Cognition 214 (2021) 104743

Contents lists available at ScienceDirect

Cognition

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

The multisensory cocktail party problem in adults: Perceptual segregation of talking faces on the basis of audiovisual temporal synchrony

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

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

ABSTRACT

Social interactions often involve a cluttered multisensory scene consisting of multiple talking faces. We investigated whether audiovisual temporal synchrony can facilitate perceptual segregation of talking faces. Participants either saw four identical or four different talking faces producing temporally jittered versions of the same visible speech utterance and heard the audible version of the same speech utterance. The audible utterance was either synchronized with the visible utterance produced by one of the talking faces or not synchronized with any of them. Eye tracking indicated that participants exhibited a marked preference for the synchronized talking face, that they gazed more at the mouth than the eyes overall, that they gazed more at the eyes of an audiovisually synchronized than a desynchronized talking face, and that they gazed more at the mouth when all talking faces were audiovisually desynchronized. These findings demonstrate that audiovisual temporal synchrony plays a major role in perceptual segregation of multisensory clutter and that adults rely on differential scanning strategies of a talker's eyes and mouth to discover sources of multisensory coherence.

When listeners hear multiple and competing speech streams, how do they segregate them to access one particular speech stream? This was the question posed by Cherry (1953) in his famous Cocktail Party Problem. Although Cherry's question spawned many subsequent studies of auditory stream segregation (Bregman, 1990; McDermott, 2009), very few studies to date have investigated perceptual segregation of audiovisual streams (Senkowski, Saint-Amour, Gruber, & Foxe, 2008; Zion Golumbic & Shavit-Cohen, 2019). This is surprising, considering that most of our daily experiences are multisensory in nature (Marks, 1978; Stein & Meredith, 1993). For example, whenever we find ourselves at a party, a busy restaurant, or a packed train station, we are confronted with multiple people producing competing streams of audiovisual speech. To access any particular speech stream, to extract meaning from that speech stream, and to use the extracted information to communicate with the talker who produced it, we must be able to perceptually segregate the multiple audiovisual speech streams.

The type of perceptual segregation necessary for solving the multisensory version of the Cocktail Party Problem requires, first and foremost, a search of the scene to identify the talker of interest. Here, principles discovered in studies of visual search can be helpful in identifying mechanisms of perceptual segregation (Treisman, 2006). As Wolfe (2020) has noted, however, these principles are based on search of simple visual scenes and may not fully explain search of real-world events. Indeed, multiple-talker scenes are a perfect example of real-world events that involve search based on unisensory as well as multisensory cues. Both types of cues improve search accuracy by reducing the uncertainty that results from the constant onslaught of real-world sensory information. Crucially, however, search of multisensory scenes may differ from search of unisensory scenes because multisensory cues provide the type of redundant information that unisensory cues do not provide. Moreover, when the redundant information is integrated, perceptual salience increases and this, in turn, leads to enhanced perceptual processing, learning, and memory (Murray, Lewkowicz, Amedi, & Wallace, 2016; Partan & Marler, 1999; Rowe, 1999; Senkowski et al., 2008; Stein & Meredith, 1993; Stein & Stanford, 2008; Sumby & Pollack, 1954; Summerfield, 1979; Thelen, Matusz, & Murray, 2014; Thelen, Talsma, & Murray, 2015; Van Atteveldt, Murray, Thut, & Schroeder, 2014; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008b).

Several processes are likely to be involved in the search and segregation of a multiple talker scene. These include: (a) a rapid assessment of the scene to obtain an inventory of constituent faces and voices, (b) identification of the visible and audible speech articulations of a talker of interest, (c) intersensory binding, involving that talker's visible and audible articulations, and (d) segregation of that talker's visible and audible articulations from those of other talkers. Intersensory binding

* 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.2021.104743
Received 11 September 2020; Received in revised form 16 April 2021; Accepted 21 April 2021

and perceptual segregation are key because the former contributes to the identification of unitary and coherent multisensory targets while the latter is essential for solving the general binding problem (Stevenson, Baum, Krueger, Newhouse, & Wallace, 2018; Treisman, 2006). Ordinarily, we accomplish search and segregation easily and seemingly automatically. Of course, in reality, the ease with which we do so belies the complex attentional, perceptual, and neural processes underlying them (Murray et al., 2016; Stein & Meredith, 1993; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010; Van Atteveldt et al., 2014; Wallace & Stevenson, 2014; Wolfe, Võ, Evans, & Greene, 2011).

The search and segregation of real-world multiple talker scenes is likely to be facilitated by a number of different multisensory cues. Arguably, one of the most powerful cues is the temporal synchrony that naturally binds visual and auditory sensory inputs (Spence & Squire, 2003; Vroomen & Keetels, 2010). This is especially true in the audiovisual speech domain where the dynamic variations of a talker's visible and audible speech streams are normally tightly correlated over time (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Summerfield, 1987, 1992; Yehia, Kuratate, & Vatikiotis-Bateson, 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998). This tight temporal correlation facilitates the binding of a particular talker's visible and audible speech streams and their segregation from the visible and audible speech streams of other talkers. Importantly, the segregation is not only facilitated by audiovisual binding but also by the fact that once a talker's visible and audible speech streams are bound together, the resulting bound and redundantly specified audiovisual speech stream becomes perceptually more salient than a unisensory speech stream. This increased salience attracts greater attention and augments processing.

The beneficial effects of redundantly specified audiovisual speech have been found in a number of studies. For example, it has been found that infants and adults attend more to the source of highly salient audiovisual speech cues - the talker's mouth – than eyes when exposed to a talking face (Hillairet de Boisferon, Tift, Minar, & Lewkowicz, 2017; Lewkowicz & Hansen-Tift, 2012; Pons, Bosch, & Lewkowicz, 2015; Võ, Smith, Mital, & Henderson, 2012). Similarly, infants and adults have been found to attend more to audiovisual speech when they are exposed to speech in an unfamiliar than a familiar language (Barenholtz, Mavica, & Lewkowicz, 2016; Birulés, Bosch, Pons, & Lewkowicz, 2020), adults have been shown to exhibit better comprehension of temporally coherent audiovisual speech than auditory-only speech (Lansing & McConkie, 2003; Senkowski et al., 2008; Sumby & Pollack, 1954; Summerfield, 1992), and adults have been found to exhibit better detection of auditory speech in noise when corresponding visual speech is available (Grant & Seitz, 2000; MacLeod & Summerfield, 1987; Shahin & Miller, 2009; Sumby & Pollack, 1954). Crucially, the multisensory redundancy benefits observed in audiovisual speech processing reflect a domain-general aspect of multisensory perceptual functioning. This is evident from studies showing that redundancy effects emerge very early in life and that responsiveness to non-speech events also benefits from multisensory redundancy (Bahrick & Lickliter, 2012; Hillairet de Boisferon et al., 2017; Hillock, Powers, & Wallace, 2011; Lewkowicz, 1996, 2000a, 2010; Lewkowicz & Hansen-Tift, 2012; Lewkowicz, Leo, & Simion, 2010; Lewkowicz, Minar, Tift, & Brandon, 2015; Spence & Squire, 2003; Stevenson et al., 2018; Wallace & Stevenson, 2014).

