In all panels black colored points indicate Speaker 1 and gray indicates Speaker 2. Circular markers indicate the vowel /u/, triangles indicate the vowel /i/.

Match and Mismatch videos

To avoid potential confounds due to only one condition being dubbed, both match and mismatch stimuli were created via dubbing. The dubbed audios were recorded in a sound-attenuated room with a Marantz PMD1671 recorder and a Sennheiser noise-reducing microphone. To ensure that the duration of the mouth opening corresponded with the length of the heard vowel, speakers dubbed while watching their own silent videos. To create match stimuli, speakers dubbed by uttering the same vowel that they produced in the silent video. For mismatch stimuli, speakers dubbed by uttering a different vowel (i.e., for visually articulated /i/, speakers uttered /u/; for visually articulated /u/, speakers uttered /i/). To confirm that the auditory vowels recorded for the match and mismatch condition are both perceived within the same intended vowel category (/i/ or /u/), the audio files (without video) were presented to 18 adult Spanish-Basque speakers in a categorization experiment. In this task, participants heard all of the /i/ and /u/ productions from both speakers (i.e., both those produced while watching the matching vowel and the mismatching vowel), as well as tokens of vowel /a/ produced by the same speakers, which were included as filler stimuli. Participants were instructed to categorize the heard vowel as "/i/", "/u/", or "some other vowel". Regardless of whether the vowels had been recorded in the context of matching or mismatching videos, they were reliably categorized correctly (minimum 98% for each vowel category). Reaction time data confirmed that the matched and mismatched vowels were processed similarly (M_{match}=1066, SD_{match}= 453; $M_{mismatch}$=1073, $SD_{mismatch}$= 418; $t(1,17) = -0.1$, $p = .9$). No differences in categorization accuracy or reaction times were observed across speakers, indicating that any inter-speaker differences should not be due to one speaker simply being better at producing dubbed vowels than the other.

To ensure that each speaker's visual vowels are distinguishable from one another, we conducted a visual discrimination task with 10 adult participants. These participants were

presented with muted versions of the videos that the infants watched in our experiment. Participants saw two muted videos in a sequence. Speakers in the two videos either uttered the same vowel category or different vowel categories. Speakers were equally distributed across categories. Video pairs were presented within speakers. Participants judged whether the presented video pair represented the same or different vowels. They succeeded with 98.7% accuracy in discriminating the vowels (/i/ vs. /u/) based on visual cues alone within and across speakers. Importantly, no difference between speakers was observed, indicating that any observed inter-speaker differences should not be due to one speaker's visual /i/ vs. /u/ being indistinguishable.[3]

Finally, the visual and auditory signals were mixed using a video and sound editing software (Adobe Premier Pro), to create the match and mismatch trials. Each video contained nine unique tokens of the given vowel with an approximately 2 second interval between the tokens, creating a video of about 30 seconds long. Importantly, we ensured that the auditory and visual signals were temporally synchronized. Specifically, for each token we aligned the onset of the dubbed auditory signal with the onset of the original auditory signal from the recorded video using the Adobe Premier Pro software. Details on auditory measures across speakers are presented in **Table A1** in the Appendix. Duration, intensity, pitch and inter-stimulus-interval (ISI) across vowel tokens within one speaker were selected to be similar, while these measures varied between speakers allowing for natural inter-speaker variation. The mixed AV videos were edited to make them similar with respect to each speaker's size on the screen, brightness, and saturation. The dubbed match and mismatch videos are available at https://osf.io/n4zwv/.

[3] However, these data do not address how *easily* discriminable the auditory and visual stimuli were for the infants, and that this is exactly what we wish to test– whether infants' ability to discriminate (along any number of dimensions) may vary so much from one speaker to another that generalization across speakers requires qualification.

Participants

In total, data from 42 infants were included in the analyses: 20 infants completed the experiment with Speaker 1 (average age 4.5 months, range 123-144 days, 9 female infants), and 22 with Speaker 2 (average age 4.5 months, range 128-146 days, 9 female infants). An additional 20 infants were tested but their data were excluded from analyses due to crying (7), fussiness (2), extreme movement causing lost pupil tracking (3), poor calibration (7), and not being attentive to the task–the infant looked away immediately after the video was presented (1). All infants were healthy, full-term, and without reported history of vision or hearing problems. Participants were recruited from the Spanish-Basque region of San Sebastian, Spain. Exposure to Spanish and/or Basque was evaluated via a detailed language exposure questionnaire that estimates infants' proportion of exposure to each language over time (the same questionnaire was used in Molnar, Gervain, & Carreiras, 2014). Only monolingual infants (Spanish N=25; 12 presented with Speaker 1; and Basque N=17; 8 presented with Speaker 1) exposed to one of the languages at least 95% of the time (M= 99.4%, SD=1.5%) were included.

