Cognition 228 (2022) 105226

Contents lists available at ScienceDirect

Cognition

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

# The multisensory cocktail party problem in children: Synchrony-based segregation of multiple talking faces improves in early childhood

David J. Lewkowicz [a,b,*], Mark Schmuckler [c], Vishakha Agrawal [a]

[a] Haskins Laboratories, New Haven, CT, USA
[b] Yale University, New Haven, CT, USA
[c] Department of Psychology, University of Toronto Scarborough, Toronto, Canada

ARTICLE INFO

*Keywords:*
Multisensory processing
Attention
Audiovisual speech integration
Perceptual segregation
Development
Cocktail party problem

ABSTRACT

Extraction of meaningful information from multiple talkers relies on perceptual segregation. The temporal synchrony statistics inherent in everyday audiovisual (AV) speech offer a powerful basis for perceptual segregation. We investigated the developmental emergence of synchrony-based perceptual segregation of multiple talkers in 3–7-year-old children. Children either saw four identical or four different faces articulating temporally jittered versions of the same utterance and heard the audible version of the same utterance either synchronized with one of the talkers or desynchronized with all of them. Eye tracking revealed that selective attention to the temporally synchronized talking face increased while attention to the desynchronized faces decreased with age and that attention to the talkers' mouth primarily drove responsiveness. These findings demonstrate that the temporal synchrony statistics inherent in fluent AV speech assume an increasingly greater role in perceptual segregation of the multisensory clutter created by multiple talking faces in early childhood.

Whenever children find themselves at a social gathering (e.g., a party, a family dinner, or in a classroom), they often see and hear multiple people talking at the same time. From a perceptual point of view, such social gatherings represent cluttered multisensory scenes. Children must be able to perceptually segregate such cluttered multisensory scenes if they are to successfully extract the communicative signals emanating from any one particular talker and if they are to respond appropriately to such signals. Perceptual segregation of such scenes involves several distinct processes. These include a rapid search of the scene, perceptual extraction of the multiple auditory and visual speech streams, integration and binding of spatiotemporally congruent auditory and visual speech streams into coherent AV entities, and finally selective attention to the most salient and relevant attributes of the multisensory scene.

The multisensory perceptual segregation problem illustrated here is theoretically similar to Cherry's (1953) Cocktail Party Problem. For Cherry, the question was: what perceptual cues enable listeners to extract one particular audible speech utterance from a set of multiple and competing audible utterances? Translated to the multisensory domain, the question is: what enables perceivers to extract one particular AV speech utterance from a set of multiple and competing AV speech utterances? To date, very few studies have examined the multisensory version of Cherry's question at the behavioral level. These studies have found that adults' and 3–5 year-old children's attention is selectively recruited by redundantly specified (i.e., temporally synchronized) AV speech and that this redundancy facilitates their perceptual segregation of multiple talkers (Alsius & Soto-Faraco, 2011; Bahrick, Soska, & Todd, 2018; Lewkowicz, Schmuckler, & Agrawal, 2021; Senkowski, Saint-Amour, Gruber, & Foxe, 2008; Zion Golumbic & Shavit-Cohen, 2019). Findings such as these raise interesting and critical questions about the development of this ability. The current study addressed these questions by investigating the development of perceptual segregation of multiple talking faces across early childhood by examining children's selective attention to such faces as a proxy for perceptual segregation.

We focused on early childhood because this is an especially interesting period for examining perceptual segregation of multisensory inputs. One of the hallmarks of this period in development is that multisensory processing is still relatively immature and that it continues to develop into late childhood (Bremner, Lewkowicz, & Spence, 2012). This is the case in the audio-visual (Barutchu, Crewther, & Crewther, 2009; Hillock, Powers, & Wallace, 2011; Hillock-Dunn & Wallace, 2012; Innes-Brown et al., 2011; Lewkowicz, 2014; Lewkowicz & Flom, 2014; Lewkowicz & Hansen-Tift, 2012; Lewkowicz, Minar, Tift, & Brandon,

* Corresponding author at: Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA.
*E-mail address:* david.lewkowicz@yale.edu (D.J. Lewkowicz).

https://doi.org/10.1016/j.cognition.2022.105226
Received 13 August 2021; Received in revised form 9 July 2022; Accepted 11 July 2022
Available online 23 July 2022

2015; Neil, Chee-Ruiter, Scheier, Lewkowicz, & Shimojo, 2006; Ross et al., 2011; Scheier, Lewkowicz, & Shimojo, 2003), the tactile-visual (Begum Ali, Spence, & Bremner, 2015; Cowie, McKenna, Bremner, Aspell, & Sterling, 2016), and the vestibular/proprioceptive-visual processing domain (Nardini, Jones, Bedford, & Braddick, 2008). As a result, the ability to segregate multisensory clutter represented by multiple talkers is likely to be immature in early childhood and improve with development. Whether this is the case is currently not known.

As noted earlier, perceptual segregation of multisensory clutter represented by multiple talkers requires a search of the scene for the talker of interest. Previous research on visual search has provided some useful general principles (Treisman, 2006; Wolfe, 2020) but, unfortunately, these principles cannot fully explain perceptual segregation of multisensory clutter. This limitation arises because such a search can benefit from the greater perceptual salience of objects and events specified by redundant multisensory cues. Such cues fall into three separate classes: amodal, modality-specific, and spatiotemporal congruence cues. Amodal cues specify equivalent auditory and visual perceptual attributes such as, for example, intensity, duration, tempo, and rhythm/prosody. Modality-specific perceptual cues specify attributes unique to their modality and include such attributes as color, temperature, timbre, or pitch. Finally, spatiotemporal congruence cues specify a common source for multisensory perceptual attributes (e.g., the sight of a hand knocking on a door and the accompanying sound of knocking). Under normal circumstances, multisensory objects or events are usually represented by various combinations of these three classes of multisensory cues. For example, a talker is usually represented by equivalent auditory and visual articulation cues (i.e., the audible and visible articulations have the same durations, tempo, rhythm, and prosody). In addition, the talker's face is always specified by a specific shape, size, and color, and the talker's voice is normally specified by a particular pitch and timbre. Finally, whenever a talker speaks, the various amodal and modality-specific multisensory cues originate, by default, from the same place and are always temporally synchronized.

Earlier it was noted that the usual multisensory redundancy of talking faces facilitates adult's perceptual segregation of multiple talking faces. These redundancy effects are largely a reflection of the fact that nervous systems have evolved to take advantage of the greater perceptual salience of redundant multisensory cues. This is illustrated by findings indicating that redundant multisensory cues augment attention, learning, and memory and that this is a species-general and age-general phenomenon (Bahrick & Lickliter, 2012; Hillairet de Boisferon, Tift, Minar, & Lewkowicz, 2017; Lewkowicz & Hansen-Tift, 2012; Murray, Lewkowicz, Amedi, & Wallace, 2016; Partan & Marler, 1999; Rowe, 1999; Senkowski et al., 2008; Stein & Stanford, 2008; Sumby & Pollack, 1954; Summerfield, 1979; Thelen, Matusz, & Murray, 2014; Van Atteveldt, Murray, Thut, & Schroeder, 2014).

Of course, redundancy benefits can only accrue if multisensory inputs are integrated and/or bound. Thus, the second process involved in perceptual segregation of multisensory clutter is integration and binding. Developmental studies have found that some relatively primitive but, nonetheless, perceptually powerful multisensory integration and binding abilities are present at birth and that they improve gradually during infancy and beyond. For example, the perception of the multisensory coherence of audible and visible sensory inputs improves substantially during infancy. This growth is illustrated by the following developmental sequence: Newborns can match non-human audible and visible vocalizations based on their temporal synchrony (Lewkowicz, Leo, & Simion, 2010), 2-month-old infants can perceive the amodal nature of audible and visible speech syllables (Kuhl & Meltzoff, 1982; Patterson & Werker, 2003), 4–10 month-old infants can perceive the temporal synchrony of audible and visible speech syllables (Lewkowicz, 2010), and 12–14 month-old infants, but not younger ones, can match fluent audible and visible speech streams (Lewkowicz et al., 2015).

The early multisensory integration abilities are also evident in infants' growing tendency to take advantage of AV redundancy. This developmental sequence is illustrated by the fact that, starting around 6 months of age, infants begin to deploy their selective attention preferentially to the talker's mouth, the source of temporally synchronized AV speech (Hillairet de Boisferon et al., 2017; Hunnius & Geuze, 2004; Lewkowicz & Hansen-Tift, 2012; Pons, Bosch, & Lewkowicz, 2015; Tenenbaum et al., 2015; Tenenbaum, Shah, Sobel, Malle, & Morgan, 2013) and that infants as well as young children rely on the greater perceptual salience of AV speech to overcome the challenge of learning more than one language (Birulés, Bosch, Brieke, Pons, & Lewkowicz, 2019; Pons et al., 2015). Crucially, the early emerging dependence on the beneficial effects of redundantly specified AV speech continues into adulthood. This is exemplified by the fact that an initial face-voice pairing facilitates adults' subsequent voice recognition (Von Kriegstein & Giraud, 2006), that adults comprehend audible speech better when it is co-specified by congruent visible speech (Lansing & McConkie, 2003; Summerfield, 1992), that adults exhibit better detection of auditory speech presented in noise when it is accompanied by corresponding visual speech (Grant & Seitz, 2000; MacLeod & Summerfield, 1987; Rennig, Wegner-Clemens, & Beauchamp, 2020; Shahin & Miller, 2009; Sumby & Pollack, 1954), and that adults deploy more attention to a talker's mouth when speech processing becomes more challenging (Barenholtz, Mavica, & Lewkowicz, 2016; Birulés, Bosch, Pons, & Lewkowicz, 2020).

The hypothesis tested here is that children's perceptual segregation of multiple competing talking faces is likely to be facilitated by the normally tight temporal synchrony statistics linking the dynamic variations of a talker's audible speech stream and the accompanying visible speech stream (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Summerfield, 1987, 1992; Yehia, Kuratate, & Vatikiotis-Bateson, 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998). This hypothesis is based on the fact that temporal synchrony statistics are a powerful multisensory binding cue that are easy to detect and highly effective in specifying the perceptual coherence of our multisensory world (Spence & Squire, 2003; Vroomen & Keetels, 2010; Wallace, Woynaroski, & Stevenson, 2020). In addition, such statistics are unique for different talkers articulating different utterances. As a result, if a perceiver can integrate and bind the visible and audible speech streams belonging to one particular talker - and not those belonging to different people - then this perceiver will be able to appropriately segregate a cluttered multisensory scene consisting of multiple talkers. Of course, as noted at the outset, successful segregation requires that one attend selectively to one attribute of a scene or event and simultaneously ignore or suppress another attribute (Murphy, Groeger, & Greene, 2016).

Studies of adults' search and segregation of complex visual scenes composed of multiple objects and accompanying sounds have found that AV temporal synchrony is, indeed, a powerful binding and segregation cue. For instance, Van der Burg and colleagues have examined adults' search for a target object embedded in a scene consisting of many other objects in the absence of a sound versus the presence of a sound synchronized with the target's actions. Search for a target object embedded in a crowded scene composed of many other objects was markedly facilitated by a sound that was synchronized with the actions of the target object but not synchronized with the actions of the distractor objects (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008b). These findings prompted Van der Burg et al. to conclude that the visual salience of the target stimulus is augmented by the integration of the sound with the target stimulus and that this bottom-up process leads to automatic attentional capture. Consisistent with this interpretation, in other studies, Van der Burg et al. reported that top-down factors do not mediate this type of AV integration, that the AV integration is automatic, that improved search is not the result of increased alertness or top-down temporal cueing, and that the greater speed of visual search in the presence of temporally synchronized sounds occurs early in sensory processing in parieto-occipital cortices (Van der Burg et al., 2008b; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008a; Van der Burg,

Talsma, Olivers, Hickey, & Theeuwes, 2011).

Although the pop-out effects reported by Van der Burg et al. are impressive, some have suggested that those effects may not only reflect the operation of a bottom-up process but a top-down one as well. For example, Matusz and Eimer (2011) pointed out that the reaction times obtained by Van der Burg et al. (2008b) were longer to targets in tone-present trials than in typical pop-out tasks, that reaction times had non-flat search slopes, and that the likelihood that the sounds increased search efficiency depended on the probability of their co-occurrence with targets or distractors. Based on these observations, Matusz and Eimer (2011) suggested that top-down factors influenced responsiveness and that search reflected an effortful process. To test their claims, they employed a multi-stimulus spatial cueing task and found evidence of multisensory enhancement of attentional capture driven by a salience-driven bottom-up process. In this particular case, the bottom-up cue was the temporal synchrony of the sound and the target. Together, the Van der Burg et al. and Matusz and Eimer findings demonstrate that AV temporal synchrony can facilitate the perceptual segregation of relatively simple objects accompanied by simple sounds and raise the obvious question of whether the temporal synchrony statistics inherent in fluent AV speech might also facilitate the perceptual segregation of more complex events such as those consisting of multiple talkers.

