Published in final edited form as:
J Exp Child Psychol. 2015 February ; 0: 147–162. doi:10.1016/j.jecp.2014.10.006.

Perception of the Multisensory Coherence of Fluent Audiovisual Speech in Infancy: Its Emergence & the Role of Experience

David J. Lewkowicz[1], **Nicholas J. Minar**[2], **Amy H. Tift**[2], and **Melissa Brandon**[2]

[1]Northeastern University

[2]Florida Atlantic University

Abstract

To investigate the developmental emergence of the ability to perceive the multisensory coherence of native and non-native audiovisual fluent speech, we tested 4-, 8–10, and 12–14 month-old English-learning infants. Infants first viewed two identical female faces articulating two different monologues in silence and then in the presence of an audible monologue that matched the visible articulations of one of the faces. Neither the 4-month-old nor the 8–10 month-old infants exhibited audio-visual matching in that neither group exhibited greater looking at the matching monologue. In contrast, the 12–14 month-old infants exhibited matching and, consistent with the emergence of perceptual expertise for the native language, they perceived the multisensory coherence of native-language monologues earlier in the test trials than of non-native language monologues. Moreover, the matching of native audible and visible speech streams observed in the 12–14 month olds did not depend on audio-visual synchrony whereas the matching of non-native audible and visible speech streams did depend on synchrony. Overall, the current findings indicate that the perception of the multisensory coherence of fluent audiovisual speech emerges late in infancy, that audio-visual synchrony cues are more important in the perception of the multisensory coherence of non-native than native audiovisual speech, and that the emergence of this skill most likely is affected by perceptual narrowing.

Social interactions usually involve the use of audiovisual speech (Rosenblum, 2008). Such speech consists of temporally coupled, redundant, and, thus, equivalent streams of audible and visible information (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Munhall & Vatikiotis-Bateson, 2004; Yehia, Rubin, & Vatikiotis-Bateson, 1998). Because of its multisensory equivalence, adults usually perceive audiovisual speech as a coherent entity and not as two distinct streams of information (McGurk & MacDonald, 1976; Rosenblum, 2008; Sumby & Pollack, 1954; Summerfield, 1979; Yehia et al., 1998). This fact raises some obvious developmental questions: When in development might this ability emerge, does it emerge in infancy, and does experience contribute to its emergence?

Address correspondence to David J. Lewkowicz, Department of Communication Sciences & Disorders, Northeastern University, 360 Huntington Ave., 226FR, Boston, MA 02115; d.lewkowicz@neu.edu.

Several studies have investigated these questions either by asking whether infants can associate fluent audible and visible speech (Bahrick, Hernandez-Reif, & Flom, 2005; Brookes et al., 2001) or whether they can match one of two faces articulating fluent speech in two different languages with a concurrently presented audible utterance that corresponds to one of the talking faces (Dodd & Burnham, 1988; Kubicek et al., 2014; Lewkowicz & Pons, 2013). These studies have indicated that infants can associate fluent audible and visible speech and that they can match a talking face to a corresponding audible utterance but only when the two are in the infants' native language. The matching findings are especially interesting because they suggest that infants can perceive the multisensory coherence of audiovisual speech. Unfortunately, however, the interpretation of the latter findings is complicated by the fact that infants had access to cross-linguistic discriminative cues and that these may have facilitated audio-visual (A-V) matching. If so, this raises two questions: (1) can infants perceive the multisensory coherence of audiovisual speech in the absence of cross-linguistic cues, and (2) if they can, at what point does this ability first emerge?

Obviously, infants should be able to perceive the multisensory coherence of fluent speech at some point - even in the absence of cross-language discriminative cues - because the perception of the multisensory coherence of their world and, especially, of their native language is fundamental to cognition (Gibson, 1969; Piaget, 1952; Rosenblum, 2008; Thelen & Smith, 1994). Most likely, however, this ability emerges relatively late in infancy for two reasons. First, speech and language perception skills emerge slowly and gradually during infancy. This is illustrated by the fact that it is not until the end of the first year of life that infants become relatively sophisticated perceivers of their native language (Saffran, Werker, & Werner, 2006; Werker, Yeung, & Yoshida, 2012). Second, multisensory processing skills also emerge slowly and gradually during infancy (Bremner, Lewkowicz, & Spence, 2012; Lewkowicz, 2014; Lewkowicz & Ghazanfar, 2009). This is illustrated by the fact that even though from birth on infants can perceive the coherence of human auditory and visual speech (Dodd, 1979; Lewkowicz, 1996a, 2000a, 2010), non-human communicative signals (Lewkowicz, Leo, & Simion, 2010), and non-speech auditory and visual information (Bahrick, 1983; Brookes et al., 2001; Lewkowicz, 1986, 1992a, 1992b, 1996b), they do so only based on whether the signals in the two modalities occur together or not. It is not until the second half of the first year of life that infants begin to perceive the multisensory coherence of their audiovisual world based on more specific and more complex attributes such as gender (Patterson & Werker, 2002; Walker-Andrews, Bahrick, Raglioni, & Diaz, 1991), affect (Walker-Andrews, 1986), and identity (Lewkowicz & Pons, 2013).

The role of A-V synchrony cues in perception is especially interesting because of their fundamental importance to perception throughout the lifespan and their complex interaction with other usually concurrent multisensory relational cues. For example, some studies have found that young infants can perceive the equivalence of the facial and vocal attributes of isolated speech syllables even when the audible syllable is temporally synchronized with both visible syllables (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2002, 2003; Walton & Bower, 1993). This suggests that, at least in the case of single syllables, infants are able to extract phonetic multisensory invariance. Studies of older (6- and 11-month-old)

infants have found similar evidence except that by then infants can even map previously heard syllables onto subsequently presented visible articulations of the same syllables (Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009).

Although the fact that infants can perceive the multisensory coherence of isolated audiovisual syllables is interesting, studies show that A-V synchrony cues not only play a role in infant processing of fluent audiovisual speech but that such cues continue to play a role in multisensory perception into adulthood. For instance, findings show that infants attend more to synchronized than desynchronized audiovisual fluent speech (Dodd, 1979) and that they can only learn specific face-voice associations when talking faces and accompanying voices are temporally synchronized (Bahrick et al., 2005). Moreover, findings indicate that children as well as adults not only can detect the temporal alignment of auditory and visual information but that this ability improves with development and that detection of A-V temporal relations continues to play a key role in the perception of multisensory coherence into adulthood (Dixon & Spitz, 1980; Grant, van Wassenhove, & Poeppel, 2004; Hillock-Dunn & Wallace, 2012; Lewkowicz, 1996b; Lewkowicz & Flom, 2013). Finally, studies have found no correlation in adults' responsiveness to audiovisual nonsense syllables, on the one hand, and adults' responsiveness to audiovisual sentences, on the other, (Grant & Seitz, 1998), suggesting that infants' responsiveness to audiovisual syllables may not generalize to their responsiveness to fluent audiovisual speech.

