frontiers in Psychology

MINI REVIEW
published: 15 December 2015
doi: 10.3389/fpsyg.2015.01844

# Sources of Confusion in Infant Audiovisual Speech Perception Research

*Kathleen E. Shaw*[1] *and Heather Bortfeld*[2,3]*

[1] Department of Psychology, University of Connecticut, Storrs, CT, USA, [2] Psychological Sciences, University of California, Merced, Merced, CA, USA, [3] Haskins Laboratories, New Haven, CT, USA

Speech is a multimodal stimulus, with information provided in both the auditory and visual modalities. The resulting audiovisual signal provides relatively stable, tightly correlated cues that support speech perception and processing in a range of contexts. Despite the clear relationship between spoken language and the moving mouth that produces it, there remains considerable disagreement over how sensitive early language learners—infants—are to whether and how sight and sound co-occur. Here we examine sources of this disagreement, with a focus on how comparisons of data obtained using different paradigms and different stimuli may serve to exacerbate misunderstanding.

**Keywords: audiovisual perception, multimodal integration, infant perception, temporal binding window, sine wave speech, speech perception, speech disorders**

OPEN ACCESS

***Edited by:***
*Andriy Myachykov, Northumbria University, UK*

***Reviewed by:***
*Jean Vroomen, University of Tilburg, Netherlands*
*Clemens Wöllner, University of Hamburg, Germany*

***Correspondence:**
*Heather Bortfeld*
*hbortfeld@ucmerced.edu*

***Specialty section:***
*This article was submitted to Cognition, a section of the journal Frontiers in Psychology*

***Received:*** *15 February 2015*
***Accepted:*** *13 November 2015*
***Published:*** *15 December 2015*

***Citation:***
*Shaw KE and Bortfeld H (2015) Sources of Confusion in Infant Audiovisual Speech Perception Research. Front. Psychol. 6:1844. doi: 10.3389/fpsyg.2015.01844*

## INTRODUCTION

Although the development of early speech perception abilities is often framed as an auditory-only process, speech is a sensory-rich stimulus, with information provided across multiple modalities. Our focus here is on the auditory (i.e., spoken language) and visual (i.e., moving mouth) modalities, which together provide relatively stable, tightly correlated cues about the resulting speech. If we focus only on the articulators, both their visual form and the corresponding auditory stream they produce share onsets and offsets, intensity changes, amplitude contours, durational cues, and rhythmic patterning (Chandrasekaran et al., 2009). This reliable co-occurrence of cues serves to support speech comprehension (Sumby and Pollack, 1954), particularly in noisy environments (Massaro, 1984; Middelweerd and Plomp, 1987) and during language learning, whether first (Teinonen et al., 2008) or subsequent (Navarra and Soto-Faraco, 2007). Yet despite the clear relationship between spoken language and the moving mouth that produces it, there remains considerable disagreement about how sensitive early language learners—particularly infants—are to whether and how sight and sound co-occur. Here we examine the bases for this disagreement, with a particular focus on how data obtained using different methodologies and different stimuli may actually serve to exacerbate it.

One issue to consider is whether infants have initial biases toward attending to one or the other modality in the first place. On the one hand, infants have considerable prenatal experience with sound (DeCasper and Spence, 1986). Although the tissue and liquid barriers of the womb filter out frequencies greater than 5000 Hz, external acoustic stimuli are heard *in utero* beginning early in gestation (Jardri et al., 2008). Indeed, both behavioral data (Hepper and Shahidullah, 1994) and physiological data (Rubel and Ryals, 1983; Pujol et al., 1991) demonstrate that the fetal auditory system begins to process sounds between about 16 and 20 weeks. From that time forward, the cochlea matures anatomically during gestation such that its frequency response broadens