If temporal synchrony plays such a powerful and domain-general role in attention and perceptual processing, might it also facilitate search of a cluttered multisensory scene? Studies by Van der Burg and colleagues suggest that this is likely to be the case. For example, Van der Burg et al. (2008b) investigated whether the search for a target object embedded in a crowded scene composed of many other objects can be facilitated by a spatially uninformative sound which is synchronized with the actions of the target object but not with the other objects. Subjects were asked to search for a horizontal or a vertical line segment in a scene consisting of 48 oblique line segments of various orientations. A random number of segments continuously changed color between red and green at random intervals and the target segment also changed color every 900 ms but, when it changed color, no other segments changed color. The subjects' task was to find the target and identify its orientation as rapidly as possible. Three important results were obtained. First, observers found the target and identified its orientation significantly faster when it was accompanied by a short pip sound than when it was not accompanied by this sound. Second, top-down factors were not responsible for audiovisual integration. Finally, the audiovisual integration was automatic. The authors concluded that the pip sounds automatically increased the perceptual salience of visual target changes, causing the visual target to pop out. They dubbed this phenomenon the *pip and pop effect*. In other studies, these researchers have replicated this effect and have shown that it is not due to increases in alertness or top-down temporal cueing (Van der Burg et al., 2008b; Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008a), that the auditory facilitation of visual search occurs early in sensory processing, and that facilitation modulates activity in parieto-occipital cortices (Van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011).

Consistent with the *pip and pop* effect, studies also have found that events specified by temporally synchronized auditory and visual attributes are perceived as categorically different than the same events specified by desynchronized auditory and visual attributes. For example, infants and adults experience two identical objects moving on the same path but in opposite directions as bouncing against one another when a spatially non-specific sound occurs at the point of their overlap but as streaming past one another when the sound occurs either prior to or after the objects overlap (Scheier, Lewkowicz, & Shimojo, 2003; Sekuler, Sekuler, & Lau, 1997; Shimojo, Watanabe, & Scheier, 2001; Watanabe & Shimojo, 1998, 2001).

The findings from the aforementioned studies provide *a priori* support for the possibility that temporal synchrony is likely to facilitate search and segregation of multiple talking faces. Nonetheless, it should be emphasized that the findings from studies documenting the importance of temporal synchrony cues in responsiveness to non-speech events reflect responsiveness to simple and punctate auditory and visual event attributes. These attributes differ from the more continuous types of temporal audiovisual relations inherent in fluent audiovisual speech. As discussed earlier, it is the dynamically varying temporal correlation between a talker's vocalizations and lip movements that specifies the multisensory unity of audiovisual speech. In addition, if different talkers produce semantically different speech utterances, these differences are represented by different patterns of audiovisual temporal correlation. For example, the pattern that specifies the correlation between the vocalizations and lip movements produced by someone saying "This party is so much fun" differs from the pattern produced by someone saying "The food at this party is delicious". Thus, the semantic differences of different talkers' speech productions require the processing of unique patterns of multisensory temporal statistics.

Importantly, it should be noted that the physical and neural transmission times of sensory signals differ across modalities. As a result, the binding of temporally synchronous auditory and visual inputs usually occurs within a temporal binding window. This binding window is relatively large in infancy (Lewkowicz, 1996, 2010), narrows gradually during childhood (Chen, Shore, Lewis, & Maurer, 2016; Lewkowicz & Flom, 2014), and reaches its smallest size in adolescence (Hillock et al., 2011; Hillock-Dunn & Wallace, 2012). Moreover, the binding window is relatively small for simple punctate audiovisual events (e.g., a bouncing ball), but it is larger for complex continuous events such as audiovisual speech (Lewkowicz, 1996, 2000b, 2010; Stevenson & Wallace, 2013; Vroomen & Keetels, 2010).

The purpose of the current study was to investigate whether the temporal synchrony of fluent audiovisual speech might facilitate the segregation of multiple talking faces. To do so, we conducted two experiments in which we manipulated the audiovisual temporal relations between a single audible speech utterance and multiple talking faces. In Experiment 1, we presented four identical talking faces producing the same visible speech utterance in a temporally jittered fashion and the same audible speech utterance which was either synchronized with the visible speech utterance produced by one of the talkers' faces or not. In Experiment 2, we replicated Experiment 1 except that this time we presented four different talking faces to determine whether identity cues might contribute to segregation. To investigate perceptual segregation, we measured selective attention with an eye tracker. Our expectation was that participants would prefer the audiovisually synchronized talker's face, that identity cues would contribute to this preference, and that attention to the talker's eyes and mouth would depend on the temporal coherence of audiovisual speech.

1. Experiment 1

To determine whether an audiovisually synchronized talking face might elicit greater attention when it competes for attention with other talking but audiovisually desynchronized faces, we presented composite videos of four identical talking faces silently articulating the same utterance in a temporally jittered fashion. At the same time, we presented the audible version of the same utterance. Because the visible utterance articulated by the four faces was temporally jittered, the audible utterance was temporally synchronized with one of the talking faces while it was desynchronized with the other three talking faces. The participants' only task was to attend to the composite video. To encourage them to pay attention, the participants were asked to identify the "talking" face when a static version of the four faces was presented at the end of each trial. Throughout the experiment, we tracked the participants' eye gaze to measure selective attention to each of the four talking faces as well as to the eyes and mouth of each face.

1.1. Method

1.1.1. Participants

We tested 40 adults (32 females) who ranged between 20 and 39 years of age (mean age = 23 years, SD = 3.58 years). All participants were monolingual English speakers who volunteered for the study and who gave their informed consent.

1.1.2. Apparatus and stimuli

We used a REDn SensoMotoric Instruments (SMI, Teltow, Germany) remote eye tracker running at a sampling rate of 60 Hz on a Dell Precision M4800 laptop computer. The eye tracker's camera was attached to the bottom of the computer's screen. We used SMI's iViewRed software to control the eye tracker camera and to process the eye gaze data and SMI's Experiment Center software to control stimulus presentation and data acquisition. The eye tracker was placed on a table in front of the participants in a quiet room and the participants' eyes were approximately 60 cm from the eye tracker camera. The initial instructions and all visual stimuli were presented on the computer's 11 × 13 in screen while the auditory stimuli were presented through a pair of Sony Professional headphones (Model # MDR-7506) at a comfortable listening level.

The experiment consisted of a calibration phase, two 15 s practice trials, and thirty-two 15 s test trials. A small yellow star was used to calibrate eye gaze and was presented in the center of the screen as well as in each of the four corners of the screen. Composite videos were created in Premiere (Adobe Systems, Inc., San Jose, CA) for the practice and test trials. Each composite video consisted of four equally sized videos of the same female face presented in each of the four quadrants of the screen. One female actor appeared in the composite videos presented during the practice trials while two other female actors appeared in the composite videos presented during the test trials. The visible speech utterance articulated by each face, as well as the concurrently presented audible utterance, were all the same in each respective composite video.

Fig. 1 shows a still picture of one of the composite videos presented during the test trials. As can be seen, participants saw the same actor's face in the four equally sized quadrants. Participants were given 16 synchrony and 16 asynchrony test trials. In the synchrony test trials, the audible utterance was temporally synchronized with the visible speech utterance produced by one of the talkers' faces (the target) and desynchronized from the visible speech utterances produced by the other three talking faces (the distractors). In contrast, in the asynchrony test trials, the audible utterance was desynchronized from all four visible utterances.