Studies of selective attention using eye tracking methodology have provided additional evidence that infants rely on A-V synchrony cues when processing fluent speech (Hunnius & Geuze, 2004; Lewkowicz & Hansen-Tift, 2012). For example, in one of these studies, videos of a person speaking either in the native language or in a non-native language were presented to monolingual, English-learning infants of different ages while their point of gaze to the talker's eyes and mouth was monitored (Lewkowicz & Hansen-Tift, 2012). Findings yielded striking developmental shifts in selective attention. Specifically, when presented with a person speaking in their native language, 4-month-olds attended more to her eyes, 6-month-olds attended equally to her eyes and mouth, 8- and 10-month-olds attended more to her mouth, and 12-month-olds attended equally to her eyes and mouth. The first attentional shift to the talker's mouth observed in the 8- and 10-month-olds happens to correspond with the onset of speech production (i.e., canonical babbling) and, as such, enables infants to gain direct access to the source of audiovisual speech. This way, infants can profit maximally from the greater perceptual salience of the multisensory redundancy of the signal which is, in part, due to the synchronous nature of the audible and visible speech streams. The second attentional shift away from a talker's mouth observed in response to native audiovisual speech by 12 months of age happens to correspond with the emergence of an initial expertise for the native language. This shift suggests that by this age infants may no longer need to rely on audiovisual redundancy for speech processing. This conclusion is supported by the fact that when infants were exposed to a person speaking in a non-native language (Spanish), they not only attended more to her mouth at eight and 10 months of age but that they continued to do so at 12 months of age.

Lewkowicz and Hansen-Tift (2012) interpreted the continued attentional focus on the mouth at 12 months as a reflection of a decline in infants' ability to perceive the perceptual

attributes of a non-native language due to emerging expertise for native speech and a concurrent narrowing of the ability to perceive non-native speech. The latter process renders non-native speech unfamiliar (Lewkowicz, 2014; Lewkowicz & Ghazanfar, 2009; Werker & Tees, 2005) and because of this, increased attention the synchronous and, thus, redundant perceptual cues available in a talker's mouth presumably provides infants with maximally discriminable cues that can help them disambiguate what has now become unfamiliar speech.

From the current perspective, and with specific regard to the importance of A-V synchrony cues in infant speech perception, the Lewkowicz and Hansen-Tift (2012) findings lead to two conclusions. First, the findings from infants' responsiveness to native speech indicate that once initial native-language perceptual expertise emerges by the end of the first year of life, infants no longer depend as much on audiovisual redundancy and, thus, presumably on the tight temporal correlation of the audible and visible streams of native speech. Second, the findings from infants' responsiveness to non-native speech indicate that infants do continue to depend on audiovisual redundancy and, thus, on A-V temporal correlation when exposed to what has now become unfamiliar speech.

The specific role of synchrony cues in infant perception of fluent audiovisual speech has not been investigated in A-V matching studies to date. In the three studies in which infants were tested with speech in different languages and which reported that infants can perceive multisensory speech coherence, synchronous auditory and visual information was presented in two of them (Dodd & Burnham, 1988; Kubicek et al., 2014) whereas asynchronous auditory and visual information was presented in the third one (Lewkowicz & Pons, 2013). Moreover, Dodd & Burnham (1988) found that 20-week-old infants can match the faces and voices of their native speech (English) but not non-native speech (Greek) and Kubicek and colleagues (2014) reported that German-learning 12-month-old infants can perceive the identity of their native language as opposed to a non-native language (French). Similarly, Lewkowicz & Pons (2013) found that 10–12 month-old English-learning infants, but not 6–8 month-old infants, can perceive the multisensory identity of a native as opposed to a non-native language (Spanish) when the audible and visible information was not presented concurrently. As noted earlier, although these findings demonstrate that infants can perceive the multisensory coherence of audiovisual speech, the language pairs used in them (English-Greek, German-French, and English-Spanish) are prosodically distinct. This makes it possible that the prosodic differences contributed to the detection of multisensory coherence in those studies. This, in turn, raises the question of whether infants also can perceive the coherence of audible and visible speech in the absence of cross-linguistic prosody cues. Furthermore, given that the previous studies only obtained A-V matching of native audible and visible speech, this finding begs the question of whether this reflects monolingual infants' exclusive experience with their native language and, if so, whether this affects their ability to perceive the multisensory coherence of non-native speech?

We carried out four experiments to answer these questions by testing infants' ability to match an audible monologue with one of two different and concurrently visible monologues. Crucially, here, both of the monologues were spoken in the same language. Thus, we presented two identical faces talking in the same language and asked whether infants would

look longer at the face whose articulations corresponded to a concurrently presented audible utterance. To determine whether experience may play a role in responsiveness, we tested some infants with audible and visible monologues spoken in their native language (English) and others with monologues spoken in a non-native language (Spanish). In Experiments 1–3 we tested 4-, 8–10, and 12–14 month-old infants' responsiveness to synchronous audible and visible native and non-native fluent speech. In Experiment 4, we tested 12–14 month-olds' response to the same stimuli except that this time the audible and visible speech streams were desynchronized.

Experiment 1

This experiment investigated whether 4-month-old infants can match synchronous audible and visible speech streams. We made three specific *a priori* predictions that were based on theoretical and empirical grounds. On theoretical grounds, it is reasonable to expect that infants will at some point begin to detect multisensory coherence because this is essential for the acquisition of a unified conception of the world (Gibson, 1969; Piaget, 1952; Thelen & Smith, 1994). On empirical grounds, it is also reasonable to expect that at some point in development infants should look longer at the face whose visible articulations correspond to audible articulations. Empirical evidence indicates that infants look more at a visual stimulus that corresponds to an auditory stimulus than either at the same visual stimulus presented in silence or at another visual stimulus that does not correspond to the auditory stimulus once they begin to perceive multisensory coherence (Bahrick et al., 2005; Lewkowicz, 1986, 1992a; Lewkowicz & Ghazanfar, 2006; Lewkowicz et al., 2010; Walker-Andrews, 1986).

Thus, our first *a priori* prediction was based on the specific design of the current study and on a comparison of looking at two talking faces first presented in silence and then in the presence of a concurrent soundtrack that corresponded to one of the talking faces. We predicted that infants would look more at the face that corresponds to a synchronously presented audible speech stream during its presentation than in its absence if they perceived the multisensory coherence of the visible and audible speech streams. Our second specific *a priori* prediction was that responsiveness during the audiovisual test trials was likely to change rapidly across repeated and identical test blocks. This is because prior studies using the same multisensory matching procedure as used here have found that, as infants gain increasing experience with the same visual stimuli during the course of an experiment, they cease to exhibit evidence of multisensory matching (Bahrick et al., 2005; Bahrick, Moss, & Fadil, 1996; Bahrick, Netto, & Hernandez-Reif, 1998; Walker-Andrews, 1986). Our final *a priori* prediction, which was closely related to the second one, was that changes in responsiveness across the test trials may differ for the native as opposed the non-native language. That is, infants may cease performing A-V matching when exposed to native audiovisual speech but may not when exposed to non-native speech because the latter becomes harder to process once perceptual narrowing has occurred (Werker et al., 2012).

As we indicated earlier, the ability to perceive the multisensory coherence of audiovisual speech probably does not emerge until relatively late in infancy. Therefore, we did not expect to obtain evidence of A-V matching at this age. Nonetheless, testing infants as young

as four months of age is essential to define more precisely the age when this ability begins to emerge.

Method

Participants—Forty-eight 4-month-old infants were tested (17 girls; M age = 17.1 weeks, range = 16.0 – 18.7 weeks). All infants came from monolingual homes. To determine the degree of language exposure, we administered a language questionnaire to the parents. The questionnaire included questions concerning (a) the infant's primary language and any other additional languages, (b) the number of hours of exposure to each language during awake time per each day of the week, and (c) the source of the speech heard by the infant per day (i.e., mother, father, grandparents, relatives, caregiver, other). Based on the results of this questionnaire, we calculated the percent of exposure to each language per week and only included infants whose language exposure to English exceeded 81%. Fifteen additional infants were tested but were excluded from data analysis due to fussiness (n = 6), inattentiveness/parent interaction (n = 5), or health concerns such as eye or ear infection (n = 4).