Apparatus

Infants' eye-gaze was collected with a monocular EyeLink 1000 LCD Arm Mount remote eye-tracker (SR Research) with integrated LCD screen. A 16mm camera lens was used with a 940nm infrared illuminator. An Acer AL1717 17″ monitor with 1024x768 resolution, and a 60 Hz refreshing rate was used for the visual stimuli presentation. Auditory stimuli were played over two JBL-duet speakers placed behind and on the sides of the screen, with 65-70 dB intensity.

Experiment design

Half of the infants were exposed to Speaker 1 and the other half were exposed to Speaker 2. Each infant was presented with both vowels, both in a match and mismatch condition. In the match condition, auditory and visual signals corresponded (i.e., the visual vowel /i/ was paired with auditory /i/, and the visual vowel /u/ was paired with auditory /u/). In the mismatch condition auditory and visual signals did not correspond (i.e., visual vowel /i/ was paired with auditory /u/, and visual vowel /u/ was paired with auditory /i/). The trials were grouped into two blocks: (1) vowel /i/, and (2) vowel /u/. Each block consisted of three sequentially presented matched and three mismatched trials. In total, each infant was presented with 12 trials (see

Figure *2*). We counterbalanced across infants whether the mismatch trials were formed based on auditory mismatch (i.e., a matched /i/ block alternated with a visual /i/-auditory /u/ mismatched block; a matched /u/ block alternated with a visual /u/-auditory /i/ mismatched block) or based on visual mismatch (i.e., a matched /i/ block alternated with a mismatch visual /u/-auditory /i/ block; a matched /u/ block alternated with a mismatch visual /i/-auditory /u/ block). The order of the matched and mismatched trials and of the vowel blocks was also counterbalanced.

