Journal of Experimental Child Psychology 130 (2015) 147–162

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

# Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience

David J. Lewkowicz [a,*], Nicholas J. Minar [b], Amy H. Tift [b], Melissa Brandon [b]

[a] Department of Communication Sciences and Disorders, Northeastern University, Boston, MA 02115, USA
[b] Department of Psychology, Florida Atlantic University, Boca Raton, FL 33431, USA

ARTICLE INFO

*Article history:*
Received 15 May 2014
Revised 11 October 2014

*Keywords:*
Multisensory perception
Auditory
Visual
Speech
Infancy
Development

ABSTRACT

To investigate the developmental emergence of the perception of the multisensory coherence of native and non-native audiovisual fluent speech, we tested 4-, 8- to 10-, and 12- to 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 8- to 10-month-old infants exhibited audiovisual matching in that they did not look longer at the matching monologue. In contrast, the 12- to 14-month-old infants exhibited matching and, consistent with the emergence of perceptual expertise for the native language, perceived the multisensory coherence of native-language monologues earlier in the test trials than that of non-native language monologues. Moreover, the matching of native audible and visible speech streams observed in the 12- to 14-month-olds did not depend on audiovisual 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 audiovisual synchrony cues are more important in the perception of the multisensory coherence of non-native speech than that of native audiovisual speech, and that the emergence of this skill most likely is affected by perceptual narrowing.



* Corresponding author.
*E-mail address:* d.lewkowicz@neu.edu (D.J. Lewkowicz).

http://dx.doi.org/10.1016/j.jecp.2014.10.006
0022-0965/

## Introduction

Social interactions usually involve the use of audiovisual speech (Rosenblum, 2008). Such speech consists of temporally coupled and redundant 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 redundancy, 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? Does experience contribute to its emergence?

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 infants' native language. The matching findings are especially interesting because they suggest that infants can perceive the multisensory coherence of audiovisual speech. Unfortunately, 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 audiovisual matching. If so, this raises two questions. First, can infants perceive the multisensory coherence of audiovisual speech in the absence of cross-linguistic cues? Second, if they can, at what age 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 in 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 in infancy (Bremner, Lewkowicz, & Spence, 2012; Lewkowicz, 2014; Lewkowicz & Ghazanfar, 2009). This is illustrated by the fact that even though from birth onward infants can perceive the coherence of human auditory and visual speech (Dodd, 1979; Lewkowicz, 1996a, 2000, 2010), nonhuman communicative signals (Lewkowicz, Leo, & Simion, 2010), and nonspeech 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 audiovisual synchrony (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 even in the absence of synchrony cues. 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 in the absence of synchrony cues is interesting, studies show that A–V synchrony cues play a role in infant processing of fluent audiovisual speech and that such cues continue to play a role in multisensory perception into adulthood. For instance, findings show that infants attend more to synchronized than to desynchronized audiovisual fluent speech (Dodd, 1979) and that they can learn specific face–voice associations only when talking faces and accompanying voices are temporally synchronized (Bahrick et al., 2005). Moreover, findings indicate that both children and adults can detect the temporal alignment of auditory and visual information, that this ability improves with development, and that detection of audiovisual 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, 2014). 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 might 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 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 corresponds with the onset of speech production (i.e., canonical babbling) and, as such, enables infants to gain direct access to the source of synchronous audiovisual speech. This way, infants can profit maximally from the greater perceptual salience of the multisensory redundancy of the signal. The second attentional shift away from a talker's mouth observed in 12-month-olds in response to native audiovisual speech corresponds with the emergence of an initial expertise for the native language. This shift suggests that by this age infants might no longer need to rely on audiovisual redundancy when processing speech that is familiar. 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 8 and 10 months of age but that they continued to do so at 12 months.