Apparatus, Stimuli, & Design—Infants were tested in a sound-attenuated booth. During the experiment, most of the infants were seated in an infant seat. If parents requested to have the infant on the lap or if the infant refused to sit in an infant seat, the parents were permitted to hold their infant in their lap. When they did, they wore headphones through which they listened to music, were not aware of the hypothesis under test, and were asked to sit still and refrain from any interactions with their infant. The infants were seated 50 cm from two side-by-side 17-inch (43.2 cm) LCD display monitors that were spaced 6.7 cm apart. A video camera was located mid-way between the two monitors and was used to record the infants' visual fixations of the talking faces on the two monitors. The experimenter was seated outside the booth and could see the infant and parent through a one-way mirror as well as via the video camera focused on the infants' face.

The stimulus materials consisted of four videos. In two of them, a female actor could be seen and heard speaking in her native English whereas in the other two another female actor could be seen and heard speaking in her native Spanish. Each of the four videos consisted of three blocks of test trials and each block consisted of two preference trials. Thus, each video consisted of a total of six 20 s paired-preference test trials. During each preference trial, infants saw side-by-side faces of the same female speaking two different monologues[1], with the side on which the two different monologues were presented switched during the second trial in each block.

[1]English monologue 1: "Good morning! Get up! Come on now, if you get up right away we'll have an hour to putter around. I love these long mornings, don't you? I wish they could last all day. Well at least it's Friday." English monologue 2: "Except, of course, for the party. Are you going to help me fix up the house? Are you? We need to buy flowers, prepare the food, vacuum the house, dust everything, and clean the records." Spanish monologue 1: "¡Desperate ya! ¡Vamos! ¡Si te levantas ahora, tendremos una hora para jugar en la casa! Me encantan estas mañanas largas, ¿y a tí? Ojalá pueden durar todo el día. Bueno, por lo menos es viernes y tenemos todo el sábado para descansar." Spanish monologue 2: "Bueno, por lo menos es viernes y tenemos todo el sábado para descansar, excepto por lo de la fiesta. ¿Me vas a ayudar arreglar la casa? ¿Si? Tenemos que comprar las flores, preparar la comida, limpiar el polvo, aspirar la casa y limpiar los discos."

The first block of trials was the silent block. Here, infants saw the two faces talking in silence. The data from this block provided a measure of responsiveness to each visible monologue in the absence of the audible monologue and, thus, served as a baseline measure. The second and third blocks of trials were the audiovisual test trials. Here, infants saw the same talking faces again and also heard one of the corresponding audible monologues during one block and the other audible monologue during the other block. The order of presentation of the two audible monologues was counterbalanced across the two audiovisual blocks of trials. Table 1 shows the experimental design used to construct the two versions of the videos for each language, respectively, including the way in which stimulus presentation was counterbalanced within and across the two videos. As indicated above and as can be seen in Table 1, the side of visual monologue presentation was counterbalanced within each block of trials, the side of visual monologue presentation was counterbalanced across the two videos for each language, and the order of audible monologue presentation was counterbalanced across the two videos for each language. Half the infants were assigned to the two English videos and the other half were assigned to the Spanish videos.

The sound pressure level of the audible monologue was 60±5 dB (A-scale). The actor smiled and spoke in a highly-prosodic style, meaning that she spoke in a slow and highly exaggerated manner with large pitch variations similar to the way adults usually talk to infants (Fernald, 1989).

Procedure—The experimenter's only task during the test session was to start the presentation of one of the four videos. Thus, the experimenter had no other control over the presentation of the stimuli nor over the infant's behavior. To center the infants' eye gaze in-between the test trials, a rotating multicolored disk was presented in the middle of the infants' visual field (the disc was split in half, with each half presented on the lower portion of each monitor, closest to the center point between the two monitors). The video recording of the infants' looking behavior was coded off-line by trained observers who were blind with respect to the stimuli presented as well as to the hypothesis under test. Inter-coder reliability between two independent coders scoring 20% of the infants yielded an agreement rate greater than 95% based on a Pearson *r*. The same was the case for the subsequent experiments.

Results and Discussion

We calculated the proportion-of-total-looking-time (PTLT) that each infant directed at the matching face for each of the three blocks of trials. This was done by dividing the total amount of looking at the matching face by the total amount of looking at both faces over the two trials of each block, respectively. If infants perceived the match between the visible and audible monologues then they were expected to exhibit increased looking at the corresponding talking face in the presence of the audible monologue than in its absence. Thus, the comparison of interest was the difference between baseline-PTLT scores obtained in the silent block of trials and the test-PTLT scores obtained within each of the two audiovisual blocks of trials, respectively. As indicated in Table 1, the specific audible monologue presented in each audiovisual block of trials differed. Therefore, the baseline-PTLT that served as a comparison for each of the respective audiovisual blocks of trials

differed. Specifically, the baseline-PTLT for one block of audiovisual test trials consisted of the proportion of looking at the visible monologue corresponding to the audible monologue presented in that block whereas the baseline-PTLT for the other block of audiovisual test trials consisted of the proportion of looking at the other visible monologue because it corresponded to the other audible monologue.

Figure 1 shows the PTLT scores for this experiment. We conducted a preliminary repeated-measures analysis of variance (ANOVA) first to determine whether responsiveness during the audiovisual blocks of test trials was affected by trial block and/or language. This ANOVA included Trial Type (2; silent speech, audiovisual speech) and Blocks (2; 1st audiovisual, 2nd audiovisual block of test trials) as the within-subjects factors and Language (2) as a between-subjects factor. There were no significant main effects of Trial Type, F (1, 46) = .008, p = .93, η_p^2 = 0, Block, F (1, 46) = 1.24, p = .27, η_p^2 = 0.26, nor Language, F (1, 46) = .003, p = .94, η_p^2 = 0. In addition, there were no significant interactions between Trial Type and Block, F (1, 46) = 1.26, p = .27, η_p^2 = .027, Trial Type and Language, F (1, 46) = .003, p = .95, η_p^2 = 0, nor between Trial Type, Block, and Language, F (1, 46) = .002, p = .96, η_p^2 = 0.

Despite the absence of any significant effects, we considered the overall ANOVA to be, at best, a conservative measure of responsiveness. This is because it does not take into account the clear *a priori* directional predictions that we described in the Introduction to this experiment. As a result, we performed planned contrast analyses to test the three *a priori* predictions offered earlier. To reiterate, we predicted that infants would look longer at the sound-specified visible speech than at the same but silent speech if they perceived the multisensory coherence of visible and audible speech. The second was that responsiveness during the audiovisual test trials was likely to change rapidly as infants gain increasing experience with the same visual stimuli as the experiment progresses (Bahrick et al., 2005; Bahrick et al., 1996; Bahrick et al., 1998; Walker-Andrews, 1986). The third was that changes in responsiveness across the test trials may be differ for the native as opposed the non-native language. To test the *a priori* predictions, we used one-tailed t-tests to compare the test-PTLT versus the baseline-PTLT directed at the matching visible monologue in each block of audiovisual test trials, respectively, separately for each language condition.