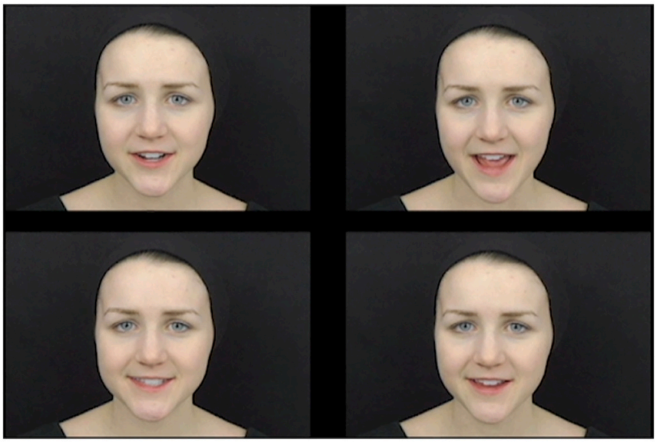

Fig. 1. Screen-shot of one of the composite videos presented in Experiment 1.

We filmed each of the two test-phase female actors speaking two different sets of two different utterances[1], yielding a total of four different videos (see Video S1 for example). We then used each of these four videos to construct four different sets of test trials. Each set consisted of four synchrony and four asynchrony test trials. The synchrony test trials included one target stimulus and three distractor stimuli and the target stimulus was presented in each of the four quadrants across trials. The asynchrony trials were identical to the synchrony trials except that the target face was now audiovisually desynchronized and thus was now a "virtual" target. Crucially, the virtual target was located in the same quadrant as in the synchrony trials. This enabled us to compare responsiveness to a particular talking face when its visible articulations either were synchronized or desynchronized with the audible utterance.

To construct the composite videos, we began with four identical audiovisually synchronized videos. Then, we desynchronized the visible articulations produced by the three distractor talking faces with respect to the audible utterance in a temporally jittered fashion. This procedure resulted in a composite video in which the visible articulations of one of the four faces were synchronized with the audible utterance while the articulations of the other three faces were desynchronized from it. This procedure also created the perceptual impression that each face was saying something different and, thus, rendered the task of detecting the audiovisually synchronized talking face more challenging. To temporally jitter the videos, we started the visible articulations produced by each distractor face at an increasingly later point into the utterance relative to the start of the visible articulation produced by the target

[1] The four utterances 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."

face. This meant that the visible speech stream articulated by each distractor face was temporally delayed with respect to the auditory speech stream by a fixed interval of time.[2] The net result of this temporal jittering was that the visible speech articulations produced by all four talking faces began simultaneously at the start of each test trial but that only the visible articulations of the target face were synchronized with the audible utterance.

It is important to note that desynchronization of fluent audiovisual speech differs from desynchronization of punctate events such as a bouncing ball or a flashing/beeping light. For punctate events, desynchronization can be specified precisely by the temporal interval separating either the visible impact of a bouncing object and its impact sound or a flash and its accompanying sound. Even for isolated speech syllables, it is possible to specify precisely the interval separating the opening of the mouth and phonation. This is not the case for fluent audiovisual speech. Desynchronization of audiovisual speech means that the dynamic variations in a talker's visible mouth movements and the accompanying vocalizations are no longer zero lag correlated across time. This temporal shift means that there are multiple points of intersensory discordance of the dynamic variations of the physical characteristics of visible and audible speech streams and of the dynamic variations of the visible and audible phonetic and semantic cues inherent in it. To make matters even more complex, natural synchronized fluent audiovisual speech is characterized by a delay in the onset of the voice relative to the visible movements of the mouth because movements of the vocal tract precede phonation by anywhere between 100 and 300 ms (Chandrasekaran et al., 2009). As a result, the audible and visible speech streams making up a fluent audiovisual speech stream must be temporally separated by more than 300 ms if they are to be perceived as desynchronized. Accordingly, as seen above, we jittered the audible and visible speech streams for each distractor by more than 300 ms to ensure that participants perceived them as audiovisually desynchronized.

1.1.3. Procedure

An experimenter was seated on one side of the participant to monitor the experiment. Unless asked a specific question, the experimenter did not interact with the participant during the course of the experiment. All data were acquired from the right eye and the experiment began with the calibration routine. Calibration was deemed acceptable if the point of fixation fell within less than one degree of visual angle of the star's position. The calibration phase was followed by an instruction phase. During this phase, written instructions were presented on the computer screen and, once participants read these instructions, they were asked if they had any questions. If they had no questions, participants proceeded to the practice phase during which they were familiarized with the procedure and their task. In this phase, participants saw a continuously looming/receding yellow disc in the center of the screen and were told that a test trial would start whenever they looked at the disc. The participants were also given the following instructions: "You will see four faces on the screen and hear a voice talking. Please look carefully to determine which face is talking."

The practice phase consisted of two 15 s trials. During these trials, participants saw composite videos composed of four videos of the same person (please recall that this person was different from the two people who served as actors for the test trials). During the first practice trial, participants saw the four faces articulating the same utterance and heard the same audible utterance which was synchronized with the visible speech utterance produced by one of the talking faces. During the second practice trial, participants saw the same four talking faces except that this time the audible utterance was temporally desynchronized with all four visible speech utterances. After each practice trial, participants saw a composite video of the four faces that they saw in the previous composite talking video except that now the faces were largely still except for the occasional blink. The participants were asked to look at these faces and indicate which one of them was the talking face in the preceding composite video. To indicate their choice, participants pressed a 1, 3, 7, or 9 on the numerical keypad. These numbers were chosen because they are spatially congruent with the four quadrants in which the faces appeared. Once the practice trials were completed, the participants were given a chance to ask questions again and then the experiment proper began.

The 32 test trials were presented according to one of four randomly generated test trial orders, with participants randomly assigned to one of these orders. Trial randomization was used to minimize the participants' ability to predict the specific actor, the utterance, the quadrant in which the target face was presented, and whether the audible and visible speech of the target face was synchronized or not. As in the practice trials, immediately following each test trial participants saw a composite still image of the four faces from the preceding test trial, and were asked to indicate "the talking face" by pressing one of the keys on the numerical key pad. Please note that the sole purpose of this task was to induce the participants to attend to the displays throughout the experiment and, thus, we did not record their choices.

To quantify selective attention, we created a face area-of-interest (AOI) for each of the four faces as well as an eye and mouth AOI for each of the four faces (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 at each talking face. This measure was 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. This measure was computed by dividing the total amount of looking at the eyes and mouth, respectively, by the total amount of looking at that particular face.

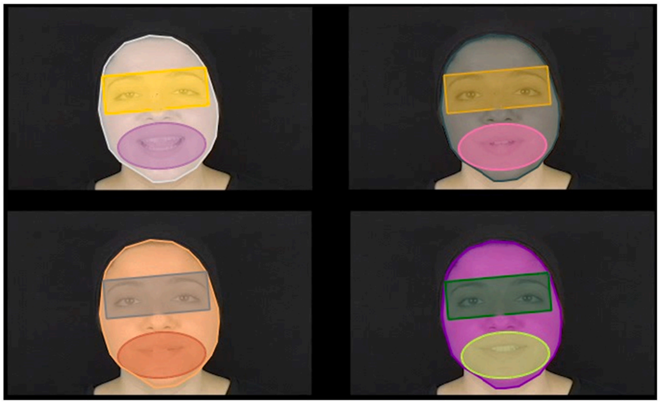

Fig. 2. Screen-shot of one of the composite videos of the four talking faces and the AOIs corresponding to the face, eyes, and mouth.