More direct evidence of the power of AV temporal synchrony to facilitate search and segregation comes from a study by Lewkowicz et al. (2021) who examined adults' perceptual segregation of multiple talking faces. Participants either viewed four identical or four different talking faces speaking the same utterance in a temporally jittered fashion. The talking faces were accompanied by the audible version of the visible utterance. During half the trials, the audible utterance was synchronized with one of the talking faces (synchrony condition) while during the other half of the trials it was desynchronized with all of the faces (asynchrony condition). Eye tracking measures of attention, operationalized as the amount of time spent looking at each of the four talking faces, indicated that participants directed more of their selective attention to the audiovisually synchronized talking face than to the desynchronized talking faces in the synchrony condition and that they exhibited no preference in the asynchrony condition. Overall, in the synchrony condition, participants gazed an average of 80% of their time at the audiovisually synchronized talking face and they gazed more at the mouth than the eyes. In addition, they gazed more at the eyes of an audiovisually synchronized talking face in the synchrony condition but they gazed more at the mouth in the asynchrony condition. These eyes/mouth findings are consistent with results from other studies showing that adults gaze more at the eyes of an audiovisually synchronized talking face when they are not engaged in speech processing per se (Lewkowicz & Hansen-Tift, 2012; Võ, Smith, Mital, & Henderson, 2012) but that they gaze more at the mouth when they are engaged in speech processing (Barenholtz et al., 2016; Birulés et al., 2020).

The Lewkowicz et al. (2021) findings raise the following developmental questions: (a) when does the ability to perceptually segregate multiple talking faces on the basis of AV temporal synchrony statistics emerge, (b) is such perceptual segregation reflected in a preference for the synchronized talking face and, if so, to what extent, and (c) is selective attention to the eyes versus the mouth affected by the temporal relation between the audible and visible speech streams associated with competing talking faces and does any such selective attention change with development? We attempted to answer these questions in the current study by using the same experimental approach employed by Lewkowicz et al. (2021) to study 3–7-year-old children's perceptual segregation of multiple talking faces. Prior studies of children's perception of AV temporal synchrony have found that children at these ages can detect temporal synchrony relations in isolated AV speech syllables (Lewkowicz & Flom, 2014) as well as in AV non-speech events (Hillock et al., 2011; Powers, Hillock, & Wallace, 2009) and that sensitivity to synchrony is poorer earlier than later in development. Therefore, we made four a priori primary predictions and three corrolary predictions arising from the third primary prediction. The four main predictions were as follows: (1) children can perceptually segregate multiple talking faces based on AV temporal synchrony statistics and, therefore, should exhibit a preference for the audiovisually synchronized talking face, (2) even though children's preference for the audiovisually synchronized talking face is likely to be less robust than that observed in adults, the magnitude of this preference should increase with age, (3) selective attention to the eyes and mouth is likely to change with age and should depend on whether a canonical, temporally synchronized talking face is present in the search array or not, and (4) response latency to the canonical, temporally synchronized talking face should be faster than to the desynchronized faces if responsiveness is mediated by an automatic, bottom-up, pop-out process. The three corrolary predictions of Prediction (3) were that selective attention: (i) should decrease to the talker's eyes and increase to the mouth as a function of age because AV speech processing is known to improve with age, (ii) should be greater to the mouth than to the eyes given that the experimental task required explicit AV speech processing, and (iii) should be greater to the mouth in the asynchrony than synchrony condition because the task requires an active search for the canonical, temporally synchronized, talking face.

We conducted two experiments to test our predictions. In Experiment 1, children watched four identical faces visibly articulating the same utterance and heard the same audible utterance that was either synchronized with one of the visible utterances or not synchronized with any of the utterances. In Experiment 2, to determine whether audible and visible identity cues might play a role in responsiveness, children watched four different faces visibly articulating the same utterance and heard the same audible utterance whose acoustic attributes differed depending on which person was talking during a given trial. As in Experiment 1, the audible utterance was either temporally synchronized with one of the four visible utterances or not synchronized with any of them. The primary aim of both experiments was to investigate synchrony-based perceptual segregation of multiple talking faces. We operationalized perceptual segregation as a preference for one of the talking faces and measured it with an eye tracker. The eye tracker provided a measure of selective attention as a proxy of preference for the canonical, temporally synchronized talking face. In addition, the eye tracker permitted us to measure selective attention to the eyes and mouth of each of the faces and, thus, made it possible for us to gain greater insight into the perceptual mechanisms underlying perceptual segregation.

## 1. Experiment 1

To test whether children would perceptually segregate multiple talking faces based solely on AV temporal synchrony, we temporally jittered the visible articulations of four talking faces and synchronized the concurrently presented audible utterance with one of these faces. Two crucial aspects of the experimental design are noteworthy. First, the task instructions provided to the children made no explicit reference to the temporal relation between the visible and audible utterances. The children were only told that they would see four people talking while they heard one person talking and were asked to watch carefully to see if they could match the face to the voice. To encourage children to actively search the scene composed of the four talking faces, the children were told that they would see a static picture of the four faces at the end of each trial and that their task was to point to the face that was talking during the immediately preceding video presentation. Second, the fact that the four talking faces were identical, that their visible articulations were temporally jittered, and that all four faces articulated the same utterance meant that the only perceptual cue that linked one of the talking faces to the concurrent audible utterance was temporal synchrony. This feature of our experimental design enabled us to control for the possibility that distinctive facial cues and/or distinct phonetic, semantic, and/or prosodic cues might contribute to responsiveness.

### 1.1. Method

#### 1.1.1. Participants

A total of 189 children contributed usable data in the current experiment. We tested separate groups of 3-year-old ($N = 26$; 13 girls), 4-year-old ($N = 43$; 26 girls), 5-year-old ($N = 57$; 25 girls), 6-year-old ($N = 32$; 19 girls), and 7-year-old ($N = 31$; 19 girls) children. We tested an additional 28 children but these participants either did not contribute usable data or their data were not used. Exclusion occurred for multiple reasons. These included technical problems ($N = 6$), children's failure to finish the experiment ($N = 10$), children not meeting our age requirement but tested anyway because of their desire to "play the game" ($N = 5$), a disqualifying disability ($N = 5$), or repeated participation in the study ($N = 2$). All children were recruited and tested in the Living Laboratory at the Boston Museum of Science except two who were tested in the first author's university laboratory. A parent or guardian provided consent prior to the child participating in the study and each child gave his or her verbal assent.

#### 1.1.2. Apparatus and stimuli

Children sat at a table in front of a REDn SensoMotoric Instruments remote eye tracker (SMI, Teltow, Germany) located approximately 60 cm from their eyes. The eye tracker was controlled by a Dell Precision M4800 laptop computer and ran at a 60 Hz sampling rate. The eye tracker's camera was mounted at the bottom of the computer screen and SMI's iViewRed software controlled the camera and processed the eye gaze data. SMI's Experiment Center software controlled stimulus presentation and data acquisition. The visual stimuli were presented on the computer's 11 × 13 in screen. The initial instructions and auditory stimuli were presented through a pair of Sony Professional headphones (Model # MDR-7506) at a comfortable listening level. The headphones also made it possible to mask a good bit of the extraneous noise normally present in the Living Laboratory environment and allowed the children to focus on the task at hand.

The experiment consisted of a calibration phase, two 15 s practice trials, and eight 15 s test trials. During the calibration phase, children saw a sequence of small yellow stars presented one at a time, first in the center of the screen and then in each of the four corners of the screen. The stimuli for the practice and test trials were composite videos created with Adobe's Premiere software (Adobe Systems, Inc., San Jose, CA). As can be seen in Fig. 1, each composite video consisted of four equally sized videos of the same female face presented in each quadrant of the screen (please note that the face and voice presented in the composite videos during the practice trials were of a different female actor than the one presented during the test trials). All four faces in the composite video for each test trial articulated the same utterance and the audible utterance heard during each trial was the same as the visible utterance.

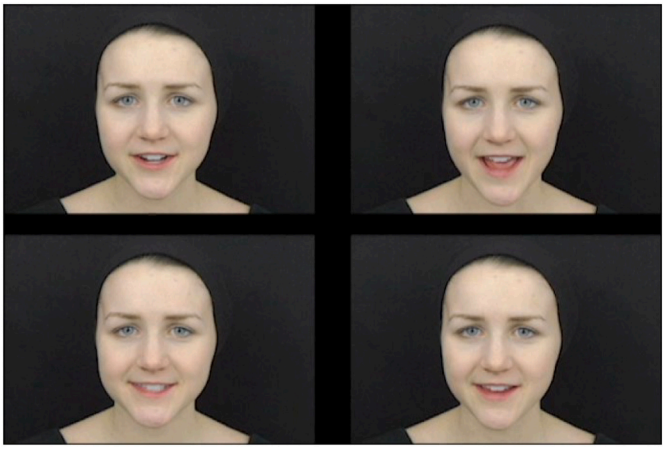

**Fig. 1.** Screenshot of one of the composite videos presented in Experiment 1.

The test trials consisted of four synchrony and four asynchrony trials. In the synchrony test trials, the audible utterance was temporally synchronized with the visible articulations of one of the four talking faces (the target) but desynchronized with the visible articulations of the other three talking faces (the distractors). In the asynchrony test trials, the audible utterance was desynchronized with the visible articulations of all four talking faces. For comparison purposes, the face that was presented in the same quadrant as the target face in the synchrony trials was labelled the "virtual" target. At the start of each of the two types of test trials, all four faces began articulating simultaneously. Crucially, however, these visible articulations were temporally jittered with respect to one another. This temporal jitter was implemented by starting the visible articulations produced by each respective distractor face increasingly later into the utterance relative to the start of the visible utterance produced by the actual target face in the synchrony test trials, and relative to the start of the virtual target face in the asynchrony test trials. The result of temporally jittering the four faces was that the visible speech stream produced by each distractor face was temporally delayed with respect to the auditory speech stream by some fixed interval of time (i.e., the auditory speech stream led the visible speech stream produced by each face by some fixed interval of time). [1] The main purpose of the temporal jitter procedure was to ensure that the visible articulations produced by all four faces began simultaneously at the start of each test trial and that only the visible articulations of the target face were synchronized with the audible utterance in the synchrony test trials. It should be noted that, even though some of the jitter intervals differed from one another by less than 1 s, the videos were clearly perceptually dissociable. In addition, the jitter procedure created the impression that the four talking faces were articulating different utterances and thus made it more difficult to detect the face that was temporally synchronized.

Importantly, desynchronization of fluent AV speech disrupts the zero-lag temporal correlation between the dynamic variations in a talker's visible mouth movements and the accompanying audible vocalizations. The practical effect of desynchronization is that it creates multiple points of intersensory discordance between the physical, phonetic, prosodic, and semantic attributes of the visible and audible speech streams which, under normal conditions, are congruent. In addition, it should be noted that the movements of the vocal tract normally precede phonation by 100–300 ms (Chandrasekaran et al., 2009). This means that the onset of the audible component of a synchronized AV speech utterance is normally delayed relative to the onset of the visible component of a speech utterance. This, in turn, means that the audible and visible speech streams of a fluent AV speech stream must be temporally separated by more than 300 ms if they are to be perceived as desynchronized. Accordingly, to ensure that children perceived AV desynchronization, the audible and visible speech streams for each distractor talking face were jittered by more than 300 ms with respect to one another.

To produce the composite videos, we filmed each of the two test-phase female actors uttering two different sets of two different

[1] The intervals for the two monologues spoken by one of the actors were 2200, 3300, and 4400 ms in both the synchrony and asynchrony test trials while the interval for the asynchronous version of the target stimulus in the asynchrony trials was 1800 ms. The intervals for one of the monologues spoken by the second actor were 1966, 2966, and 3899 ms in the synchrony and asynchrony test trials while the interval for the asynchronous version of the target stimulus in the asynchrony trials was 1799 ms. The intervals for the second monologue spoken by the second actor were 966, 1766, and 2633 ms in the synchrony and asynchrony test trials while the interval for the asynchronous version of the target stimulus in the asynchrony trials was 2933 ms.

utterances. [2] This yielded a total of four different videos reflecting the four different actor/utterance combinations (see Video S1 for an example of one of the videos). We then used each of these four videos to construct four pairs of test trials. Each pair consisted of a synchrony and an asynchrony test trial. The two types of test trials making up each pair were identical except that the utterance spoken by the target face was temporally synchronized with the audible utterance in the synchrony test trial but desynchronized with it in the asynchrony test trial. Pairing the composite videos in this way provided two advantages. First, it enabled us to compare responsiveness to an audiovisually synchronized target face in one specific quadrant vs. responsiveness to the audiovisually desynchronized virtual target face in the identical quadrant. Second, it helped maximize children's ability to detect the perceptual difference between the synchronized and asynchronized test trials.