(Graven and Browne, 2008). Likewise, fetal abilities to discriminate among simultaneous frequencies, to separate rapid sequences of sounds (as in ordinary speech), and to perceive very quiet sounds all improve during the remaining gestational period (for reviews of empirical work see Busnel and Granier-Deferre, 1983; Lecanuet, 1996). As infants near term, their sensitivity to more complex auditory stimuli improves, allowing them to perceive details such as variations in music (Kisilevsky et al., 2004) and contrasting prosodic cues in familiar and novel rhymes (DeCasper et al., 1994). From this, one might conclude that development of auditory perceptual abilities has an initial advantage over vision, at least chronologically. On the other hand, and despite processing of visual stimuli beginning only postnatally (Turkewitz and Kenny, 1982; Slater, 2002), newborns' preference for faces (or face-like patterns) relative to any other visual stimulus is well documented (Goren et al., 1975; Morton and Johnson, 1991). This combination of early exposure in the auditory domain and precocious preference for faces—the source of spoken language—in the visual one would seem to position the newborn to easily recognize the relationship between spoken language and visual speech.

Not surprisingly, a talking face is more salient to a newborn than is a still face (Nagy, 2008), due at least in part to its inherent multimodality (Watson et al., 2014). But even when presented with a talking face with no accompanying sound (i.e., to visual speech alone), by the second half of the first year infants show greater sensitivity to the patterns of mouth movements found in their native language than in an unfamiliar language (Weikum et al., 2007). This suggests that they already recognize how specific movements of the visual articulators shape the speech signal, and a strong case has been made that the perception of the visual component of audiovisual speech facilitates the development of speech production abilities (Tenenbaum et al., 2015). Indeed, babbling infants tend to focus on the mouth of a speaker more than pre-babbling infants (Tenenbaum et al., 2013). Infants' own vocal productions interact with this as well, such that their real time attention to audiovisual speech changes as a function of their own articulatory modulations (Yeung and Werker, 2013); when presented with audiovisually produced vowels, infants imitate presentations more often when the audio and visual tokens are congruent than when they are incongruent (Legerstee, 1990). These and other findings inevitably lead to questions about what role, if any, the motor system plays in speech processing (e.g., Liberman and Mattingly, 1985). However, where perception of audiovisual speech clearly engages regions of sensorimotor cortex in both children and adults (Dick et al., 2010), other data indicate that motor activation is not necessary for audiovisual speech integration (Matchin et al., 2014). Therefore, we will set that debate aside to focus on the issue of integration itself.

Although a growing body of evidence demonstrates that substantial fine-tuning for various forms of audiovisual processing continues throughout childhood and well into adolescence (Baart et al., 2015; Tomalski, 2015), suffice it to say that at least some primitive form of multimodal perception emerges in early infancy (Bahrick et al., 2004). This can be characterized as guided by both *modal* cues (i.e., those that are specific to a single modality, such as color information in the visual domain or the timbre of someone's voice in the auditory domain) and *amodal* ones (i.e., those that are available across modalities and are thus redundant; Bahrick, 1988). These amodal cues provide perceptual evidence that distinct sensory events can share a point of origin. By gaining experience with the correlated cues in audiovisual speech (or their intersensory redundancy, Lickliter and Bahrick, 2000), infants should come to identify information shared between them.

## ASSOCATION IS NOT INTEGRATION

What remains unclear is when in the course of development association of these cues becomes actual integration of them. This is because, generally speaking, research techniques that are compatible with testing infants do not allow researchers to distinguish between these two processes. While this may seem like a subtle distinction, it is not a trivial one, in that it differentiates between those neural systems that evaluate cross-modal coincidence of physical stimuli (association) and those that actually mediate perceptual binding (integration; Miller and D'Esposito, 2005). Substantial animal research indicates that cumulative perceptual experience is critical to the development of the neural foundation for integration (Wallace and Stein, 2007; Yu et al., 2010), where presumably the cortical regions that contribute to such perceptual coding are fed by those regions engaged in initial associations between stimuli. It follows, then, that infants' perception of the relationship between the auditory and visual signals, as measured by looking procedures, contributes to the development of those neural underpinnings that will eventually support adult-like audiovisual integration. But implicit in that is the view that association precedes integration. The primary challenge to our understanding of the time course of this developmental process is that we have limited research methodologies for probing infants' perceptual experiences in a way that differentiates between behavioral evidence of association (e.g., looking behavior) and integration (e.g., some measure of perceptual fusion; c.f., Rosenblum et al., 1997). Although advances in infant-friendly neurophysiological testing techniques are allowing researchers new ways of tackling this issue (e.g., Kushnerenko et al., 2013), there remain many constraints on what can be reasonably asked of (and therefore concluded about) infant perception, whether with behavioral or neurophysiological techniques.