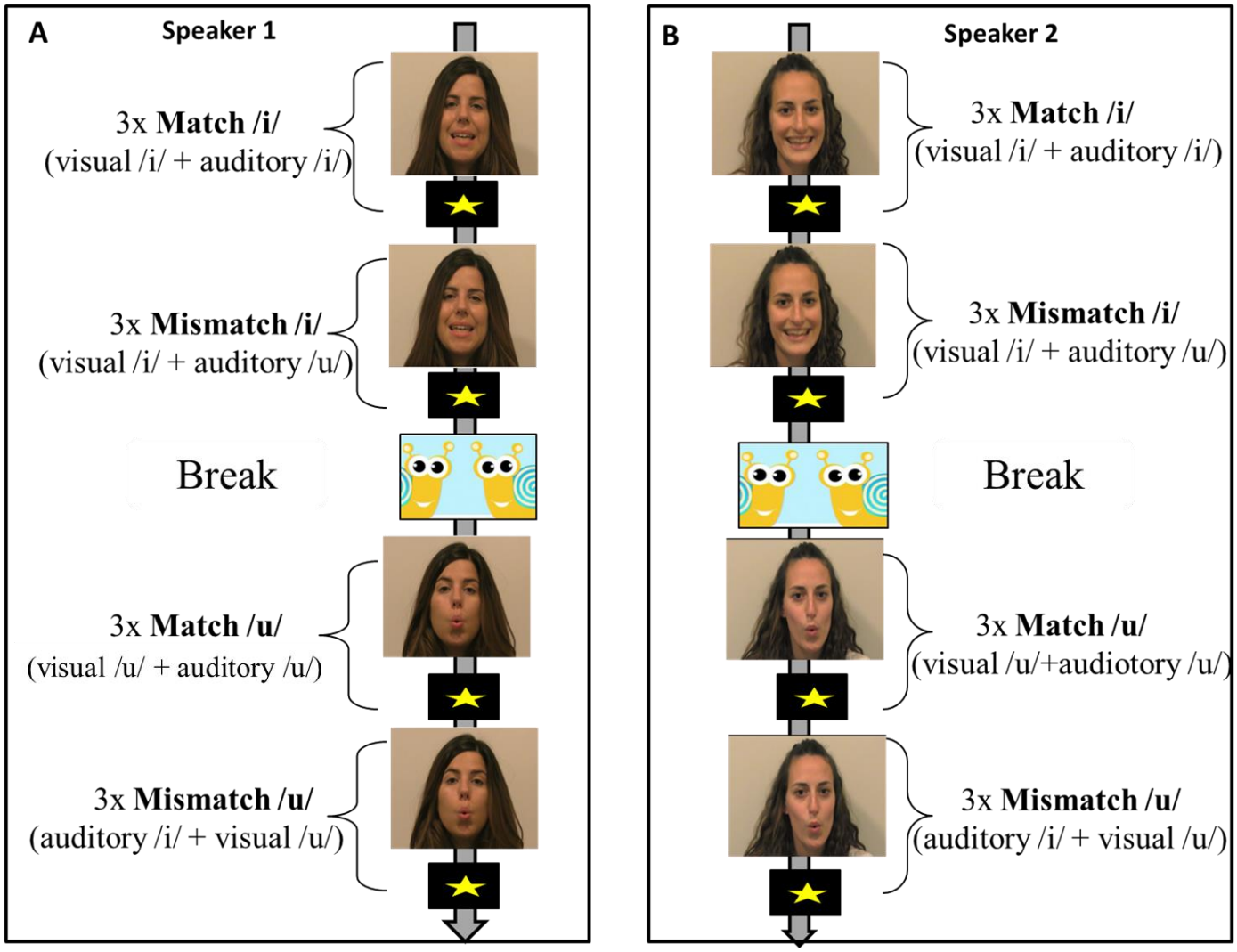

Figure 2. Experiment design. Infants were presented with blocks of three match and three mismatch trials for one vowel, followed by a short break, after which they were presented with another block of match and mismatch trials for the other vowel. The order of match and mismatch, /i/, and /u/, as well auditory or visual mismatch was counterbalanced. Note that Figure 2 illustrates stimuli presentation based on the auditory mismatch. Every trial began with an attention-getter. Trials were infant-controlled. Two speakers (example frames given) were presented across infants. The trials are available at https://osf.io/n4zwv/.

Procedure

Infants were seated in their caregivers' lap, facing a monitor placed 55-60 cm away. Parents wore noise-cancelling headphones and dark glasses to prevent them influencing their infants' behavior.

At the beginning of each session, the infant's eye-gaze was calibrated using a 5-point calibration and validation system with a 1000 ms interval between calibration points. Then, each experimental trial started with an infant-friendly, small-in-size attention-getter displayed centrally on the screen, accompanied by infant-friendly sounds. The attention-getter also functioned as a drift correction for the eye-tracking system (correcting for small drifts in calculation of the gaze position), by which we maintained high eye-tracking accuracy throughout the session. When infants' gaze at the attention-getter was registered and the drift correction was performed, the trial began. Trial presentation was fully infant-controlled; when infants looked away for more than two seconds, the trial ended and the attention-getter appeared on the screen. The maximum trial duration was 30 seconds. The entire experiment lasted about 20 minutes.

Results

The total looking time for each trial (12 in total) was calculated for each infant separately as the sum of all fixations on the entire screen recorded by the eye tracker (as in previous infant AV matching studies; Altvater-Mackensen & Grossmann, 2015; Altvater-Mackensen et al., 2015; Yeung & Werker, 2013)[4].

First, to test whether infants in the current study exhibited any AV matching ability, we compared mean looking times between match and mismatch conditions. Looking times

[4] We also collected data on infants' processing of face features (i.e., the eyes vs. the mouth). However, that data is part of a larger project on the development of infants' selective attention and is reported in a separate manuscript (Pejovic, Yee, and Molnar, in prep).

(in milliseconds) for each infant were averaged across match (6 trials) and mismatch (6 trials) conditions. A paired *t*-test revealed no significant difference between conditions ($t_{(41)}$ = -1.4, p = .14, d = .25), suggesting that infants spent the same amount of time looking at AV match (M=11,461; SD=5,082) and mismatch (M=12,826; SD=5,780) events across the two speakers.