Lewkowicz and Hansen-Tift (2012) interpreted the continued attentional focus on the mouth at 12 months of age 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 to the synchronous and, thus, highly salient cues available in a talker's mouth presumably enables infants to 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, Lewkowicz and Hansen-Tift's (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 audiovisual 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 audiovisual matching studies to date. In the three studies where infants were tested with speech in different languages and where it was 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 study (Lewkowicz & Pons, 2013). Moreover, Dodd and Burnham (1988) found that 20-week-old infants can match the faces and voices of their native speech (English) but not those of 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 and Pons (2013) found that 10- to 12-month-old English-learning infants, but not 6- to 8-month-old infants, could perceive the multisensory identity of a native language, 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 obtained audiovisual matching only 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, 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 tested others with monologues spoken in a non-native language (Spanish). In Experiments 1 to 3, we tested 4-, 8- to 10-, and 12- to 14-month-old infants' responsiveness to synchronous audible and visible native and non-native fluent speech. In Experiment 4, we tested 12- to 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 at some point in their development infants will 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 their development infants should look longer at the face whose visible articulations correspond to audible articulations. Empirical evidence indicates that once infants begin to perceive multisensory coherence they 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 (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 in 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 in the course of an experiment, they cease to exhibit evidence of multisensory matching (Bahrick, Moss, & Fadil, 1996; Bahrick et al., 2005; 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 language as opposed the non-native language. That is, infants may cease performing audiovisual matching when exposed to native audiovisual speech but might 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 audiovisual matching at this age. Nonetheless, testing infants as young as 4 months is essential to define more precisely the age when this ability begins to emerge.

## *Method*

### *Participants*

The sample consisted of 48 4-month-old infants (17 girls; $M_{age}$ = 17.1 weeks, range = 16.0–18.7). 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 day of the week, and (c) the source of the speech heard by the infant per day of the week (i.e., mother, father, grandparents, relatives, caregiver, or other). Based on the results of this questionnaire, we calculated the percentage of exposure to each language per week and included only infants whose language exposure to English exceeded 81%. An additional 15 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, and 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 their lap or if the infant refused to sit in an infant seat, parents were permitted to hold the infant on their lap. When they did, parents 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. 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 midway between the two monitors and was used to record infants' visual fixations on the talking faces on the two monitors. The experimenter was seated outside the booth and could see the infants and parents through a one-way mirror as well as via the video camera focused on infants' faces.

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. In each preference trial, infants saw side-by-side faces of the same female actor speaking two different monologues,[1] with the side on which the two different monologues were presented switched in the second trial of each block.

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

[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".

**Table 1**
Design of Experiment 1 showing the two versions of the QuickTime movies constructed for each language.

| | 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-first block* | | | |
| Trial 3 | Monologue 1 | Monologue 1 | Monologue 2 |
| Trial 4 | Monologue 2 | Monologue 1 | Monologue 1 |
| *Audiovisual speech-second 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-first block* | | | |
| Trial 3 | Monologue 2 | Monologue 2 | Monologue 1 |
| Trial 4 | Monologue 1 | Monologue 2 | Monologue 2 |
| *Audiovisual speech-second block* | | | |
| Trial 5 | Monologue 2 | Monologue 1 | Monologue 1 |
| Trial 6 | Monologue 1 | Monologue 1 | Monologue 2 |

*Note.* 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 block of test trials.

were the audiovisual test trials. Here, infants saw the same talking faces again and also heard one of the corresponding audible monologues in one block and heard the other audible monologue in 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, 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 of 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 in which 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 or over infants' behavior. To center infants' eye gaze in between the test trials, a rotating multicolored disk was presented in the middle of infants' visual field (the disk 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 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. If infants perceived the match between the visible and audible monologues, they were expected to exhibit increased looking at the corresponding talking face in the presence of the audible monologue rather 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. 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.

Fig. 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 consisted of trial type (2: silent or audiovisual speech) and block (2: first or second audiovisual block of test trials) as the within-participants factors and language (2) as a between-participants factor. There were no significant main effects of trial type, $F(1,46) = 0.008$, $p = .93$, $\eta_p^2 = 0$, of block, $F(1,46) = 1.24$, $p = .27$, $\eta_p^2 = .26$, or of language, $F(1,46) = 0.003$, $p = .94$, $\eta_p^2 = 0$. In addition, there were no significant interactions between trial type and block, $F(1,46) = 1.26$, $p = .27$, $\eta_p^2 = .027$, between trial type and language, $F(1,46) = 0.003$, $p = .95$, $\eta_p^2 = 0$, or among trial type, block, and language, $F(1,46) = 0.002$, $p = .96$, $\eta_p^2 = 0$.

