Cognitive Development 70 (2024) 101453

Contents lists available at ScienceDirect

Cognitive Development

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

Sensitivity to temporal synchrony in audiovisual speech in early infancy: Current issues and future avenues

Itziar Lozano[a,b,c,*,1], Ruth Campos[a,2], Mercedes Belinchón[a,3]

[a] Department of Basic Psychology, Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain
[b] Department of Cognitive Psychology and Neurocognitive Science. Faculty of Psychology, University of Warsaw, Warsaw, Poland
[c] Neurocognitive Development Lab, Institute of Psychology, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Keywords:
Audiovisual speech
Sensitivity to temporal synchrony
Audiovisual processing
Infants

ABSTRACT

Audiovisual speech integration during infancy is crucial for socio-cognitive development. A key perceptual cue infants use to achieve this is temporal synchrony detection. Although the current developmental literature on this ability is rich, unsolved disagreements obscure the interpretation of findings. Here, we propose conceptual and methodological issues that may have contributed to a still unclear picture of the developmental trajectory of sensitivity to temporal synchrony, particularly when studied in audiovisual *fluent* speech. We discuss several sources of confusion, including a lack of terminological precision, heterogeneity in the experimental manipulations conducted, and in the paradigms and stimuli used. We propose an approach that clarifies the definition and operationalization of sensitivity to temporal synchrony and explores its developmental course, emphasizing the role of infants' linguistic experiences. Ultimately, we expect that our analytical review will contribute to the field by aligning theoretical constructs, proposing more fine-grained designs, and using stimuli closer to infants' experiences.

From early in life, infants encounter a vast amount of information coming from multiple sensory modalities. In social interactions, caregivers provide a variety of multimodal experiences, with audiovisual events being particularly present (Lewkowicz, 2010a). Infants are frequently exposed to speaking faces, thus concurrently hearing speech sounds and seeing mouth movements, along with other audiovisual co-occurrent cues, such as head and facial feature movements (e.g., eye-widening, raising eyebrows, and moving cheeks; see Guellaï et al., 2016; Kitamura et al., 2014). Audiovisual speech is thus an 'inherently multimodal event' (Rosenblum, 2008) that infants need to appropriately integrate to make perceptual sense of their environment (Bahrick, 2010). This ability to link what they see with what they hear is known as *audiovisual processing.*

Adults effortlessly merge auditory and visual information into a unified percept during conversation (McGurk & MacDonald, 1976). For infants, achieving this ability is one of the greatest perceptual challenges in early development. Its acquisition is critical because it enables infants to later develop more complex socio-cognitive abilities, such as temporally coordinating interactions with adults, joint attention, or early language acquisition (e.g., word learning; Bahrick & Lickliter, 2012). Thus, the benefits of mastering the domain-general ability of audiovisual integration extend beyond the first few months of life and the domain of speech perception,

* Correspondence to: Neurocognitive Development Lab, Institute of Psychology, Polish Academy of Sciences, Jaracza, 1, 00-378 Warsaw, Poland.
E-mail addresses: ilozano@psych.pan.pl (I. Lozano), ruth.campos@uam.es (R. Campos), mercedes.belinchon@gmail.com (M. Belinchón).

[1] ORCID: 0000-0003-4280-4747

[2] ORCID: 0000-0003-0209-6002

[3] ORCID: 0000-0001-5367-279X

https://doi.org/10.1016/j.cogdev.2024.101453
Received 10 November 2023; Received in revised form 19 February 2024; Accepted 8 May 2024
Available online 27 May 2024

ultimately facilitating engagement with and learning from the social world (Wallace et al., 2019).

1. Audiovisual speech processing and sensitivity to temporal synchrony in early infancy

For about four decades, a large body of evidence on audiovisual processing in the context of speech has shown that the ability to integrate auditory and visual information develops very early in life (e.g., Altvater-Mackensen et al., 2016; Brookes et al., 2001; Imafuku & Myowa, 2016; Imafuku et al., 2019; Patterson & Werker, 1999, 2002; 2003; Spelke, 1979), with some studies observing it even in neonates (e.g., Aldridge et al., 1999). By 4–5 months, infants already detect the correspondence between auditory and visual speech (Dorn et al., 2018; Kuhl & Meltzoff, 1982, 1984; Kushnerenko et al., 2008; Walton & Bower, 1993) and show the McGurk effect (Burnham & Dodd, 1996, 2004; Rosenblum et al., 1997; but see Mercure et al., 2019 for a slightly later onset and Desjardins & Werker, 2004 for evidence that this effect is not a mandatory phenomenon).

As postulated by one of the main contemporary hypotheses of perceptual development—the Intersensory Redundancy Hypothesis (*IRH*, henceforth; Bahrick & Lickliter, 2002, 2014)—and its theoretical predecessor (Gibson, 1969, 2000; Gibson & Pick, 2000), to successfully bind auditory and visual events into a single percept, infants develop sensitivity to amodal properties available within multisensory information (e.g., rhythm, tempo, intensity). These properties are highly salient to infants and provide information not specific to one sensory modality but common across senses (Bahrick, 2010). Among them, temporal synchrony seems to be the most important for organizing perceptual experience during the first few months of development (Bahrick et al., 2004; Bahrick & Lickliter, 2012).

Temporal synchrony is defined as “changes in events that occur at the same moment in time” across different sensory modalities (Bahrick & Hollich, 2008, p. 164). For example, in natural audiovisual speech, there is temporal synchrony between lip movements and speech sounds. Cumulatively, and consistent with the IRH, evidence suggests that infants’ sensitivity to detecting this property (*sensitivity* to temporal synchrony, henceforth) guides the perceptual integration of auditory and visual events from early in life (e.g., Lewkowicz, 2010a; Hillairet de Boisferon et al., 2017). For example, at 7.5 months, infants use audiovisual synchrony to segment the speech stream (Hollich et al., 2005). Crucially, infants’ sensitivity to disruptions in audiovisual temporal synchrony varies depending on their prior experience with the stimuli presented in the tasks, highlighting the importance of exposure to audiovisual speech in their specialization of their neurocognitive system to detect amodal properties (Addabbo et al., 2022; Fava et al., 2014; Lewkowicz, 2014). Just like audiovisual processing, sensitivity to temporal synchrony is also a domain-general perceptual ability critical for the detection of multisensory holistic perception of our everyday world (Pons & Lewkowicz, 2012). It plays a role in associative multisensory learning (Bahrick et al., 2005; Brookes et al., 2001), and the development of higher-order nonlinguistic (e.g., learning object’ behavior; Spelke et al., 1983) and linguistic abilities (e.g., detecting object-word correspondences; Werker et al., 1998).

Despite the extensive literature exploring sensitivity to temporal synchrony and audiovisual processing (see Supplementary Table S1 for a summary), the few available reviews from a developmental perspective have highlighted the complexity and lack of clarity of the trajectory of audiovisual mismatch detection—defined here as matching the mouth articulatory movements seen with the speech sounds heard—, which extends from birth to the first year’s second half (Streri, et al., 2016; Tomalski, 2015). As Cox et al., (2022) pointed out in a meta-analysis, the developmental course of infants’ ability to detect audiovisual congruence in speech is not straightforward, even when studies are systematically explored. The role of experience (i.e., exposure to native versus non-native audiovisual speech) in the emergence and changes of this ability over time is also unclear. Accordingly, here we argue that current findings on the development of infants’ sensitivity to temporal synchrony in audiovisual speech are still mixed, leading to an unclear picture of the trajectory. First, it remains unknown *when* this ability arises in development. Potentially, by the end of the first year, at least *in the domain of audiovisual (fluent) speech perception,* while in non-speech and non-social events—beyond the scope of this review—it is observable much earlier (Bahrick, 1983; 2001). However, the results do not converge. Some studies report infants detecting temporal asynchronies of 400 ms in audiovisual speech as early as 4 months (Dodd, 1979), while others of 666 and 500 ms at 8 months (Pons & Lewkowicz, 2014, Exp. 1). Second, the developmental trajectory of sensitivity to temporal synchrony in the domain of audiovisual speech perception over the first year is not clear enough, partly due to the lack of longitudinal studies tracking potential changes in this ability.

Note that although we are aware that audiovisual processing and sensitivity to temporal synchrony are fundamental domain-general abilities in infancy (see, e.g., Bahrick, 1987; 1992; Lewkowicz, 1986; 1996), we will focus this review only on studies of audiovisual speech for two reasons.

First, audiovisual speech events are far more complex than audiovisual non-speech events (e.g., dynamic sounding objects striking a surface). Naturalistic auditory and visual continuous speech streams are highly unpredictable, dynamic, and temporally complex (Chandrasekaran et al., 2009), and usually represented by multiple and concurrent perceptual features that interact intricately (e.g., coarticulation; Hardcastle & Hewlett, 1999). Indeed, infants encounter more difficulty in audio-visually processing and detecting temporal asynchronies in audiovisual speech (particularly *fluent* audiovisual speech) than dynamic-sounding objects, evidenced by a later onset of sensitivity to this cue and a larger temporal binding window in audiovisual speech than dynamic-sounding objects (Bahrick, 1983; 1987; 1988; Lewkowicz, 1996; Pons & Lewkowicz, 2014, Exp. 1). This complexity makes the unsolved conceptual and methodological issues we will discuss here likely exclusive to audiovisual speech events while not applicable to all audiovisual non-speech events. The closer the perceptual properties of an audiovisual non-speech event are to those of audiovisual speech events, the more likely the conceptual and methodological issues we will discuss would also apply to them. For example, some challenges related to the experimental manipulations’ accuracy may extend to audiovisual non-speech events sharing perceptual cues with audiovisual speech, such as facial expressions and gestures (which are also highly unpredictable, dynamic, and temporally complex), but not to events like dynamic objects impacting surfaces, which are dynamic but often more predictable and temporally simpler.

Second, unlike the literature on audiovisual speech events, the developmental picture of audiovisual processing and sensitivity to temporal synchrony in non-speech events in infancy seems much clearer and convergent (see Lewkowicz, 2000, for a comprehensive review), thus making it unnecessary to critically review it. Overall, evidence in these audiovisual events shows that temporal synchrony detection emerges first and early, by 2 months (Lewkowicz, 1996), which then declines and is followed by responsiveness to duration, rate, and rhythm, in that order (Lewkowicz, 2000). Unlike in audiovisual non-speech events, as well as other non-speech communicative events, in audiovisual speech events, we cannot assume that the importance of detecting temporal synchrony declines in the hierarchy of perceptual features with development, while higher-level features become increasingly important (e.g., amodal features such as duration, tempo, rhythm/prosody, but also identity-specific features; Lewkowicz & Ghazanfar, 2009; Lewkowicz & Hansen-Tift, 2012). For example, infants do not detect temporal synchrony in audiovisual *fluent* speech until 8 months (Pons & Lewkowicz, 2012) and continue using this cue along with prosody at the end of the first year (Roth et al., 2022). Not being able to assume the same hierarchy of perceptual features for both types of events makes it necessary to think critically, particularly about studies on audiovisual speech events.