[2] The intervals for the two utterances 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 utterances 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 utterance 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. Importantly, please note that despite the fact that some of intervals separating the visible and audible streams differed by less than 1 s, the individual videos were perceptually different from one another.

1.2. Results

The principal question was whether the segregation of a cluttered visual scene composed of multiple talking faces is facilitated by concurrent audible speech when the audible speech is synchronized with the visible speech articulations of one of the talking faces. Importantly, the talking faces in the current experiment were identical which made the task relatively difficult due to the absence of distinctive visual discriminative cues. Nonetheless, the express purpose of this experiment was to assess the role of audiovisual synchrony in perceptual segregation in its purest form in terms of facial discriminative cues. Evidence of successful segregation would be manifest in longer gaze to the audiovisually synchronized talking face than to the audiovisually desynchronized talking faces.

1.2.1. Face AOIs

To determine whether participants preferentially fixated the audiovisually synchronized talking face, we compared the PTLT scores for the target face with the average of the PTLT scores for the three distractor faces. As a first step, we performed a preliminary repeated-measures analysis of variance (ANOVA), with Synchrony Condition (2; Synchrony, Asynchrony), Actor (2), Utterance (2), Quadrant (4), and Stimulus Type (2; Target, Distractor) as within-subjects variables. The purpose of this analysis was to determine whether the specific actor and/or utterance affected responsiveness. Results of this analysis yielded main effects of Synchrony Condition, $F(1, 39) = 1206.53$, $p < .001$, $\eta_p^2 = 0.97$, and Stimulus Type, $F(1, 39) = 776.59$, $p < .001$, $\eta_p^2 = 0.95$, and a Synchrony Condition x Stimulus Type interaction, $F(1, 39) = 795.00$, $p < .001$, $\eta_p^2 = 0.95$. Neither the specific actor nor utterance affected responsiveness. As a result, we collapsed the data across actor and utterance and re-analyzed them with a repeated-measures ANOVA, with Synchrony Condition (2), Quadrant (4), and Stimulus Type (2) as within-subjects variables. Again, we compared the PTLT scores for the target face versus the average of the PTLT scores for the three distractor faces. This analysis yielded significant main effects of Synchrony Condition, $F(1, 39) = 1206.53$, $p < .001$, $\eta_p^2 = 0.97$, Quadrant, $F(1, 39) = 5.25$, $p < .01$, $\eta_p^2 = 0.12$, and Stimulus Type, $F(1, 39) = 776.6$ $p < .001$, $\eta_p^2 = 0.95$. In addition, this analysis yielded a significant Quadrant x Stimulus Type, $F(1, 117) = 7.26$, $p < .001$, $\eta_p^2 = 0.16$, a Synchrony Condition x Stimulus Type, $F(1, 39) = 795.00$, $p < .001$, $\eta_p^2 = 0.95$, interaction as well as a non-significant Synchrony Condition x Stimulus Type x Quadrant interaction, $F(3, 117) = 1.51$, $p = .214$, $\eta_p^2 = 0.037$.

The most relevant effect obtained in the foregoing analysis is the Synchrony Condition x Stimulus Type interaction. As Fig. 3 shows, in the synchrony condition, participants gazed longer at the target face than the distractor faces, whereas in the asynchrony condition, participants gazed equally at the two types of faces. Planned comparison tests indicated that, in the synchrony condition, gaze duration at the target face was longer than at the distractor faces, $F(1, 39) = 844.67$, $p < .001$, but that in the asynchrony condition, this was not in the case, $F(1, 39) = 0.31$, $p = .58$. Additional planned comparisons indicated that participants gazed longer at the distractor faces in the asynchrony condition than in the synchrony condition, $F(1, 39) = 521.19$, $p < .001$, and that they gazed longer at the target face in the synchrony condition than in the asynchrony condition, $F(1, 39) = 917.62$, $p < .001$.

The design of the current experiment involved presenting each actor x utterance combination four times to counterbalance the quadrant of target presentation. This design makes it possible to ask whether the preference for the audiovisually synchronized talking face was a stable characteristic of responsiveness regardless of quadrant of target-stimulus presentation. Indeed, the results of the principal ANOVA described earlier indicated that the Synchrony Condition x Stimulus Type x Quadrant interaction was not statistically significant, meaning that responsiveness was similar regardless of quadrant of target-stimulus presentation. As Fig. 4 shows, the preference for the audiovisually synchronized target face was observed in each of the four quadrants in the synchrony condition while its absence was found in all four quadrants in the asynchrony condition. When this non-significant effect of Quadrant is combined with the highly significant Synchrony Condition x Stimulus Type interaction reported earlier, it becomes clear that quadrant of stimulus presentation did not affect responsiveness. This result is further evidence of the robust nature of the preference for the audiovisually synchronized talking face.

1.2.2. Eyes/Mouth AOIs

Next, we investigated the relative deployment of eye gaze to the talker's eyes and mouth with a repeated-measures ANOVA with AOI (2), Synchrony Condition (2), and Stimulus Type (2) as within-subjects factors. This analysis yielded several significant main effects, including AOI, $F(1, 39) = 63.09$, $p < .001$, $\eta_p^2 = 0.62$, Synchrony Condition, $F(1, 39) = 47.95$, $p < .001$, $\eta_p^2 = 0.55$, and Stimulus Type, $F(1, 39) = 35.78$, $p < .001$, $\eta_p^2 = 0.48$. It also yielded a significant AOI x Synchrony Condition interaction, $F(1, 39) = 61.07$, $p < .001$, $\eta_p^2 = 0.61$, a Synchrony Condition x Stimulus Type interaction, $F(1, 39) = 39.10$, $p < .001$, $\eta_p^2 = 0.50$, and an AOI x Synchrony Condition x Stimulus Type interaction, $F(1, 39) = 4.42$, $p < .05$, $\eta_p^2 = 0.10$.

The two most interesting and theoretically relevant findings are the main effect of AOI and the AOI x Synchrony Condition x Stimulus Type interaction. Both findings are depicted in Fig. 5. As can be seen, the AOI effect reflects the fact that participants gazed more at the mouth than the eyes in each condition. As can also be seen in Fig. 5, the triple interaction reflects the fact that participants exhibited different patterns of selective

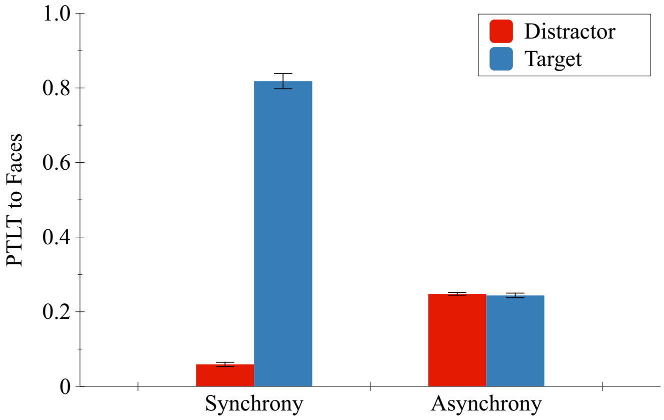

Fig. 3. Mean proportion of total looking time at the distractor and target faces across the two synchrony conditions in Experiment 1. Error bars represent the standard errors of the mean.