Nonetheless, infants clearly demonstrate sensitivity to audiovisual relations (see Shaw et al., 2015, for an example of how familiarity and coherence differentially influence infants' perception of audiovisual speech). Interest in the topic stemmed initially from a now classic study, in which 4-month-olds matched auditory vowels to videos of their corresponding articulation (Kuhl and Meltzoff, 1982). Follow-up studies replicated that original finding and extended it to male speakers (Patterson and Werker, 1999), as well as to infants of younger ages (Patterson and Werker, 2003). However, when the structured spectral elements of speech were replaced with simple tones, 5-month-olds struggled to recognize the

appropriate cross-modal match (Kuhl and Meltzoff, 1984; Kuhl et al., 1991). Because of this, much of the theoretical discussion of these early findings focused on whether and to what degree infants show privileged processing of speech and whether that indicates they have early access to phonetic representations. In the process, infants' ability to simply *match* auditory and visual streams was often mischaracterized as their ability to *integrate* audiovisual speech, leading to the loss of this important distinction. This formed the basis for much of the subsequent disagreement about early perceptual integration abilities. In more recent years, although this source of confusion has been recognized (see Stein et al., 2010, for a review), the broadly held view that infants integrate (rather than associate) has prevented the establishment of a more mechanistic account of how, for example, early association happens, and how it relates to the development of integration at a neural level.

## NON-COMPARABLE STIMULI

Another source of confusion stems from generalizations made based on findings obtained using stimuli that vary in complexity. For example, much of the early infant research employed the simplest form of audiovisual speech possible: single vowels or consonant-vowel combinations (e.g., Kuhl and Meltzoff, 1984). And, although these stimuli were characterized as audiovisual speech, it is well understood that the cues that support comprehension are both spatial and temporal in nature. For example, one of the strongest available cues is timing (i.e., temporal correlations between duration, onsets, offsets, and rate of the auditory and visual streams; Parise et al., 2012), so the truncated speech stimuli used in many of the early studies inadvertently limited infants' access to that class of cues. In other words, the infant data demonstrate their sensitivity to how visual spatial cues relate to auditory spectral cues (and vice versa) but say nothing about their ability to map articulator motion to the unfolding temporal information in continuous speech. Infants *are* sensitive to timing relationships in a variety of simple non-speech, multimodal events (Lewkowicz, 1992, 1994, 2003), but their ability to deal with timing relationships between streams of continuous auditory and visual speech has only recently become the focus of systematic research (e.g., Baart et al., 2014; Kubicek et al., 2014; Lewkowicz et al., 2015; Shaw et al., 2015).

Beyond inconsistencies in stimulus complexity, there are other sources of variability in infant audiovisual research, such as which dimension (spectral or temporal) is manipulated to create the non-matching (i.e., control) stimuli. Although these are not entirely orthogonal sources of information, spectral integration generally relies more on stimulus congruence and temporal integration generally relies more on stimulus timing. Much of the behavioral research with infants has been conducted using some form of a multimodal preferential looking technique in which one of two side-by-side visual displays matches the auditory stream while the other does not. The non-matching stimulus might differ in congruence (i.e., a different stimulus, such as visual /e/ and visual /a/ presented side-by-side with auditory /e/) or in timing (i.e., the identical stimulus but offset in time relative to the audio). Congruence traditionally has been the more commonly manipulated dimension, as reflected by the matching/non-matching vowel stimuli used by Kuhl and colleagues in their early work. The McGurk effect (McGurk and MacDonald, 1976) also motivated a substantial line of research on perceptual fusion, typically with a single screen, and auditory and visual streams of single consonant-vowel pairs that are either congruent or non-congruent. In recent years, researchers have made substantial progress in using these sorts of stimuli in combination with electrophysiological measures with infants to identify neural indictors of perceptual fusion (e.g., Kushnerenko et al., 2008), but the former approach is far more commonly used.