This review critically assesses research on infants' sensitivity to temporal synchrony in audiovisual speech. We first discuss current conceptual and methodological issues hindering clear conclusions on its developmental trajectory, especially in audiovisual *fluent* speech. Conceptual issues involve lack for terminological precision, while methodological ones include heterogeneity in experimental manipulations, paradigms, and stimuli used. We then propose future avenues to overcome these sources of confusion, including a clearer definition of audiovisual processing and sensitivity to temporal synchrony in audiovisual speech, a way of operationalizing them separately, and a theoretical framework placing infants' language experiences and developmental processes at the core.

2. Conceptual and methodological issues in the study of sensitivity to temporal synchrony in audiovisual speech in early infancy

A closer look at the main studies published reveals several factors that may be creating at times confusion in research on infants' development of sensitivity to temporal synchrony in audiovisual speech. Although most developmental scientists have surely taken into account various 'confounds' linked to the complexity of experimentally manipulating audiovisual speech, unsolved ones remain. In this paper, we discuss two main controversial issues by reviewing the primary studies published: the lack of specificity in defining audiovisual processing and sensitivity to temporal synchrony, and the use of mixed paradigms and stimuli when investigating this latter perceptual ability.

First, we argue that there have been diverse definitions of the theoretical constructs assessed, as sensitivity to temporal synchrony and audiovisual processing have often been treated interchangeably. This lack of terminological precision may have influenced how sensitivity to temporal synchrony has been manipulated in the tasks used to measure it. Particularly, it may have led to the lack of fine-

Table 1
Methodological Details of Some Exemplar Studies Investigating Audio-visual Processing (AVP) in Typically Developing Infants.

Study	Age (months)	Paradigm	Stimuli	Comparisons	Fragment of the stimuli used
Lewkowicz et al., (2015) - Exps. 1-3.	4 8-10 12-14	IS matching procedure (V-V-AV-AV)	AV fluent speech	-Matching face -Non-matching face	*"¡Buenos días! ¡Despiértate ya! ¡Vamos!"/"Good morning! Get up! Come on now…"*
Lewkowicz & Pons (2013)	6-8 10-12	IS matching procedure (baseline-fam-test-fam-test) (V-V-A-V-A-V)	AV fluent speech	-Matching visible language -Non-matching visible language	ING- *"Good morning! Get up. Come on now. If you get up right away…"*/ ESP- *"¡Buenos días, despiértate ya! ¡Si te levantas ahora …"*
Kubicek, Gervain et al.,(2014) - Exp 1a.	4.5 6	Variant of IS matching procedure (V-V-A-V-A-V)	AV fluent speech	-Matching face (baseline) -Matching face (test)	*"Bébé , coucou…"*
Kubicek, Gervain et al.,(2014) - Exp. 2	6 12	Variant of IS matching procedure (Baseline-test) (V-V-AV-AV)	AV fluent speech	-Matching face (baseline) -Matching face (test)	*"Bébé , coucou…"*
Kuhl & Meltzoff (1982)	4.5-5	*Familiarization (V)-test (AV)*	Vowels	-Matching face -Mismatching face	*/a/* and */i/*
Pons et al., (2009)	6 and 11	IS matching procedure (V-V-A-V-A-V); fam-test	Syllables	-Matching face (baseline) -Matching face (test)	*/ba/* and */va/*
Shaw et al., (2015) - Exp. 1 -	5-10	AV preferential looking task	AV fluent speech	-Matching side -Mismatching side	"*A little white dog…*"/"*El perrito blanco…*"

Notes: V = Visual. A = Auditory; AV = Audio-visual; Exp.= Experiment.

grained designs that allow to separately address whether infants (1) integrate audiovisual information, and (2) can detect temporal synchrony. In the first subsection, we will explain this point. Second, and perhaps due to unclearly defined constructs, we argue that there has also been a high heterogeneity in the paradigms and stimuli used to explore the development of sensitivity to temporal synchrony. In the second subsection, we will briefly describe and critically discuss the methodological details of some of the main studies conducted in this perceptual ability in audiovisual speech in infancy. An overview of these two issues is necessary. It will ultimately contribute to a more precise definition of sensitivity to temporal synchrony and facilitate conducting aligned, more fine-grained designs to explore this perceptual ability more accurately.

The idea that variable paradigms, stimuli, and terminology have exacerbated the misunderstanding in the field of infant audiovisual speech processing has been previously laid out by Shaw & Bortfeld (2015). While our argument parallels the authors', our review adds two new meaningful contributions. First and foremost, our focus solely on behavioral paradigms of audiovisual integration (rather than both neurophysiological and behavioral) enables a more careful dissection of the current state of research on sensitivity to temporal asynchrony relative to broader audiovisual processing research. Second, our review goes beyond identifying sources of disagreement and focuses on more clearly defining and operationalizing terminology.

2.1. One or Two Abilities?

A long list of terms has been used to refer to the ability to integrate events from different modalities (i.e. audiovisual processing). Some examples are *audiovisual processing* (Shaw & Bortfeld, 2015), *audiovisual integration* (Tomalski, 2015), *intersensory processing* (Edgar et al., 2023), *audiovisual congruency detection* (Shaw et al., 2015), *cross-modal matching* (Kubicek et al., 2014), and *perception of multisensory coherence* (Lewkowicz et al., 2015). The mention of coherence, congruency, matching, and integration is shared by almost all, emphasizing that what is key is the co-occurrence of auditory and visual information. However, while these terms may seem equivalent and are often used in literature as if they were, nuances differentiate them.

In the context of audiovisual speech processing literature in infancy, the idea that audiovisual processing and sensitivity to temporal synchrony constitute either one or two different abilities is controversial (see Shaw & Bortfeld, 2015). Most researchers have argued, more or less explicitly, that audiovisual processing (illustrated in Table 1) and sensitivity to temporal synchrony (illustrated in Table 2) are two *different* perceptual abilities (e.g., Hillairet de Boisferon et al., 2017; Stevenson et al., 2018). However, others have

Table 2
Methodological Details of the Main Studies Investigating Sensitivity to Temporal Synchrony (STS) in Typically Developing Infants.

Study	Age (months)	Type of procedure	Paradigm	Stimuli	Comparisons	Fragment of a trial
Bahrick et al., (2018) - Exp. 2 * -	12	Simultaneous	AV preferential looking task	AV fluent speech	-Matching -Non-matching	*"I can't find my elephant…"/ "It is just a house for me, me, me…"*
Lewkowicz (2010a) - Exp. 1 -	4, 8 and 10	Sequential	Habituation/test	AV syllable	-SYNC (habit) -ASYNC A-V 366 ms (test) -ASYNC A-V 500 ms (test) -ASYNC A-V 666 ms) (test)	*/ba/*
Pons & Lewkowicz (2014)	8	Sequential	Habituation/test	AV fluent speech	-SYNC (habit) -ASYNC A-V 366 ms (test) -ASYNC A-V 500 ms (test) -ASYNC A-V 666 ms (test)	ING- *"Good morning! Get up. Come on now. If you get up right away…"*/ ESP- *"¡Buenos días, despiértate ya! ¡Si te levantas ahora …"*
Hillairet de Boisferon et al., (2017) * *	4, 6, 8, 10 and 12	———	Free viewing	AV fluent speech	-Eyes/mouth in ASYNC	ING- *"Good morning! Get up. Come on now. If you get up right away…"*/ ESP- *"¡Buenos días, despiértate ya! ¡Si te levantas ahora …"*
Dodd (1979)	2.5 – 4	Sequential	Free viewing	AV live fluent speech	-SYNC trials -ASYNC trials	*Non-available (nursery rhymes)*
Edgar et al., (2022) *	12, 18, and 24	Simultaneous	AV preferential looking task	AV fluent speech	-Matching -Non-matching	*"I can't find my elephant…"/ "It is just a house for me, me, me…"*

Notes: ASYNC = Asynchronous. SYNC = Synchronous. Exp.= Experiment.
* Based on the description given in the procedure of these studies, the authors explored sensitivity to temporal synchrony, though they referred to the aim of the study as the assessment of '*intersensory AV matching*'. * * In this study the authors only presented an 'asynchronous condition' and then compared their findings with those of the 'synchronous condition' from the study by Lewkowicz & Hansen-Tift (2012). Thus, since these two conditions were separated in two different studies, it is not possible to apply the categories of sequential and simultaneous.

implicitly assumed that they are *equivalent*.

2.1.1. Two different perceptual abilities

The study by Shaw et al., (2015, Exp. 1) exemplifies a procedure testing audiovisual processing. In a preferential looking task, infants saw two speaking faces simultaneously articulating a story (one in English, another in Spanish), with the audio corresponding to one face (auditory Spanish). Thus, infants' looking times were compared between a 'congruent' speaking face ('audiovisual matching condition') and a speaking face visually uttering the same story in another language ('audiovisual mismatching condition'). Therefore, temporal synchrony cues were absent in the 'non-matching' condition because there was no temporal correspondence between the onsets and offsets of the lip movements and speech sounds, but other amodal properties were also disrupted. Successful audiovisual matching was interpreted if infants preferred looking at the audiovisual matching condition (vs. the mismatching condition).

Another example comes from Lewkowicz et al., (2015, Exps. 1–3). They showed infants a block of visual silent trials (baseline) where two identical faces articulated two different monologues in the same language without audio. This was followed by two test blocks (preference test) showing one auditory monologue and two articulating faces of the same female side-by-side. One face audio-visually matched the audio ('audiovisual matching condition'), while the other visually uttered a different monologue ('audiovisual mismatching condition'). The authors operationalized infants' success in audiovisual matching as an increase in looking time to the talking face matching its corresponding audio during the test trials, compared to the same face when the audio was absent in baseline trials.