The planned comparisons indicated that those infants who were tested with English did not look longer at the sound-specified visible monologue than at the same silent monologue in the first block of audiovisual test trials, t (23) = −0.52, p = .30, Cohen's d = .22, nor in the second block of audiovisual test trials, t (23) = 0.59, p = .28, Cohen's d = .25. Similarly, the planned comparisons showed that those infants who were tested with Spanish did not look longer at the sound-specified visible monologue than at the same silent monologue in the first block of audiovisual test trials, t (23) = −0.46, p = .32, Cohen's d = .19, nor in the second block of audiovisual test trials, t (23) = 0.64, p = .26, Cohen's d = .27. Overall, these findings show that 4-month-old infants do not perceive the coherence of audible and visible fluent speech.

Experiment 2

The results from Experiment 1 indicated that 4-month-old infants did not match the audible and visible streams of fluent speech. This failure might be attributable to their young age and/or their relative inexperience with audiovisual speech. To test this possibility, in Experiment 2 we tested a group of 8–10 month-old infants. We chose this specific age range because it is during this time in development that infants begin to attend specifically to audiovisual speech by focusing on a talker's mouth (Lewkowicz & Hansen-Tift, 2012). This attentional focus may facilitate the detection of the overlapping and time-locked dynamic variations in auditory and visual speech streams in the 8–10 month age range. Alternatively, it may be that infants require additional experience with this aspect of fluent speech and, because of this, may not be able to perceive the multisensory coherence of audiovisual speech in this specific age range.

Method

Participants—Fifty-five 8–10 month-old infants were tested (30 girls; M age = 37.92 weeks, range = 33.29 – 44.29 weeks). All infants came from monolingual homes (81% or more of the language exposure was in English). Five additional infants were tested but were not included in the data analysis due to fussiness (n = 4) or equipment failure (n = 1).

Apparatus & Stimuli—The apparatus and stimuli used in this experiment were identical to those used in Experiment 1.

Procedure—The procedure in this experiment was identical to the procedure used in Experiment 1.

Results and Discussion

Figure 2 depicts the PTLT scores from this experiment. As in Experiment 1, first we conducted a repeated-measures ANOVA on the PTLT scores, with Trial Type (2) and Blocks (2) as within-subjects factors and Language (2) as a between-subjects factor. There were no significant main effects of Trial Type, F (1, 53) = .06, p = .81, η_p^2 = 0.001, Block, F (1, 53) = 0.12, p = .73, η_p^2 = 0.002, nor Language, F (1, 53) = .087, p = .77, η_p^2 = 0.002. There was a marginally significant interaction between Trial Type and Block, F (1, 53) = 3.22, p = .07, η_p^2 = .058, but there were no significant interactions between Trial Type and Language, F (1, 53) = .087, p = .77, η_p^2 = 0.002, nor between Trial Type, Block, and Language, F (1, 53) = .006, p = .94, η_p^2 = 0.

The planned comparisons indicated that those infants who were tested with English did not look longer at the sound-specified visible monologue than at the same silent monologue in the first block of audiovisual test trials, t (27) = −0.99, p = . 16, Cohen's d = .26, and that they exhibited marginally significant greater looking at the non-sound-specified visible monologue in the second block of audiovisual test trials, t (27) = 1.34, p = .08, Cohen's d = . 34. For those infants who were tested with Spanish, the planned comparisons showed that they did not look longer at the sound-specified visible monologue than at the same silent monologue in the first block of audiovisual test trials, t (26) = −0.90, p = .19, Cohen's d = .

26, nor in the second block of audiovisual test trials, t (26) = 0.99, p = .16, Cohen's d = .28. Overall, these findings show that, like the 4-month-olds, the 8–10 month-old infants did not perceive the multisensory coherence of audiovisual fluent speech.

Experiment 3

The results from Experiments 1 and 2 indicated that neither the 4-month-old nor the 8–10 month-old infants matched the audible and visible streams of fluent speech. This is probably due to a combination of factors including the infants' relative inexperience, immaturity, and/or the greater complexity of fluent audiovisual speech as opposed to isolated speech syllables. As a result, in Experiment 3 we tested 12–14 month-old infants. We expected that by this age, infants should be able to perceive multisensory speech coherence given their greater experience with speech and given that by this age they have attained a degree of auditory-only (Werker et al., 2012) and audiovisual expertise (Lewkowicz, 2014; Lewkowicz & Ghazanfar, 2009) in their native language. Based on the specific *a priori* predictions outlined in the Introduction to Experiment 1, here we expected that infants would exhibit perception of multisensory coherence early in the test trials when exposed to native speech and only later in the test trials when exposed to nonnative speech.

Methods

Participants—Forty-eight 12–14 month-old infants were tested (24 girls; M age = 56.0 weeks, range = 51.0 – 61.1 weeks). All infants came from monolingual homes (81% or more of the language exposure was in English). Five additional infants were tested but were not included in the data analysis due to inattentiveness/parent interaction (n = 4) or ear infection (n = 1).

Apparatus & Stimuli—The apparatus and stimuli used in this experiment were identical to those used in Experiment 1.

Procedure—The procedure in this experiment was identical to the procedure used in Experiment 1.

Results & Discussion

Figure 3 shows the PTLT scores from this experiment. As in Experiment 1, first we conducted a repeated-measures ANOVA on the PTLT scores, with Trial Type (2) and Blocks (2) as within-subjects factors and Language (2) as a between-subjects factor. This analysis yielded a significant Trial Type effect, F (1, 46) = 8.13, p = .007, η_p^2 = .15, but no Block effect, F (1, 46) = .00, p = .98, η_p^2 = .00, nor Language effect, F (1, 46) = 1.65, p = .21, η_p^2 = .035. In addition, there were no interactions between Trial Type and Block, F (1, 46) = .22, p = .64, η_p^2 = .005, Trial Type and Language, F (1, 46) = 1.88, p = .18, η_p^2 = .039, nor between Trial Type, Block, and Language, F (1, 46) = 2.17, p = .15, η_p^2 = .045. The Trial Type effect indicates that, overall, infants looked significantly longer at the matching visible monologue in the presence of the audible monologue than in its absence.

To further probe the main effect of Trial Type, and to test our *a priori* hypotheses, once again we performed planned comparison analyses. These analyses indicated that those

infants who were tested with English looked longer at the sound-specified visible monologue than at the same silent monologue in the first block of audiovisual test trials, t (23) = 1.72, p = .049, one-tailed, Cohen's d = .72, and that they no longer did so in the second block of audiovisual test trials, t (23) = 0.83, p = .21, one-tailed, Cohen's d = .35. The cessation of matching in the second block of audiovisual test trials probably reflects the effects of habituation due to repeated exposure to the same visual stimuli over the course of the experiment and is consistent with similar effects reported in other studies (Bahrick et al., 2005; Bahrick et al., 1996; Bahrick et al., 1998; Walker-Andrews, 1986).

The planned comparison analyses also indicated that those infants who were tested with Spanish did not look longer at the sound-specified visible monologue than at the same silent visible monologue in the first block of audiovisual test trials, t (23) = 0.60, p = .27, one-tailed, Cohen's d = .25, but that they did in the second block of audiovisual test trials, t (23) = 1.92, p = .03, one-tailed, Cohen's d = .80. The later onset of matching in this condition probably reflects the combined effects of perceptual narrowing and native-language specialization. That is, by the end of the first year of life, monolingual English-learning infants have had exclusive exposure to their native language. As a result, their ability to recognize the perceptual attributes of a non-native language has declined while their expertise for the native language has increased. Because of these two developmental processes, it takes infants longer to discover the multisensory coherence of Spanish audiovisual speech.