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 prediction was that responsiveness in the audiovisual test trials was likely to change rapidly as infants

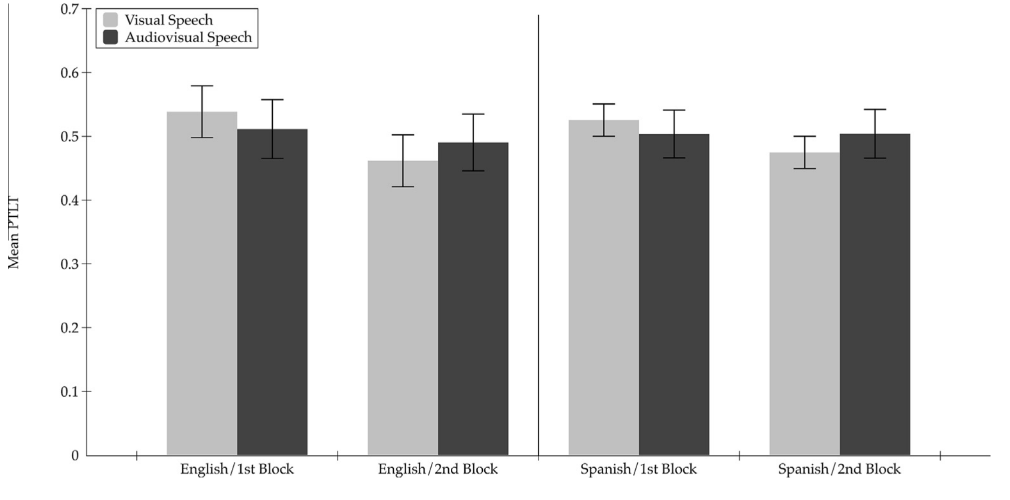


**Fig. 1.** Mean proportions of total looking time directed at the matching visible monologue during the silent and audiovisual blocks of test trials in the 4-month-old infants in Experiment 1. The data are shown separately for each block of audiovisual test trials in each language condition. Error bars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

gain increasing experience with the same visual stimuli as the experiment progresses (Bahrick et al., 1996, 1998, 2005; Walker-Andrews, 1986). The third was that changes in responsiveness across the test trials may differ for the native and non-native languages. To test the a priori predictions, we used one-tailed *t* tests to compare the test PTLT and baseline PTLT directed at the matching visible monologue in each block of audiovisual test trials 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 $d = 0.22$, or in the second block of audiovisual test trials, $t(23) = 0.59$, $p = .28$, Cohen's $d = 0.25$. 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 $d = 0.19$, or in the second block of audiovisual test trials, $t(23) = 0.64$, $p = .26$, Cohen's $d = 0.27$. 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- to 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- to 10-month age range. Alternatively, it may be that infants require additional experience with this aspect of fluent speech and, because of this, might not be able to perceive the multisensory coherence of audiovisual speech in this specific age range.

### *Method*

#### *Participants*

The sample consisted of 55 8- to 10-month-old infants (30 girls; $M_{age} = 37.92$ weeks, range = 33.29–44.29). All infants came from monolingual homes (81% or more of the language exposure was in English). An additional 5 infants were tested but were not included in the data analysis due to fussiness ($n = 4$) or equipment failure ($n = 1$).

#### *Apparatus and stimuli*

The apparatus and stimuli used in this experiment were identical to those used in Experiment 1.

#### *Procedure*

The procedure used in this experiment was identical to that used in Experiment 1.