#### 1.1.3. Procedure and design

An experimenter was seated next to the child and monitored the experiment. Unless the child posed a specific question, the experimenter did not interact with the child. The experiment began with the calibration routine and calibration was deemed acceptable if the point of fixation fell within less than one degree of visual angle of the star's actual position. All data were acquired from the right eye. The calibration phase was followed by an instruction phase. During this phase, children were given a series of pre-recorded audible instructions via the headphones and were shown explanatory illustrations on the computer screen at the same time. The instructions were as follows: "Today, we are going to play a game. The game will only continue if you are looking at the screen. If you look away, it will stop. You will see four people talking but you will only hear one person talking. Watch carefully and see if you can match the face to the voice. Then, point to the face that you thought was talking. Sometimes it will be easy and sometimes hard. Watch carefully!" During these instructions, children saw a still and silent image of four identical faces.

The practice phase consisted of two trials during which children saw composite videos of a female actor who was not the same as the actors presented during the test trials. During the first practice trial, they saw the four faces articulating the same utterance and they heard the audible version of the same utterance. The audible utterance was synchronized with one of the talking faces and desynchronized with the other three talking faces. During the second practice trial, children saw the same four talking faces again except that this time the audible utterance was desynchronized with all four talking faces. After each practice trial, children saw a composite video that consisted of the same four faces that they saw in the composite talking-video presented in the preceding trial. In this composite video, however, the four faces that were presented in each quadrant were now largely still and only blinked occasionally. The children were now asked to look at these faces and point to the face that was talking in the preceding composite video and were not given any feedback. Once the practice trials were completed, the children were allowed to ask questions and then the experiment proper began.

We used a Latin Square design to order the four pairs of synchrony/asynchrony test trials into four different "Quadrant" groups. The Quadrant groups differed in terms of the location of target-face presentation for each unique pair of synchrony/asynchrony test trials, with two constraints. The first constraint was that, within each group, the quadrant of target-face presentation for each pair of synchrony/asynchrony test trials had to be counterbalanced across the four quadrants. The second constraint was that, across the four groups, the specific quadrant in which the targets in a given pair of test stimuli were presented had to be counterbalanced. Assignment to one of the four Quadrant groups was determined randomly.

Immediately following each test trial, children saw a still image of the four faces from the preceding test trial and were asked to point to the face that they thought was talking in the preceding test trial. They were given no feedback. The sole purpose of asking the children to point to the talking face was to induce them to attend to the displays and to actively engage in a search for the talking face during the test trial. Given that this was the only purpose for asking the children to point and that our primary aim was to investigate children's selective attention to the talking faces in the presence of an audible utterance, we did not record the children's pointing choices.

To measure selective attention, we created three areas-of-interest (AOIs). One AOI was for each of the four faces while the other two AOIs were for the eye and mouth regions of each face (see Fig. 2). We used the total amount of looking at each AOI to derive two sets of dependent measures for each test trial. The first set of dependent measures consisted of the proportion of total looking time (PTLT) directed to each of the four talking faces. These PTLT scores were computed by dividing the total amount of looking at each respective face AOI by the total amount of looking at the four face AOIs. The second set of dependent measures consisted of the PTLT directed to the eyes and mouth of each respective face. These PTLT scores were computed by dividing the total amount of looking at the eyes and mouth, respectively, by the total amount of looking at that particular face.

### 1.2. Results

As indicated earlier, adults exhibit robust perceptual segregation of the same types of talking-face arrays that were presented in the current experiment (Lewkowicz et al., 2021). Thus, one aim of the current experiment was to investigate whether children also might be able to perceptually segregate these types of talking-face arrays and the second aim was to examine the mechanisms underlying perceptual segregation. As indicated in the Introduction, we made four primary predictions and three corrolary ones associated with Prediction (3). Prediction (1) was that children should exhibit a preference for the audiovisually synchronized talking face when such a face is available in the array. Prediction (2) was that the magnitude of the preference is likely to increase

**Fig. 2.** Screenshot of one of the composite videos of the four talking faces and the AOIs corresponding to the face, eyes, and mouth.

[2] The four utterances spoken by the actors were as follows: (1) "But your favorite will be the elephants. They're big and gray and have large floppy ears. Maybe we'll see a baby elephant too? What do you think about that? If not, we could go to story time at the library. All your friends will be there"; (2) "They like to ice skate, right? But, before we can go anywhere, what do we have to do? Change your clothes and eat breakfast, of course. It's cold outside, so you need to wear a sweater. How about the green one with the duck? For breakfast, you can have oatmeal with blueberries."; (3) "Good morning, get up, come on now. If you get up right away, we'll have an hour to play in the house. I love these long mornings, don't you? I wish they could last all day."; (4) "We can hang around all day Saturday. Except, of course, for the party. Are you going to help me fix up the house? Are you? We need to buy flowers, prepare the food, vacuum the house, dust."

with age. Prediction (3) and its corrolaries were that children would exhibit developmental changes in selective attention to the eyes and mouth and that they would attend more to the mouth than eyes given that the experimental task involved AV speech processing. Finally, Prediction (4) was that response latency to the canonical, temporally synchronized talking face should be faster than to the desynchronized faces if responsiveness is mediated by an automatic, bottom-up, pop-out process

To test our predictions, we conducted two separate sets of analyses. To test Predictions (1) and (2), the first set consisted of analyses of the Face PTLT scores. To test Prediction (3), the second set of analyses consisted of analyses of the Eye and Mouth PTLT scores. Each set consisted of an initial overall analysis of variance (ANOVA) of the PTLT scores to investigate where children deployed their selective attention and how this was affected by the theoretically relevant factors. This was then followed by specific statistical comparisons intended to test our predictions and to clarify the significant effects found in the overall ANOVA. Finally, to test Prediction 4, we compared the amount of time it took the children to deploy their first look to the target stimulus in the synchrony versus the asynchrony condition. If the temporally synchronized target stimulus is perceptually more salient than the temporally desynchronized one, it is reasonable to expect that the former might attract selective attention faster than the latter.

### 1.2.1. Face AOIs

To determine whether children preferred the audiovisually synchronized face, we compared the PTLT scores for the target face with the average of the PTLT scores for the three distractor faces. First, we performed a preliminary mixed, repeated-measures analysis of variance (ANOVA) with Synchrony Condition (2; Synchrony, Asynchrony), Actor (2), Utterance (2), and Stimulus Type (2; Target, Distractor) as within-subjects factors and Age (5) and Quadrant Group (4) as between-subjects factors. Results of this preliminary analysis [3] indicated that Actor, Utterance, and Quadrant Group interacted with some of the other factors but that none of these interactions were theoretically meaningful. As a result, we collapsed the data across these three factors and re-analyzed the data with a new mixed, repeated-measures ANOVA, with Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Age (5) as a between-subjects factor. The results of this analysis yielded a main effect of Age, $F(4, 184) = 14.12$, $p < .001$, $\eta_p^2 = 0.23$, Synchrony Condition, $F(1, 184) = 188.00$, $p < .001$, $\eta_p^2 = 0.51$, and Stimulus Type, $F(1, 184) = 242.08$, $p < .001$, $\eta_p^2 = 0.57$. The analysis also yielded a Synchrony Condition x Age interaction, $F(4, 184) = 14.92$, $p < .001$, $\eta_p^2 = 0.24$, a Stimulus Type x Age interaction, $F(4, 184) = 16.29$, $p < .001$, $\eta_p^2 = 0.26$, a Synchrony Condition x Stimulus Type interaction, $F(1, 184) = 191.07$, $p < .001$, $\eta_p^2 = 0.51$, and a Synchrony Condition x Stimulus Type x Age interaction, $F(4, 184) = 14.55$, $p < .001$, $\eta_p^2 = 0.24$.

From an a priori theoretical standpoint, and based on the results from studies demonstrating age-related changes in children's responsiveness to multisensory inputs (see Introduction), the most relevant finding is the statistically significant Stimulus Type x Synchrony Condition x Age interaction. This triple interaction is depicted in Fig. 3. As can be seen, in the synchrony condition, children gazed more at the target face than at the distractor faces at all ages and the magnitude of the difference in gazing at the target versus the distractor faces increased with age. In contrast, in the asynchrony condition, children gazed equally at the target and distractor faces at all ages. Separate repeated-measures ANOVAs indicated that the Stimulus Type x Age interaction was significant in the synchrony condition, $F(4, 184) = 19.96$, $p < .001$, $\eta_p^2 = 0.30$, but that it was not significant in the asynchrony condition, $F(4, 184) = 0.25$, *ns*. Simple effects ANOVAs showed that the age-related increase in gazing at the target face was significant, $F(4, 184) = 19.86$, $p < .001$, $\eta_p^2 = 0.30$, and that the age-related decrease in gazing at the distractor faces also was significant, $F(4, 184) = 18.91$, $p < .001$, $\eta_p^2 = 0.29$.

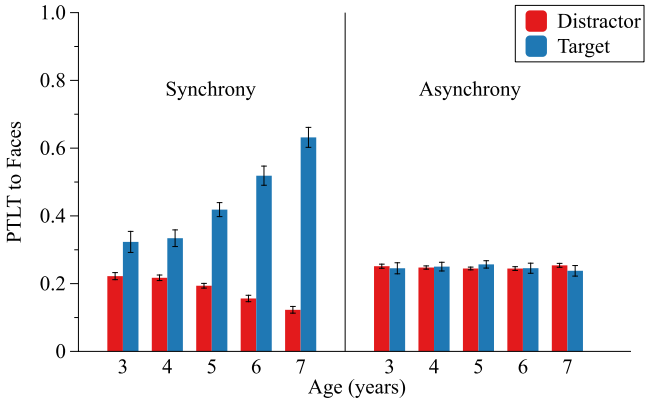


**Fig. 3.** Mean proportion of total looking time at the distractor and target faces as a function of age across the two synchrony conditions in Experiment 1. Error bars represent the standard errors of the mean.

To further probe the data, we conducted a set of planned comparisons of the data from each synchrony condition, respectively, to determine at what age children gazed longer at the target face than the distractor faces. Results of these comparisons indicated that, in the synchrony condition, children gazed more at the target face than at the distractor face at each age (3-year-olds: $F(1, 184) = 6.23$, $p < .05$; 4-year-olds: $F(1, 184) = 16.69$, $p < .001$; 5-year-olds: $F(1, 184) = 67.76$, $p < .001$; 6-year-olds: $F(1, 184) = 97.33$, $p < .001$; and 7-year-olds: $F(1, 184) = 173.97$, $p < .001$). In contrast, in the asynchrony condition, the results of the planned comparisons showed that children gazed equally at the target and distractor faces at each age (3-year-olds: $F(1, 184) = 0.039$, *ns*; 4-year-olds: $F(1, 184) = 0.093$, *ns*; 5-year-olds: $F(1, 184) = 0.17$, *ns*; 6-year-olds: $F(1, 184) = 0.16$, *ns*; and 7-year-olds: $F(1, 184) = 0.56$, *ns*).

### 1.2.2. Eye and mouth AOIs

To examine deployment of eye gaze to the eyes and mouth, first we conducted a preliminary repeated-measures ANOVA of the PTLT scores for the mouth and eye AOIs, with AOI (2), Synchrony Condition (2), Utterance (2), Actor (2), and Stimulus Type (2) as within-subjects variables and Age (5) and Quadrant Group (4) as between-subjects variables. As was the case with the face AOIs, the purpose of this analysis was to determine whether the utterance spoken, the actor speaking it, and/or the quadrant in which the target face was presented affected responsiveness. Results of this analysis indicated that none of these three factors, either alone or in interaction with other factors, played any theoretically meaningful role in responsiveness.