Experiment 4

Experiments 1 and 2 demonstrated that neither 4- nor 8–10 month-old infants detected the multisensory coherence of audiovisual fluent speech regardless of whether it was English or Spanish. In contrast, Experiment 3 showed that 12–14 month-old infants successfully detected the multisensory coherence of fluent English audiovisual speech in the first block of test trials and of fluent Spanish audiovisual speech in the second block of test trials. Given the previous discussion of the complex role of A-V synchrony cues in perception in early development, the present findings beg the question of whether the older infants relied on A-V synchrony cues to detect multisensory coherence. Recall that prior studies have found that infants can perceive A-V synchrony but most of those studies have tested infants with isolated speech syllables (Lewkowicz, 2000, 2010; Lewkowicz et al., 2010). The two exceptions are the Dodd (1979) study which showed that 10–26 week-old infants actually attend more to synchronous than to asynchronous fluent audiovisual speech and a study by Pons and Lewkowicz (2014) which found that 8-month-old infants actually can discriminate synchronous from asynchronous audiovisual speech. Unfortunately, neither study assessed the possible role of A-V temporal synchrony in infants' detection of the multisensory coherence of audible and visible fluent speech.

Despite the fact that infants are responsive to A-V temporal synchrony cues in both isolated and fluent audiovisual speech and that such cues continue to play an important role into adulthood, it is interesting to note that the relative importance of such cues appears to decline to some extent in early development (Lewkowicz & Ghazanfar, 2009; Lewkowicz & Hansen-Tift, 2012). Specifically, studies indicate that whereas younger infants tend to rely

on synchrony for the detection of multisensory coherence in most cases, older infants rely less on it as they discover more complex multisensory relations. For example, younger but not older infants bind auditory and visual inputs on the basis of synchrony and, importantly, they do this regardless of whether the inputs are part of their native ecology or not (Lewkowicz & Ghazanfar, 2006; Lewkowicz et al., 2010). Similarly, 5-month-old infants require synchrony to match human affective visual and auditory expressions whereas 7-month-old infants do not (Walker-Andrews, 1986). Finally, infants younger than six months of age do not perceive the equivalence of auditory and visual gender attributes in the absence of synchrony cues whereas 6-month-old infants do (Walker-Andrews et al., 1991).

Given the apparent relative decline of the importance of A-V temporal synchrony cues across infancy, we hypothesized that infants may rely less on such cues in their perception of the multisensory coherence of native audiovisual speech but that they may continue to rely on such cues in their responsiveness to non-native audiovisual speech. Therefore, in Experiment 4 we asked whether the temporal synchrony of the audible and visible speech streams may have contributed to the successful matching observed in the 12–14 month-old infants in Experiment 3. Thus, we repeated Experiment 3 except that this time we desynchronized the audible and visible speech streams.

Method

Participants—Fifty-two 12–14 month-old infants were tested (23 girls; M age = 51.4 weeks, range = 51.0 – 61.1 weeks). All infants came from monolingual homes (81% or more of the language exposure was in English). Nine additional infants were tested but were not included in the data analysis due to fussiness (n = 2), inattentiveness/parent interaction (n = 3), or health concerns such as eye or ear infection (n = 4).

Apparatus and Stimuli—The apparatus used in this experiment and the stimuli presented were identical to those in Experiment 2. The only difference was that the audible speech stream was desynchronized vis-à-vis the visible speech stream. We chose an A-V asynchrony of 666 ms because this degree of asynchrony has previously been found to be discriminable to infants tested with, both, isolated syllables (Lewkowicz, 2010) and fluent audiovisual speech (Pons & Lewkowicz, 2014). To achieve desynchronization in the current experiment, we delayed the initial onset of mouth motion at the beginning of the test trial by 666 ms (20 video frames) vis-à-vis the initial onset of the audible monologue. This resulted in a misalignment of the audible and visible speech streams with respect to one another for the entire duration of the test trial.

Procedure—The procedure used in this experiment was the same as in Experiment 3. The only difference was that 28 of the infants were tested in the English condition and 24 were tested in the Spanish condition.

Results and Discussion

Figure 4 shows the PTLT scores for this experiment. As in the other experiments, first we conducted an overall repeated-measures ANOVA on the PTLT scores, with Trial Type (2) and Blocks (2) as within-subjects factors and Language (2) as a between-subjects factor.

There were no significant main effects of Trial Type, F (1, 50) = .20, p = .65, η_p^2 = .073, Block, F (1, 50) = 1.17, p = .28, η_p^2 = .186, nor Language, F (1, 50) = .03, p = .86, η_p^2 = .053. In addition, there were no significant interactions between Trial Type and Block, F (1, 50) = 3.00, p = .09, η_p^2 = .396, Trial Type and Language, F (1, 50) = .29, p = .86, η_p^2 = .053, nor between Trial Type, Block, and Language, F (1, 50) = .07, p = .79, η_p^2 = .058.

The planned comparison analyses showed that those infants who were tested with English exhibited longer looking at the sound-specified visible monologue than at the same but silent monologue in the first block of audiovisual test trials, t (27) = 1.96, p = .03, one-tailed, Cohen's d = .75, but that they did not in the second block of these trials, t (27) = 1.05, p = .15, one-tailed, Cohen's d = .40. The planned comparison analyses also indicated that those infants who were tested with Spanish did not look longer at the sound-specified visible monologue than at the same silent visible monologue in either the first block of the audiovisual test trials, t (23) = .79, p = .22, one-tailed, Cohen's d = .33, nor in the second block of the audiovisual test trials, t (23) = −.86, p = .20, one-tailed, Cohen's d = −.36.

To further determine whether desynchronization of the audible and visible speech streams affected multisensory responsiveness, we compared the visual preferences obtained in the current experiment with those obtained in Experiment 3 in those cases where the 12–14 month-old infants exhibited a preference for the sound-specified visible monologue in Experiment 3. To compare the visual preferences across the two experiments directly, first we computed difference PTLT scores for each experiment (test-PTLT minus baseline-PTLT). Then, we compared the difference scores directly across the two experiments. We found that when the infants were tested with English during the first block of audiovisual test trials, their looking at the sound-specified visible monologue did not differ across the two experiments, t (50) = 0.30, p = .77, 2-tailed, Cohen's d = .085. In contrast, when the infants were tested with Spanish during the second block of the audiovisual test trials, their looking at the sound-specified visible monologue was significantly lower in the Experiment 4 than in Experiment 3, t (46) = 2.01, p = .05, 2-tailed, Cohen's d = .59.

Overall, the findings from this experiment indicated that desynchronization of the English audible and visible speech streams did not disrupt multisensory matching. In contrast, the findings showed that desynchronization of the Spanish audible and visible speech streams did disrupt multisensory matching.

General Discussion

This study investigated whether the ability to perceive the multisensory coherence of fluent audiovisual speech (in the absence of cross-linguistic cues) emerges in infancy and, if it does, whether temporal A-V synchrony plays a role in this ability. Experiments 1 and 2 showed that 4- and 8–10 month-old infants did not exhibit evidence of multisensory matching. In contrast, Experiment 3 showed that 12–14 month-old infants did exhibit evidence of matching when presented with both native and non-native audiovisual speech, although evidence of matching emerged later in the test trials for non-native speech. Finally, Experiment 4 demonstrated that multisensory matching in the 12–14 month-old infants did not depend on the audible and visible speech streams being synchronized when native

audiovisual speech was presented but that they did when non-native audiovisual speech was presented. Together, these findings show that infants become capable of perceiving the multisensory coherence of fluent audiovisual native and non-native speech by the end of the first year of life and that synchrony-based multisensory redundancy is only critical for the perception of the multisensory coherence of non-native audiovisual speech.