### *Results and discussion*

Fig. 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 block (2) as within-participants factors and language (2) as a between-participants factor. There were no significant main effects of trial type, $F(1,53) = 0.06$, $p = .81$, $\eta_p^2 = .001$, of block, $F(1,53) = 0.12$, $p = .73$, $\eta_p^2 = .002$, or of language, $F(1,53) = 0.087$, $p = .77$, $\eta_p^2 = .002$. There was a marginally significant interaction between trial type and block, $F(1,53) = 3.22$, $p = .07$, $\eta_p^2 = .058$, but there were no significant interactions between trial type and language, $F(1,53) = 0.087$, $p = .77$, $\eta_p^2 = .002$, or among trial type, block, and language, $F(1,53) = 0.006$, $p = .94$, $\eta_p^2 = 0$.

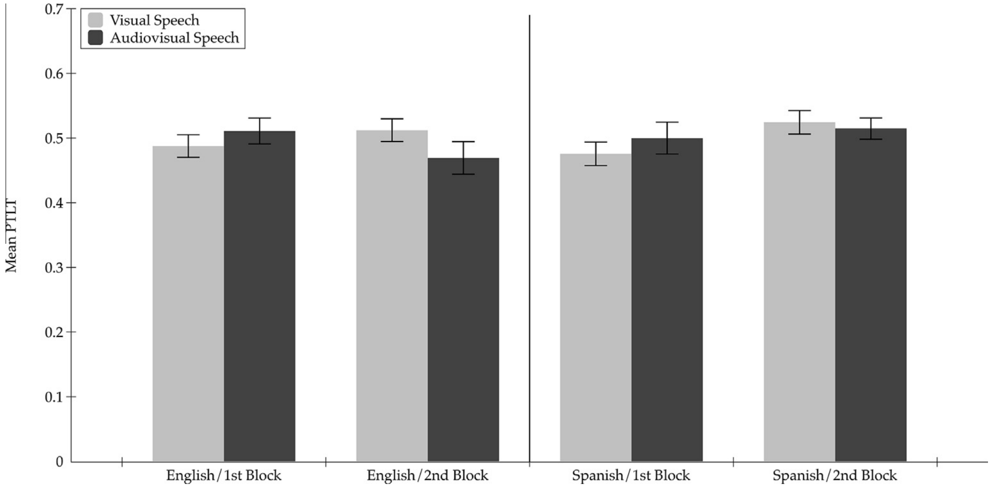


**Fig. 2.** Mean proportions of total looking time directed at the matching visible monologue during the silent and audiovisual blocks of test trials in the 8- to 10-month-old infants in Experiment 2. The data are shown separately for each block of audiovisual test trials in each language condition.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

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 $d = 0.26$, 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 $d = 0.34$. 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 $d = 0.26$, or in the second block of audiovisual test trials, $t(26) = 0.99$, $p = .16$, Cohen's $d = 0.28$. Overall, these findings show that, like the 4-month-olds, the 8- to 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 4-month-old nor 8- to 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 and immaturity and/or the greater complexity of fluent audiovisual speech syllables as opposed to isolated speech syllables. As a result, in Experiment 3 we tested 12- to 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 non-native speech.

### *Method*

#### *Participants*

That sample consisted of 48 12- to 14-month-old infants (24 girls; $M_{age} = 56.0$ weeks, range = 51.0–61.1). All infants came from monolingual homes (81% or more of the language exposure was in English). An additional 5 infants were tested but were not included in the data analysis due to inattentiveness/parent interaction ($n = 4$) or ear infection ($n = 1$).

*Apparatus and stimuli*

The apparatus and stimuli used in this experiment were identical to those used in Experiment 1.

*Procedure*

The procedure used in this experiment was identical to that used in Experiment 1.