Given that the theoretically unimportant factors did not affect responsiveness, we collapsed the PTLT scores for the mouth and eye AOIs over these factors and performed a new repeated-measures ANOVA, with AOI (2), Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Age (5) as a between-subjects factor. This analysis yielded main effects of Age, $F(4, 184) = 3.78$, $p < .01$, $\eta_p^2 = 0.07$,

[3] Results of the preliminary analysis yielded significant main effects of Age, $F(4, 169) = 15.42$, $p < .001$, $\eta_p^2 = 0.267$, Synchrony Condition, $F(1, 169) = 188.11$, $p < .001$, $\eta_p^2 = 0.53$, and Stimulus Type, $F(1, 169) = 239.06$, $p < .001$, $\eta_p^2 = 0.58$, significant two-way interactions including an Utterance x Quadrant Group interaction, $F(3, 169) = 17.58$, $p < .001$, $\eta_p^2 = 0.24$, a Stimulus Type x Age interaction, $F(4, 169) = 17.60$, $p < .001$, $\eta_p^2 = 0.29$, and a Stimulus Type x Synchrony Condition interaction, $F(1, 169) = 187.38$, $p < .001$, $\eta_p^2 = 0.52$, and three-way interactions including an Actor x Utterance x Quadrant Group interaction, $F(3, 169) = 18.19$, $p < .001$, $\eta_p^2 = 0.24$, an Utterance x Stimulus Type x Quadrant Group interaction, $F(3, 169) = 16.14$, $p < .001$, $\eta_p^2 = 0.22$, and a Stimulus Type x Synchrony Condition x Age interaction, $F(4, 169) = 15.49$, $p < .001$, $\eta_p^2 = 0.27$.

AOI, $F(1, 184) = 23.32$, $p < .001$, $\eta_p^2 = 0.11$, and Stimulus Type, $F(1, 184) = 4.97$, $p < .05$, $\eta_p^2 = 0.03$. The analysis also yielded an AOI x Age, $(1, 184) = 7.60$, $p < .05$, $\eta_p^2 = 0.14$, an AOI x Synchrony Condition, $F(1, 184) = 60.61$, $p < .001$, $\eta_p^2 = 0.25$, and a Synchrony Condition x Stimulus Type, $F(1, 184) = 5.06$, $p < .05$, $\eta_p^2 = 0.03$, interaction. Finally, the analysis yielded an AOI x Synchrony Condition x Stimulus Type, $F(1, 184) = 5.76$, $p < .05$, $\eta_p^2 = 0.03$, interaction and an AOI x Synchrony Condition x Age, $F(4, 184) = 3.92$, $p < .01$, $\eta_p^2 = 0.08$, interaction.

*Children's selective attention to the mouth dominates responsiveness, especially when a canonical, temporally synchronized talking face is absent.* Theoretically speaking, one of the two most relevant findings from the overall ANOVA was the AOI x Synchrony Condition x Stimulus Type (see Fig. 4). Planned comparisons of the data seen in Fig. 4 indicated that children gazed more at the mouth than the eyes in the synchrony, $F(1, 184) = 5.62$, $p < .025$, as well as in the asynchrony condition, $F(1, 184) = 46.15$, $p < .001$, that they gazed more at the mouth in the asynchrony than in the synchrony condition, $F(1, 184) = 34.59$, $p < .001$, and that they gazed more at the eyes in the synchrony than the asynchrony condition, $F(1, 184) = 40.35$, $p < .001$. Finally, planned comparisons indicated that, in the synchrony condition, children gazed marginally more at the eyes of the target than the distractor, $F(1, 184) = 3.43$, $p = .065$, and more at the mouth of the target than the distractor, $F(1, 184) = 4.28$, $p < .05$. In the asynchrony condition, children gazed more at the eyes of the target than the distractor, $F(1, 184) = 6.65$, $p < .05$, but they gazed more at the mouth of the distractor than the target, $F(1, 184) = 4.82$, $p < .05$.

*As children grow, they increase their selective attention to the mouth while they decrease their selective attention to the eyes.* Theoretically speaking, the second of the two most relevant findings from the overall ANOVA was the AOI x Synchrony Condition x Age interaction (see Fig. 5). As can be seen, the preference for the mouth appears to emerge by 6 years of age in the synchrony condition and by 5 years of age in the asynchrony condition. Indeed, planned comparisons in the synchrony condition, showed that the 3-, 4-, and 5-year-olds did not prefer the mouth [$F(1, 184) = 2.18$, $p = .14$, *ns*; $F(1, 184) = 0.09$, $p = .76$, *ns*; $F(1, 184) = 0.024$, $p = .88$, *ns*, respectively) but that the 6- and 7-year-olds did prefer the mouth [$F(1, 184) = 13.58$, $p < .001$; $F(1, 184) = 9.69$, $p < .01$, respectively]. The planned comparisons in the asynchrony condition indicated that neither the 3- nor the 4-year-olds preferred the mouth [$F(1, 184) = 0.03$, $p = .86$, *ns*; $F(1, 184) = 0.00$, $p = .98$, *ns*, respectively] but that the 5-, 6-, and 7-year-olds did prefer the mouth [$F(1, 184) = 9.00$, $p < .01$; $F(1, 184) = 34.97$, $p < .001$; $F(1, 184) = 36.61$, $p < .001$, respectively].

*Except for the 4-year-olds, children at each age attended more to the mouth when a canonical talking face was absent and more to the eyes when it was present.* To further explore the apparent age-related patterns seen in

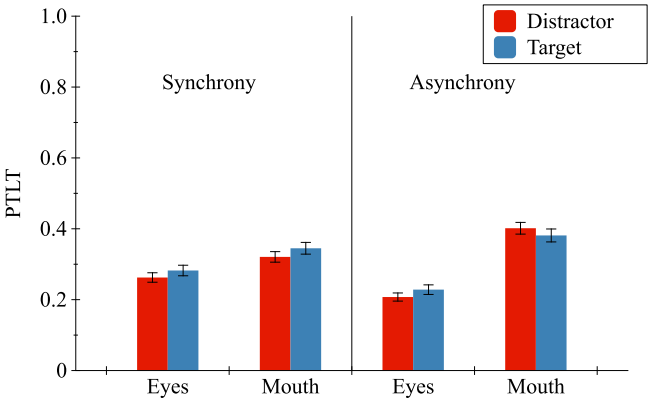


**Fig. 4.** Mean proportion of total looking time to the distractor- and target-face eyes and mouth across the two synchrony conditions in Experiment 1. Error bars represent the standard errors of the mean.

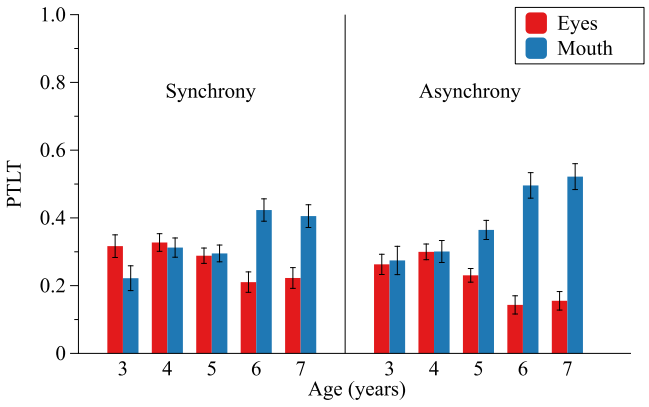


**Fig. 5.** Mean proportion of total looking time at the eyes and mouth as a function of age across the two synchrony conditions in Experiment 1. Error bars represent the standard errors of the mean.

Fig. 5, we conducted separate analyses of the PTLT scores for the eye and mouth AOIs, respectively, across the two synchrony conditions separately. Each of these analyses were repeated-measures ANOVAs, with Stimulus Type (2) and Synchrony Condition (2) as within-subjects factors and Age as a between-subjects factor.

The ANOVA of the mouth AOI scores showed that even though gazing at the mouth differed as a function of Age, $F(4, 184) = 8.08$, $p < .001$, $\eta_p^2 = 0.15$, and Synchrony Condition, $F(1, 184) = 34.59$, $p < .001$, $\eta_p^2 = 0.16$, it also depended on the joint effects of Age and Synchrony Condition, $F(4, 184) = 4.53$, $p < .01$, $\eta_p^2 = 0.09$, and Stimulus Type and Synchrony Condition, $F(1, 184) = 8.82$, $p < .01$, $\eta_p^2 = 0.04$. The most theoretically interesting interaction is the Age x Synchrony Condition interaction. This interaction indicates that children's gazing at the mouth differed as a function of age and synchrony condition. Planned comparisons showed that this difference was due to the fact that the 3-year-olds, 5-year-olds, 6-year-olds, and 7-year-olds gazed more at the mouth in the asynchrony condition [$F(1, 184) = 3.93$, $p < .05$, $F(1, 184) = 15.12$, $p < .001$, $F(1, 184) = 9.30$, $p < .01$, and $F(1, 184) = 23.18$, $p < .001$, respectively] but that the 4-year-olds did not, $F(1, 184) = 0.32$, *ns*.

The ANOVA of the eye AOI scores yielded significant main effects of Age, $F(4, 184) = 5.55$, $p < .001$, $\eta_p^2 = 0.11$, Synchrony Condition, $F(1, 184) = 40.35$, $p < .001$, $\eta_p^2 = 0.18$, and Stimulus Type, $F(1, 184) = 9.50$, $p < .01$, $\eta_p^2 = 0.05$. Fig. 5 shows the age-related decrease in selective attention to the eyes as well as the difference in selective attention to the eyes across the synchrony conditions. Although the latter difference did not appear to be affected by age, we nonetheless carried out a set of planned comparisons at each respective age to provide a parallel analysis to the one conducted on the mouth AOI data. These planned comparisons showed that the 3-year-olds, 5-year-olds, 6-year-olds, and 7-year-olds gazed more at the eyes in the synchrony than asynchrony condition [$F(1, 184) = 5.77$, $p < .05$, $F(1, 184) = 14.70$, $p < .001$, $F(1, 184) = 11.11$, $p < .01$, and $F(1, 184) = 10.81$, $p < .01$, respectively] and that the 4-year-olds did not, $F(1, 184) = 2.50$, *ns*. Finally, even though Fig. 5 does not depict the Stimulus effect, inspection of the data revealed that this effect was due to greater selective overall attention to the target's than the distractor's eyes.

### 1.2.3. *Latency of response*

This final analysis tested Prediction (4) and, thus, asked whether a talking face in a particular quadrant elicited faster initial attention when it was audiovisually synchronized than when it was desynchronized. In this analysis, we compared response latency to a target talking face in the same quadrant across the two synchrony conditions. We defined response latency as the length of time between trial onset and first fixation of the target face. Prior to examining the latency data, however, first we asked whether children exhibited an initial quadrant preference

at the onset of the trial. We did this based on the findings from a similar study with adults in which the same stimuli were presented and a similar testing procedure was used (Lewkowicz et al., 2021). In that study, findings indicated that adults directed 70.5% of their initial fixations to the top left quadrant at the start of each test trial regardless of whether the face presented there was audiovisually synchronized or not. To perform a similar analysis in the current experiment, we considered the possibility that initial fixations to the top left quadrant might have differed as a function of age. Indeed, we found that this was the case. The following proportions of initial fixations were directed to the top left quadrant: 38.4% at 3 years of age, 43.9% at 4 years of age, 49.1% at 5 years of age, 54.3% at 6 years of age, and 53.6% at 7 years of age. A Chi Square goodness-of-fit test indicated that these proportions were statistically different, $X^2$ (4, $N = 727$) = 442.88, $p < .001$.

Because the initial fixation was not always directed at the target or the virtual target, the data that contributed to the response latency analysis represent the time to the first fixation of the target talking face in a given trial regardless of whether the participant looked elsewhere first or not. Importantly, for this analysis, we excluded the data of any child for whom we did not obtain a latency score for the target stimulus in a particular synchrony trial and/or for the virtual target in an asynchrony trial. [4] Application of this exclusion criterion yielded usable data for 131 out of the 200 children tested in this experiment. Based on the findings from the adult study (Lewkowicz et al., 2021) where no latency differences were found, we expected that children would not fixate the audiovisually synchronized target face faster than the desynchronized one. Indeed, a comparison of response latency scores to the synchronized vs. desynchronized target talking faces indicated that it took children an average of 2301.3 ms to look at the synchronized target and 2208.1 ms to look to the desynchronized target, $F$ (1,130) = 0.44, $p$ = .84.