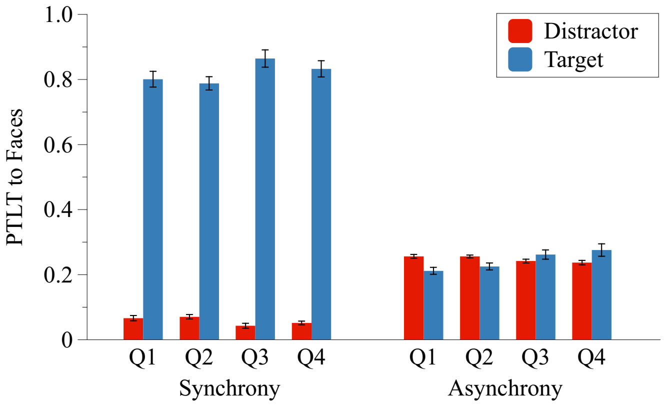

Fig. 4. Mean proportion of total looking time at the distractor and target faces in the four quadrants (Q) across the two synchrony conditions in Experiment 1. Error bars represent the standard errors of the mean.

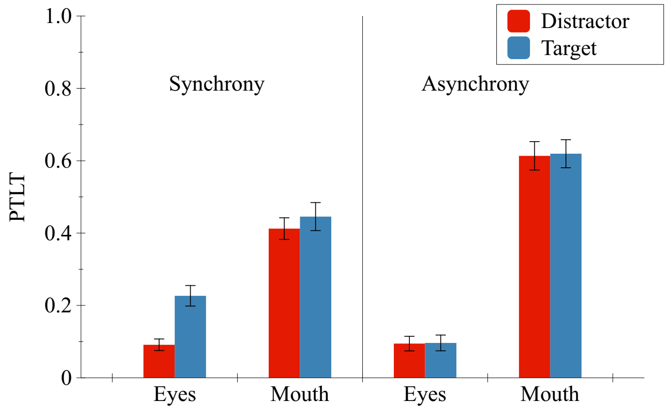

Fig. 5. Mean proportion of total looking time to the distractor- and target-face eyes and mouth across the two synchrony conditions in Experiment 1. Error bars represent the standard errors of the mean.

attention to the eyes and mouth of the distractor and target faces in the two synchrony conditions. This finding is interesting in the context of studies showing that adults direct more of their gaze at the mouth when they are actively processing audiovisual speech (Barenholtz et al., 2016; Birulés et al., 2020) but that they direct more of their gaze at the eyes when they are not actively processing speech (Võ et al., 2012). Thus, given that the participants' task was to engage in audiovisual speech processing, it is not surprising that they gazed more at the talker's mouth.

To further investigate the triple interaction depicted in Fig. 5, we performed a series of planned comparisons. These comparisons yielded several effects. First, they indicated that participants gazed more at the mouth than the eyes in both the synchrony, $F(1, 39) = 30.95$, $p < .001$, and the asynchrony, $F(1, 39) = 87.01$, $p < .001$, conditions. Second, they indicated that participants gazed more at the eyes of the target face than at the eyes of the distractor face in the synchrony condition, $F(1, 39) = 31.89$, $p < .001$, but not in the asynchrony condition, $F(1, 39) = 0.04$, $p = .83$. Finally, they showed that participants gazed equally at the mouth of the target and distractor face in the synchrony condition, $F(1, 39) = 1.32$, $p = .26$, as well as in the asynchrony condition, $F(1, 39) = 0.20$, $p = .66$, but that they gazed more at the mouth in the Asynchrony Condition than in the Synchrony Condition, $F(1, 39) = 64.48$, $p < .001$.

1.2.3. Latency of response

In a final analysis, we examined response latency scores to determine whether an audiovisually synchronized talking face in a given quadrant elicited faster initial attention in the synchrony condition than did the same but desynchronized talking face presented in the same quadrant in the asynchrony condition. The scores used for this analysis represented the length of time between the onset of the composite video and first fixation on the target face. Crucially, a preliminary examination of the data showed that the most frequent first look was directed at the face presented in the top left quadrant regardless of whether the face presented there was audiovisually synchronized or not. Specifically, we found that, out of 1280 trials (32 trials/participant x 40 participants), 902 initial fixations (70.45%) were to the top left face. Given that the first look was not always directed at the target stimulus, the data that contributed to the latency-score analysis represent the time to the first fixation of the target talking face regardless of whether the participant looked elsewhere first or not. The mean response latency to the audiovisually synchronized talking face was 1387.8 ms while it was 1491.4 ms to the desynchronized face, $F(1, 39) = 0.83$, $p = .37$. This indicates that participants did not orient their initial gaze to the audiovisually synchronized talking face faster than to the same but desynchronized talking face.

1.3. Discussion

As expected, we found that participants looked far longer at the audiovisually synchronized talking face than at audiovisually desynchronized distractor faces. Importantly, we also found that the preference for the audiovisually synchronized talking face did not depend on the specific person nor the specific utterance spoken by that person. This shows that the marked preference for the audiovisually synchronized talking face reflects a general perceptual phenomenon. This conclusion is further buttressed by the finding that the preference for the audiovisually synchronized target face was highly stable regardless of its spatial position in the synchrony condition and by the concurrent finding that the absence of such a preference was highly stable in each of the quadrants in the asynchrony condition. This overall pattern of responsiveness and its highly stable nature suggests that the preference obtained in the synchrony condition reflects a relatively automatic process.

Analyses of gaze directed at the eyes and mouth shed additional light on the relative perceptual salience of these two parts of the talking face. The most salient and attractive part of the talking faces was the talker's mouth as evidenced by the fact that participants deployed more than twice as much time gazing at the mouth than at the eyes. Moreover, participants gazed more at the mouth when the audible speech utterance was not synchronized with any talking faces than they did when the audible speech utterance was synchronized with one of them. These findings suggest that participants were engaged in audiovisual speech processing. At the same time, the findings indicate that participants gazed more at the eyes when detection of a talking face was relatively easy (i.e., in the synchrony condition) than when detection was more challenging (i.e., in the asynchrony condition). Conversely, when the speech processing task was challenging due to the fact that none of the talking faces were audiovisually synchronized, participants gazed equally to the eyes of the virtual target face and the eyes of the distractor faces. When considered together, the eye and mouth gaze data suggest that the relative allocation of selective attention to each region is determined by the task at hand and the audiovisual coherence of the talking face.

2. Experiment 2

Experiment 1 demonstrated that participants exhibited a marked preference for an audiovisually synchronized talking face when it competed for attention with other audiovisually desynchronized talking faces. Importantly, however, it should be noted that the talking faces presented in Experiment 1 were identical and, thus, it is not clear whether the marked preference obtained in Experiment 1 reflects the absence of unique visible and audible identity cues that normally accompany multiple talkers. In other words, might identity cues play a role in perceptual segregation of multiple talking faces? Prior studies have found that adults not only perceive audible and visible speech cues of specific talkers but that they also link such cues to represent individual talkers (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a, 2004b). As a result, it is likely that individual identity cues play a role in perceptual segregation of multiple talking faces.