*Results and discussion*

Fig. 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 block (2) as within-participants factors and language (2) as a between-participants factor. This analysis yielded a significant trial type effect, $F(1,46) = 8.13$, $p = .007$, $\eta_p^2 = .15$, but no block effect, $F(1,46) = 0.00$, $p = .98$, $\eta_p^2 = .00$, or language effect, $F(1,46) = 1.65$, $p = .21$, $\eta_p^2 = .035$. In addition, there were no interactions between trial type and block, $F(1,46) = 0.22$, $p = .64$, $\eta_p^2 = .005$, between trial type and language, $F(1,46) = 1.88$, $p = .18$, $\eta_p^2 = .039$, or among trial type, block, and language, $F(1,46) = 2.17$, $p = .15$, $\eta_p^2 = .045$. 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 $d = 0.72$, 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 $d = 0.35$. 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., 1996, 1998, 2005; 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 $d = 0.25$, but that they did in the second block of audiovisual test trials, $t(23) = 1.92$, $p = .03$, one-tailed, Cohen's

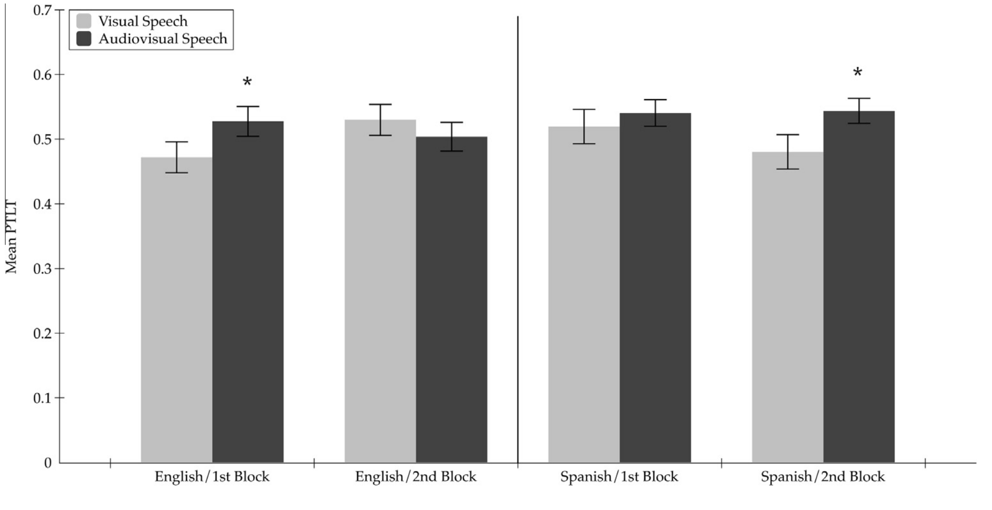


**Fig. 3.** Mean proportions of total looking time directed at the matching visible monologue during the silent and audiovisual blocks of test trials in the 12- to 14-month-old infants in Experiment 3. The data are shown separately for each block of audiovisual test trials in each language condition.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

$d = 0.80$. 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, whereas 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-month-old nor 8- to 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- to 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 during early development, the current 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 tested infants with isolated speech syllables (Lewkowicz, 2000, 2010; Lewkowicz et al., 2010). The two exceptions are Dodd's (1979) study, which showed that 10- to 26-week-old infants actually attend more to synchronous fluent audiovisual speech 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 audiovisual speech 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 during 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, 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 6 months 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- to 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*

The sample consisted of 52 12- to 14-month-old infants (23 girls; $M_{age} = 51.4$ weeks, range = 51.0–61.1). All infants came from monolingual homes (81% or more of the language exposure was in English). An additional 9 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 used in Experiment 3. 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 isolated syllables (Lewkowicz, 2010) and with 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 that used 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*

Fig. 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 block (2) as within-participants factors and language (2) as a between-participants factor. There were no significant main effects of trial type, $F(1,50) = 0.20$, $p = .65$, $\eta_p^2 = .073$, of block, $F(1,50) = 1.17$, $p = .28$, $\eta_p^2 = .186$, or of language, $F(1,50) = 0.03$, $p = .86$, $\eta_p^2 = .053$. In addition, there were no significant interactions between trial type and block, $F(1,50) = 3.00$, $p = .09$, $\eta_p^2 = .396$, between trial type and language, $F(1,50) = 0.29$, $p = .86$, $\eta_p^2 = .053$, or among trial type, block, and language, $F(1,50) = 0.07$, $p = .79$, $\eta_p^2 = .058$.