To address our primary question, whether naturally occurring inter-speaker differences in the distinctiveness of the vowels (i.e., that Speaker 2 produced the two vowels more distinctively than did Speaker 1) modulates infants' AV matching ability, we conducted a 2x2 ANOVA on mean looking times with *Speaker* (Speaker1/Speaker2) as a between-subject factor, and *Condition* (match/mismatch) as a within-subject factor. As predicted, this analysis revealed a significant *Speaker x Condition* interaction ($F_{(1, 40)}$ = 7.3, p = .01, η^2_G = .05)[5]. A post hoc power analysis with the program *G* Power* (Buchner, Erdfelder, Faul, & Lang, 2009) revealed adequate power (power = .82) given the sample size. No other effects reached significance (all Fs < 2.1, all *ps* > .15, η^2_G < .01). Post hoc paired *t*-tests revealed that infants spent more time looking at mismatch (M=12,826; SD=5,789) than match (M=9,751; SD=5,223) trials for Speaker 2 ($t_{(21)}$ =2.9, p < .01, d = .61), but not for Speaker 1 ($t_{(19)}$ = 0.9, p = .4, d = .22), see **Figure 3**. Notably, the main effect of speaker did not reach significance ($M_{speaker\ 1}$ = 12,141, $M_{speaker\ 2}$ = 12,163), indicating that infants' visual attention was not modulated by an *overall* preference for one speaker over the other. Note that block order (match/mismatch) or whether the mismatch was based on auditory or visual stimuli did not affect the results (more details on the analysis can be found in the Supplemental Material).

[5] Note that we observed similar results if only the first trial of each block was analyzed.