Likewise, the synchrony of auditory and visual timing was manipulated early on (e.g., Dodd, 1979), revealing that older children (between 10 and 20 months of age) prefer synchronous over asynchronous running speech. More recently, questions have been raised about the extended developmental time course of such timing sensitivities and whether the temporal binding window continues to adjust further on in development. This refers to the period during which two sensory events can be separated in time yet still be perceptually bound into a unified event (see Wallace and Stevenson, 2014). Critically, testing this sensitivity requires temporally manipulating stimuli (i.e., comparing synchronous to non-synchronous audiovisual signals) rather than spatially manipulating them (i.e., comparing visual speech that matches the auditory speech to that which does not). If individuals have a temporal binding window that is too large, they may erroneously bind those events together (Van Wassenhove et al., 2007). In contrast, if the window is too narrow, individuals may be overly sensitive to whatever temporal discontinuity exists between two events and fail to recognize a cause-effect relationship between them (Dogge et al., 2012; Stevenson et al., 2012). Growing evidence of age-related differences in this form of temporal sensitivity is adding support to the view that data on infant association does not necessarily reflect integration of the sort that the temporal binding measures. For example, adolescents and pre-adolescents have larger temporal binding windows for audiovisual non-speech displays than older adolescents and adults (Hillock et al., 2011; Innes-Brown et al., 2011), and infants fail to indicate any sensitivity to temporal asynchrony unless the component signals are offset by over half a second (Lewkowicz, 2010; Pons et al., 2012).

While the research on timing sensitivities in typical development is still limited, there is even less data from atypical populations. Nevertheless, interest has grown recently in the role that temporal binding plays in a variety of developmental disorders such as autism (Bebko et al., 2006; Foss-Feig et al., 2010; de Boer-Schellekens et al., 2013) and dyslexia (Hairston et al., 2005), as well as with speech processing by cochlear implant users (Bergeson et al., 2005). Temporal-order-judgment tasks reveal that individuals with dyslexia, even when given non-linguistic audiovisual signals, tend to provide simultaneity judgments at longer lags than typical readers (Hairston et al., 2005). In this case, wider temporal binding windows may underlie reading deficits, reflecting poor temporal sensitivity to the auditory signal, visual signal, or both. By better understanding audiovisual

integration and the factors that lead to appropriate binding of events across senses, we will better understand the pathways leading to different developmental disorders and whether atypical perceptual integration may be at their base (Wallace and Stevenson, 2014).

## FURTHER ISOLATING SPECTRAL AND TEMPORAL INFLUENCES ON PROCESSING

While the correlation between the spectral and temporal information in the visual and auditory components of audiovisual speech makes it difficult to determine the influence of each, researchers have begun trying to isolate these components by degrading stimuli, for example, by using vocoded or sine wave speech (e.g., Tuomainen et al., 2005; Möttönen et al., 2006; Vroomen and Baart, 2009). Sine wave speech is natural speech that is synthetically reduced to three sinusoids replicating the frequency and amplitude of the first three formants (Remez et al., 1981). Unlike typical speech signals, sine wave speech is stripped of most extraneous spectral cues yet retains the temporal qualities of natural speech. Adults have difficulty recognizing the underlying phonetic content of sine wave speech unless they have been trained to hear it as language, or put into "speech-mode" (Vroomen and Baart, 2009). Because of this, sine wave speech is an ideal tool for examining the relative influence of top-down and bottom-up information on speech perception, and it is proving useful in isolating the relative influences of spectral and temporal information in infants' processing of audiovisual speech (e.g., Baart et al., 2014).

In typical experiments, participants are first exposed to sine wave speech without prior knowledge of its relationship to natural speech. After a training phase in which participants are put into speech mode, they are tested again to ascertain whether phonetic knowledge provides a top-down processing advantage in speech perception. Differences between naïve and informed sine wave speech perception demonstrate that the top-down forces (e.g., phonetic representations) underlie a variety of perceptual phenomena, including phonetic recalibration (Vroomen and Baart, 2009), McGurk responses (Vroomen and Stekelenburg, 2011), and enhanced neural responsiveness (Stekelenburg and Vroomen, 2012). So what happens when participants do not have access to the phonetic representation corresponding to the sine-wave signal, as is the case with young infants?