Although in these two studies the authors acknowledged the disruption of temporal synchrony in the 'audiovisual mismatching condition', they also clearly reported when temporal synchrony was disrupted *besides other perceptual cues* (e.g., prosody, rhythm, spectral detail, and energetic modulations; see Shaw et al., 2015, p. 8), or *independently disrupted* in another task (see Lewkowicz et al., 2015, Exp. 4). Consistently, the authors labelled the abilities measured as congruency detection (Shaw et al., 2015) and perception of multisensory coherence (Lewkowicz, 2015 et al., Exps. 1–3) in audiovisual fluent speech, but not as temporal synchrony detection. We interpret this distinction as the authors treating audiovisual processing and sensitivity to temporal synchrony as different perceptual abilities. One could argue that the authors might view temporal synchrony detection to be a nested component of audiovisual processing. While this may be the case, the key point is that they explicitly considered synchrony detection as *distinct* from audiovisual speech processing and, consequently, they aligned the labelling of the constructs with their experimental manipulations.

In contrast, Lewkowicz (2010a, Exp. 1) illustrates a study exclusively operationalizing sensitivity to temporal synchrony in audiovisual speech. In a habituation study, 4- to 10-month-old infants' attention to a synchronous speaking face ('synchronous condition'—habituation trials) was compared with their attention to a condition that exclusively differed in the temporal misalignment (366, 500, or 666 ms—test trials) between the auditory and visual events ('asynchronous condition'). With temporal misalignment, here we mean the discrepancy in the timing of presentation of events from the visual and auditory sensory modalities. Hence, only temporal synchrony was manipulated in this study. An increase in infants' looking durations to the test trials compared to habituation ones was interpreted as successful perception of synchrony relation changes (and, therefore, detecting audiovisual temporal synchrony relations, which occurred only for the 666 ms test trials at all ages). Another way of successfully measuring this ability has been two-displays procedures (e.g., preferential-looking paradigm; Bebko et al., 2006; Pons et al., 2013; Righi et al., 2018; Zhou et al., 2022—although these studies involve early childhood and clinical populations), further discussed in the next section.

Other examples in the literature are studies comparing the two abilities within the same participants. For example, Stevenson et al., (2014), explicitly used two different tasks to examine audiovisual processing and sensitivity to temporal synchrony: one measured children's perceptual binding of audiovisual speech signals (using a McGurk paradigm), while the other audiovisual temporal synchrony in both non-speech stimuli (i.e., flashes, beeps, and tool sounds) and speech stimuli (i.e., syllables). Stevenson et al., (2015) also argued in a review that atypicalities in audiovisual processing may be specifically linked to difficulties in audiovisual temporal processing, consistent with both abilities being closely related yet distinct.

In line with Stevenson et al., (2018) and Shaw et al., (2015), we argue that the two operationalization scenarios above correspond to measuring two different abilities. In each, infants faced two different demands relying on two distinct perceptual cues. In Lewkowicz (2010a), only the temporal misalignment between the visual and the auditory events was manipulated, allowing for testing infants' ability to detect temporal correspondence between auditory speech and lip articulatory movements (i.e., based on the onsets and offsets, a *low-level* cue). In contrast, Shaw et al., (2015, Exp. 1) and Lewkowicz et al., (2015, Exps. 1–3) also disrupted other speech properties. Specifically, the correspondence between phonemes and visemes (i.e., the equivalent of phonemes in the visual domain) and several other amodal properties, such as tempo, or prosody, also differ between the 'audiovisual matching' and the 'audiovisual mismatching' conditions described above. Therefore, this manipulation would test infants' capacity to match visual and auditory speech streams (i.e., to rely on the phonetic cues of audiovisual speech; see also Baart et al., 2014), based on '*higher-order*' perceptual cues.

2.1.2. Assumed equivalent perceptual abilities

In contrast to the most common approach in the literature, other researchers have implicitly argued that audiovisual processing and sensitivity to temporal asynchrony are *equivalent skills*, or at least they have studied them as if they were. Surprisingly, this conceptual confusion at times holds even in robust hypotheses on perceptual development (such as the IRH; Bahrick & Lickliter, 2000; but see also Edgar et al., 2023, where the same authors use audiovisual processing and sensitivity to temporal synchrony appropriately, namely, as if were *distinct* capacities).

An example of treating the two capacities as equivalent skills is provided by Bahrick et al., (2018, Exp. 2). In this study, the authors interchangeably claimed that they measured *sensitivity to temporal synchrony* and *intersensory matching* (a term often used in literature to refer to audiovisual processing)—see also Edgar et al., (2022) for an equivalent example. However, when one examines their task in detail, in practice, the authors manipulated the stimuli by temporarily misaligning the onset and offset of the movements of the lips and sounds of speech. Therefore, it seems that what is being compared is whether infants discriminate between 'audiovisual synchronous' vs. 'audiovisual asynchronous' conditions. We argue that this manipulation is not equivalent to the one performed in studies that compare between 'audiovisual matching' vs. 'audiovisual mismatching' conditions and claim to measure audiovisual congruency (e.g., Imafuku et al., 2019; Shaw et al., 2015, Exp. 1). Whereas in the formers the audio and video streams of the audiovisual mismatching condition are temporally desynchronized in 3 s (the two faces utter the same sentence, but one is delayed), in the latters the audiovisual mismatching condition is made by a face uttering an entirely different sentence. This means that although disruptions in amodal properties other than temporal synchrony are inevitably present in both audiovisual mismatching conditions, they are not equally present in both.

Instead, we propose that in a temporal synchrony detection task, such as in Bahrick et al., (2018) or Edgar et al., (2022), the temporal synchrony between the onset and offsets of the lip movements and speech sounds is mainly altered. In contrast, in an audiovisual congruency task (as Imafuku et al., 2019; Shaw et al., 2015, Exp. 1), both the onset and offsets correspondences between mouth movements and speech *and* a more global level of temporal synchrony between specific movements of the face and changes of the speech sounds (characterizing phonemes, prosody, rhythm, and intonation) are disrupted. Therefore, we suggest that these two manipulations (and the perceptual abilities they measure) should not be labelled interchangeably in the literature.

Of course, Bahrick et al., (2018) and Edgar et al., (2022) did not only assess temporal synchrony detection, as audiovisual matching and sensitivity to temporal synchrony are not mutually exclusive perceptual abilities. Baart et al., (2014) pointed out the challenge of manipulating them entirely separately, as the 'practical effect' of temporally desynchronizing the audible and visible streams of fluent audiovisual speech unavoidably impacts other intersensory cues (e.g., prosody; Lewkowicz et al., 2022). This makes it highly difficult to separately manipulate sensitivity to temporal synchrony and intersensory matching or, in other words, to isolate the 'temporal macrostructure' of audiovisual speech (Bahrick, 1988; Lewkowicz et al., 2010b). We agree with this perspective.

However, despite this challenge, we argue that researchers should align how they implement this operationalization (e.g., by 'exclusively' misaligning the temporal synchrony between mouth movements and auditory speech using software like Adobe Premiere Pro, even though it affects other amodal properties) with how they label it. The operationalization performed by Bahrick et al., (2018) matches that of other authors who nevertheless labelled it as a manipulation of temporal synchrony (e.g., Bebko et al., 2006; Righi et al., 2018), thus measuring sensitivity to temporal synchrony and not intersensory matching.

One could argue that temporal synchrony detection is a prerequisite for audiovisual integration and, thus, Bahrick et al., (2018) correctly interpreted infants' audiovisual matching based on temporal synchrony cues as successful intersensory matching. However, we argue that, at least in audiovisual speech events, infants can audiovisually integrate without relying on temporal synchrony, instead using other amodal cues like prosody (Kitamura et al., 2014). Thus, despite temporal synchrony detection and audiovisual integration being two embedded perceptual abilities, the former is not a prerequisite for the latter but rather a *distinct* ability.

Overall, the studies reviewed in this section highlight the need for more conceptual clarity in defining and labelling these two perceptual abilities and for implementing experimental manipulations aligned with the assigned labels. To gain both conceptual and methodological clarity, it seems therefore desirable for researchers to provide more accurate labelling of the constructs and limit their conclusions to the specific ones they manipulate in their studies.

2.2. Methodological sources of heterogeneity: paradigms and stimuli

Heterogeneity in the research on sensitivity to temporal synchrony in the context of audiovisual speech processing can also be found in at least two aspects (summarized in Table 2): the paradigms, and the stimuli used.

2.2.1. The paradigms

Sensitivity to temporal synchrony has been assessed in the literature in a variety of ways, including, at least: 1) habituation studies, where the same events are shown in synchrony versus out of synchrony (e.g., Pons & Lewkowicz, 2014), 2) two-displays intermodal matching studies, where infants must match a video with a soundtrack based on temporal synchrony –one video is in synchrony with the sound, and another is asynchronous in some way (by either playing this second video backwards, delayed, or ahead). In habituation studies, infants' attention recovery on test trials relative to habituation trials is interpreted as successful discrimination of temporal asynchrony, whereas in intermodal matching studies this is assessed by infants' preference for the synchronous face (vs. the asynchronous).

These operationalizations could be classified into two broader categories: 'simultaneous' and 'sequential' procedures. Some studies have presented to infants the 'synchronous' and the 'asynchronous' conditions at once (see, e.g., Bahrick et al., 2018, Exp. 2, using preferential looking paradigm), while others subsequently (i.e., one after the other; see, e.g., Pons & Lewkowicz, 2014). Note that different procedures have been combined with different paradigms. For example, in Dodd (1979), stimuli changed from being in-synchrony to out-of-synchrony every 60 s, which could be categorized as a free-viewing paradigm combined with a sequential procedure. In contrast, other studies have implemented sequential procedures using habituation-test paradigms (e.g., Lewkowicz, 2010a).

Deciding which procedures and paradigms to use often depends on whether the experimental question focuses on discrimination (typically measured through a habituation-test paradigm) or behavioral preference (usually tested using a preferential-looking paradigm). However, the literature also contains some examples of no precise alignment between research questions, paradigms used, and conclusions drawn. For instance, in Dodd's (1979) study, infants showed 'higher inattention to the asynchronous vs. synchronous speech'. The author interpreted this as 'an indication that young infants are aware of the *congruence* between lip movements and speech sounds' (Dodd, 1979, p. 478).