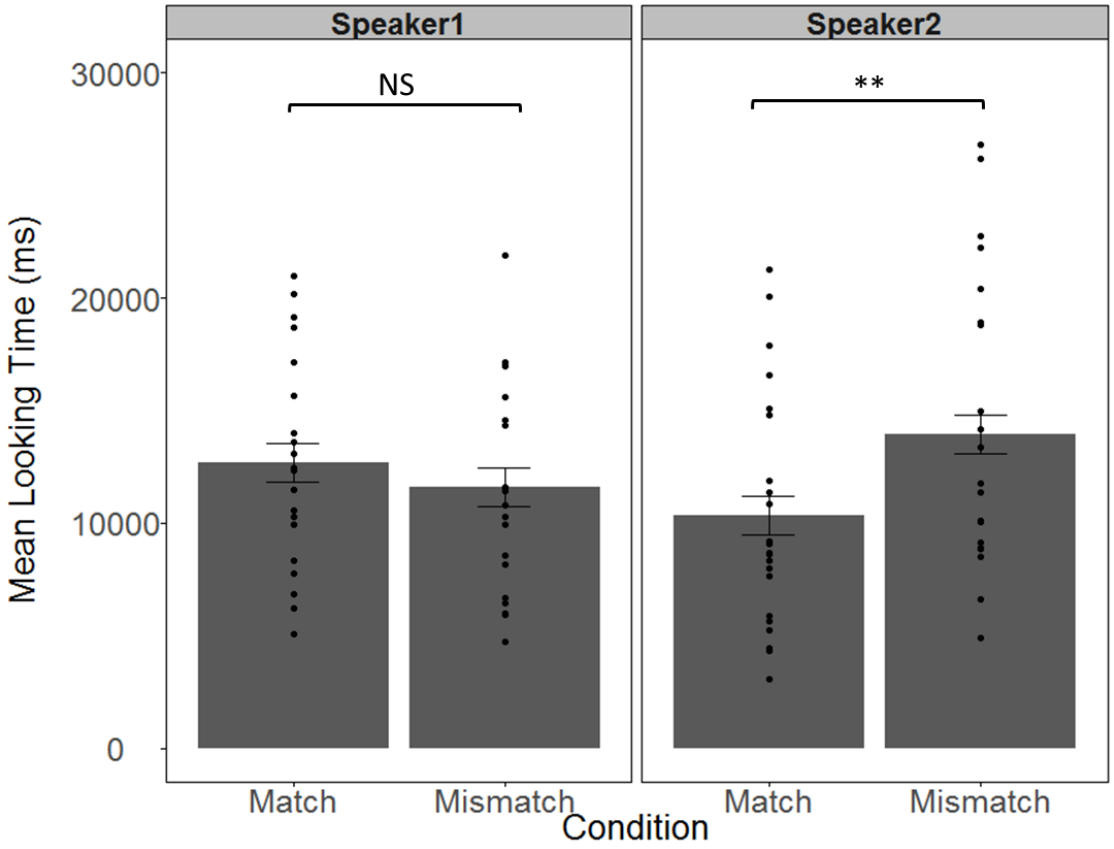

Figure 3. Mean looking times (in milliseconds) for match and mismatch conditions across the two speakers. Points represent individual infants' looking times averaged across trials. Error bars represent +/- 1 SE, asterisks indicate a significance level of ** $p \leq .01$. Note that the same overall findings were obtained when the outlier in the mismatch condition for Speaker 1 was excluded.

Finally, to test whether visual vowel type modulates the AV matching ability, as has been suggested by Altvater-Mackensen et al. (2015), we conducted a 2x2x2 ANOVA on the mean looking times with *Visual Vowel* (visual /i/ vs. /u/) and *Condition* (match/mismatch) as within-subject factors, and *Speaker* (Speaker1/Speaker2) as a between-subject factor (**Figure**

4).[6] The analysis confirmed a significant *Speaker x Condition* interaction ($F_{(1, 40)} = 7.3$, $p < .01$, $\eta^2_G = .02$), a close to significant *Speaker x Visual Vowel x Condition* interaction ($F_{(1, 40)} = 3.6$, $p = .06$, $\eta^2_G = .02$), and a close to significant *Visual Vowel* effect ($F_{(1, 40)} = 3.4$, $p = .07$, $\eta^2_G = .02$), indicating an overall tendency to attend less when visual /i/ was presented (M = 10,960, SD = 7,710) than visual /u/ (M = 13,328, SD = 7,906). Considering that the most evident inter-speaker articulation difference was for the vowel /i/, we also explored the *Speaker x Visual Vowel x Condition* interaction (even though this interaction was not quite significant, $p = .06$). Post hoc *t*-tests revealed that for Speaker 2 infants detected AV mismatch for visual /i/ ($t_{(21)} = -3.2$, $p = .003$, $d = .91$), but not for visual /u/ ($t_{(21)} = -0.4$, $p = .6$, $d = .12$). For Speaker 1, infants were not able to detect AV mismatch for either of the vowels (for /i/, $t_{(19)} = 1.3$, $p = .2$, $d = .35$; for /u/, $t_{(19)} = -.4$, $p = .6$, $d = .10$).

Overall, the results suggest that infants' AV matching ability differs across the two speakers. This suggests that differences in vowel production between speakers (i.e., producing vowels more or less distinctly) modulates infant AV matching ability. Specifically, we observed infants' AV matching ability only in Speaker 2, who produced her vowels more distinctively with respect to horizontal lip-movements and F3 values.

[6] Note that because of the symmetry of the design, we necessarily observe the same *Speaker x Condition interaction (F(1, 40) = 7.3, p < .01)* when the 2x2x2 ANOVA is conducted with *auditory* vowel as a factor, instead of the *visual* vowel.