### 1.3. Discussion

This experiment yielded clear evidence of a preference for one of four identical and concurrently talking faces when the visible articulations of the preferred face were synchronized with the same audible utterance while the visible articulations of the other three faces were not synchronized. This was consistent with our predictions. There were three notable aspects to the observed preference for the audiovisually synchronized talking face: (a) it was observed in the synchrony condition at all ages, (b) its magnitude increased with age, and (c) the age-related increase in preference was due to an increase in preference for the audiovisually synchronized talking face and a concomitant decrease in preference for the audiovisually desynchronized talking faces.

The observed preference was spontaneous because no explicit instructions were given to attend to the temporal relation between the audible and visible articulations per se. The only instructions given were to match the face to the voice and then, at the end of each test trial, to point to the face that was talking. Given that all four faces were always seen talking, the choice of *the* talking face could not be biased. In other words, based on the instructions given, the children in this experiment could have gazed at any of the four talking faces. The fact that they deployed more attention to the audiovisually synchronized talking face in the synchrony condition and that they deployed equal attention to all four talking faces in the asynchrony condition demonstrates that the temporally based multisensory coherence of a talking face is a powerful driver of selective attention.

[4] Please note that this latency analysis is independent of the quadrant-of-initial-look analysis. Unlike the latency analysis in which we had to exclude any child for whom we did not obtain a latency score, this was not a consideration in the quadrant-of-initial-look analysis. All that mattered was the number of first looks to the top left quadrant regardless of whether the target was presented there or not.

The eye and mouth gaze data provided insights into the underlying perceptual mechanism driving the observed preference. They showed that selective attention to the eyes and mouth differed as a function of age and synchrony condition. The mouth data indicated that children attended more to the mouth than eyes in both synchrony conditions and more to the mouth in the asynchrony than synchrony condition. Furthermore, children began to prefer the talker's mouth by 6 years of age when an audiovisually synchronized talking face was present in the stimulus array but by 5 years of age when no audiovisually synchronized talking face was present. The eye data indicated that children's attention to the eyes declined as a function of age in both synchrony conditions, that their attention to the target's eyes was greater than to the distractor's eyes, and that their attention to the eyes was greater in the synchrony condition than in the asynchrony condition.

## 2. Experiment 2

Experiment 1 demonstrated that children between 3 and 7 years of age prefer an audiovisually synchronized talking face when that face competes for attention with other identical but audiovisually desynchronized talking faces. Crucially, the only multisensory redundancy cue available in Experiment 1 was AV temporal synchrony. This was a deliberate design feature of Experiment 1 intended to demonstrate that this perceptual cue is, indeed, a powerful perceptual segregation cue. As noted earlier, however, most naturalistic social situations involve different talkers and, thus, offer perceivers a significantly wider variety of AV redundancy cues as well as individual audible and visible identity cues. Accordingly, it is important to explore whether children might also prefer an audiovisually synchronized talking face when that face competes for attention with different, rather than same, audiovisually desynchronized talking faces.

How might auditory and visual identity cues that normally serve to distinguish different individuals' talking faces contribute to perceptual segregation? Based on theoretical and empirical grounds, it is reasonable to expect that whenever multisensory inputs are temporally redundant and integrated into unitary entities, whatever modality-specific multisensory attributes are present as well are also bound together and constitute part-and-parcel of the integrated entity (Atilgan et al., 2018). In other words, multisensory integration processes and multisensory binding processes both contribute to the perceptual experience of multi-dimensionally specified, unitary multisensory entities. As indicated earlier, such multisensory entities are generally perceptually more salient than those specified by unisensory attributes. Given this, individual identity cues might augment perceptual segregation of multiple talking faces. Two findings are consistent with this prediction. First, adults are known to detect and link various identity cues to create representations of individual talkers (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a, 2004b). Second, in an identical perceptual segregation task using the identical stimuli used in the current study, Lewkowicz et al. (2021) found that adults took advantage of individual identity cues to segregate multiple talking faces.

Whether individual identity cues might contribute to children's perceptual segregation is currently not known and difficult to predict a priori. On the one hand, some evidence indicates that children do not rely on individual identity cues to discriminate static faces because their ability to perceive the spatial relations among the internal features of a face is immature (Mondloch, Le Grand, & Maurer, 2002) and because they rely more on misleading cues inherent in paraphernalia (Baenninger, 1994; Carey & Diamond, 1977; Freire & Lee, 2001). On the other hand, some evidence shows that children are sensitive to featural and facial contour cues (Mondloch et al., 2002), Thus, whether visible and audible individual identity cues may contribute to children's segregation of multiple talking faces is an open question and was investigated in Experiment 2. To do so, we used the same method as in Experiment 1 except that now children saw four different talkers and heard their

unique voices across test trials.

Although a priori predictions regarding the specific influence of identity cues were difficult to make, we expected the overall pattern of findings from the current experiment to be fully consistent with the developmental improvement of perceptual segregation found in Experiment 1. In other words, we expected the findings from this experiment to essentially mirror those obtained in Experiment 1 and, thus, to be in line with the predictions outlined in the Introduction.

### 2.1. Method

#### 2.1.1. Participants

A total of 102 children contributed usable data in this experiment. As in Experiment 1, we tested separate groups of 3-year-olds ($N = 23$; 15 girls), 4-year-olds ($N = 29$; 10 girls), 5-year-olds ($N = 24$; 13 girls), 6-year-olds ($N = 17$; 9 girls), and 7-year-olds ($N = 10$; 6 girls). We tested an additional 22 children, but they either did not contribute usable data, or their data were not used. Exclusion criteria included technical problems ($N = 1$), a failure to complete the experiment ($N = 15$), a failure to meet our age requirements (i.e., they were too young or too old but were tested anyway because of a desire to 'play the game' ($N = 4$), or a disqualifying disability ($N = 2$). All the children were recruited and tested in the Living Laboratory at the Boston Museum of Science except for three children who were tested in the first author's university laboratory. A parent or guardian for all children granted written consent prior to participating in the study.

#### 2.1.2. Apparatus and stimuli

All aspects of apparatus and stimuli were the same as in Experiment 1 except that here the composite videos consisted of four different, rather than same, female actors (see Fig. 6). As in Experiment 1, during each test trial, the actors simultaneously articulated the same utterance in a temporally jittered fashion and the audible version of the same utterance was played at the same time (for an example see Video S2). Also, like in Experiment 1, each trial began with all four faces articulating the visible utterance in a temporally jittered fashion. The temporal jitter was produced by delaying the start of the visible articulations for each of the distractor faces, respectively, increasingly later into the utterance relative to the start of the visible articulation produced by the target face. [5] The only difference between the synchrony and asynchrony test trials was that the audible utterance was temporally synchronized with the visible articulations of the target face and desynchronized with the three distractor talking faces in the synchrony trials but that it was desynchronized with all four talking faces in the asynchrony trials.

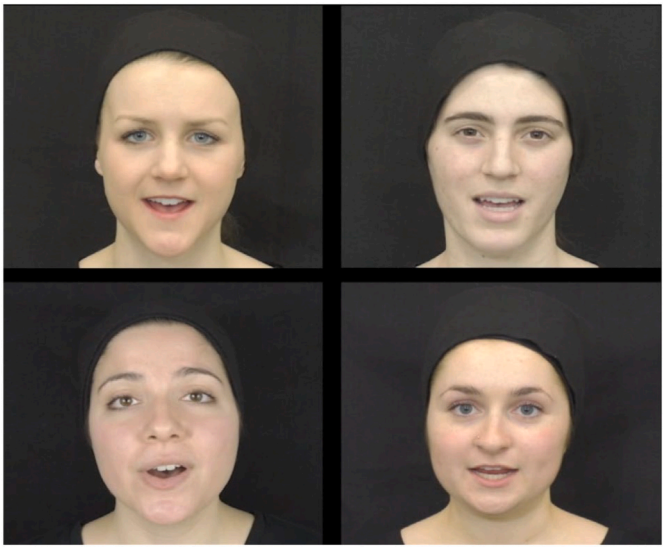

**Fig. 6.** Screenshot of one of the composite videos presented in Experiment 2.

To create the composite videos for this experiment, we filmed the four female actors while each spoke the same two different utterances at roughly the same pace with similar intonation and prosody. The resulting eight audiovisually synchronized videos corresponded to eight unique actor/utterance combinations. Then, we created two sets of test trials with the constraint that all four actors be presented in each set. In each set, two of the actors spoke one utterance while the other two actors spoke the other utterance, with the specific utterance spoken by each actor counterbalanced across the two sets. The quadrant of target-stimulus presentation was counterbalanced across the synchrony and asynchrony test trials, respectively. As a result, the target stimulus appeared once in each of the four quadrants across the four synchrony test trials and once in each of the four corresponding quadrants across the four asynchrony trials. As in Experiment 1, the same-quadrant synchrony and asynchrony test trials were presented in pairs, with the synchrony test trial presented first followed by the corresponding asynchrony test trial. During the asynchrony test trials, the audible utterance was desynchronized with respect to the target face by the same temporal interval as in Experiment 1.

#### 2.1.3. Design and procedure

The design and procedure in this experiment was the same as in Experiment 1. Thus, the experiment consisted, in turn, of a calibration phase, two 15 s practice trials, and eight 15 s test trials comprised of four pairs of synchrony and asynchrony test trials. Like in Experiment 1, we used a Latin Square design to generate four Quadrant groups for the four pairs of synchrony/asynchrony test trials for each of the two sets of eight test trials each. Membership in a given Quadrant group determined the specific quadrant in which the target face for each respective pair of synchrony and asynchrony test trials was presented. The specific quadrant of target-face presentation for each respective pair of test trials was counterbalanced across the four quadrants within each Quadrant group. Furthermore, the specific quadrant in which the targets in a given pair of test stimuli were presented was counterbalanced across the four Quadrant groups. Children were randomly assigned to one of the four Quadrant groups within each one of the two 8-test trial sets.

### 2.2. Results

#### 2.2.1. Face AOIs

First, we performed a preliminary analysis to determine whether the specific actor/utterance combination might have influenced gaze behavior given that each actor produced a different utterance across the two sets of test trials. Consequently, we compared the PTLT scores for the target talking face with the average of the PTLT scores for the three distractor talking faces with a mixed, repeated-measures ANOVA, with Synchrony Condition (2), Utterance (2), Actor (2), and Stimulus Type (2) as within-subjects factors and Age (5), Quadrant Group (4), and Test Trial Set (2) as between-subjects factors. This analysis only yielded a Synchrony Condition x Stimulus Type x Age interaction, $F(1, 68) = 4.13$,

[5] The delay intervals for both the synchrony and asynchrony test trials were the same for each of the four actors and for each of the two utterances that they spoke (2233, 3366, and 4433 ms). Similarly, the delay interval to create the asynchronous version of the target stimulus was the same for all actors and for both utterances (1833 ms).

$p < .05$, $\eta_p^2 = 0.057$, indicating that Actor, Utterance, and Quadrant did not affect responsiveness. Thus, we collapsed the PTLT scores over these three factors and performed a new repeated-measures ANOVA with Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Age (5) as a between-subjects factor. This analysis yielded a main effect of Age, $F(4, 98) = 10.23$, $p < .001$, $\eta_p^2 = 0.29$, Synchrony Condition, $F(1, 98) = 161.93$, $p < .001$, $\eta_p^2 = 0.62$, and Stimulus Type, $F(1, 98) = 119.91$, $p < .001$, $\eta_p^2 = 0.55$. It also yielded several two-way interactions, including a Synchrony Condition x Age interaction, $F(4, 98) = 16.57$, $p < .001$, $\eta_p^2 = 0.40$, a Stimulus Type x Age interaction, $F(4, 98) = 11.08$, $p < .001$, $\eta_p^2 = 0.31$, and a Synchrony Condition x Stimulus Type interaction, $F(1, 98) = 168.89$, $p < .001$, $\eta_p^2 = 0.63$. Finally, the analysis yielded a Synchrony Condition x Stimulus Type x Age interaction, $F(4, 98) = 16.26$, $p < .001$, $\eta_p^2 = 0.40$.

As in Experiment 1, the most theoretically relevant finding was the three-way Synchrony Condition x Stimulus Type x Age interaction. This interaction is depicted in Fig. 7 and as can be seen, these findings are similar to those in Experiment 1. That is, in the synchrony condition, children of all ages gazed more at the target face than at the distractor faces. In contrast, in the asynchrony condition, children of all ages gazed equally at the two types of faces. Moreover, the PTLT directed at the target face relative to the PTLT directed at the distractor faces increased as a function of age in the synchrony condition but not in the asynchrony condition. These observations were confirmed by separate repeated-measures ANOVAs which yielded a statistically significant Stimulus Type x Age interaction in the synchrony condition, $F(4, 98) = 20.52$, $p < .001$, $\eta_p^2 = 0.45$, but not in the asynchrony condition, $F(4, 98) = 0.44$, *ns*. Furthermore, simple effects ANOVAs in the synchrony condition indicated that the age-related increase in gazing at the target face was significant, $F(4, 98) = 20.48$, $p < .001$, $\eta_p^2 = 0.45$, and that the age-related decrease in gazing at the distractor faces was also significant, $F(4, 98) = 19.8$, $p < .001$, $\eta_p^2 = 0.45$.