In our view, Dodd's phrasing is consistent with the definitions of audiovisual speech matching, coherence, and congruency that usually underlie the broader construct of audiovisual processing in the literature. However, we believe Dodd's design prevents deriving her interpretation of infants being able to detect temporal synchrony. Pons & Lewkowicz (2014, p. 143) also interpreted Dodd's results of higher inattention to the asynchronous vs. synchronous speech as: 'young infants can perceive audiovisual speech synchrony in fluent speech'. Nevertheless, in a free-viewing paradigm with a sequential procedure like Dodd's, it is not possible to attribute infants' distinction between lip and voice congruence vs. incongruence to the detection of onsets and offsets (temporal synchrony detection) or the congruence of the dynamics of the two speech streams (audiovisual congruency detection).

Alternatively, we suggest that it would be more accurate and cautious to interpret Dodd's results as infants showing greater interest or preference for temporally synchronous vs. asynchronous fluent speech. Indeed, this aligns with the prediction of the IRH (Bahrick, 2010; Bahrick et al., 2004) that early on audiovisual redundancy is more salient to infants than non-redundancy. To draw conclusions about infants' ability to detect congruence between lip movements and speech sounds, we argue that using a method allowing testing audiovisual matching in fluent speech would have been necessary (see Table 1). Potential candidates include the intersensory matching procedures (e.g., Kubicek et al., 2014, Exp. 1a & Exp. 2; Lewkowicz et al., 2015, Exps. 1–3; Lewkowicz & Pons, 2013) or an audiovisual preferential looking task (e.g., Shaw et al., 2015, Exp. 1).

Distinguishing between 'simultaneous' and 'sequential' procedures is crucial beyond aligning research questions, paradigms, and conclusions. As mentioned previously, simultaneous procedures often disrupt other amodal properties than temporal synchrony in audiovisual *fluent* speech. For example, in a two-displays paradigm (e.g., preferential looking, Bebko et al., 2006), if the same woman speaks side-by-side, one synchronously with the sound and the other asynchronously, infants can match the speech sounds to the synchronous face using temporal synchrony, but also rhythm, duration, intensity shifts, and prosody. In contrast, sequential procedures (e.g., habituation; Pons & Lewkowicz, 2014) offer more precise manipulation of temporal synchrony, minimizing disruption of other amodal properties. This seeming nuance in operationalization may partly explain mixed evidence on the exact age(s) infants show sensitivity to temporal synchrony, overall observable earlier in sequential (8 months; Pons & Lewkowicz, 2014) than simultaneous paradigms (10 months; Hillairet de Boisferon et al., 2017), even when the level of asynchrony or the stimuli remain constant across experiments.

2.2.2. *The stimuli*

The second source of heterogeneity in the literature has been the type of stimuli used in the experimental tasks. Studies range from those assessing responsiveness to speech at the lowest level of the hierarchy (vowels, isolated syllables) to the highest (fluent speech). While most have measured sensitivity to temporal synchrony in audiovisual fluent speech, others have used audiovisual syllables (Lewkowicz, 2010a). This extant range of stimuli provides a rich database of findings revealing a complex picture but also obscures the developmental pattern of sensitivity to temporal synchrony in audiovisual speech.

The choice of stimuli is often but not consistently justified by the specific research question being addressed, while it is critical for at least two reasons. First, these two types of events differ in their closeness to infants' experiences since they are not exposed to audiovisual syllables but to fluent speech in their daily interactions. However, these two events are often treated interchangeably, perhaps partly because research on audiovisual face-voice integration (and, by extension, on sensitivity to temporal synchrony) has not been approached in the context of person perception, but rather audiovisual speech perception (Campanella & Belin, 2007). Under this latter view, speech elements are hierarchically related, with sentences and syllables seen as more complex and difficult than consonants and vowels. Consequently, studies adopting this view are approached regardless of infants' language experiences in their natural contexts. Second, in audiovisual *fluent* speech, silences between words are less predictable. In contrast, in audiovisual syllables, silences are systematic, possibly aiding infants to detect asynchronies between auditory and visual information (Pons & Lewkowicz, 2012).

If the type of speech stimuli used can modulate infants' likelihood of detecting temporal asynchronies, we argue that more ecological validity is needed. In the context of audiovisual speech studies in infancy, more ecological validity refers to using events closely resembling human natural speech infants are exposed to. The closer the stimuli are to that, the more reliably the experimental task can measure infants' ability to detect temporal asynchrony. Therefore, when investigating this ability in the specific domain of audiovisual *fluent* speech processing, research should test infants with the speech type they actually hear during typical interactions, which is audiovisual *fluent* speech. Alternatively, researchers choosing less naturalistic speech (e.g., syllables or vowels) should be cautious not to over-generalize the conclusions derived from these studies to infants' ability to process audiovisual *fluent* speech, instead limiting their conclusions to the specific stimuli types presented in their studies.

Other sources of stimuli variability include using infant-directed speech vs. adult-directed speech (Roth et al., 2022), the number of actors involved, or the degree of temporal asynchrony (ranging from 366 ms—Pons & Lewkowicz, 2014—to 3 s—Bebko et al., 2006). Together, they may affect how many amodal properties differ between the synchronous and asynchronous conditions. For example, using two (vs. one) actresses may influence whether identity is additionally manipulated. Similarly, as the temporal misalignment between auditory and visual events increases, more additional amodal properties beyond synchrony are manipulated. Prosody constitutes a good example. For instance, if using 1 s of temporal misalignment, the 'visual' prosody is less similar between the synchronous and asynchronous conditions than if using 400 ms. Therefore, while in the former manipulation infants are more likely to rely on prosody to audio-visually match, this amodal cue would be less helpful in the latter. Crucially, if one understands the degree of asynchrony as a continuum, using very extreme degrees of asynchrony can even shift the task from measuring temporal synchrony detection (e.g., when using a low degree, such as 366 ms) to audiovisual congruency detection (e.g., when using a high degree, such as 5 s).

To sum up, particular experimental conditions may modulate infants' chances of detecting temporal asynchronies, suggesting that the stimuli choice could affect task demands and, consequently, infants' performance. The studies reviewed in this section reveal significant gaps in knowledge. It remains unclear whether the timing of the developmental trajectory of temporal synchrony detection is influenced by the interaction between specific experimental paradigms and stimuli. Sequential paradigms overall appear more sensitive than simultaneous ones for measuring temporal synchrony detection. However, audiovisual syllables contain in-between-silent-cues absent in audiovisual *fluent* speech, which together may impact task sensitivity and, thus, infants' performance. Clarifying these interactions is crucial for advancing our understanding of infants' sensitivity to temporal synchrony in audiovisual speech.

3. Future avenues: towards increasing clarity

Our review suggests that, despite four decades of cautious developmental research on audiovisual speech perception, the field requires more conceptual and methodological clarity to advance. The issues reviewed may explain why the developmental trajectory of sensitivity to temporal synchrony in audiovisual speech processing remains unclear. While we advocate for greater aligning constructs and experimental manipulations, and for higher homogeneity in stimuli and paradigms used, we do not propose adopting a single and uniform experimental approach. Instead, we highlight the need not to over-interpret literature findings across constructs, stimuli, and paradigms when they are not sufficiently homogenized.

To overcome our fragmented understanding, we propose key conceptual and methodological directions. We first attempt to clarify how the constructs of sensitivity to temporal synchrony and audiovisual processing should be defined conceptually and operationally. Subsequently, we suggest framing future research under a theoretical approach that explores the course of sensitivity to temporal synchrony by focusing on the nature of infants' linguistic experiences. Next, we argue for investigating this perceptual ability under theoretical frameworks that emphasize the developmental processes underlying its changes over time and its functional role in development. Finally, we highlight some limitations of our review.

3.1. Defining the constructs

As outlined above, we believe that sensitivity to temporal synchrony and audiovisual processing are *distinct* and should be studied as such. Although much of the careful research to date aligns with this distinction, our review indicates remarkable exceptions in the literature. At the level of improving conceptual clarity, if one assumes—as we and other authors do (e.g., Hillairet de Boisferon et al., 2017; Stevenson et al., 2018)—that sensitivity to temporal synchrony and audiovisual processing are different constructs, then it is crucial to clarify first whether these two perceptual abilities are (or are not) mutually exclusive. Our interpretation of the literature is that they are *not* mutually exclusive, but rather nested concepts, meaning that sensitivity to temporal synchrony is embedded in audiovisual processing, which would be a broader term encompassing also other various more specific perceptual abilities. In addition to temporal synchrony, the detection of common duration, tempo, intensity patterns, spatial location, and phonemic cues across auditory and visual speech events would also be part of this broader construct. This general conceptual framework is not entirely new and has previously been used in reviews of audiovisual processing of non-speech (e.g., Lewkowicz, 2000) and emotional events (Walker-Andrews, 1997). However, based on the evidence reviewed here, we suggest it has been overlooked in the context of processing audiovisual *speech* events (and, particularly, audiovisual *fluent* speech; see Soto-Faraco et al., 2012, for an exception).

In this latter context, two arguments support the distinctiveness of the two perceptual abilities despite being nested. First, infants do not successfully learn them at the same time-points in development. Instead, infants' onset trajectory is earlier for audiovisually matching (2 months in vowels, and 6 months in *fluent* speech; Imafuku et al., 2019; Kuhl & Meltzoff, 1982; Patterson & Werker, 2003) than for detecting temporal asynchronies (presumably, 4 months in vowels, Lewkowicz, 2010a; and 8 months in *fluent* speech; Pons & Lewkowicz, 2014, Exp. 1). Second, infants can audiovisually process by relying on perceptual cues that are not temporal synchrony (e.g., prosody; Kitamura et al., 2014; Roth et al., 2022).

In addition to how these constructs are organized hierarchically, the most crucial point to clarify is the definition of these constructs themselves. To investigate sensitivity to temporal synchrony as a perceptual ability different from audiovisual processing, researchers first need to agree on appropriate definitions that allow coherent and aligned experimental manipulations. We suggest that a solution to this puzzle could be revisiting the original historical terms defined by ecological hypotheses of perceptual development, such as the IRH (Bahrick & Lickliter, 2002, 2014), its theoretical precursor (Gibson, 2000; Gibson & Pick, 2000), and Lewkowicz's works from 2010 (Lewkowicz, 2010a; 2010b). These authors distinguished between two types of relations redundant across senses supporting infants' ability to perceive auditory and visual events as a unit: temporal macrostructure and temporal microstructure (Bahrick, 1987; 1988; Lewkowicz, 2010a; 2010b). Whereas 'temporal macrostructure relations' refer to the temporal synchrony between the energy onsets and offsets of auditory and visual stimulations (i.e., in audiovisual speech, the onsets and offsets of the lip movements and speech sounds), 'temporal microstructural relations' refer to a more global, nested level of temporal synchrony between specific facial movements and sound changes (i.e., in audiovisual speech, between specific facial movements and changes in speech sounds at the level of prosody, rhythm, intonation, and phonemic cues; Bahrick, 1987).