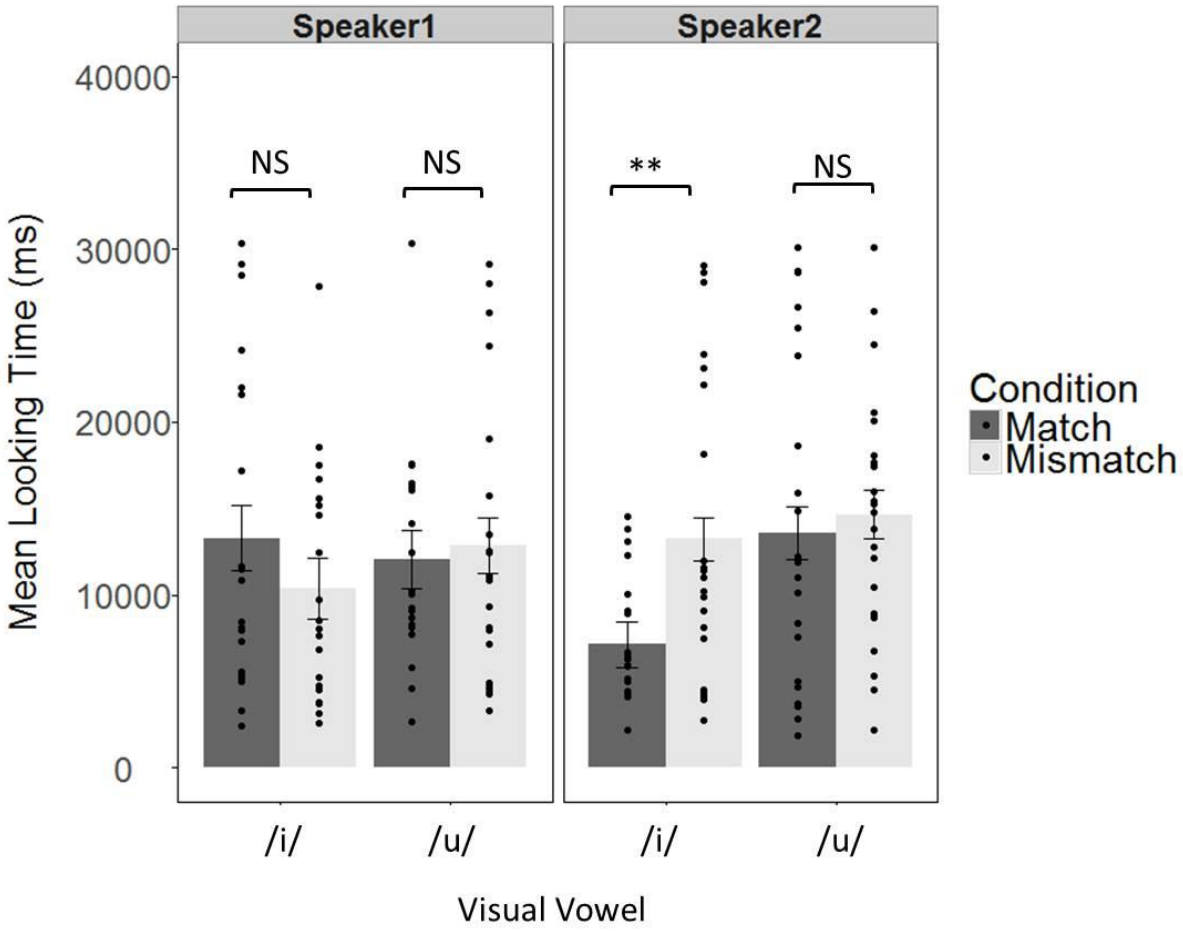

Figure 4. Mean looking times (in milliseconds) for match and mismatch condition in each of the two speakers in response to the two visually presented vowels. Points represent individual infants' scores. Error bars represent +/- 1 SE, asterisks indicate a significance level of ** $p \leq .01$. Note that the same overall findings were obtained when the outliers in the mismatch /i/ and match /u/ conditions for Speaker 1 were excluded.

Discussion

In the current experiment we assessed whether visual-articulatory differences (or accompanying acoustic differences—we consider these later) produced across two speakers for the same vowel categories affect preverbal infants' auditory-visual (AV) speech matching abilities. We selected the videos of two speakers who showed clear evidence for inter-speaker

variability in their visual articulation of /i/ and /u/. Then we tested 4.5-month-old infants in their AV matching ability for these vowels using AV speech from these two speakers. The study revealed three important findings. First, when data across both speakers were considered in a between-subjects design, infants did not demonstrate sensitivity to AV matching. Next, sensitivity to AV match/mismatch information was present only in infants who were presented with the speaker whose visual-articulatory cues were more salient. Finally, the AV mismatch was more pronounced in this speaker for the visual vowel /i/.

This finding is in line with previous studies reporting that visual-articulatory cues related to specific vowel categories affect AV matching abilities in infants (Altvater-Mackensen et al., 2015). In particular, Altvater-Mackensen et al. suggested that the contrast between the lip-rounding feature of the vowel /o/ and the lip-spreading feature of the vowel /a/ might provide a more perceptually prominent cue for detecting the AV mismatch in the /a/-/o/ contrast, in comparison to similarly spread lips in the /a/-/e/ contrast. In the current study we provide converging evidence that visual distinctiveness is relevant for infants' AV matching abilities, but we also extend the prior work by showing that whether or not the visual-articulatory distinctiveness of a vowel pair is salient enough for infants to detect depends upon the speaker. That is, we observed no evidence of AV mismatch detection in Speaker 1, despite that one of the vowels was produced with rounded lips (/u/) and the other with spread lips (/i/). Hence, the difference between spread and rounded lips is not always produced by speakers in a way that is salient enough for infants to detect an AV mismatch. Only when the visual-articulatory features differ more dramatically (as in Speaker 2), are infants able to detect AV mismatch.