Finally, planned comparisons of the PTLT scores in the synchrony condition showed that children gazed more at the target face at each age (3-year-olds: $F(1, 98) = 7.00$, $p < .01$; 4-year-olds: $F(1, 98) = 8.81$, $p < .01$; 5-year-olds: $F(1, 98) = 37.13$, $p < .001$; 6-year-olds: $F(1, 98) = 107.97$, $p < .001$; and 7-year-olds: $F(1, 98) = 87.79$, $p < .001$). Planned comparisons of the PTLT scores in the asynchrony condition showed that children did not gaze more at the target than at the distractor faces at any age (3-year-olds: $F(1, 98) = 0.21$, *ns*; 4-year-olds: $F(1, 98) = 0.02$, *ns*; 5-year-olds: $F(1, 98) = 1.19$, *ns*; 6-year-olds: $F(1, 98) = 1.61$, *ns*; and 7-year-olds: $F(1, 98) = 0.01$, *ns*).

### 2.2.2. Eye and mouth AOIs

The initial analysis of selective attention of the PTLT scores for the eye and mouth AOIs was a repeated-measures ANOVA, with AOI (2), Synchrony Condition (2), Utterance (2), Actor (2), and Stimulus Type (2) as within-subjects factors and Age (5), Quadrant Group (4), and Test Trial Set (2) as between-subjects factors. This analysis indicated that neither the utterance spoken, the actor speaking it, the quadrant in which the target face was presented, nor the specific test trial set played a statistically significant or theoretically interesting role in attention to the eyes or mouth. Thus, the PTLT scores for the mouth and eye AOIs were collapsed over these factors and a new repeated-measures ANOVA was conducted, with AOI (2), Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Age (5) as a between-subjects factor. This analysis yielded an AOI, $F(1, 98) = 8.08$, $p < .01$, $\eta_p^2 = 0.08$, and a Stimulus Condition, $F(1, 98) = 7.39$, $p < .01$, $\eta_p^2 = 0.07$, main effect, and a Synchrony Condition x Age interaction, $F(4, 98) = 3.94$, $p < .01$, $\eta_p^2 = 0.14$.

Even though the AOI x Synchrony Condition x Age interaction was not statistically significant in the current experiment ($F(4, 98) = 0.35$, *ns*), we present the eye and mouth AOI data as a function of synchrony condition and age (see Fig. 8) to facilitate a direct comparison of the AOI data from this experiment with the AOI data from Experiment 1. Inspection of Fig. 8 suggests that children generally attended more to the mouth than the eyes in each synchrony condition. This observation is consistent with the significant main effect of AOI mentioned earlier. Separate repeated-measures ANOVAs of the AOI data in each condition, respectively, with AOI (2) and Stimulus Type (2) as within-subjects factors and Age (5) as a between-subjects factor were consistent with the main effect of AOI. That is, overall, children attended more to the mouth than eyes in the synchrony condition, $F(1, 98) = 4.55$, $p < .05$, $\eta_p^2 = 0.04$, as well as in the asynchrony condition, $F(1, 98) = 10.78$, $p < .01$, $\eta_p^2 = 0.10$.

Further inspection of Fig. 8 suggests that the age-related growth of the attentional preference for the mouth is greater in the asynchrony than synchrony condition. Planned comparisons of the eye and mouth AOI data in each respective synchrony condition supported this observation. Specifically, in the synchrony condition, the 3-, 4-, 5, and 6-year-olds did not exhibit a mouth preference [$F(1, 98) = 0.00$, $p = .99$, *ns*; $F(1, 98) = 0.17$, $p = .68$, *ns*; $F(1, 98) = 2.75$, $p = .10$, *ns*, $F(1, 98) = 1.46$, $p < .23$, *ns*, respectively], and the 7-year-olds mouth preference was only marginally significant [$F(1, 98) = 3.23$, $p = .075$]. In contrast, planned comparisons in the asynchrony condition showed that whereas the 3- and the 4-year-olds did not exhibit a mouth preference [$F(1, 98) = 0.05$, $p = .83$, *ns*; $F(1, 98) = 0.06$, $p = .80$, *ns*, respectively], the 5-, 6-, and 7-year-olds did exhibit a mouth preference [$F(1, 98) = 3.73$, $p = .05$; $F(1, 98) = 5.29$, $p < .025$; $F(1, 98) = 6.10$, $p < .025$, respectively].

Finally, as in Experiment 1, we compared the eye and mouth AOI scores, respectively, across the two synchrony conditions. We did so by way of repeated-measures ANOVAs, with Stimulus Type (2) and Synchrony Condition (2) as within-subjects factors and Age as a between-

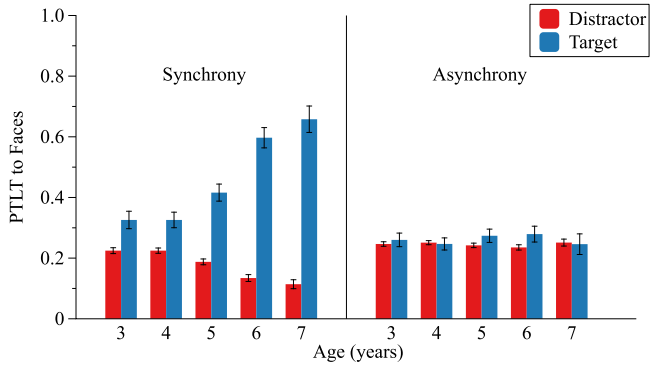


**Fig. 7.** Mean proportion of total looking time at the distractor and target faces as a function of age across the two synchrony conditions in Experiment 2. Error bars represent the standard errors of the mean.

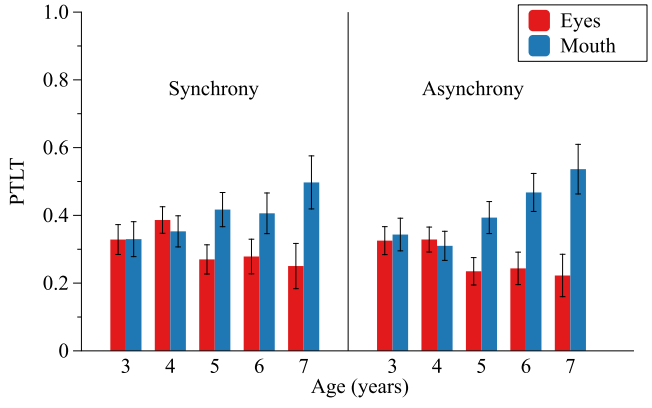


**Fig. 8.** Mean proportion of total looking time at the eyes and mouth as a function of age across the two synchrony conditions in Experiment 2. Error bars represent the standard errors of the mean.

subjects factor. The ANOVA of the mouth AOI scores yielded no statistically significant effects, indicating that looking at the mouth did not differ across the two conditions. In contrast, the ANOVA of the eye AOI showed that there was a significant effect of synchrony condition, $F(1, 98) = 4.06$, $p < .05$, indicating that, overall, children attended more to the eyes in the synchrony than in the asynchrony condition.

### 2.2.3. Latency of response

As in Experiment 1, first we asked whether children directed most of their initial looks to the top left quadrant. The proportions of initial fixations to the top left quadrant were as follows: 48.9% at 3 years of age, 50.0% at 4 years of age, 39.0% at 5 years of age, 61.0% at 6 years of age, and 45.0% at 7 years of age. A Chi Square goodness-of-fit test indicated that these proportions were statistically different, $X^2$ (4, $N = 397$) = 438.74, $p < .001$.

The analysis of response latency indicated that the 63 of the 102 children tested who did not have any missing latency scores took an average of 2443.7 ms to look at the synchronized target and 2390.1 ms to look at the desynchronized virtual target, $F$ (1, 62) = 0.07, $p = .79$.

### 2.2.4. Comparison of Experiments 1 and 2

In a final set of analyses, we compared the results of Experiments 1 and 2. The goal here was to assess the comparability of the magnitude of the effect across the experiments and to determine whether individual visual and auditory identity cues contributed to perceptual segregation of the multisensory clutter created by the multiple talking faces.

*2.2.4.1. Face AOIs.* We compared the face gaze data by way of a repeated-measures ANOVA, with Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Age (5) and Experiment (2) as a between-subjects factors. Not surprisingly, given the results of the individual experiments, this analysis yielded main effects for Synchrony Condition, $F(1, 282) = 304.49$, $p < .001$, $\eta_p^2 = 0.52$, Stimulus Type, $F(1, 282) = 337.86$, $p < .001$, $\eta_p^2 = 0.54$, and Age, $F(4, 282) = 22.70$, $p < .001$, $\eta_p^2 = 0.24$. Also, not surprisingly, this analysis yielded significant Synchrony Condition x Age, $F(4, 282) = 26.52$, $p < .001$, $\eta_p^2 = 0.27$, Stimulus Type x Age, $F(4, 282) = 26.02$, $p < .001$, $\eta_p^2 = 0.27$, Stimulus Type x Synchrony Condition, $F(1, 282) = 311.43$, $p < .001$, $\eta_p^2 = 0.24$, and Stimulus Type x Synchrony Condition x Age, $F(4, 282) = 25.78$, $p < .001$, $\eta_p^2 = 0.27$ interactions. Despite these significant effects, the principal finding of interest here is the absence of any effects related to the Experiment factor. This indicates that identity cues did not play a role in children's overall ability to perceptually segregate multiple talking faces and, thus, in their marked preference for audiovisually synchronous talking faces.

*2.2.4.2. Eye and mouth AOIs.* We compared the eye and mouth AOI data from both experiments via a mixed, repeated-measures ANOVA, with Synchrony Condition (2), Stimulus Type (2), and AOI (2) as within-subjects factors and Age (5) and Experiment (2) as between-subjects factors. Again, and not surprisingly, this analysis yielded main effects of Synchrony Condition, $F(1, 282) = 12.80$, $p < .001$, $\eta_p^2 = 0.04$, and AOI, $F(1, 282) = 27.16$, $p < .001$, $\eta_p^2 = 0.09$, as well as Synchrony Condition x AOI, $F(1, 282) = 28.26$, $p < .001$, $\eta_p^2 = 0.09$, AOI x Age, $F(4, 282) = 6.85$, $p < .001$, $\eta_p^2 = 0.09$, and Synchrony Condition x Age, $F(4, 282) = 3.81$, $p < .01$, $\eta_p^2 = 0.05$, interactions. Most importantly, we found a significant Synchrony Condition x AOI x Experiment, $F(1, 282) = 6.23$, $p < .025$, $\eta_p^2 = 0.02$ interaction. This three-way interaction is of particular interest and is depicted in Fig. 9.

As can be seen in Fig. 9, gazing at the eyes across the synchrony conditions appears to be similar in each experiment. This observation was confirmed by a follow-up, mixed, repeated-measures ANOVA of gaze at the eye AOI, with Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Experiment (2) as a between-subjects factor. This analysis revealed a non-significant Synchrony Condition x Experiment interaction, $F(1, 290) = 1.65$, *ns*. In contrast to the eye gaze data, Fig. 9 shows that the gaze to the mouth differed across experiments. A follow-up, mixed, repeated-measures ANOVA of gaze at the mouth AOI, with Synchrony Condition (2) and Stimulus Type (2) as within-subjects factors and Experiment (2) as a between-subjects uncovered a significant Synchrony Condition x Experiment interaction, $F(1, 290) = 10.31$, $p < .01$, $\eta_p^2 = 0.03$. This interaction is due to a significant difference between synchrony conditions in Experiment 1, $F(1, 282) = 31.04$, $p < .001$, and the absence of such a difference in Experiment 2, $F(1, 282) = 0.41$, *ns*. This suggests that identity cues may have had a more subtle influence on children's ability to identify the talking face based on AV synchrony by requiring them to focus more on the talker's mouth to determine which face was talking.