There are clues from an early series of studies in which infants' audiovisual perception was tested using stimuli that, though not sine wave speech, were quite similar to it. In an effort to assess which cues infants were relying on to cross-modally match audio and visual vowels in their initial study (Kuhl and Meltzoff, 1982), Kuhl and colleagues (Kuhl and Meltzoff, 1984; Kuhl et al., 1991) then asked whether modulating the spectral content of the acoustic signal impaired this ability. Four- to five-month-old infants were presented with audiovisual displays of a model silently articulating target vowels, but the auditory vowels were replaced by either pure tones, tones that matched the fundamental frequencies of the vowels, or three-tone vowel analogs somewhat akin to sine wave speech (i.e., tones were matched to the first three formants of the naturally spoken vowels). As before, when given the natural acoustic speech signal, infants matched the auditory vowels to the appropriate articulating face. However, across all three spectral manipulations, they failed to attend to the matching face relative to the mismatching face.

Although not interpreted by the authors as such, these results suggest that temporal correlations between the auditory and visual signals did not provide enough information for infants to match stimuli across the auditory and visual modalities. Instead, Kuhl and colleagues suggested that the phonetic identity of the component signals served as the basis for early audiovisual sensitivity and that infants needed the natural speech stimulus (with its full phonetic realization) to process these cross-modal relationships. Moreover, they argued that audiovisual speech perception is a holistic process whereby infants are relatively insensitive to low-level cues. Therefore, when the phonetic content of the stimulus is reduced, any top-down processing advantages for infants are eliminated. In other words, their argument was that spectral information above and beyond the first three formants must be available for infants to combine heard and seen speech.

Critically, however, this study suffers from both of the stimulus problems we have outlined (i.e., very short stimuli; congruency manipulation rather than timing manipulation). Given a single vowel, it is not surprising that infants were unable to use the degraded spectral information to match the auditory to the visual vowel because there was virtually no corresponding temporal information to support them in the process. In recent research (Baart et al., 2014), we have addressed this problem by giving infants longer stimuli. In this study, we presented infants and adults with trisyllabic non-words in natural speech or the sine wave tokens of that speech, together with two visual displays of the same woman articulating each of the two non-words. In both the natural speech and sine wave speech conditions, only one display matched the auditory signal. Adults performed significantly worse with sine wave speech than natural speech across trials, suggesting that they were unable to match the articulatory information in the degraded auditory signal to the corresponding visual speech. In contrast, infants performed identically for both sine wave speech and natural speech, apparently able to access whatever cues existed across both signals to appropriately match the audio to the visual display. It is important to note, however, that infants performed significantly worse than adults did with natural speech; after all, adults have full access to the detailed phonetic representations that being a native speaker of a language entails. Not surprisingly, they performed near ceiling in this simple matching task when the full spectral and temporal information is made available. Without it, however, they were not able to use the temporal cues any more than the infants. Critically, there was no difference in infants' performance in the natural speech and sine wave speech conditions, indicating that the temporal correlation between the auditory and visual signals was the basis for their performance rather than the spectral content of the speech itself. In other

words, infants' audiovisual association—at least in this case—was driven by relatively low level timing cues rather than by any form of phonetic representation. Importantly, this was only revealed by providing infants with the relevant temporal information in the form of sufficiently long stimuli, as well as by varying their access to the spectral information.

We are the first to admit that much remains unclear about how infants use spectral and temporal cues in audiovisual speech and how this contributes to their development of mature audiovisual integration. Nonetheless, we would argue that the factors we have identified here (i.e., lack of terminological precision, paradigmatic differences, variable stimulus length, and inconsistent manipulation of spectral and temporal dimensions of test stimuli) underlie much of the disagreement about infants' audiovisual perceptual abilities. Attention to such factors will improve the quality of the research and the clarity of the discussion.

## FUNDING

This work was supported by NIH R01 DC10075 and National Science Foundation IGERT Training Grant 114399.