Furthermore, our results also suggest that between-vowel differences in horizontal lip-opening (which were larger in Speaker 2) are more relevant for AV matching than between-vowel differences in vertical lip-opening (which were larger in Speaker 1)—although a

caveat is that the horizontal between-vowel differences were, in pixels, larger than the vertical between-vowel differences. Future work would be needed to determine whether infants are still more sensitive to horizontal differences when size of between-vowel difference is controlled.

Finally, in Speaker 2 we observed AV mismatch detection only for the visual vowel /i/. This also supports the idea that visual distinctiveness of the sounds is relevant in this task, as it is Speaker 2's mouth shape when producing /i/ that is most visually distinct from the mouth shape that would be expected for an auditory /u/ (**Figure 1**). Beyond the parameters discussed above, other information conveyed by visual means can be also considered. For instance, a recent report suggest that adults' visual vowel discrimination depends on the lip-kinematics (i.e., more or less dynamic mouth movements; Masapollo et al., 2019). Future studies could also focus on whether lip-kinematics affect infants' AV vowel processing within and/or across speakers. Furthermore, although we selected speakers whose expressions we judged to be similarly infant-friendly (and we found no difference in infants' overall looking time between the two speakers), there may have been subtle differences in the speakers' overall affect. Future studies should address whether infants' AV matching ability is influenced by speakers' affect.

It is also important to note that there were inter-speaker differences in the acoustic, as well as visual properties of the stimuli (**Table A1 and A2 – Appendix**). Particularly, we observed greater vowel distinctiveness in Speaker 2 than Speaker 1 with respect to F3 values, reflecting the difference in the mouth rounding—Speaker 2 produced the vowel /u/ with a more rounded mouth shape than Speaker 1 did. In addition, Speaker 1 produced the two vowel categories slightly more distinctively on F2. Thus, although we have focused on visual differences in our interpretation (in part considering the findings of Altvater-Mackensen et al., 2015 and in part because when categorizing the auditory vowels, adults showed no

evidence of sensitivity to inter-speaker differences), it is possible that acoustic differences also affected the infants' performance. Regardless of whether the inter-speaker differences we have observed in the current study are based on visual distinctiveness, acoustic distinctiveness, or a combination of the two, the larger point, that infants' AV matching ability is modulated by inter-speaker differences, remains.

Interestingly, unlike Altvater-Mackensen et al. (2015), who used a paradigm similar to ours, we found that infants looked longer to AV mismatching over AV matching trials. One difference between the two studies that might explain this difference is related to the different vowel pairs used across the studies. In the current study, infants were presented with vowels that are more distinct from one another than in Altvater-Mackensen et al., and there is evidence that AV mismatches that are particularly large (i.e., are perceived as impossible by adults) elicit longer looking times than AV matched events (Tomalski et al., 2013). Therefore, it is possible that AV mismatch trials with more distinct vowels, such as the ones presented in the current study, elicit behaviors similar to the AV impossible trials presented in Tomalski and colleagues. In addition, the infants tested in the current study were a little younger than the infants tested in Altvater-Mackensen et al. (2015). Age is also a contributing factor for preference directions in infants (e.g., Hunter & Ames, 1988).

In summary, this study demonstrates that infants' AV matching ability is modulated by inter-speaker differences. Future experiments should consider speaker-related, and not only phonetic category-related effects when it comes to evaluating young infants' AV processing. It may be that an ability that infants are not thought to possess at a given age could be evident if they were tested on a different speaker, or conversely, that an ability infants *are* thought to possess at a given age is only evident with sufficiently distinctive cues. It is even possible that taking speaker differences into account may help resolve discrepancies in the literature about the age at which infants develop various AV speech perception abilities. Perhaps most

interestingly, if future work reveals that the inter-speaker differences we observed were due to visual articulatory distinctiveness, then the distinctiveness of caregivers' visual articulations may even play a role in phonetic learning (as has been suggested for the auditory domain, e.g., Liu et al., 2003). We hope that our findings will stimulate research into some of these questions.