The finding that the 4- and 8–10 month-old infants did not exhibit evidence of multisensory matching - even though they had access to synchronous auditory and visual inputs - might, at first blush, seem at odds with previous findings showing that infants are sensitive to AV synchrony relations (Lewkowicz, 2010). The fact is, however, that the previous findings come from studies in which infants only had to detect the onsets and offsets of audible and visible syllables. Thus, findings from studies that have investigated infant perception of A-V synchrony relations inherent in fluent speech are more relevant here. Two such studies have been carried out. One investigated whether 3.5-month-old infants can detect the temporal synchrony of the audible and visible streams of fluent audiovisual speech by presenting synchronized and desynchronized audiovisual speech. Findings showed that infants looked less at desynchronized speech (Dodd, 1979). A more recent study investigated whether 8-month-old infants can discriminate synchronized from desynchronized fluent audiovisual speech and found that they can and that this is the case regardless of whether the speech is native or not (Pons & Lewkowicz, 2014). Although both of these studies show that the A-V synchrony cues inherent in fluent speech are perceived by infants, they do not indicate whether infants rely on them in their detection of audiovisual speech coherence in a task that requires them to detect which of two visible speech utterances corresponds to a concurrent audible speech utterance. This sort of task requires that infants be able to perceive the statistics of continuous and dynamically varying temporal correlations between corresponding audible and visible speech streams (Chandrasekaran et al., 2009; Yehia et al., 1998). Experiments 1 and 2 indicated that neither the 4 nor the 8–10 month-old infants detected such statistics because they exhibited no evidence of multisensory matching. Experiments 3 and 4 did show, however, that the 12–14 month-old infants detected such statistics and that they did so even when those statistics were not defined by A-V synchrony relations in the native-language condition.

The fact that the 12–14 month-old infants exhibited multisensory matching even though A-V synchrony was disrupted in the native-language condition indicates that infants one year of age and older can rely on some other perceptual attributes for multisensory matching. The most likely such attribute is prosody. This conclusion is consistent with findings that adults can use prosody alone to perceive the relation between the acoustic variation in speech and the motion of a corresponding referent (Jesse & Johnson, 2012). The adult findings suggest that our oldest infants also may have relied on prosody to perceive multisensory speech coherence and that they may be more proficient at this when presented with their native language. By the same token, the finding that our oldest infants did not exhibit multisensory matching when the A-V temporal synchrony of the audible and visible speech streams of non-native speech was disrupted suggests that infants of this age still rely on synchrony for the detection of the multisensory coherence of non-native audiovisual speech.

The conclusion that the oldest infants must have relied on perceptual cues other than A-V synchrony to perceive the multisensory coherence of native audiovisual speech sheds new light on the fundamental binding role that A-V synchrony plays in infancy. Lewkowicz (2014) has argued that A-V synchrony plays an especially crucial role as a binding cue early in life because young infants do not yet perceive more complex multisensory perceptual cues and, thus, do not yet bind multisensory inputs based on such cues. This argument is predicated on the assumption that, in most cases, determining whether multisensory inputs are synchronous only requires the detection of the onsets and offsets of such inputs. Given this assumption, Lewkowicz & Ghazanfar (2009) argued that as infants gradually acquire the ability to detect increasingly more complex multisensory perceptual cues, the role of A-V synchrony cues diminishes. Of course, the argument becomes more complicated when A-V synchrony cues specify the dynamic and continuous temporal correlation between audible and visible speech streams. In this case, the question is whether and when infants can detect this type of temporal correlation. The current results indicate that infants one year of age and older relied on this more complex type of temporal correlation when exposed to non-native speech but that they dispensed with it when exposed to native speech. This response pattern demonstrates that by the end of the first year of life infants can track the complex temporal statistics that link audible and visible speech streams. When audiovisual speech is in their native language, infants no longer need to track such statistics because they are now presumably able to detect other multisensory binding cues (e.g., prosody). When, however, audiovisual speech is in a non-native language, infants continue to rely on the precise temporal alignment of the audible and visible speech streams simply because other binding cues have now become unfamiliar, presumably due to perceptual narrowing (Lewkowicz, 2014; Lewkowicz & Ghazanfar, 2009; Scott, Pascalis, & Nelson, 2007; Werker & Tees, 2005).

In conclusion, the current findings are some of the first to demonstrate that infants acquire the ability to perceive the multisensory coherence of native speech at the suprasegmental level by 12–14 months of age. This enables infants to perceive the coherent nature of everyday fluent audiovisual speech and, as a result, enables them to profit maximally from its multisensory redundancy and, thus, its greater perceptual salience. This, in turn, facilitates the extraction of meaning and the subsequent acquisition of increasingly greater linguistic expertise.

Acknowledgments

We thank Kelly Henning for her assistance. This study was supported by Grant R01HD057116 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development to DJL. This work was performed when the first author was at Florida Atlantic University.

References

Bahrick LE. Infants' perception of substance and temporal synchrony in multimodal events. Infant Behavior & Development. 1983; 6:429–451.

Bahrick LE, Hernandez-Reif M, Flom R. The development of infant learning about specific face-voice relations. Developmental Psychology. 2005; 41:541–552. [PubMed: 15910161]

Bahrick LE, Moss L, Fadil C. Development of visual self-recognition in infancy. Ecological Psychology. 1996; 8:189–208.

Bahrick LE, Netto D, Hernandez-Reif M. Intermodal perception of adult and child faces and voices by infants. Child Development. 1998; 69:1263–1275. [PubMed: 9839414]

Bremner, AJ.; Lewkowicz, DJ.; Spence, C. Multisensory Development. Oxford: Oxford University Press; 2012.

Brookes H, Slater A, Quinn PC, Lewkowicz DJ, Hayes R, Brown E. Three-month-old infants learn arbitrary auditory-visual pairings between voices and faces. Infant & Child Development. 2001; 10:75–82.

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. PLoS Computational Biology. 2009:5. e1000436.

Dixon NF, Spitz LT. The detection of auditory visual desynchrony. Perception. 1980; 9:719–721. [PubMed: 7220244]

Dodd B. Lip reading in infants: Attention to speech presented in- and out-of-synchrony. Cognitive Psychology. 1979; 11:478–484. [PubMed: 487747]

Dodd B, Burnham D. Processing speechread information. Volta Review. 1988; 90:45–60.

Fernald A. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? Child Development. 1989; 60:1497–1510. [PubMed: 2612255]

Gibson, EJ. Principles of perceptual learning and development. New York: Appleton; 1969.

Grant KW, Seitz PF. Measures of auditory–visual integration in nonsense syllables and sentences. The Journal of the Acoustical Society of America. 1998; 104:2438–2450. [PubMed: 10491705]

Grant KW, van Wassenhove V, Poeppel D. Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. Speech Communication. Special Issue: Audio Visual Speech Processing. 2004; 44:43–53.

Hillock-Dunn A, Wallace MT. Developmental changes in the multisensory temporal binding window persist into adolescence. Developmental Science. 2012; 15:688–696. [PubMed: 22925516]

Hunnius S, Geuze RH. Developmental Changes in Visual Scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. Infancy. 2004; 6:231–255.