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 $d = 0.75$, but that they

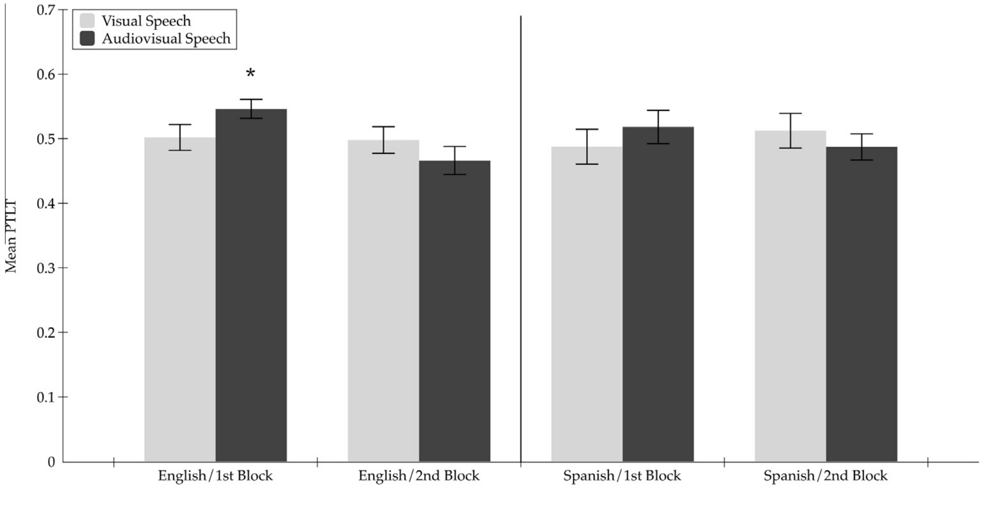


**Fig. 4.** Mean proportions of total looking time directed at the matching visible monologue during the silent and audiovisual blocks of test trials in the 12- to 14-month-old infants when the audible and visible speech streams were desynchronized in Experiment 4. The data are shown separately for each 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 in the audiovisual test trials than in the silent trials.

did not in the second block of these trials, $t(27) = 1.05$, $p = .15$, one-tailed, Cohen's $d = 0.40$. 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, $t(23) = 0.79$, $p = .22$, one-tailed, Cohen's $d = 0.33$, or the second block of the audiovisual test trials, $t(23) = -0.86$, $p = .20$, one-tailed, Cohen's $d = -0.36$.

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- to 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 in the first block of audiovisual test trials, their looking at the sound-specified visible monologue did not differ across the two experiments, $t(50) = 0.30$, $p = .77$, two-tailed, Cohen's $d = 0.085$. In contrast, when the infants were tested with Spanish in the second block of the audiovisual test trials, their looking at the sound-specified visible monologue was significantly lower in Experiment 4 than in Experiment 3, $t(46) = 2.01$, $p = .05$, two-tailed, Cohen's $d = 0.59$.

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 A–V temporal synchrony plays a role in this ability. Experiments 1 and 2 showed that 4- and 8- to 10-month-old infants did not exhibit evidence of multisensory matching. In contrast, Experiment 3 showed that 12- to 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- to 14-month-old infants did not depend on the audible and visible speech streams being synchronized when native audiovisual speech was presented but that it 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 critical only for the perception of the multisensory coherence of non-native audiovisual speech.

The finding that the 4- and 8- to 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 A–V synchrony relations (Lewkowicz, 2010). It should be noted, however, that the previous findings come from studies in which infants only needed 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 fluent audiovisual speech 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 4- nor 8- to 10-month-old infants detected such statistics because they exhibited no evidence of multisensory matching. Experiments 3 and 4 did show, however, that 12- to 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- to 14-month-old infants exhibited multisensory matching even though A–V synchrony was disrupted in the native-language condition indicates that infants 1 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) 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 requires only the detection of the onsets and offsets of such inputs. Given this assumption, Lewkowicz and 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 1 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 to 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 and Human Development to D.J.L. This work was performed when the first author was at Florida Atlantic University.