Finally, our study together with previous reports (Altvater-Mackensen et al., 2015) suggests that infants' AV perception is shaped by general articulatory-acoustic features not specific to vowel categories (e.g., the distinctiveness of the visual cues that accompany speech sounds). More research is needed to address the questions of whether visual distinctiveness related to speakers during AV processing is relevant during the processing of other speech sounds, and whether inter-speaker differences are relevant at later stages of development (e.g., after 1 year of age, when speech sound categories are more established).

Funding

This research was funded by the grant PSI2014-5452-P from the Spanish Ministry of Economy and Competitiveness to M.M. The authors also acknowledge financial support from the "Severo Ochoa" Programme for Centres/Units of Excellence in R&D" (SEV-2015-490) and from the Basque Government "Programa Predoctoral" to J.P.

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*(2), 362–378.

Altvater-Mackensen, N., Mani, N., & Grossmann, T. (2015). 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. *Developmental Psychology*, *52*(2), 191–204.

Baier, R., Idsardi, W. J., & Lidz, J. (2007, August). 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 Proceessing.*

Buchner, A., Erdfelder, E., Faul, F., & Lang, A. (2009). G* Power (version 3.1. 2)[computer program].

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*(6), 1529–1542.

Huei-Mei Liu, Patricia K. Kuhl, & Feng-Ming Tsao. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, *6*(3), F1–F10.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In *Advances in Infancy Research, Vol. 5* (pp. 69–95).

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*(2 Pt 1), 701–714.

Kuhl, P., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138–1141.

Kuhl, P., & Meltzoff, A. (1984). The Intermodal Representation of Speech in Infants. *Infant Behavior and Development*, *7*(3), 361–381.

Masapollo, M., Polka, L., Ménard, L., Franklin, L., Tiede, M., & Morgan, J. (2019). Asymmetries in Unimodal Visual Vowel Perception: The Roles of Oral-Facial Kinematics, Orientation, and Configuration. *The Journal of the Acoustical Society of America*, *44*(7), 1103–1118.

Molnar, M., Gervain, J., & Carreiras, M. (2014). Within-rhythm class native language discrimination abilities of Basque-Spanish monolingual and bilingual infants at 3.5 months of age. *Infancy*, *19*(3), 326–337.

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

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

Streri, A., Coulon, M., Marie, J., & Yeung, H. H. (2016). Developmental Change in Infants' Detection of Visual Faces that Match Auditory Vowels. *Infancy*, *21*(2), 177–198.

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*(5), 611–624.

Yeung, H. H., & Werker, J. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*(5), 603-612.

Appendix

Table A1

Acoustic properties of matched and mismatched auditory stimuli across the two speakers

	Speaker 1		Speaker 2	
Vowel	/i/	/u/	/i/	/u/
Mean duration (s)	1.46	1.42	1.39	1.40
Duration range (s)	1.26-1.66	1.10-1.65	1.13-1.60	1.23-1.60
Mean pitch (Hz)	223.18	226.68	258.67	263.19
Pitch range (Hz)	219.50-229.36	191.80-232.60	249.30-268.40	247.30-277.30
Mean intensity (dB)	64.85	65.2	65.05	65.45
Intensity range (dB)	63.90-65-7	63.9-69.20	63.10-66.90	64.90-66.10
Mean ISI (s)	2.01	2.03	1.92	1.88
ISI range (s)	1.74-2.20	1.75-2.33	1.53-2.19	1.40-2.20

Table A2

The mean values on acoustic (F1, F2, F3) and visual measures (Horizontal, Vertical lip-opening) across vowels and speakers.

	Speaker 1			Speaker 2		
	/i/	/u/	Mean difference between the vowels	/i/	/u/	Mean difference between the vowels
F1 (Hz)	400	375	25	348	384	35
F2 (Hz)	2871	711	2160	2617	764	1853
F3 (Hz)	3395	3294	101	3289	2962	327
Horizontal lip-opening (pixels)	132	69	63	172	78	94
Vertical lip-opening (pixels)	75	64	11	62	58	3

Note. The mean difference between the vowels is given as an absolute value.