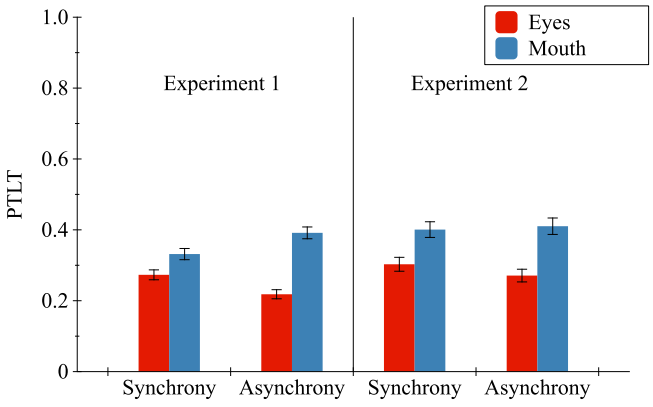


**Fig. 9.** Mean proportion of total looking time at the eyes and mouth in the two synchrony conditions across the two experiments. Error bars represent the standard errors of the mean.

## 2.3. Discussion

Like in Experiment 1, we found that when AV temporal synchrony statistics distinguished between the talking faces, children attended more to an audiovisually synchronized talking face than to competing desynchronized talking faces. When, however, the AV temporal statistics did not distinguish the four different talking faces, children exhibited equal attention to all four talking faces. These findings extend the findings from Experiment 1 by demonstrating that children between 3 and 7 years of age can successfully segregate four different talking faces accompanied by different voices across different trials based on AV temporal statistics. Also, like in Experiment 1, we found an age-related increase in attention to the audiovisually synchronized talking face and a concomitant decrease in attention to the audiovisually desynchronized talking faces. Finally, we obtained no evidence that children's responsiveness was affected by identity cues over-and-above AV temporal synchrony cues.

The eye and mouth gaze data indicated that, regardless of age, children attended more to the mouth than to the eyes in each synchrony condition. Nevertheless, the AV temporal synchrony statistics appeared to exert some differential influence on this trend in that the preference for the mouth was generally weaker in the synchrony condition. This difference appears to reflect differential interest in the eyes versus the mouth across the two synchrony conditions, with greater interest in the eyes in the synchrony condition but greater interest in the mouth in the asynchrony condition. The former probably reflects less of a need to attend to the mouth when the source of AV temporal synchrony is relatively obvious whereas the latter probably reflects search for AV temporal synchrony when it is absent.

## 3. General discussion

The current study investigated 3–7-year-old children's ability to perceptually segregate a typical cluttered multisensory scene consisting of multiple talking people. Experiment 1 examined perceptual segregation in its simplest form by investigating whether AV temporal synchrony statistics can facilitate segregation. Children saw four identical talking faces visibly articulating the same utterance in a temporally jittered fashion and heard the audible version of the same utterance. Across trials, the audible utterance was either temporally synchronized with the visible articulations of one of the four talking faces or desynchronized from all of them. Experiment 2 examined perceptual segregation in its more ecologically typical form. This time, children saw four different talking faces articulating the same utterance in a temporally jittered fashion and heard the audible version of the utterance spoken by different people across different trials. As a result, children could now potentially segregate the multiple talking faces based on AV temporal synchrony statistics, individual audible and visible identity cues, or a combination of the two.

### 3.1. Perceptual segregation of multiple talking faces

Perceptual segregation of multiple talking faces was evident in two principal findings. First, findings indicated that children of all ages allocated more selective attention to the canonical, temporally synchronized, talking face when it was present in the stimulus array and that they allocated equal amounts of selective attention when a canonical talking face was not present in the stimulus array. Second, the findings showed that the allocation of selective attention to the canonical talking face increased with age regardless of whether the faces and voices were identical (Experiment 1) or whether they differed in terms of identity cues (Experiment 2). Together, these findings indicate that the AV temporal synchrony statistics inherent in canonical talking faces are a powerful driver of young children's selective attention and that they facilitate their perceptual segregation of competing talking faces.

In certain respects, the current findings are similar to those from a study which investigated perceptual segregation of the same talking faces presented here but in adults (Lewkowicz et al., 2021). Results from that study showed that adults exhibited a marked preference for the canonical talking face and that they exhibited a greater average preference for this face than did the oldest children tested here (see below for a statistical comparison of thse data). When children's smaller preference for the canonical talking face is considered together with its age-related increase, there is little doubt that the preference for the canonical talking face develops slowly during early childhood and that the growth of this preference continues past 7 years of age. In addition to the perceptual segregation findings, the response latency data from both the adult and the current study are of particular interest because they shed important light on the possible mechanisms underlying responsiveness. The response latency data indicated that responsiveness to the canonical, temporally synchronized, talking face was not faster than to the same but temporally desynchronized talking face. These findings are consistent with results from other studies of young children's latency of response to a temporally synchronized talking face embedded in an array of other talking faces than to the same but silently talking face (Bahrick et al., 2018). The most reasonable conclusion from the response latency data is that visual search of multiple talking faces and their segregation based on AV temporal synchrony statistics is probably driven by a serial and effortful AV feature matching process (Fujisaki, Koene, Arnold, Johnston, & Nishida, 2006) rather than an automatic AV pop-out process (Matusz & Eimer, 2011). Nonetheless, it may be that visual search is also to some extent mediated by bottom-up saliency effects. Whether this is the case, remains to be determined in future studies.

It should be noted that the age-related growth of a canonical talking-face preference is accompanied by an age-related decrease in selective attention to non-canonical, audiovisually desynchronized, talking faces. This demonstrates that the marked preference for the canonical talking face found in adults is the result of two parallel developmental processes. This age-related perceptual differentiation and the resulting increase in a preference for the canonical talking face is consistent with the unity assumption which holds that adults possess a general perceptual bias for multisensory coherence (Welch, 1999; Welch & Warren, 1980). The findings from this study lead to the conclusion that the unity assumption for talking faces takes many years to reach its mature form. Although this conclusion seems reasonable on its surface, it appears to be at odds with findings from studies showing that infants are sensitive to low-level AV relations specified by such perceptual cues as intensity (Lewkowicz & Turkewitz, 1980), duration (Lewkowicz, 1986), and temporal synchrony (Bahrick, 1983; Lewkowicz, 1992, 1996, 2010; Lewkowicz et al., 2010; Lewkowicz et al., 2015). This suggests that the perception of multisensory coherence begins early in development (Lewkowicz, 2000; Lewkowicz & Ghazanfar, 2009; Lickliter & Bahrick, 2000; Murray et al., 2016; Walker-Andrews, 1997). Despite this, however, data show that the emergence of mature multisensory integration takes many years, that it depends on the specific modality pairs that must be integrated, that it depends on the gradual growth of sensory capacity in different modalities (Cowie, Sterling, & Bremner, 2016; Ernst, 2008; Ernst & Banks, 2002; Stevenson, Baum, Krueger, Newhouse, & Wallace, 2018), and that the ability to perform a visual search on the basis of AV redundancy improves between middle childhood and adulthood (Matusz et al., 2015; Matusz, Merkley, Faure, & Scerif, 2019). Thus, when all the developmental findings are considered together, it is not surprising that the development of perceptual segregation of multiple talking faces based on AV temporal synchrony is a slow and complex process.

The developmental differentiation of selective attention to canonical vs. non-canonical talking faces also demonstrates that children become increasingly better at perceiving the temporal synchrony statistics linking fluent auditory and visual speech streams as they grow and as they acquire increasingly greater perceptual experience with their multisensory world. This conclusion is consistent with the fact that even though infants as young as 2 months of age can perceive the intersensory equivalence of isolated visible and audible syllables (Kuhl & Meltzoff, 1982; Lewkowicz, 2010; Patterson & Werker, 2003), only 12–14 month-old infants can perceive the equivalence of fluent/continuous visible and audible speech streams (Lewkowicz et al., 2015). The relatively slow age-related growth in the preference for the canonical, audiovisually synchronized, talking face is also consistent with several improvements in multisensory processing, including young children's ability to detect AV temporal synchrony relations inherent in AV speech syllables (Lewkowicz & Flom, 2014), the ability to take advantage of visual speech information to facilitate detection of speech in noise (Ross et al., 2011), and the ability to perceive temporally based AV illusions starting in infancy (Scheier et al., 2003) and continuing into childhood (Barutchu et al., 2009). It is also consistent with the slow age-related improvements in the integration of haptic/tactile and visual own-body information (Cowie, Makin, & Bremner, 2013; Cowie, McKenna, Bremner, & Aspell, 2018; Ernst, 2008; Ernst & Banks, 2002; Nardini, Begus, & Mareschal, 2013).

In terms of specific temporal mechanisms, the gradual age-related growth in the magnitude of the preference for the canonical talking face probably reflects, in part, the gradually diminishing size of the temporal binding window (the period of time during which A and V inputs are automatically fused into unitary percepts). In general, the temporal binding window is larger for speech than for non-speech events, mostly because the former is a continuous event where lip motion naturally precedes phonation whereas the latter is a more punctate event characterized by the precise correspondence between sounds and abrupt changes in visual motion (e.g., a bouncing ball). Overall, regardless of type of event, sensitivity to AV temporal synchrony improves gradually from infancy up through adolescence. This is evident in

findings indicating that the size of the binding window is relatively large in infancy (Lewkowicz, 1996, 2010), that it narrows gradually during childhood (Chen, Shore, Lewis, & Maurer, 2016; Lewkowicz & Flom, 2014), and that it reaches its smallest size in adolescence (Hillock et al., 2011; Hillock-Dunn & Wallace, 2012). As the size of the temporal binding window decreases with age, intersensory temporal accuracy increases. Given such findings, it is reasonable to conclude that the shrinking binding window enables children to perceive AV synchrony with increasingly greater precision and that this makes it easier for them to perceptually segregate multiple talking faces.

Finally, the developmental changes observed in the current study are especially interesting when compared to findings from a prior study of adults performing the same segregation task. In that study, Lewkowicz et al. (2021) found that adults allocated an average of 80% of their attention to the canonical talking face and only an average of 6% of their attention to the non-canonical, desynchronized, talking faces (a 73% difference). Unlike the adults, the oldest children in the current study allocated an average of 62% of their attention to the canonical talking face and only an average of 12% of their attention to the non-canonical, desynchronized, talking faces (a 50% difference). To determine whether this difference in the magnitude of the preference for the canonical talking face is statistically significant, we compared the combined data from the adults tested by Lewkowicz et al. (2021) and the 7-year-old children in the two experiments of the current study with a repeated-measures ANOVA, with Synchrony Condition (2) and Stimulus Condition (2) as within-subjects factors and Age (2) as a between-subjects factor. Results yielded a significant Synchrony Condition x Stimulus Condition x Age interaction, $F$ (1, 106) = 17.43, $p < .001$, $\eta_p^2 = 0.14$, demonstrating that perceptual segregation of multiple talking faces based on temporal AV synchrony continues to grow beyond 7 years of age.

### 3.2. *Mechanisms of perceptual segregation: selective attention to the eyes and mouth*

The eye and mouth data from the current study provided important insights into the mechanisms underlying children's perceptual segregation of competing talking faces. They showed that children attended more to the mouth than the eyes and that they did so regardless of whether a canonical talking face was present in the stimulus array or not and regardless of whether individual identity cues distinguished between the faces and voices or not. The data also showed that children attended more to the mouth when a canonical talking face was absent in the stimulus array and when individual identity cues were absent. Importantly, this greater overall focus on the talker's mouth is consistent with the specific task assigned to the participants. To reiterate, children were instructed to match the face to the voice and to point to the face that was talking. To succeed in such a task, children had to focus their selective attention on the very source of AV coherence, namely the talker's mouth. The fact that they did suggests that the temporal synchrony statistics that characterize everyday AV speech are a fundamental feature of fluent AV speech and that they play a critical role in young children's perception of multisensory coherence. In addition, the children's greater focus on the talker's mouth is interesting because it parallels similar findings in adult studies of speech processing. These studies have found that adults also attend more to a talker's mouth than eyes when processing non-native (i.e., unfamiliar) speech and speech-in-noise (Barenholtz et al., 2016; Birulés et al., 2020; Grant & Seitz, 2000; Rennig et al., 2020; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998; Võ et al., 2012).