## REFERENCES

Baart, M., Bortfeld, H., and Vroomen, J. (2015). Phonetic matching of auditory and visual speech develops during childhood: evidence from sine-wave speech. *J. Exp. Child Psychol.* 129, 157–164. doi: 10.1016/j.jecp.2014.08.002

Baart, M., Vroomen, J., Shaw, K., and Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition* 130, 31–43. doi: 10.1016/j.cognition.2013.09.006

Bahrick, L. E. (1988). Intermodal learning in infancy: learning on the basis of two kinds of invariant relations in audible and visible events. *Child Dev.* 59, 197–209. doi: 10.2307/1130402

Bahrick, L. E., Lickliter, R., and Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Curr. Dir. Psychol. Sci.* 13, 99–102. doi: 10.1111/j.0963-7214.2004.00283.x

Bebko, J. M., Weiss, J. A., Demark, J. L., and Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *J. Child Psychol. Psychiatry* 47, 88–98. doi: 10.1111/j.1469-7610.2005.01443.x

Bergeson, T. R., Pisoni, D. B., and Davis, R. A. O. (2005). Development of audiovisual comprehension skills in prelingually deaf children with cochlear implants. *Ear Hear.* 26, 149–164. doi: 10.1097/00003446-200504000-00004

Busnel, M. C., and Granier-Deferre, C. (1983). "And what of fetal audition?" in *The Behavior of Human Infants*, eds A. Oliveirio and M. Zappella (New York, NY: Plenum), 93–126.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

de Boer-Schellekens, L., Eussen, M., and Vroomen, J. (2013). Diminished sensitivity of audiovisual temporal order in autism spectrum disorder. *Front. Integr. Neurosci.* 7:8. doi: 10.3389/fnint.2013.00008

DeCasper, A. J., Lecanuet, J. P., Busnel, M. C., Granier-Deferre, C., and Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant Behav. Dev.* 17, 159–164. doi: 10.1016/0163-6383(94)90051-5

DeCasper, A. J., and Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behav. Dev.* 9, 133–150. doi: 10.1016/0163-6383(86)90025-1

Dick, A. S., Solodkin, A., and Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain Lang.* 114, 101–114. doi: 10.1016/j.bandl.2009.08.005

Dodd, B. (1979). Lip reading in infants: attention to speech presented in-and out-of-synchrony. *Cogn. Psychol.* 11, 478–484. doi: 10.1016/0010-0285(79)90021-5

Dogge, M., Schaap, M., Custers, R., Wegner, D. M., and Aarts, H. (2012). When moving without volition: implied self-causation enhances binding strength between involuntary actions and effects. *Conscious. Cogn.* 21, 501–506. doi: 10.1016/j.concog.2011.10.014

Foss-Feig, J. H., Kwakye, L. D., Cascio, C. J., Burnette, C. P., Kadivar, H., Stone, W. L., et al. (2010). An extended multisensory temporal binding window in autism spectrum disorders. *Exp. Brain Res.* 203, 381–389. doi: 10.1007/s00221-010-2240-4

Goren, C. C., Sarty, M., and Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics* 56, 544–549.

Graven, S. N., and Browne, J. V. (2008). Sensory development in the fetus, neonate, and infant: introduction and overview. *Newborn Infant Nurs. Rev.* 8, 169–172. doi: 10.1053/j.nainr.2008.10.007

Hairston, W. D., Burdette, J. H., Flowers, D. L., Wood, F. B., and Wallace, M. (2005). Altered temporal profile of visual-auditory multisensory interactions in dyslexia. *Exp. Brain Res.* 166, 474–480. doi: 10.1007/s00221-005-2387-6

Hepper, P. G., and Shahidullah, B. S. (1994). Development of fetal hearing. *Arch. Dis. Child.* 71, F81–F87. doi: 10.1136/fn.71.2.F81

Hillock, A. R., Powers, A. R., and Wallace, M. (2011). Binding of sights and sounds: age-related changes in multisensory temporal processing. *Neuropsychologia* 49, 461–467. doi: 10.1016/j.neuropsychologia.2010.11.041