To investigate the possible role of individual audible and visible speech identity cues in perceptual segregation of multiple talking faces, we used the same method in this experiment as in Experiment 1 except that now we presented four different talking faces and an audible speech utterance that had the unique acoustic and prosodic properties of the individual speaking in a given trial. Given the robust findings from Experiment 1, we expected that participants would again exhibit a preference for the synchronized talking face in the synchrony condition and no preference in the asynchrony condition. In addition, given the aforementioned findings showing that adults can perceive and link unique audible and visible speech cues, we predicted that the unique speech cues associated with each talker were likely to yield two possible outcomes. First, the unique speech cues of each talker might increase the overall discriminability of the faces and voices and lead to an even greater preference for the audiovisually synchronized talking face than the one observed in Experiment 1. Second, given that each unique talker and her voice were presented four times over the course of the experiment, participants may associate the unique facial and vocal attributes of each individual talker. If they do, this may compel them to match even desynchronized visible and audible speech utterances and, thus, they may prefer the virtual target talking face over the other talking faces whose unique facial and vocal attributes do not match.

2.1. Method

2.1.1. Participants

Twenty seven adults (20 females), who ranged between 19 and 44 years of age (mean age = 24.5 years, SD = 7.9 years), were tested. All participants were monolingual English-speaking volunteers who gave their informed consent prior to taking part in the study.

2.1.2. Apparatus and stimuli

The apparatus used in this study was the same as that used in Experiment 1 but the stimuli were now different. As can be seen in Fig. 6, the composite videos presented in this experiment consisted of four different female faces (please note that the four female faces presented during the practice trials were not seen during the test trials). The visible utterance spoken by each actor, as well as the concurrently presented audible utterance, were all the same. Thus, in contrast to Experiment 1, here participants saw four different talking faces and heard a unique voice speaking the audible utterance each time a different target actor spoke (for an example of the composite videos presented in Experiment 2 see Video S2).

As in Experiment 1, participants were given 16 synchrony and 16 asynchrony test trials. During the synchrony trials, the audible speech utterance was temporally synchronized with the visible utterance produced by one of the four talking faces (the target) but desynchronized from the visible utterances produced by the other three talking faces (distractors). We used the same method as in Experiment 1 to temporally jitter the distractors. During the asynchrony trials, the audible utterance was desynchronized from the visible utterances produced by all four talking faces and all four were presented in a temporally jittered fashion.

We filmed each of the four test-phase female actors speaking two different utterances. During filming, we asked each actor to read the utterance during the same interval of time thus ensuring that they spoke at a similar rate of speed. We then used each of the eight videos to construct eight different sets of eight test trials each. Each 8-trial set consisted of four synchrony and four asynchrony test trials. The quadrant in which the target was presented during the four synchrony test trials was counterbalanced. This resulted in the target appearing equally often in each of the four quadrants. Like in Experiment 1, we designated the specific target quadrants used during the four synchrony trials as the target quadrants for the four asynchrony trials and desynchronized the audible utterance from the visible utterance in each of those respective quadrants by the same temporal interval as in Experiment 1.

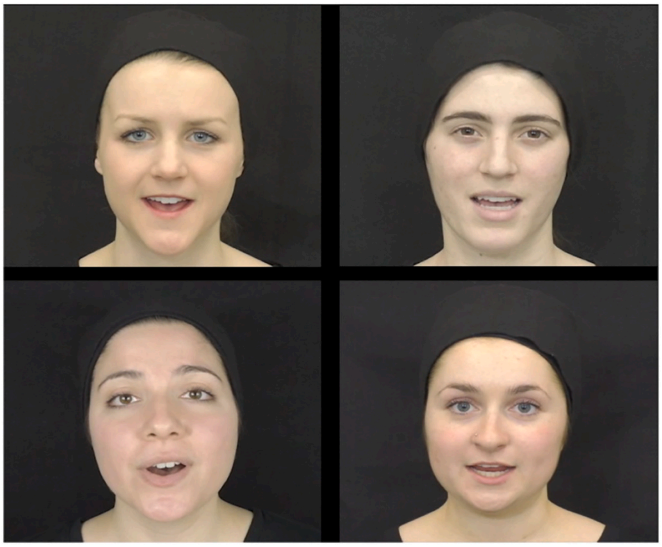

Fig. 6. Screen-shot of one of the composite videos presented in Experiment 2.

We used four of the 8-trial video sets to construct one 32-trial stimulus set and the other four of the 8-trial video sets to construct a second 32-trial stimulus set. The specific actor-utterance pairings were counterbalanced across the two 32-trial stimulus sets, with the constraint that two of the actors spoke one utterance while the other two actors spoke the other utterance in each set, respectively. As in Experiment 1, we temporally jittered the distractors by delaying the visible articulations for each of the distractor faces increasingly later into the utterance relative to the start of the visible articulation produced by the target face.[3] This way, like in Experiment 1, the visible articulations all began at the same time at the beginning of each test trial but only the visible articulations of the target face were synchronized with the audible utterance. The temporal jitter created the impression that each face was saying something different.

2.1.3. Procedure

The procedure used in this experiment was the same as that used in Experiment 1. The test trials were presented in random order and participants were randomly assigned to one of the two stimulus sets.

2.2. Results

2.2.1. Face AOIs

Because each actor spoke a different utterance across the two randomization groups, there was the possibility that the specific actor-utterance combination influenced responsiveness. To determine if this was the case, the first preliminary analysis investigated whether the specific actor and the utterance spoken by that actor affected gaze behavior. To do so, as in Experiment 1, 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; Synchrony, Asynchrony), Actor (4), Quadrant (4), and Stimulus Type (2; Target, Distractor) as within-subjects factors and Randomization Group (2) as a between-subjects factor. Results of this ANOVA indicated that the only theoretically meaningful effect involving Actor and Randomization Group was a significant Synchrony Condition x Actor x Stimulus Type interaction, $F(3, 75) = 6.28$, $p < .001$, $\eta_p^2 = 0.20$. Inspection of this effect revealed, however, that the pattern of responsiveness as a function of Actor was nearly identical across the four actors, with the primary reason for the significant effect being variation in the magnitude of the difference in gaze to the target vs. the distractor faces in the Asynchrony condition.

Because neither Actor nor Randomization Group affected responsiveness, we collapsed the data across these two factors and re-analyzed the PTLT scores with a repeated-measures ANOVA, with Synchrony Condition (2), Quadrant (4), and Stimulus Type (2) as within-subjects factors. We obtained several significant main effects, including Synchrony Condition, $F(1, 26) = 268.59$, $p < .001$, $\eta_p^2 = 0.91$, Quadrant, $F(1, 26) = 3.16$, $p < .05$, $\eta_p^2 = 0.11$, Stimulus Type, $F(1, 26) = 405.85$, $p <$

[3] 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).

.001, $\eta_p^2 = 0.94$, as well as two significant two-way interactions, including a Quadrant x Stimulus Type interaction, $F(1, 26) = 2.96$, $p < .05$, $\eta_p^2 = 0.10$, and a Synchrony Condition x Stimulus Type interaction, $F(1, 26) = 317.44$, $p < .001$, $\eta_p^2 = 0.92$. Fig. 7 shows the latter two-way interaction. As can be seen, participants gazed longer at the target than distractor faces in both conditions, but they gazed much longer at the target face in the synchrony than in the asynchrony condition. Planned comparison tests indicated that eye gaze directed at the target face was significantly greater than at the distractor faces in both the synchrony, $F(1, 26) = 475.36$, $p < .001$, and asynchrony, $F(1, 26) = 23.16$, $p < .001$, conditions. Moreover, planned comparisons showed that gaze directed at the distractor faces was greater in the asynchrony than synchrony condition, $F(1, 26) = 350.28$, $p < .001$, and that gaze at the target face was greater in the synchrony than in the asynchrony condition, $F(1, 26) = 302.17$, $p < .001$.