The developmental differences in selective attention to the mouth also are interesting when selective attention to the mouth is examined across the two experiments. Thus, in Experiment 1 - when a canonical talking face was present in the stimulus array and when no individual identity cues were available - children first exhibited greater attention to the mouth at 6 years of age. When, however, a canonical talking face was not present in the stimulus array and when no individual identity cues were available, children exhibited greater attention to the mouth at 5 years of age. In Experiment 2 – when a canonical talking face was present in the stimulus array and when individual identity cues were present - children first exhibited only marginally greater attention to the mouth at 7 years of age. When, however, no canonical talking face was present and when individual identity cues were present - children first exhibited greater attention to the mouth at 5 years of age. Together, these data demonstrate that, regardless of the presence of identity cues, greater attention to the mouth was manifest at 5 years of age when no canonical talking face was present and, thus, when face-voice matching was difficult. This suggests that it is at this age that children become more cognizant of the temporal synchrony statistics that normally specify everyday AV speech, that they begin to utilize an explicit search strategy focused on identifying such information, and that they do so especially when speech processing becomes challenging. This interpretation is supported by the fact that it is at 5 years of age that children also first begin to exhibit a larger difference in selective attention to the canonical than non-canonical talking face compared to the 3- and 4-year-old children.

The eye gaze data indicated that children's attention to the eyes declined as a function of age regardless of whether a canonical talking face was present in the stimulus array or not and that attention to the eyes of a canonical talking face was greater than to the eyes of a non-canonical talking face. The former finding most likely reflects children's gradual discovery of the importance of attending to the source of AV speech in a talker's mouth to segregate a cluttered multisensory talker scene. The latter finding probably reflects the fact that when the task of identifying who is talking is relatively easy, observers are free to explore the other key part of a talker's face in the search for the social and deictic cues located in the eye region. This interpretation is consistent with findings that as long as the task does not involve speech processing per se, perceivers tend to devote more of their attention to a talker's eyes (Buchan, Paré, & Munhall, 2007; Lewkowicz & Hansen-Tift, 2012; Võ et al., 2012).

Finally, children's responsiveness was not affected by identity cues in the same way that it was in adults. Lewkowicz et al. (2021) found that adults gazed longer at the talking face in the asynchrony test trials in the same quadrant in which they also saw the temporally synchronized version of the same talking face during the synchrony test trials. Importantly, however, adults were given 32 test trials during which they saw and heard each unique face and voice four times per each synchrony condition. This provided them with ample opportunity to associate each person's dynamic visual "signature" with that person's voice over the course of the experiment. Thus, once they formed these specific face-voice associations, they attended more to that person's face even though the visible and audible articulations were desynchronized. If this interpretation is correct, it is not surprising that children's responsiveness was not affected by identity cues. They only saw each individual person and heard that person's voice once per each synchrony condition and, thus, had less opportunity to make face-voice associations. It is also possible, however, that children may find it more difficult to learn multiple face-voice relations.

In conclusion, the current study documents for the first time the developmental emergence of a perceptual bias for canonical, temporally synchronized, talking faces in young children and offers insights into how they manage to perceptually segregate their cluttered multisensory world. The emerging bias for canonical talking faces helps young children gradually become better at surmounting the multisensory Cocktail Party Problem and, in the process, to become increasingly better at identifying and extracting coherent and meaningful AV communicative signals from the mélange of multisensory inputs that characterize their everyday social world and the interlocutors within it.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105226.

## CRediT authorship contribution statement

**David J. Lewkowicz:** Conceptualization, Methodology, Data curation, Formal analysis, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing - original draft, Writing - review & editing. **Mark Schmuckler:** Conceptualization, Methodology, Writing - review & editing. **Vishakha Agrawal:** Data curation, Project administration, Software, Validation.

## Acknowledgements

This work and manuscript preparation were supported by the National Science Foundation (Grant BCS 1749507) awarded to DJL and an NSERC Discovery Grant awarded to MAS.

## References

Alsius, A., & Soto-Faraco, S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. *Experimental Brain Research, 213*(2–3), 175–183.
Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., & Bizley, J. K. (2018). Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron, 97*(3), 640-655. e644.
Baenninger, M. (1994). The development of face recognition: Featural or configurational processing? *Journal of Experimental Child Psychology, 57*(3), 377–396.
Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior & Development, 6*(4), 429–451.
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: Oxford University Press.
Bahrick, L. E., Soska, K. C., & Todd, J. T. (2018). Assessing individual differences in the speed and accuracy of intersensory processing in young children: The intersensory processing efficiency protocol. *Developmental Psychology, 54*(12), 2226.
Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition, 147*, 100–105.
Barutchu, A., Crewther, D. P., & Crewther, S. G. (2009). The race that precedes coactivation: Development of multisensory facilitation in children. *Developmental Science, 12*(3), 464–473.
Begum Ali, J., Spence, C., & Bremner, A. J. (2015). Human infants' ability to perceive touch in external space develops postnatally. *Current Biology, 25*, R978–R979.
Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2019). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Developmental Science, 22*(3), Article e12755. https://doi.org/10.1111/desc.12755
Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition and Neuroscience, 1-12*. https://doi.org/10.1080/23273798.2020.1762905
Bremner, A. J., Lewkowicz, D. J., & Spence, C. (2012). *Multisensory development*. Oxford: Oxford University Press.
Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience, 2*(1), 1–13.
Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science, 195*(4275), 312–314.
Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*. e1000436.
Chen, Y.-C., Shore, D. I., Lewis, T. L., & Maurer, D. (2016). The development of the perception of audiovisual simultaneity. *Journal of Experimental Child Psychology, 146*, 17–33.
Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.
Cowie, D., Makin, T., & Bremner, A. J. (2013). Children's responses to the rubber hand illusion reveal dissociable pathways in body representations. *Psychological Science, 24*, 762–769.
Cowie, D., McKenna, A., Bremner, A. J., & Aspell, J. E. (2018). The development of bodily self-consciousness: Changing responses to the full body illusion in childhood. *Developmental Science, 21*(3), Article e12557.
Cowie, D., McKenna, A., Bremner, A. J., Aspell, J. E., & Sterling, S. (2016). The development of bodily self-consciousness: Changing responses to the full body illusion in childhood. *Developmental Science, 142*, 230–238.
Cowie, D., Sterling, S., & Bremner, A. J. (2016). The development of multisensory body representation and awareness continues to 10years of age: Evidence from the rubber hand illusion. *Journal of Experimental Child Psychology, 142*, 230–238. https://doi.org/10.1016/j.jecp.2015.10.003
Ernst, M. O. (2008). Multisensory integration: A late bloomer. *Current Biology, 18*, R519–R521. https://doi.org/10.1016/j.cub.2008.05.002
Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information a statistically optimal fashion. *Nature, 415*, 429–433. https://doi.org/10.1038/415429a
Freire, A., & Lee, K. (2001). Face recognition in 4-to 7-year-olds: Processing of configural, featural, and paraphernalia information. *Journal of Experimental Child Psychology, 80*(4), 347–371.
Fujisaki, W., Koene, A., Arnold, D., Johnston, A., & Nishida, S.y. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proceedings of the Royal Society B: Biological Sciences, 273*(1588), 865–874.
Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197–1208.
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), n/a. https://doi.org/10.1111/desc.12381
Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: Age-related changes in multisensory temporal processing. *Neuropsychologia, 49*(3), 461–467.
Hillock-Dunn, A., & Wallace, M. T. (2012). Developmental changes in the multisensory temporal binding window persist into adolescence. *Developmental Science, 15*(5), 688–696.
Hunnius, S., & Geuze, R. H. (2004). Gaze shifting in infancy: A longitudinal study using dynamic faces and abstract stimuli. *Infant Behavior & Development, 27*(3), 397–416.
Innes-Brown, H., Barutchu, A., Shivdasani, M. N., Crewther, D. P., Grayden, D. B., & Paolini, A. (2011). Susceptibility to the flash-beep illusion is increased in children compared to adults. *Developmental Science, 14*(5), 1089–1099.
Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice: Matching identity across modality. *Current Biology, 13*(19), 1709–1714.
Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138–1141.
Lachs, L., & Pisoni, D. (2004a). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 30*(2), 378.
Lachs, L., & Pisoni, D. (2004b). Crossmodal source identification in speech perception. *Ecological Psychology, 16*(3), 159–187.
Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65*(4), 536–552.
Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior & Development, 9*(3), 335–353.
Lewkowicz, D. J. (1992). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior & Development, 15*(3), 297–324.
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.
Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*(2), 281–308.
Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*(1), 66–77.
Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental Psychobiology, 56*(2), 292–315.
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.
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 USA, 109*(5), 1431–1436.
Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match non-human primate faces & voices. *Infancy, 15*(1), 46–60.
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., Schmuckler, M., & Agrawal, V. (2021). The multisensory cocktail party problem in adults: Perceptual segregation of talking faces on the basis of audiovisual temporal synchrony. *Cognition, 214*, Article 104743.
Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology, 16*, 597–607.
Lickliter, R., & Bahrick, L. E. (2000). The development of infant intersensory perception: Advantages of a comparative convergent-operations approach. *Psychological Bulletin, 126*(2), 260–280.
MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131–141.
Matusz, P. J., Broadbent, H., Ferrari, J., Forrest, B., Merkley, R., & Scerif, G. (2015). Multi-modal distraction: Insights from children's limited attention. *Cognition, 136*, 156–165.
Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic Bulletin & Review, 18*(5), 904–909.
Matusz, P. J., Merkley, R., Faure, M., & Scerif, G. (2019). Expert attention: Attentional allocation depends on the differential development of multisensory number representations. *Cognition, 186*, 171–177.
Mondloch, C. J., Le Grand, R., & Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception, 31*(5), 553–566.
Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic Bulletin & Review, 23*(5), 1316–1340.

Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory processes: A balancing act across the lifespan. *Trends in Neurosciences, 39*(8), 567–579.
Nardini, M., Begus, K., & Mareschal, D. (2013). Multisensory uncertainty reduction for hand localization in children and adults. *Journal of Experimental Psychology. Human Perception and Performance, 39*, 773–787.
Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology, 18*(9), 689–693.
Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2006). Development of multisensory spatial integration and perception in humans. *Developmental Science, 9*(5), 454–464.
Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science, 283*(5406), 1272–1273.
Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196.
Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism modulates infants' selective attention to the mouth of a talking face. *Psychological Science, 26*(4), 490–498.
Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *The Journal of Neuroscience, 29*(39), 12265–12274.
Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2020). Face viewing behavior predicts multisensory gain during speech perception. *Psychonomic Bulletin & Review, 27*(1), 70–77.
Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience, 33*(12), 2329–2337.
Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour, 58*, 921–931.
Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2003). Sound induces perceptual reorganization of an ambiguous motion display in human infants. *Developmental Science, 6*, 233–244.
Senkowski, D., Saint-Amour, D., Gruber, T., & Foxe, J. J. (2008). Look who's talking: The deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. *NeuroImage, 43*(2), 379–387.
Shahin, A. J., & Miller, L. M. (2009). Multisensory integration enhances phonemic restoration. *The Journal of the Acoustical Society of America, 125*(3), 1744–1750.
Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology, 13*(13), R519–R521.
Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Review Neuroscience, 9*(4), 255–266.
Stevenson, R. A., Baum, S. H., Krueger, J., Newhouse, P. A., & Wallace, M. T. (2018). Links between temporal acuity and multisensory integration across life span. *Journal of Experimental Psychology. Human Perception and Performance, 44*(1), 106–116. https://doi.org/10.1037/xhp0000424
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.
Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36*, 314–331.
Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd, & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–52). Hillsdale, NJ: Lawrence Erlbaum.
Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335*(1273), 71–78.
Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., & Morgan, J. L. (2013). Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study. *Infancy, 18*(4), 534–553.
Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Shah, R. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language, 42*(06), 1173–1190.
Thelen, A., Matusz, P. J., & Murray, M. M. (2014). Multisensory context portends object memory. *Current Biology, 24*(16), R734–R735.
Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4–8), 411–443.
Van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: Flexible use of general operations. *Neuron, 81*(6), 1240–1253.
Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008a). Audiovisual events capture attention: Evidence from temporal order judgments. *Journal of Vision, 8*(5), 2.
Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008b). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053.
Van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55*(3), 1208–1218.
Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926–940.
Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision, 12*(13), 3.
Von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4*(10), Article e326.
Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72*(4), 871–884.
Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: Differentiation of multimodal information. *Psychological Bulletin, 121*(3), 437–456.
Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2020). Multisensory integration as a window into orderly and disrupted cognition and communication. *Annual Review of Psychology, 71*, 193–219.
Welch, R. B. (1999). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perception. In G. Aschersleben, T. Bachman, & J. Musseler (Eds.), *Vol. 129. Cognitive contributions to the perception of spatial and temporal events* (pp. 371–387). Amsterdam: Elsevier.
Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638–667.
Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? *Annual Review of Vision Science, 6*.
Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1–2), 23–43.
Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555–568.
Zion Golumbic, E., & Shavit-Cohen, K. (2019). The dynamics of attention shifts among concurrent speech in a naturalistic multi-speaker virtual environment. *Frontiers in Human Neuroscience, 13*, 386.