Innes-Brown, H., Barutcha, A., Shivdasani, M. N., Crewther, D. P., Grayden, D. B., and Paolini, A. G. (2011). Susceptibility to the flash-beep illusion is increased in children compared to adults. *Dev. Sci.* 14, 1089–1099. doi: 10.1111/j.1467-7687.2011.01059.x

Jardri, R., Pins, D., Houfflin-Debarge, V., Chaffiotte, C., Rocourt, N., Pruvo, J. P., et al. (2008). Fetal cortical activation to sound at 33 weeks of gestation: a functional MRI study. *Neuroimage* 42, 10–18. doi: 10.1016/j.neuroimage.2008.04.247

Kisilevsky, B. S., Hains, S. M. J., Jacquet, A. Y., Granier-Deferre, C., and Lecanuet, J. P. (2004). Maturation of fetal response to music. *Dev. Sci.* 7, 550–559. doi: 10.1111/j.1467-7687.2004.00379.x

Kubicek, C., Hillairet de Boisferon, A., 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. doi: 10.1371/journal.pone.0089275

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

Kuhl, P. K., and Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behav. Dev.* 7, 361–381. doi: 10.1016/S0163-6383(84)80050-8

Kuhl, P. K., Williams, K. A., and Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 829–840. doi: 10.1037/0096-1523.17.3.829

Kushnerenko, E., Teinonen, T., Volein, A., and Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11442–11445. doi: 10.1073/pnas.0804275105

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E., et al. (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *Eur. J. Neurosci.* 38, 3363–3369. doi: 10.1111/ejn.12317

Lecanuet, J. P. (1996). "Prenatal auditory experience," in *Musical Beginnings: Origins and Development of Musical Competence,* eds I. Deliege and J. Slobada (New York, NY: Oxford University Press), 3–34.

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant Behav. Dev.* 13, 343–354. doi: 10.1016/0163-6383(90)90039-B

Lewkowicz, D. (1992). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Percept. Psychophys.* 52, 519–528. doi: 10.3758/BF03206713

Lewkowicz, D. J. (1994). Limitations on infants' response to rate-based auditory-visual relations. *Dev. Psychol.* 30, 880–892. doi: 10.1037/0012-1649.30.6.880

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants: the hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Dev. Psychol.* 39, 795–804. doi: 10.1037/0012-1649.39.5.795

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Dev. Psychol.* 46, 66–77. doi: 10.1037/a0015579

Lewkowicz, D. J., Minar, N. J., Tift, A. H., and Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: its emergence and the role of experience. *J. Exp. Child Psychol.* 130, 147–162. doi: 10.1016/j.jecp.2014.10.006

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lickliter, R., and Bahrick, L. E. (2000). The development of infant intersensory perception: advantages of a comparative convergent-operations approach. *Psychol. Bull.* 126, 260–280. doi: 10.1037/0033-2909.126.2.260

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Dev.* 55, 1777–1788. doi: 10.2307/1129925

Matchin, W., Groulx, K., and Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk Effect, and fMRI. *J. Cogn. Neurosci.* 26, 606–620. doi: 10.1162/jocn_a_00515

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

Middelweerd, M. J., and Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *J. Acoust. Soc. Am.* 82, 2145–2147. doi: 10.1121/1.395659

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 164–181. doi: 10.1523/JNEUROSCI.0896-05.2005

Morton, J., and Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychol. Rev.* 98, 164–181. doi: 10.1037/0033-295X.98.2.164

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage* 30, 563–569. doi: 10.1016/j.neuroimage.2005.10.002

Nagy, E. (2008). Innate intersubjectivity: newborns' sensitivity to communication disturbance. *Dev. Psychol.* 44, 1779–1784. doi: 10.1037/a0012665

Navarra, J., and Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychol. Res.* 71, 4–12. doi: 10.1007/s00426-005-0031-5

Parise, C. V., Spence, C., and Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Curr. Biol.* 22, 46–49. doi: 10.1016/j.cub.2011.11.039

Patterson, M. L., and Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Dev. Sci.* 6, 191–196. doi: 10.1111/1467-7687.00271