Finally, given that each actor x utterance combination was presented four times to counterbalance the quadrant of target presentation, it was possible to determine again whether the preference for the audiovisually synchronized talking face was a stable characteristic of responsiveness regardless of the target's spatial position. Indeed, the Synchrony Condition x Stimulus Type x Quadrant interaction was not statistically significant, $F(3, 78) = 1.23$, $p = .30$, $\eta_p^2 = 0.04$, thus indicating that the pattern of gaze behavior directed at the distractor and target stimuli did not vary as a function of Quadrant nor Synchrony Condition. As in Experiment 1, and as can be seen in Fig. 8, the marked preference for the target stimulus in the Synchrony condition was highly stable (essentially identical) across the four quadrants. Similarly, the preference in the Asynchrony condition - though much smaller in magnitude – was evident for each "virtual" target.

2.2.2. *Eyes/Mouth AOIs*

To examine the relative distribution of gaze directed at the talker's eyes and mouth, we used a repeated-measures ANOVA with AOI (2), Synchrony Condition (2), and Stimulus Type (2) as within-subjects factors. Results of this analysis yielded several main effects, including an effect of AOI, $F(1, 26) = 22.17$, $p < .001$, $\eta_p^2 = 0.46$, Synchrony Condition, $F(1, 26) = 12.35$, $p < .01$, $\eta_p^2 = 0.32$, and Stimulus Type, $F(1, 26) = 40.11$, $p < .001$, $\eta_p^2 = 0.61$. The analysis also yielded an AOI x Synchrony Condition interaction, $F(1, 26) = 20.40$, $p < .001$, $\eta_p^2 = 0.44$, a Synchrony Condition x Stimulus Type interaction, $F(1, 26) = 25.93$, $p < .001$, $\eta_p^2 = 0.50$, and an AOI x Synchrony Condition x Stimulus Type interaction, $F(1, 26) = 7.74$, $p < .01$, $\eta_p^2 = 0.23$. The triple interaction is the most interesting and relevant finding. As can be seen in Fig. 9, and as supported by the significant main effect of AOI, participants gazed more at the mouth than the eyes. Furthermore, as indicated by planned comparisons, participants gazed more at the mouth than the eyes in both the synchrony, $F(1, 26) = 12.60$, $p < .01$, and the asynchrony, $F(1, 26) = 27.84$, $p < .001$, conditions, that they gazed more at the eyes of the target than the eyes of the distractor face in the synchrony condition, $F(1, 26) = 24.14$, $p < .001$), but not in the asynchrony condition, $F(1, 26) = 0.006$, $p = .94$, that they gazed equally at the mouth of the target and distractor face in the synchrony condition, $F(1, 26) = 0.24$, $p = .63$, and in the asynchrony condition, $F(1, 26) = 0.05$, $p = .83$, and that they gazed more at the mouth in the asynchrony than in the synchrony, $F(1, 26) = 28.11$, $p < .001$, condition.

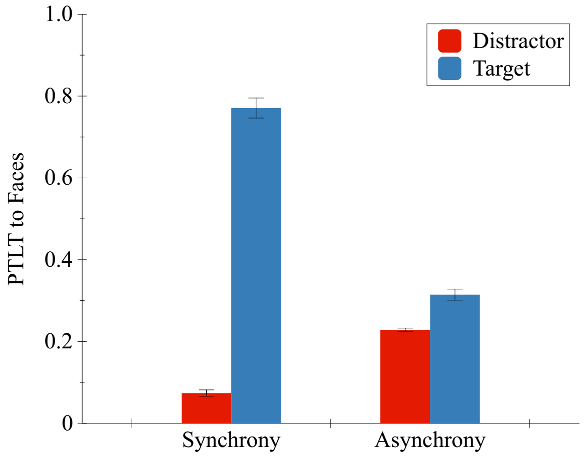

Fig. 7. Mean proportion of total looking time at the distractor and target faces across the two synchrony conditions in Experiment 2. Error bars represent the standard errors of the mean.

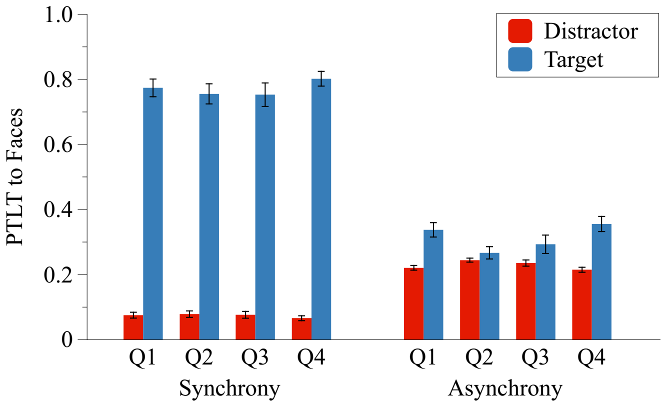

Fig. 8. Mean proportion of total looking time at the distractor and target faces in each quadrant (Q) across the two synchrony conditions in Experiment 2. Error bars represent the standard errors of the mean.

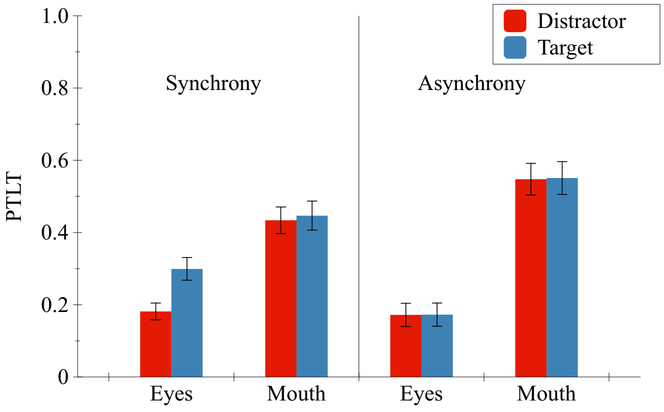

Fig. 9. Mean proportion of total looking time at the distractor- and target-face eyes and mouth across the two synchrony conditions in Experiment 2. Error bars represent the standard errors of the mean.

2.2.3. *Latency of response*

Finally, a preliminary analysis of response latency scores once again revealed that the most frequent first look was directed at the face presented in the top left quadrant regardless of whether the face presented there was audiovisually synchronized or not. That is, out of 864 trials (32 trials/participant x 27 participants), 531 initial fixations (61.4%) were directed to the top left quadrant. An analysis of response latency scores - the time to the first fixation of the target talking face regardless of whether the participant looked elsewhere first or not - indicated that the mean response latency to the audiovisually synchronized talking face was 1277 ms and 1294 ms to the same but desynchronized face, $F(1, 26) = 0.04$, $p = .84$. Thus, like in Experiment 1, participants did not orient their initial gaze to the audiovisually synchronized talking face faster than to the same but desynchronized talking face.