We suggest that, in the context of audiovisual speech perception, there is a parallel or correspondence between the two amodal relations 'temporal macrostructure' and 'temporal microstructure', and the constructs 'sensitivity to temporal synchrony' and 'audiovisual processing', respectively. While sensitivity to temporal synchrony would entail detecting disruptions between auditory and visual events at the *temporal macrostructure level* (i.e., in the onsets and offsets of lip movements and speech sounds), audiovisual processing would consist of perceiving a disruption *at both the temporal macrostructure and microstructure levels* (i.e., in the onsets and offsets of lip movements and speech sounds, and a more global nested level of temporal synchrony between facial movements and speech sound changes). Thus, we propose that the former definition aligns with what should be measured in a temporal synchrony detection task (as in Bahrick et al., 2018, Bebko et al., 2006; Righi et al., 2018), while the latter corresponds to what should be assessed in an audiovisual congruency or matching task (as in Imafuku et al., 2019; Shaw et al., 2015, Exp. 1). Ultimately, the terms 'temporal macrostructure' and 'microstructure' relations are accurate, acknowledge the non-mutual exclusivity between the two perceptual abilities, and provide clear guidance on what needs to be manipulated to measure each. Hence, we suggest that future studies could benefit from systematically adopting them.

One potential criticism of using this framework in audiovisual *fluent* speech is that there is no a priori way to determine which of the multiple and concurrent features that typically characterize naturalistic auditory and visual continuous speech fall into the macrostructure vs. microstructure categories. While we acknowledge potential limitations to conducting experimental manipulations specific to each type of 'temporal microstructural relations' (i.e., prosody, rhythm, intonation, and phonemic cues), we still find this framework useful for our broader purpose of independently manipulating disruptions at the level of 'temporal macrostructural' (temporal synchrony) and 'temporal microstructural' relations (audiovisual congruency).

3.2. Operationalizing the constructs

Acknowledging that sensitivity to temporal synchrony and audiovisual processing are nested, non-mutually exclusive perceptual abilities does not imply they should not be measured as separate ones—an idea not consistently accepted in the literature. We suggest that if one views them as distinct perceptual abilities conceptually, then methodologically, they should be operationalized in different (and, therefore, to the extent feasible, separate) ways.

However, conducting manipulations that minimize the overlap between these two abilities is challenging (see Lalonde & Werner, 2021). As outlined above, manipulating temporal synchrony in *fluent* audiovisual speech stimuli often disrupts other amodal properties. The key question is then how to measure infants' sensitivity to each type of amodal relations (i.e., 'temporal macrostructure' and 'temporal microstructure') separately, despite their naturally co-occurrence in audiovisual *fluent* speech. We argue that it is possible—and desirable to clarify the field—to isolate these two relations in different experimental tasks, thereby allowing for the accurate measurement of infants' performance in detecting each.

The key question to independently test infants' sensitivity to temporal synchrony and infants' audiovisual processing in *fluent* speech is to experimentally control for either temporal synchrony (a temporal macrostructure relation) or for phonemic cues (a key temporal microstructure relation). This ensures that 1) the two perceptual abilities are not conflated and 2) the definitions of the constructs adopted above and their operationalization align.

There are at least two options meeting these criteria. One possibility is to use sequential paradigms, which allow testing infants' audiovisual matching for *fluent* speech in the absence of temporal synchrony cues. For instance, as Kubicek et al., (2014 – Exp. 1a), we could use a variant of the intersensory matching procedure to present infants with auditory-only familiarization trials (auditory fluent speech) followed by visual-only test trials (silent fluent speech). If infants show longer looking times to the visual-only face matching the auditory-only speech during the test phase, it would be legitimate to conclude that infants can audiovisually match fluent speech. Since this paradigm does not include temporal asynchrony cues (i.e., temporal macrostructure relations), it would also be reasonable to infer that infants do not rely on this cue to succeed, but instead on prosodic and phonemic cues (i.e., temporal microstructure relations). Infants' performance in this study could then be compared to their ability to detect temporal asynchrony in a sequential task, where the only manipulation is to temporarily misalign the onsets and offsets of lip movements and speech sounds (i.e., temporal macrostructure relations; e.g., Pons & Lewkowicz, 2014). However, depending on the asynchrony level used in this task, ensuring that only temporal synchrony differs between the two conditions would be challenging. As mentioned, the higher the level of temporal asynchrony used, the more likely it is that temporal microstructure relations would also be manipulated.

One of the closest attempts to the two manipulations described above within the same study is the cross-sectional work by Lewkowicz et al., (2015). The authors investigated the ability of infants aged 4–14 months to audiovisually match *fluent* speech, and the role of temporal synchrony in achieving this. Infants viewed a baseline block of visual silent trials with two identical faces articulating two different monologues. This was followed by two test blocks of trials where an audible monologue and two articulating faces of the same female were shown simultaneously side-by-side. In one condition, one of the faces visually uttered a different monologue from the audible one ('audiovisual mismatch'), while in the other, the onsets of the mouth movements and the audible monologue were temporarily misaligned by 666 ms ('audiovisual asynchronous'). Only 12–14-month-olds matched fluent auditory and visual speech (i.e., looked longer at the asynchronous monologue relative to the same audible monologue in the visual baseline), even when the two streams were 'desynchronized' (or, we would say, 'mismatched'), presumably because infants relied on prosody and not on temporal synchrony.

However, this study's paradigm does not allow testing infants' sensitivity to temporal synchrony, only whether they rely on it for audiovisually matching *fluent* speech. It also does not allow disentangling which perceptual cues, other than prosody, infants relied on for audiovisually matching, as many others were also disrupted. Therefore, conducting the two manipulations described above within the same study is still necessary. Additionally, it remains unclear which specific cues infants rely on to audiovisually match (prosody, rhythm, tempo, identity) when controlling for temporal synchrony, whether the timing of the developmental trajectory depicted in Lewkowicz's cross-sectional study holds longitudinally, and whether within-infant associations in their performance in these two skills exist across infancy.

A second possibility is to use more sophisticated ways to separate temporal synchrony cues from phonemic cues and determine their relative contribution to audiovisual speech processing. Sine-wave speech (see Remez et al., 1981), a technique still underused in infants' behavioral studies on audiovisual *fluent* speech processing (but see Baart et al., 2014 for a study with pseudowords and Homae et al., 2014 for neural evidence) offers one such alternative. Sine-wave speech involves transforming the natural speech signal through replacing the center frequencies of the first three formants by sinusoids (Baart et al., 2014). Thus, it allows the opposite scenario to sequential paradigms: it preserves the temporal characteristics of natural speech while severely degrading phonemic cues. Comparing infants' performance in audiovisually matching an articulating face to natural auditory speech (vs. in sine-wave speech) allows inference of their reliance on temporal synchrony cues. Equal success in both conditions (i.e., infants matching the sound to the articulating face equally well for natural speech vs. sine-wave speech, as in Baart et al., 2014) would suggest that infants do not rely on phonemic cues (temporal microstructure relations) to detect audiovisual correspondences, but rather on temporal synchrony cues (temporal macrostructure relations). Conversely, better performance with natural speech vs. sine-wave speech (i.e., infants matching the sound to the articulating face better for natural speech vs. sine-wave speech) would suggest infants' greater reliance on phonemic cues for successful audiovisual speech processing.

Conducting longitudinal studies testing the same infants with both tasks is essential. Using both a sequential paradigm, which eliminates temporal synchrony cues while preserving phonemic cues (as in Kubicek et al., 2014), and a sine-wave speech task, which removes phonemic cues while keeping temporal synchrony (as in Baart et al., 2014), would accurately operationalize infants' detection of temporal macro and microstructure relations in audiovisual speech, enabling tracking both across development while minimizing overlap between constructs. Crucially, both tasks should use audiovisual *fluent* speech and keep stimuli content and actress (es) constant to gain methodological clarity. This design would also test whether the relative use of each audiovisual perceptual cue changes across early development. If infants equally succeed at audiovisually matching natural fluent speech and sine-wave speech, but do not solve a sequential paradigm that eliminates temporal synchrony, this would suggest they primarily rely on macrotemporal relations. If infants equally succeed at both tasks at some developmental time-point, this would indicate they use both micro and macrotemporal relations. A third scenario (less likely, based on existing literature; e.g., see Pons & Lewkowicz, 2014, where 8-month--olds were still sensitive to temporal asynchrony in both native and non-native audiovisual fluent speech) would be that infants cease using macrotemporal relations and rely primarily on microtemporal ones, reflected in infants' success with the sequential paradigm that eliminates temporal synchrony, while showing better performance with natural fluent speech vs. sine-wave speech.

3.3. *Theoretically framing the questions: language experience and developmental processes at the core*

A final future direction is the importance of adopting a theoretical approach that places at the core: (1) the nature of infants' everyday language experience, and (2) the functional role and developmental processes underlying sensitivity to temporal synchrony in audiovisual speech.

3.3.1. *The nature of infants' language experience*

A focus on infants' language experience has at least two methodological implications for how we investigate the early development of sensitivity to temporal synchrony in audiovisual speech.

First, if we assume that temporal synchrony is a cue especially embedded in infants' social and linguistic experiences, including audiovisual speech, then testing the development of infants' sensitivity to temporal synchrony should preferably occur within their language learning context. This implies using stimuli as close as possible to those embedded in infants' linguistic experiences (i.e., highly redundant, containing audiovisual speaking faces of their caregivers, who are primarily female, adult, and own-race; Jayaraman & Smith, 2019; Sugden et al., 2014). The closer stimuli are to infants' real-world experiences, the more accurately tasks can measure the influence of early audiovisual speech perceptual experience on infants' sensitivity to detect the temporal synchrony embedded in it.

We suggest that using audiovisual *fluent* speech (vs. syllables or vowels) is generally preferable because it represents the most complex level of language input and, more importantly, constitutes the primary event that caregivers provide infants in their early natural interactions. Using audiovisual *fluent* speech containing critical features of infants' experience, like infant-directed speech, may offer additional insights into how they become experts in the stimuli they are exposed to. Surprisingly, only a few infant studies in the extensive literature have examined sensitivity to temporal synchrony in audiovisual *fluent* speech (see Table 2), which we hope highlights the necessity for additional research with this stimulus. Ideally, the most ecologically valid study would test infants' temporal synchrony detection in this event with 'live' eye-tracking, but implementing it is technically challenging.