Patterson, M. L., and Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behav. Dev.* 22, 237–247. doi: 10.1016/S0163-6383(99)00003-X

Pons, F., Teixidó, M., Garcia-Morera, J., and Navarra, J. (2012). Short-term experience increases infants' sensitivity to audiovisual asynchrony. *Infant Behav. Dev.* 35, 815–818. doi: 10.1016/j.infbeh.2012.06.006

Pujol, R., Lavigne-Rebillard, M., and Uziel, A. (1991). Development of the human cochlea. *Acta Otolaryngol.* 111, 7–13. doi: 10.3109/00016489109128023

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–949. doi: 10.1126/science.7233191

Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (1997). The McGurk effect in infants. *Percept. Psychophys.* 59, 347–357. doi: 10.3758/BF03211902

Rubel, E. W., and Ryals, B. M. (1983). Development of the place principle: acoustic trauma. *Science* 219, 512–514. doi: 10.1126/science.6823549

Shaw, K., Baart, M., Depowski, N., and Bortfeld, H. (2015). Infants' preference for native audiovisual speech dissociated from congruency preference. *PLoS ONE* 10:e0126059. doi: 10.1371/journal.pone.0126059

Slater, A. (2002). Visual perception in the newborn infant: issues and debates. *Intellectica* 34, 57–76. Available online at: http://intellectica.org/SiteArchives/archives/n34/n34_table.htm

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Alex Meredith, M., Perrault, T. J., et al. (2010). Semantic confusion regarding the development of multisensory integration: a practical solution. *Eur. J. Neurosci.* 31, 1713–1720. doi: 10.1111/j.1460-9568.2010.07206.x

Stekelenburg, J. J., and Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. *Neuropsychologia* 50, 1425–1431. doi: 10.1016/j.neuropsychologia.2012.02.027

Stevenson, R. A., Zemtsov, R. K., and Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1517–1529. doi: 10.1037/a0027339

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Teinonen, T., Aslin, R. N., Alku, P., and Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition* 108, 850–855. doi: 10.1016/j.cognition.2008.05.009

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., and Morgan, J. L. (2013). Increased focus on the mouth among infants in the first year of life: a longitudinal eye-tracking study. *Infancy* 18, 534–553. doi: 10.1111/j.1532-7078.2012.00135.x

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Malle, B. F., and Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *J. Child Lang.* 42, 1173–1190. doi: 10.1017/S0305000914000725

Tomalski, P. (2015). Developmental trajectory of audiovisual speech integration in early infancy: a review of studies using the McGurk paradigm. *Psychol. Lang. Commun.* 19, 77–100. doi: 10.1515/plc-2015-0006

Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (2005). Audio-visual speech perception is special. *Cognition* 96, B13–B22. doi: 10.1016/j.cognition.2004.10.004

Turkewitz, G., and Kenny, P. A. (1982). Limitations on input as a basis for neural organization and perceptual development: a preliminary theoretical statement. *Dev. Psychobiol.* 15, 357–368. doi: 10.1002/dev.420150408

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vroomen, J., and Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition* 110, 254–259. doi: 10.1016/j.cognition.2008.10.015

Vroomen, J., and Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: not that special. *Cognition* 118, 75–83. doi: 10.1016/j.cognition.2010.10.002

Wallace, M., and Stein, B. (2007). Early experience determines how the senses will interact. *J. Neurophysiol.* 97, 921–926. doi: 10.1152/jn.00497.2006

Wallace, M. T., and Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia* 64, 105–123. doi: 10.1016/j.neuropsychologia.2014.08.005

Watson, T. L., Robbins, R. A., and Best, C. T. (2014). Infant perceptual development for faces and spoken words: an integrated approach. *Dev. Psychobiol.* 56, 1454–1481. doi: 10.1002/dev.21243

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

Yeung, H. H., and Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychol. Sci.* 24, 603–612. doi: 10.1177/0956797612458802

Yu, L., Rowland, B. A., and Stein, B. E. (2010). Initiating the development of multisensory integration by manipulating sensory experience. *J. Neurosci.* 30, 4904–4913. doi: 10.1523/JNEUROSCI.5575-09.2010

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.