## References


Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior and Development, 6*, 429–451.

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face–voice relations. *Developmental Psychology, 41*, 541–552.

Bahrick, L. E., Moss, L., & Fadil, C. (1996). Development of visual self-recognition in infancy. *Ecological Psychology, 8*, 189–208.

Bahrick, L. E., Netto, D., & Hernandez-Reif, M. (1998). Intermodal perception of adult and child faces and voices by infants. *Child Development, 69*, 1263–1275.

Bremner, A. J., Lewkowicz, D. J., & Spence, C. (2012). *Multisensory development*. Oxford, UK: Oxford University Press.

Brookes, H., Slater, A., Quinn, P. C., Lewkowicz, D. J., Hayes, R., & Brown, E. (2001). Three-month-old infants learn arbitrary auditory–visual pairings between voices and faces. *Infant and Child Development, 10*, 75–82.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436.

Dixon, N. F., & Spitz, L. T. (1980). The detection of auditory visual desynchrony. *Perception, 9*, 719–721.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology, 11*, 478–484.

Dodd, B., & Burnham, D. (1988). Processing speechread information. *Volta Review, 90*, 45–60.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development, 60*, 1497–1510.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*, 2438–2450.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication, 44*, 43–53.

Hillock-Dunn, A., & Wallace, M. T. (2012). Developmental changes in the multisensory temporal binding window persist into adolescence. *Developmental Science, 15*, 688–696.

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

Jesse, A., & Johnson, E. K. (2012). Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 1567–1581.

Kubicek, C., de Boisferon, A. H., Dupierrix, E., Pascalis, O., Lœvenbruck, H., Gervain, J., et al (2014). Cross-modal matching of audio–visual German and French fluent speech in infancy. *PLoS ONE, 9*, e89275.

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

Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development, 9*, 335–353.

Lewkowicz, D. J. (1992a). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior and Development, 15*, 297–324.

Lewkowicz, D. J. (1992b). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Perception & Psychophysics, 52*, 519–528.

Lewkowicz, D. J. (1996a). 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, 32*, 347–366.

Lewkowicz, D. J. (1996b). Perception of auditory–visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1094–1106.

Lewkowicz, D. J. (2000). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development, 71*, 1241–1257.

Lewkowicz, D. J. (2010). Infant perception of audio–visual speech synchrony. *Developmental Psychology, 46*, 66–77.

Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental Psychobiology, 56*, 292–315.

Lewkowicz, D. J., & Flom, R. (2014). The audio–visual temporal binding window narrows in early childhood. *Child Development, 85*, 685–694.

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 6771–6774.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences, 13*, 470–478.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 1431–1436.

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match non-human primate faces and voices. *Infancy, 15*, 46–60.

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development, 37*(2), 90–94.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 229–239.

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 177–188). Cambridge, MA: MIT Press.

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

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*, 93–115.

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

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

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audio–visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica, 149*, 142–147.
Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 10598–10602.
Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*, 405–409.
Saffran, J. R., Werker, J. F., & Werner, L. A. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In D. Kuhn, R. S. Siegler, W. Damon, & R. M. Lerner (Eds.). *Handbook of child psychology: Cognition, perception, and language* (6th ed.) (Vol. 2, pp. 58–108). Hoboken, NJ: John Wiley.
Scott, L. S., Pascalis, O., & Nelson, C. A. (2007). A domain general theory of the development of perceptual discrimination. *Current Directions in Psychological Science, 16*, 197–201.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.
Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36*, 314–331.
Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.
Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology, 22*, 373–377.
Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., & Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology, 3*, 55–75.
Walton, G. E., & Bower, T. G. (1993). Amodal representations of speech in infants. *Infant Behavior and Development, 16*, 233–243.
Werker, J. F., & Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental Psychobiology, 46*, 233–234.
Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science, 21*, 221–226.
Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*, 23–43.