Testing sensitivity to temporal synchrony in the same context where infants learn it aligns with classical theories and hypotheses of perceptual development (e.g., Bahrick & Lickliter, 2000; Gibson & Pick, 2000). A core assumption of these theories is that infants are 'perceptual learners' and temporal synchrony is a low-level property embedded in most audiovisual events (Bahrick, 2010)—and, we add, in audiovisual *fluent* speech. However, infants are not inherently sensitive to this property at birth but increasingly become across development through perceptual experience with audiovisual redundancy (Bahrick, 2010; Bahrick & Lickliter, 2014). The onset of this process of perceptual learning is well-documented, with studies showing that infants can detect temporal synchrony in non-social events (mostly dynamic sounding objects) as early as 4 weeks of age (Bahrick, 2001) and by 8 months for audiovisual *fluent* speech events (Pons & Lewkowicz, 2012; 2014, Exp. 1). Furthermore, several cross-sectional studies support this process continuing, as infants become increasingly better at detecting temporal synchrony across development. For instance, the temporal binding window (i.e., the range of asynchronies under which an auditory and a visual event are likely to be bound into a unified percept; Van Wassenhove et al., 2007) decreases during the first year of life and childhood for audiovisual syllables (Lewkowicz, 2010a; Lewkowicz & Flom, 2014), audiovisual *fluent* speech (Pons & Lewkowicz, 2014), and non-speech events (Lewkowicz, 1996).

A second implication of focusing on infants' language experience is to recognize that sensitivity to temporal synchrony in audiovisual speech is not static but develops over time, increasing significantly around the end of the first year (Pons & Lewkowicz, 2014). If one wants to track changes in this ability as infants gain perceptual experience with audiovisual speech, then the most appropriate design should be chosen. The cross-sectional research reviewed has tested numerous ages-groups, leading to mixed results and an unclear picture of the developmental trajectory of sensitivity to temporal synchrony. We suggest future studies should employ longitudinal designs to directly address the changes in this perceptual ability over time and test the hypothesis that infants' sensitivity to temporal synchrony changes across development as their audiovisual speech experience increases.

3.3.2. The functional role and underlying developmental processes

The functional role of sensitivity to temporal synchrony to audiovisual speech in development merits further investigation. If it goes beyond the perceptual domain and extends to the socio-communicative and linguistic domains (e.g., vocabulary acquisition), we should study changes in sensitivity to temporal synchrony and the potential associated changes in the developmental processes underlying these complex domains. To track these changes, it is essential to identify the most relevant time-points for testing infants' shifts in sensitivity to temporal synchrony in audiovisual speech.

Cross-sectional research on perceptual narrowing of audiovisual speech shows that infants become sensitive to temporal synchrony by the end of the first year, precisely when their perceptual system becomes attuned to native audiovisual correspondences (Pons et al., 2009). Therefore, we suggest it is necessary to longitudinally study during the first year the course of sensitivity to temporal synchrony in audiovisual speech *and* the perceptual developmental processes underlying speech perception (i.e., perceptual narrowing).

Another general developmental process resembling perceptual narrowing that is crucial for guiding future research comes from the IRH (Bahrick & Lickliter, 2002, 2014). According to Bahrick and colleagues, the development of perceptual expertise in a given domain involves *increasing specificity*, defined as the 'progressive differentiation of finer levels of stimulation as a result of perceptual experience' (Bahrick & Hollich, 2008, p. 164). This term aligns with E.J. Gibson's *differentiation process*, where infants learn from more global to more specific perceptual properties toward closer correspondences with the environment (Gibson, 2000; Gibson & Pick, 2000, p. 10). At a broader level, *increasing specificity* also recalls the mechanism of *specialization* proposed by the Neuroconstructivist approach to development (Karmiloff-Smith, 1998b; Mareschal et al., 2007), which posits that infants' brain develops following a process of progressive (neural and cognitive) specificity of processing the environmental stimuli.

Applied to audiovisual speech processing, the process of increasing specificity may reflect developmental changes in infants' prioritization of different perceptual cues. As argued by Lewkowicz (2014), changes in how infants use audiovisual cues while attuning to their native language occur in the order of increasing specificity: they first rely on more global perceptual intermodal properties (e.g., temporal cues, rhythmic cues such as 'audiovisual prosody'; see, e.g., Kitamura et al., 2014), and then transition to language-specific cues (i.e., phonetic audiovisual correspondences; Lalonde & Werner, 2021). Increasing specificity, therefore, does not mean narrowing to a specific language, but instead gradually detecting finer-grained intermodal properties.

In line with Lewkowicz (2014), we argue these two experience-dependent processes may develop coordinately during the first year. While infants become increasingly attuned to the properties of their native audiovisual speech, they may also gradually become attuned to more 'event-specific' audiovisual cues (i.e., audiovisual phonetic cues instead of temporal ones). Further research within the same infants and across the first year is needed to understand whether perceptual narrowing and increasing specificity may be intertwined developmental processes organizing audiovisual speech perception. A large longitudinal research program could investigate whether, as infants attune to the properties of their native audiovisual speech and unattune to the non-native ones (by the first year's second half; Pons et al., 2009), they rely increasingly on temporal microstructure and less on temporal macrostructure relations to succeed in audiovisual speech processing.

Some cross-sectional evidence challenges the increasing specificity principle, as infants initially do not rely mostly on low-level types of audiovisual relations in audiovisual *fluent* speech. Instead, infants do not detect temporal synchrony in this event until 8 months (Pons & Lewkowicz, 2012), right after attuning to audiovisual native speech. This striking finding suggests the relevance of further work on this topic. One possibility is that the type of audiovisual speech (complex fluent speech vs. simple syllables or vowels), the type of amodal relations on which infants rely (macro/micro-structure relations), and developmental time may interact, thus leading to more than one possible trajectory (see Addabbo et al., 2022). Yet longitudinal research is needed, our theoretical position is that reliance on temporal synchrony cues in audiovisual *fluent* speech may decrease *after* perceptual narrowing but persist when phonemic cues are not useful for infants' successful audiovisual speech processing (e.g., when seeing unfamiliar audiovisual fluent speech). Thus, the relative use of these perceptual cues would change across development and be experience and stimuli-dependent (Baart et al., 2014; Navarra et al., 2010). Current research status, yet heterogeneous, fits our view (Supplementary Table S1). During and post-perceptual narrowing (8–10 months onwards), language familiarity effects are observable in infants' audiovisual processing of *fluent* speech, but not in temporal synchrony detection.

Low-level abilities beyond perception (e.g., attention) may also shape the developmental trajectory of perceptual narrowing of audiovisual speech and sensitivity to temporal synchrony. Correspondingly, the timing of changes in infants' selective attention to speakers' mouths during the first year mirrors that in audiovisual speech perceptual narrowing (Lewkowicz & Hansen-Tift, 2012; Tomalski et al., 2013). To fully understand this developmental picture, it is necessary to first longitudinally track changes in infants' attention to the mouth, attunement to native audiovisual *fluent* speech, and the onset of sensitivity to temporal synchrony in this same event across early infancy. And, second, to explore the relationship between these perceptual and attentional abilities and their underlying mechanisms across development (e.g., specialization; Karmiloff-Smith, 1998b; Mareschal et al., 2007). Addressing these gaps warrants future implications for models of perceptual development (Gibson & Pick, 2000) and domain-general language acquisition (D'Souza et al., 2017).

Investigating these questions in neurodevelopmental disorders, particularly Autism Spectrum Disorder (ASD), is also fundamental, as early atypicalities in the basic perceptual and attentional abilities mentioned above may underlie later atypical language development, usually observed in ASD. Two clinical studies support this argument. Bradshaw et al. (2019) showed that applying socio-communicative interventions to young children with ASD (aged 18–48 months) improved their language development, which was associated with reduced mouth-looking to a static face. Yet, further investigation is needed to determine if these associations generalize to dynamic speaking faces and real-life contexts, since in 'live' eye-tracking studies with dynamic faces children with ASD show less mouth-looking than typically developing peers (Zhao et al., 2023). Newman et al., (2021) found that young children with ASD show difficulties comprehending audiovisual speech in noisy environments, but those who looked longer to the dynamic speaker's face appeared less disadvantaged. However, mouth-looking was not measured, which could be a promising future research avenue. Overall, this evidence supports our argument of close developmental intertwin of perceptual, attentional abilities, and language development also in ASD.

3.4. Limitations

This review has several limitations. First, we group all research not focusing on sensitivity to temporal synchrony as audiovisual processing research to simplify this complex field. We acknowledge this may lead to gross generalizations across non-sensitivity to temporal synchrony literature. However, this decision allowed us taking a broader perspective of the status of developmental research in audiovisual speech processing while identifying caveats to be overcome. Second, we excluded studies on non-social events and neural research, which limits the full picture of sensitivity to temporal synchrony development. We chose to focus on audiovisual *fluent* speech at a *behavioral* level for seeking homogeneity, clarity, and stimuli relevance. It remains to be seen, therefore, how the sources of confusion reviewed generalize to other stimuli and levels of analyses.

Third, while more nuanced categorizations and discussing variations within the literature would be desirable, we believe that this goal is overly ambitious for this review's scope. The main reason for this is that the developmental literature lacks a clear hierarchical organization of the nature of the stimuli used to investigate sensitivity to temporal synchrony. Several commonly antagonized dimensions such as social vs. non-social, speech vs. non-speech, simple vs. complex speech, communicative vs. non-communicative, and naturalistic vs. non-naturalistic are not easily separable and may instead overlap (see Falck-Ytter et al., 2023 for a similar problem when measuring 'social attention'). For example, audiovisual speech likely shares rhythmic complexity with other non-speech but social events (e.g., music). However, if we consider event predictability, audiovisual syllables might be closer to dynamic sounding objects (a non-social event) than to audiovisual *fluent* speech (a social event). This lack of clear categorization partly justifies the unsolved question of whether sensitivity to temporal synchrony follows independent or common developmental trajectories across the various events to which infants are exposed daily. Thus, more work is needed to categorize and manipulate these dimensions in a way that is useful for organizing our research questions but not problematic in its definitions.