Jesse A, Johnson EK. Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. Journal of Experimental Psychology: Human Perception and Performance. 2012; 38:1567. [PubMed: 22545598]

Kubicek C, de Boisferon AH, Dupierrix E, Pascalis O, Lœvenbruck H, Gervain J, Schwarzer G. Cross-Modal Matching of Audio-Visual German and French Fluent Speech in Infancy. PLoS ONE. 2014; 9:e89275. [PubMed: 24586651]

Kuhl PK, Meltzoff AN. The bimodal perception of speech in infancy. Science. 1982; 218:1138–1141. [PubMed: 7146899]

Lewkowicz DJ. Developmental changes in infants' bisensory response to synchronous durations. Infant Behavior & Development. 1986; 9:335–353.

Lewkowicz DJ. Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. Infant Behavior & Development. 1992a; 15:297–324.

Lewkowicz DJ. Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. Perception & Psychophysics. 1992b; 52:519–528. [PubMed: 1437484]

Lewkowicz DJ. Infants' response to the audible and visible properties of the human face. I: Role of lexical-syntactic content, temporal synchrony, gender, and manner of speech. Developmental Psychology. 1996a; 32:347–366.

Lewkowicz DJ. Perception of auditory-visual temporal synchrony in human infants. Journal of Experimental Psychology: Human Perception & Performance. 1996b; 22:1094–1106. [PubMed: 8865617]

Lewkowicz DJ. Infants' perception of the audible, visible and bimodal attributes of multimodal syllables. Child Development. 2000; 71:1241–1257. [PubMed: 11108094]

Lewkowicz DJ. Infant perception of audio-visual speech synchrony. Developmental Psychology. 2010; 46:66–77. [PubMed: 20053007]

Lewkowicz DJ. Early experience and multisensory perceptual narrowing. Developmental Psychobiology. 2014; 56:292–315.10.1002/dev.21197 [PubMed: 24435505]

Lewkowicz DJ, Flom R. The audio-visual temporal binding window narrows in early childhood. Child Development. 201310.1111/cdev.12142

Lewkowicz DJ, Ghazanfar AA. The decline of cross-species intersensory perception in human infants. Proceedings of the National Academy Sciences USA. 2006; 103:6771–6774.

Lewkowicz DJ, Ghazanfar AA. The emergence of multisensory systems through perceptual narrowing. Trends in Cognitive Sciences. 2009; 13:470–478. [PubMed: 19748305]

Lewkowicz DJ, Hansen-Tift AM. Infants deploy selective attention to the mouth of a talking face when learning speech. Proceedings of the National Academy of Sciences. 2012; 109:1431–1436.

Lewkowicz DJ, Leo I, Simion F. Intersensory perception at birth: Newborns match non-human primate faces & voices. Infancy. 2010; 15:46–60.

Lewkowicz DJ, Pons F. Recognition of amodal language identity emerges in infancy. International Journal of Behavioral Development. 2013; 37(2):90–94. [PubMed: 24648601]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:229–239.

Munhall, KG.; Vatikiotis-Bateson, E. Spatial and Temporal Constraints on Audiovisual Speech Perception. In: Calvert, GA.; Spence, C.; Stein, BE., editors. The handbook of multisensory processes. Cambridge, MA: MIT Press; 2004. p. 177-188.

Patterson ML, Werker JF. Matching phonetic information in lips and voice is robust in 4.5-month-old infants. Infant Behavior & Development. 1999; 22:237–247.

Patterson ML, Werker JF. Infants' ability to match dynamic phonetic and gender information in the face and voice. Journal of Experimental Child Psychology. 2002; 81:93–115. [PubMed: 11741376]

Patterson ML, Werker JF. Two-month-old infants match phonetic information in lips and voice. Developmental Science. 2003; 6:191–196.

Piaget, J. The origins of intelligence in children. New York: International Universities Press; 1952.

Pons F, Lewkowicz DJ. Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. Acta Psychologica. 2014; 149:142–147. [PubMed: 24576508]

Pons F, Lewkowicz DJ, Soto-Faraco S, Sebastián-Gallés N. Narrowing of intersensory speech perception in infancy. Proceedings of the National Academy of Sciences USA. 2009; 106:10598–10602.

Rosenblum LD. Speech perception as a multimodal phenomenon. Current Directions in Psychological Science. 2008; 17:405. [PubMed: 23914077]

Saffran, JR.; Werker, JF.; Werner, LA. The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language. In: Kuhn, D.; Siegler, RS.; Damon, W.; Lerner, RM., editors. Handbook of child psychology: Vol 2, Cognition, perception, and language. 6. Hoboken, NJ, US: John Wiley & Sons Inc; 2006. p. 58-108.

Scott LS, Pascalis O, Nelson CA. A domain general theory of the development of perceptual discrimination. Current Directions in Psychological Science. 2007; 16:197–201. [PubMed: 21132090]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America. 1954; 26:212–215.

Summerfield AQ. Use of visual information in phonetic perception. Phonetica. 1979; 36:314–331. [PubMed: 523520]

Thelen, E.; Smith, LB. A dynamic systems approach to the development of cognition and action. Cambridge, MA: MIT Press; 1994.

Walker-Andrews AS. Intermodal perception of expressive behaviors: Relation of eye and voice? Developmental Psychology. 1986; 22:373–377.

Walker-Andrews AS, Bahrick LE, Raglioni SS, Diaz I. Infants' bimodal perception of gender. Ecological Psychology. 1991; 3:55–75.

Walton GE, Bower TG. Amodal representations of speech in infants. Infant Behavior & Development. 1993; 16:233–243.

Werker JF, Tees RC. Speech perception as a window for understanding plasticity and commitment in language systems of the brain. Developmental Psychobiology. Special Issue: Critical Periods Re-examined: Evidence from Human Sensory Development. 2005; 46:233–234.

Werker JF, Yeung HH, Yoshida KA. How Do Infants Become Experts at Native-Speech Perception? Current Directions in Psychological Science. 2012; 21:221–226.

Yehia H, Rubin P, Vatikiotis-Bateson E. Quantitative Association of Vocal-Tract and Facial Behavior. Speech Communication. 1998; 26:23–43.

- We investigated infant perception of fluent audiovisual speech coherence
- 12–14 month-old but neither 4- nor 8–10 month-old infants perceived audiovisual speech coherence
- Perception of native-speech coherence is easier than of non-native speech coherence
- Audio-visual synchrony is not necessary for perception of native-speech coherence but is necessary for the perception of non-native audiovisual speech coherence

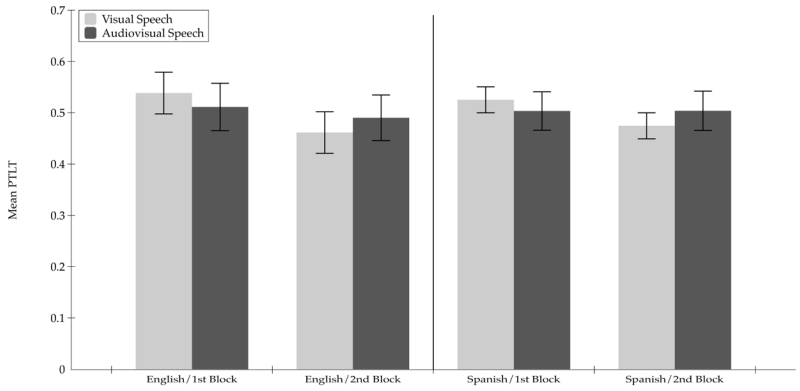

Figure 1.

Mean proportion of total looking time directed at the matching visible monologue during the silent and the audiovisual blocks of test trials in the 4-month-old infants in Experiment 1. The data are shown separately for each block of audiovisual test trials in each language condition. Error bars indicate the standard errors of the mean.