4. Conclusions

This review has critically overviewed developmental literature on infants' sensitivity to temporal synchrony in audiovisual speech, identifying conceptual and methodological challenges. To advance in the field, we have introduced a hierarchical framework that distinguishes between sensitivity to temporal synchrony and audiovisual processing, along with separate operationalizations for each. Future research could benefit from taking a theoretical framework that emphasizes infants' language experiences and developmental processes. Our analytical and more streamlined approach might improve clarity, improving the refinement of future research questions and methodological approaches. It will also avoid building a developmental picture of sensitivity to temporal asynchrony that assumes linearity between the constructs and the designs, and homogeneity across the stimuli used, when this does not seem to be systematically the case (see Cox et al., 2022; Roth et al., 2022). Overall, we hope our proposal encourages a more holistic and developmental understanding of infant perceptual development in audiovisual speech processing.

Our review also suggests clinical implications for individuals with ASD, whose language difficulties may stem from atypical temporal synchrony detection and audiovisual speech processing (Righi et al., 2018; Woynaroski et al., 2013). If, as we argued, sensitivity to temporal synchrony is a basic ability nested within audiovisual processing and fundamental for language development, clinicians could intervene early at this lowest level, leading to developmental cascading effects on higher-level audiovisual functioning and language outcomes. For example, in the general population, brief multisensory training narrows the temporal binding window, enhancing long-term multimodal speech perception (Zerr et al., 2019). Single case studies published in school-aged children with ASD show promise. While explicit instructions of looking at the synchronous talking face (top-down approach) seem less efficient than in control peers (Grossman et al., 2015), some individuals show malleable temporal binding window (bottom-up approach; Feldman et al., 2020). Early bottom-up interventions may ameliorate later, more severe audiovisual speech processing challenges (Foxe et al., 2015), supporting successful navigation of educational settings.

Funding

This work was funded by a predoctoral contract awarded to IL by the Spanish Ministry of Education, Culture and Sports (Grant Number FPU13/03508; 2015–2018) and two mobility grants to IL supportive of short-term international stays (FPU 2016 mobility subprogram and Universidad Autónoma de Madrid 2018 mobility subprogram). It also benefited from the projects TRABERITEA: "Developmental trajectories in infants at risk for autism: processes of neurocognitive specialization and early interaction and phenotypic outcomes." [Grant Number PID2020–117087 GB-I00] – financed by the Spanish Ministry of Science and Innovation and the State Research Agency and TRABERITEA Project [Grant Number. PSI2015–66509-P MINECO/FEDER; 2016–2019] financed by the Spanish Ministry of Economy and Competitiveness and the European Regional Developmental Fund (ERDF) , both of which the author RC is PI. Additional work of this research has been funded by a National Science Centre of Poland grant Sonatina 7 [Grant number 2023/48/C/HS6/00264, awarded to IL]. The funding sources had no role in any of the research steps or in the preparation of this article.

CRediT authorship contribution statement

ITziar Lozano Sánchez: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Ruth Campos:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Mercedes Belinchón:** Conceptualization, Supervision, Writing – review & editing.

Declarations of Competing Interest

None.

Acknowledgements

Thanks to Zuzanna Laudańska for her valuable help on preparing this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cogdev.2024.101453.

References

Addabbo, M., Colombo, L., Picciolini, O., Tagliabue, P., & Turati, C. (2022). Newborns' ability to match non-speech audio-visual information in the absence of temporal synchrony. *European Journal of Developmental Psychology, 19*(4), 547–565. https://doi.org/10.1080/17405629.2021.1931105

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. (1999). The intermodal representation of speech in newborns. *Developmental Science, 2*(1), 42–46. https://doi.org/10.1111/1467-7687.00052

Altvater-Mackensen, N., Mani, N., & Grossmann, T. (2016). Audiovisual speech perception in infancy: The influence of vowel identity and infants' productive abilities on sensitivity to (mis) matches between auditory and visual speech cues. *Developmental Psychology, 52*(2), 191. https://doi.org/10.1037/a0039964

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition, 130*(1), 31–43. https://doi.org/10.1016/j.cognition.2013.09.006

Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior and Development, 6*(4), 429–451. https://doi.org/10.1016/S0163-6383(83)90241-2

Bahrick, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development, 10*(4), 387–416. https://doi.org/10.1016/0163-6383(87)90039-7

Bahrick, L. E. (1988). Intermodal learning in infancy: learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development, 59*(1), 197–209. https://doi.org/10.2307/1130402

Bahrick, L. E. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *Journal of Experimental Child Psychology, 53*(2), 180–199. https://doi.org/10.1016/0022-0965(92)90048-B

Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology, 79*, 253–270. https://doi.org/10.1006/jecp.2000.2588

Bahrick, L. E. (2010). Intermodal perception and selective attention to intersensory redundancy: Implications for typical social developmental and autism. In J. G. Bremner, & T. D. Wachs (Eds.), *The Wiley- Blackwell handbook of infant development* (2nd ed..,, pp. 120–166). Oxford, UK: Wiley-Blackwell. https://doi.org/10.1002/9781444327564.ch4.

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The Development of Infant Learning About Specific Face-Voice Relations. *Developmental Psychology, 41*(3), 541–552. https://doi.org/10.1037/0012-1649.41.3.541

Bahrick, L. & Hollich, G. (2008). Intermodal perception. In M. Haith y J. Benson (Eds.) Encyclopedia of Infant and Early Childhood Development (pp. 164–176). San Diego, California, USA: Academic Press.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*(2), 190. https://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior, 30*, 153–189. https://doi.org/10.1016/S0065-2407(02)80041-6

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory development* (pp. 183–206). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199586059.003.0008.

Bahrick, L. E., & Lickliter, R. (2014). Learning to attend selectively: The dual role of intersensory redundancy. *Current Directions in Psychological Science, 23*(6), 414–420. https://doi.org/10.1177/0963721414549

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13*(3), 99–102. https://doi.org/10.1111/j.0963-7214.2004.00283.x

Bahrick, L. E., Todd, J. T., & Soska, K. C. (2018). The Multisensory Attention Assessment Protocol (MAAP): Characterizing individual differences in multisensory attention skills in infants and children and relations with language and cognition. *Developmental Psychology, 54*(12), 2207. https://doi.org/10.1037/dev0000594

Bebko, J. M., Weiss, J. A., Demark, J. L., & Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *Journal of Child Psychology and Psychiatry, 47*(1), 88–98. https://doi.org/10.1111/j.1469-7610.2005.01443.x

Bradshaw, J., Shic, F., Holden, A. N., Horowitz, E. J., Barrett, A. C., German, T. C., & Vernon, T. W. (2019). The use of eye tracking as a biomarker of treatment outcome in a pilot randomized clinical trial for young children with autism. *Autism Research, 12*(5), 779–793. https://doi.org/10.1002/aur.2093

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: An International Journal of Research and Practice, 10*(1-2), 75–82. https://doi.org/10.1002/icd.249

Burnham, D., & Dodd, B. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. *Speechreading by humans and machines* (pp. 103–114). Berlin, Heidelberg: Springer,. https://doi.org/10.1007/978-3-662-13015-5_7

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 45*(4), 204–220. https://doi.org/10.1002/dev.20032

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535–543. https://doi.org/10.1016/j.tics.2007.10.001

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational Biology, 5*(7). https://doi.org/10.1371/journal.pcbi.1000436

Cox, C. M. M., Keren-Portnoy, T., Roepstorff, A., & Fusaroli, R. (2022). A Bayesian meta-analysis of infants' ability to perceive audio–visual congruence for speech. *Infancy, 27*(1), 67–96. https://doi.org/10.1111/infa.12436

D'Souza, D. E. A. N., D'Souza, H., & Karmiloff-Smith, A. (2017). Precursors to language development in typically and atypically developing infants and toddlers: the importance of embracing complexity. *Journal of Child Language, 44*(3), 591–627. https://doi.org/10.1017/S030500091700006X

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 45*(4), 187–203. https://doi.org/10.1002/dev.20033

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in-and out-of-synchrony. *Cognitive Psychology, 11*(4), 478–484. https://doi.org/10.1016/0010-0285(79)90021-5

Dorn, K., Weinert, S., & Falck-Ytter, T. (2018). Watch and listen -A cross-cultural study of audio-visual- matching behavior in 4.5-month- old infants in German and Swedish talking faces. *Infant Behavior & Development, 52*, 121–129. https://doi.org/10.1016/j.infbeh.2018.05.003

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (2022). Intersensory matching of faces and voices in infancy predicts language outcomes in young children. *Developmental Psychology, 58*(8), 1413. https://doi.org/10.1037/dev0001375

Edgar, E. V., Todd, J. T., & Bahrick, L. E. (2023). Intersensory processing of faces and voices at 6 months predicts language outcomes at 18, 24, and 36 months of age. *Infancy, 28*(3), 569–596. https://doi.org/10.1111/infa.12533

Falck-Ytter, T., Kleberg, J. L., Portugal, A. M., & Thorup, E. (2023). Social attention: Developmental foundations and relevance for autism spectrum disorder. *Biological Psychiatry, 94*(1), 8–17. https://doi.org/10.1016/j.biopsych.2022.09.035

Fava, E., Hull, R., & Bortfeld, H. (2014). Dissociating cortical activity during processing of native and non-native audiovisual speech from early to late infancy. *Brain Sciences, 4*(3), 471–487. https://doi.org/10.3390/brainsci4030471

Feldman, J. I., Dunham, K., Conrad, J. G., Simon, D. M., Cassidy, M., Liu, Y., & Woynaroski, T. G. (2020). Plasticity of temporal binding in children with autism spectrum disorder: A single case experimental design perceptual training study. *Research in Autism Spectrum Disorders, 74*, Article 101555. https://doi.org/10.1016/j.rasd.2020.101555

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H. P., Russo, N. N., Blanco, D., & Ross, L. A. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cerebral cortex, 25*(2), 298–312. https://doi.org/10.1093/cercor/bht213

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

Gibson, E. J. (2000). Perceptual learning in development: Some basic concepts. *Ecological Psychology, 12*(4), 295–302. https://doi.org/10.1207/S15326969ECO1204_04

Gibson, E. J., & Pick, A. D. (2000). *An ecological approach to perceptual learning and development.* USA: Oxford University Press,.

Grossman, R. B., Steinhart, E., Mitchell, T., & McIlvane, W. (2015). "Look who's talking!" gaze patterns for implicit and explicit audio-visual speech synchrony detection in children with high-functioning autism. *Autism Research, 8*(3), 307–316. https://doi.org/10.1002/aur.1447

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: Evidence from point-line displays of talking faces. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1275. https://doi.org/10.1037/xhp0000208

Hardcastle, W. J., & Hewlett, N. (1999). *Coarticulation: Theory, Data and Techniques.* New York, NY: Cambridge University Press,.