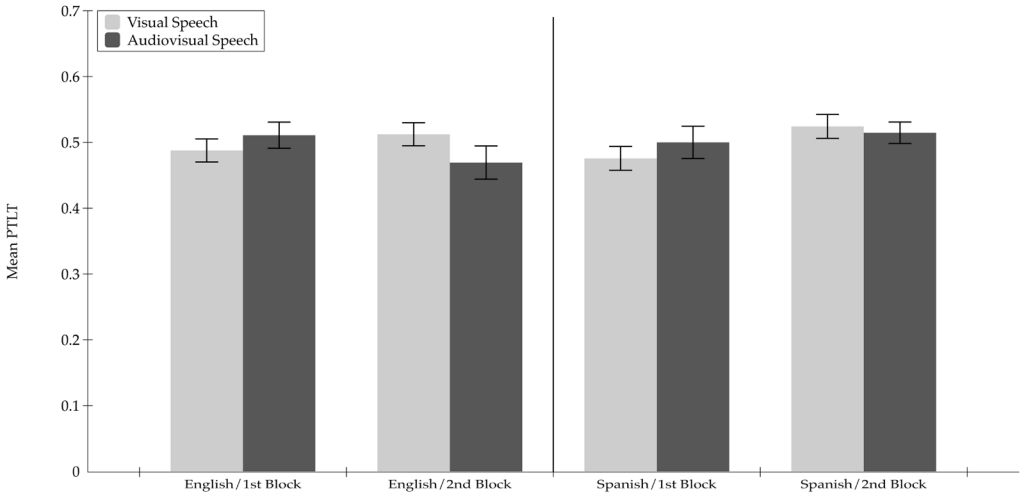

Figure 2.

Mean proportion of total looking time directed at the matching visible monologue during the silent and the audiovisual blocks of test trials in the 8–10 month-old infants in Experiment 2. The data are shown separately for each block of audiovisual test trials in each language condition. Error bars indicate the standard errors of the mean.

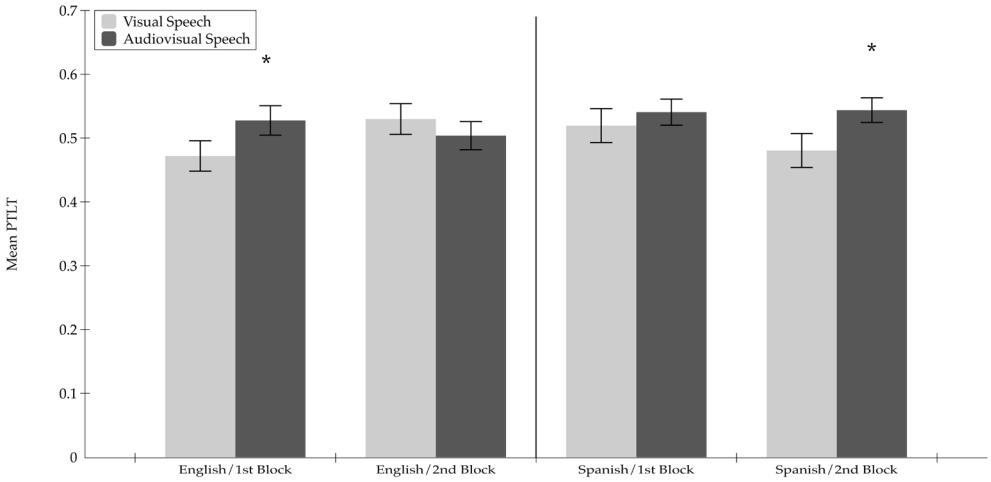

Figure 3.

Mean proportion of total looking time directed at the matching visible monologue during the silent and the audiovisual blocks of test trials in the 12–14 month-old infants in Experiment 3. The data are shown separately for each block of audiovisual test trials in each language condition. Error bars indicate the standard errors of the mean. Asterisks indicate statistically greater looking at the matching visible speech monologue during the audiovisual test trials than during the silent trials.

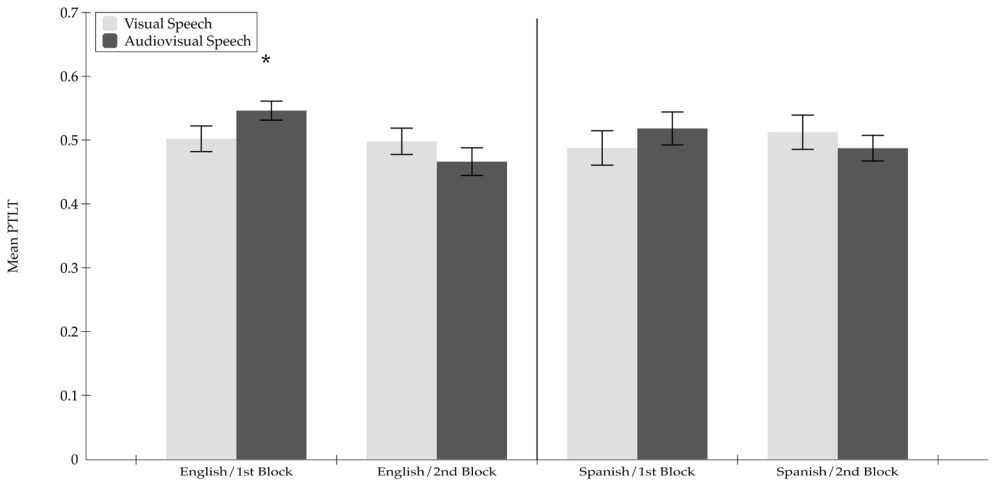

Figure 4.

Mean proportion of total looking time directed at the matching visible monologue during the silent and the audiovisual blocks of test trials in the 12–14 month-old infants when the audible and visible speech streams were desynchronized in Experiment 4. The data are shown separately for each block of audiovisual test trials in each language condition. Error bars indicate the standard errors of the mean. Asterisks indicate statistically greater looking at the matching visible speech monologue during the audiovisual test trials than during the silent trials.

Table 1

Design of Experiment 1 showing the two versions of the Quick Time movies constructed for each language. Also, shown are the side on which the two visible monologues were presented in each block of trials as well the order of audible monologue presentation during the audiovisual blocks of test trials.

	Movie Version 1		
	Left Visible Monologue	**Audible Monologue**	**Right Visible Monologue**
Silent-Speech Block			
Trial 1	Monologue 1		Monologue 2
Trial 2	Monologue 2		Monologue 1
Audiovisual-Speech - 1st Block			
Trial 3	Monologue 1	Monologue 1	Monologue 2
Trial 4	Monologue 2	Monologue 1	Monologue 1
Audiovisual-Speech - 2nd Block			
Trial 5	Monologue 1	Monologue 2	Monologue 2
Trial 6	Monologue 2	Monologue 2	Monologue 1

	Movie Version 2		
	Left Visible Monologue	**Audible Monologue**	**Right Visible Monologue**
Silent-Speech Block			
Trial 1	Monologue 2		Monologue 1
Trial 2	Monologue 1		Monologue 2
Audiovisual-Speech - 1st Block			
Trial 3	Monologue 2	Monologue 2	Monologue 1
Trial 4	Monologue 1	Monologue 2	Monologue 2
Audiovisual-Speech - 2nd Block			
Trial 5	Monologue 2	Monologue 1	Monologue 1
Trial 6	Monologue 1	Monologue 1	Monologue 2