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2017). Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and linguistic experience. *Developmental Science, 20*(3). https://doi.org/10.1111/desc.12381

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*(3), 598–613. https://doi.org/10.1111/j.1467-8624.2005.00866.x

Homae, F., Watanabe, H., & Taga, G. (2014). The neural substrates of infant speech perception. *Language Learning, 64*(s2), 6–26. https://doi.org/10.1111/lang.12076

Imafuku, M., & Myowa, M. (2016). Developmental change in sensitivity to audiovisual speech congruency and its relation to language in infants. *Psychologia, 59*(4), 163–172. https://doi.org/10.2117/psysoc.2016.163

Imafuku, M., Kawai, M., Niwa, F., Shinya, Y., & Myowa, M. (2019). Audiovisual speech perception and language acquisition in preterm infants: A longitudinal study. *Early Human Development, 128*, 93–100. https://doi.org/10.1016/j.earlhumdev.2018.11.001

Jayaraman, S., & Smith, L. B. (2019). Faces in early visual environments are persistent not just frequent. *Vision Research, 157*, 213–221. https://doi.org/10.1016/j.visres.2018.05.005

Karmiloff-Smith, A. (1998b). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences, 2*(10), 389–398. https://doi.org/10.1016/S1364-6613(98)01230-3

Kitamura, C., Guellaï, B., & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS One, 9*(10), Article e111467. https://doi.org/10.1371/journal.pone.0111467

Kubicek, C., Hillairet de Boisferon, A., Dupierrix, E., Pascalis, O., Loevenbruck, H., et al. (2014). Cross-Modal Matching of Audiovisual German and French Fluent Speech in Infancy. *PLoS One, 9*(2), Article e89275. https://doi.org/10.1371/journal.pone.0089275

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1144. https://doi.org/10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3), 361–381. https://doi.org/10.1016/S0163-6383(84)80050-8

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences, 105*(32), 11442–11445. https://doi.org/10.1073/pnas.0804275105

Lalonde, K., & Werner, L. A. (2021). Development of the Mechanisms Underlying Audiovisual Speech Perception Benefit. *Brain Sciences, 11*(1), 49. https://doi.org/10.3390/brainsci11010049

Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development, 9*(3), 335–353. https://doi.org/10.1016/0163-6383(86)90008-1

Lewkowicz, D. J. (1996). Perception of auditory–visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance, 22*(5), 1094–1106. https://doi.org/10.1037/0096-1523.22.5.1094

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*(2), 281–308. https://doi.org/10.1037/0033-2909.126.2.281

Lewkowicz, D. J. (2010a). Infant perception of audiovisual speech synchrony. *Developmental Psychology, 46*(1), 66. https://doi.org/10.1037/a0015579

Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental psychobiology, 56*(2), 292–315. https://doi.org/10.1002/dev.21197

Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development, 85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences, 13*(11), 470–478. https://doi.org/10.1016/j.tics.2009.08.004

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. https://doi.org/10.1073/pnas.111478310

Lewkowicz, D. J., Leo, I., & Simion, F. (2010b). Intersensory perception at birth: newborns match nonhuman primate faces and voices. *Infancy, 15*(1), 46–60. https://doi.org/10.1111/j.1532-7078.2009.00005.x

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

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development, 37*(2), 90–94. https://doi.org/10.1177/0165025412467582

Lewkowicz, D. J., Schmuckler, M., & Agrawal, V. (2022). The multisensory cocktail party problem in children: Synchrony-based segregation of multiple talking faces improves in early childhood. *Cognition, 228*, Article 105226. https://doi.org/10.1016/j.cognition.2022.105226

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., Thomas, M. S., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cognition* (Vol. 1). New York, NY: Oxford University Press,.

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

Mercure, E., Kushnerenko, E., Goldberg, L., Bowden-Howl, H., Coulson, K., Johnson, M. H., & MacSweeney, M. (2019). Language experience influences audiovisual speech integration in unimodal and bimodal bilingual infants. *Developmental Science, 22*(1), Article e12701. https://doi.org/10.1111/desc.12701

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., & Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Research, 1323*, 84–93. https://doi.org/10.1016/j.brainres.2010.01.059

Newman, R. S., Kirby, L. A., Von Holzen, K., & Redcay, E. (2021). Read my lips! Perception of speech in noise by preschool children with autism and the impact of watching the speaker's face. *Journal of Neurodevelopmental Disorders, 13*, 1–20. https://doi.org/10.1186/s11689-020-09348-9

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development, 22*(2), 237–247. https://doi.org/10.1016/S0163-6383(99)00003-X

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*(1), 93–115. https://doi.org/10.1006/jecp.2001.2644

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196. https://doi.org/10.1111/1467-7687.00271

Pons, F., Andreu, L., Sanz-Torrent, M., Buil-Legaz, L., & Lewkowicz, D. J. (2013). Perception of audio-visual speech synchrony in Spanish-speaking children with and without specific language impairment. *Journal of Child Language, 40*(3), 687–700. https://doi.org/10.1017/S0305000912000189

Pons, F., & Lewkowicz, D. J. (2012). Infant perception of audiovisual synchrony in fluent speech, 36-36 *Seeing and Perceiving, 25*. https://doi.org/10.1163/187847612×646587.

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audiovisual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica, 149*, 142–147. https://doi.org/10.1016/j.actpsy.2013.12.013

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*(26), 10598–10602. https://doi.org/10.1073/pnas.0904134106

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

Righi, G., Tenenbaum, E. J., McCormick, C., Blossom, M., Amso, D., & Sheinkopf, S. J. (2018). Sensitivity to audio-visual synchrony and its relation to language abilities in children with and without ASD. *Autism Research, 11*(4), 645–653. https://doi.org/10.1002/aur.1918

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current directions in psychological Science, 17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347–357. https://doi.org/10.3758/BF03211902

Roth, K. C., Clayton, K. R., & Reynolds, G. D. (2022). Infant selective attention to native and non-native audiovisual speech. *Scientific Reports, 12*(1), 1–12. https://doi.org/10.1038/s41598-022-19704-5

Shaw, K., Baart, M., Depowski, N., & Bortfeld, H. (2015). Infants' preference for native audiovisual speech dissociated from congruency preference. *PLoS One, 10*(4), Article e0126059. https://doi.org/10.1371/journal.pone.0126059

Shaw, K. E., & Bortfeld, H. (2015). Sources of confusion in infant audiovisual speech perception research. *Frontiers in Psychology, 6*, 1844. https://doi.org/10.3389/fpsyg.2015.01844

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J., & Lewkowicz, D. J. (2012). The development of audiovisual speech perception. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory Development* (pp. 207–228). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199586059.003.0009.

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology, 15*(6), 626–636. https://doi.org/10.1037/0012-1649.15.6.626

Spelke, E. S., Born, W. S., & Chu, F. (1983). Perception of moving, sounding objects by four-month-old infants. *Perception, 12*(6), 719–732. https://doi.org/10.1068/p120719

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., Camarata, S., & Wallace, M. T. (2015). Keeping time in the brain: Autism spectrum disorder and audiovisual temporal processing. *Autism Research, 9*(7), 720–738. https://doi.org/10.1002/aur.1566

Stevenson, R. A., Segers, M., Ncube, B. L., Black, K. R., Bebko, J. M., Ferber, S., & Barense, M. D. (2018). The cascading influence of multisensory processing on speech perception in autism. *Autism, 22*(5), 609–624. https://doi.org/10.1177/1362361317704413

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., & Wallace, M. T. (2014). Multisensory temporal integration in autism spectrum disorders. *The Journal of Neuroscience, 34*(3), 691–697. https://doi.org/10.1523/JNEUROSCI.3615-13.2014

Streri, A., Coulon, M., Marie, J., & Yeung, H. H. (2016). Developmental change in infants' detection of visual faces that match auditory vowels. *Infancy, 21*(2), 177–198. https://doi.org/10.1111/infa.12104

Sugden, N. A., Mohamed-Ali, M. I., & Moulson, M. C. (2014). I spy with my little eye: Typical, daily exposure to faces documented from a first-person infant perspective. *Developmental Psychobiology, 56*, 249–261. https://doi.org/10.1002/dev.21183

Tomalski, P. (2015). Developmental trajectory of audiovisual speech integration in early infancy. A review of studies using the McGurk paradigm. *Psychology of Language and Communication, 19*(2), 77–100. https://doi.org/10.1515/plc-2015-0006

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology, 10*(5), 611–624. https://doi.org/10.1080/17405629.2012.728076

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: Differentiation of multimodal information. *Psychological Bulletin, 121*(3), 437–456. https://doi.org/10.1037/0033-2909.121.3.437

Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2019). Multisensory Integration as a Window into Orderly and Disrupted Cognition and Communication. *Annual Review of Psychology, 71*, 193–219. https://doi.org/10.1146/annurev-psych-010419-051112

Walton, G. E., & Bower, T. G. R. (1993). Amodal representation of speech in infants. *Infant Behavior and Development, 16*(2), 233–243. https://doi.org/10.1016/0163-6383(93)80019-5

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental Psychology, 34*(6), 1289–1309. https://doi.org/10.1037/0012-1649.34.6.1289

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *J Autism Dev Disord, 43*(12), 2891–2902. https://doi.org/10.1007/s10803-013-1836-5

Zerr, M., Freihorst, C., Schütz, H., Sinke, C., Müller, A., Bleich, S., Münte, T. F., & Szycik, G. R. (2019). Brief sensory training narrows the temporal binding window and enhances long-term multimodal speech perception. *Frontiers in Psychology, 10*, 2489. https://doi.org/10.3389/fpsyg.2019.02489

Zhao, Z., Tang, H., Zhang, X., Zhu, Z., Xing, J., Li, W., & Lu, J. (2023). Characteristics of visual fixation in Chinese children with autism during face-to-face conversations. *Journal of Autism and Developmental Disorders, 53*(2), 746–758. https://doi.org/10.1007/s10803-021-04985-y

Zhou, H., Yang, H., Wei, Z., Wan, G., Lui, S. S. Y., & Chan, R. C. K. (2022). Audiovisual synchrony detection for fluent speech in early childhood: An eye-tracking study. *PsyCh Journal, 11*(3), 409–418. https://doi.org/10.1002/pchj.538418