2.2.4. Comparison of experiments 1 and 2

2.2.4.1. Face AOIs. Although the response pattern to the talking faces obtained in the two experiments was similar, a visual comparison of Figs. 3 and 7 reveals that participants gazed more at the target face in the synchrony condition in Experiment 1 but that they did so in both conditions in Experiment 2. To determine whether this difference was statistically significant, we compared the data from the two experiments by way of a repeated measures ANOVA, with Synchrony Condition (2), Quadrant (4), and Stimulus Type (2) as within-subjects factors and Experiment (2) as a between-subjects factor. This analysis yielded several significant main effects, including Synchrony Condition, $F(1, 65) = 1150.4$, $p < .001$, $\eta_p^2 = 0.95$, Quadrant, $F(3, 195) = 4.63$, $p < .01$, $\eta_p^2 = 0.07$, and Stimulus Type, $F(1, 65) = 1123.76$, $p < .001$, $\eta_p^2 = 0.94$, several significant two-way interactions, including a Synchrony Condition x Experiment interaction, $F(1, 65) = 17.0$, $p < .001$, $\eta_p^2 = 0.21$, a Quadrant x Experiment interaction, $F(3, 195) = 3.61$, $p < .05$, $\eta_p^2 = 0.05$, a Quadrant x Stimulus Type interaction, $F(3, 195) = 5.83$, $p < .001$, $\eta_p^2 =$ 0.08, and two three-way interactions, including a Quadrant x Stimulus Type x Experiment interaction, $F(3, 195) = 3.45$, $p < .05$, $\eta_p^2 = 0.05$, and a Synchrony Condition x Stimulus Type x Experiment interaction, $F(1, 65) = 12.36$, $p < .001$, $\eta_p^2 = 0.16$. The final three-way interaction confirms the visual impression noted above and a post-hoc Tukey test confirmed that the locus of the difference between the two experiments was the asynchrony condition in Experiment 2 where gaze duration to the target stimulus was greater than to the distractor stimuli ($p < .025$).

2.2.4.2. Eyes/Mouth AOIs. A visual comparison of Figs. 5 and 9 suggests that the patterns of gaze directed at the eyes and mouth did not differ across the two experiments. To determine if this was the case, we compared the data from the two experiments by way a repeated-measures ANOVA with AOI (2), Synchrony Condition (2), and Stimulus Type (2) as within-subjects factors and Experiment (2) as a between-subjects factor. This analysis yielded significant main effects, including AOI, $F(1, 65) = 75.38$, $p < .001$, $\eta_p^2 = 0.54$, Synchrony Condition, $F(1, 65) = 47.86$, $p < .001$, $\eta_p^2 = 0.42$, and Stimulus Type, $F(1, 65) = 60.52$, $p < .001$, $\eta_p^2 = 0.48$, two-way effects, including a Synchrony Condition x Experiment interaction, $F(1, 65) = 12.16$, $p < .001$, $\eta_p^2 =$ 0.16, a Synchrony Condition x AOI interaction, $F(1, 65) = 71.43$, $p <$ .001, $\eta_p^2 = 0.52$, a Stimulus Type x AOI interaction, $F(1, 65) = 6.37$, $p <$ .05, $\eta_p^2 = 0.09$, a Synchrony Condition x Stimulus Type interaction, $F(1, 65) = 59.1$, $p < .001$, $\eta_p^2 = 0.48$, and a three-way effect consisting of a Synchrony Condition x Stimulus Type x AOI interaction, $F(1, 65) = 9.5$, $p < .01$, $\eta_p^2 = 0.13$. Crucially, the Synchrony Condition x AOI x Experiment interaction was not statistically significant, $F(1, 65) = 2.12$, $p =$ .15. This confirms that the differential distribution of selective attention to the eyes and mouth did not differ across the two experiments.

2.3. Discussion

Like in Experiment 1, participants deployed the bulk of their attention to an audiovisually synchronized talking face when it was presented together with three audiovisually desynchronized talking faces. This finding replicates the main finding in Experiment 1 and provides additional evidence of the power of temporal audiovisual synchrony cues to selectively recruit attention to holistic multisensory events. Interestingly, unlike in Experiment 1, here we found that participants deployed more attention to the virtual target talking face during the asynchrony test trials despite the fact that the visible articulations of the talking face were desynchronized with respect to the audible speech utterance. This finding suggests that associative learning contributed to responsiveness in this experiment. That is, it appears that participants quickly associated each person's face with that person's voice over the course of the experiment. As a result, the next time they saw that same face talking, presumably they directed more attention to it simply because they remembered that person's voice. Although post hoc, this interpretation is in line with the fact that perceivers can rapidly learn the visual identity cues of talking faces (Jesse & Bartoli, 2018) and that they can link the audible and visible identity cues of specific talkers (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b). Overall, the results from the asynchrony test trials suggest that participants take advantage of individual identity cues in their perceptual segregation of the multisensory clutter created by multiple talking faces.

The pattern of responsiveness to the eyes and mouth was similar to the pattern obtained in Experiment 1. In terms of eye gaze, we found that during the synchrony test trials participants gazed more at the eyes of the audiovisually synchronized talking face than at the eyes of the desynchronized talking faces but that during the asynchrony test trials they exhibited no preference for the eyes of the target talking face. This finding suggests that when the audiovisual speech processing task is relatively easy, participants are free to explore other aspects of the talking faces and focus on the other salient aspect of faces, namely the eyes. Under normal circumstances, the eyes provide deictic and other socially relevant cues and it is known that adults focus on these cues when not processing speech (Võ et al., 2012). Consistent with this interpretation, when speech processing becomes more challenging – as is the case in the asynchrony test trials in which no audiovisually synchronized talking face is present - participants gaze longer at the talker's mouth. This finding replicates the same finding from Experiment 1 and, again, suggests that participants rely on the audiovisual temporal synchrony cues located in a talker's mouth to determine who is talking. When such cues are absent, presumably participants focus more of their attention on the mouth to gain direct access to the audiovisual speech cues that are essential to determining who is talking in the hope of confirming or not that a particular person is, indeed, talking.

3. General discussion

We investigated whether the temporal synchrony that normally binds fluent audible and visible speech utterances affects selective attention and, thereby, perceptual segregation of competing talking faces. Adult participants watched four simultaneously talking faces articulating the same utterance while they listened to the auditory version of the same utterance. The participants' only assigned task was to indicate which face was talking at the end of each test trial. This task was employed explicitly to simulate the usual task of having to pick out a talking face that corresponds to a particular person's audible utterance from among multiple, concurrently talking faces. During half the test trials, the audible speech utterance was temporally synchronized with the visible speech articulations of one of the four faces and, during the other half of the test trials, the audible utterance was desynchronized from all four talking faces. In Experiment 1, the four talking faces were identical and the voice of the talker belonged to the person seen talking. In contrast, in Experiment 2, the four talking faces were different and, as a result, the voice corresponding to each of the target talking faces differed across the test trials. This difference between Experiments 1 and 2 enabled us to investigate the relative contribution of audiovisual temporal cues and identity cues to selective attention and to perceptual segregation of multiple talking faces.

Despite the different types of perceptual cues available in the two experiments, the results were strikingly similar across them. First, participants exhibited a marked preference for the audiovisually synchronized talking face when such a face was present in an array of four simultaneously talking faces. By contrast, participants exhibited either no preference (Experiment 1) or a significantly weaker preference (Experiment 2) when an audiovisually synchronized talking face was not present in the array. Crucially, the preference was remarkably similar across all four quadrants of target-face presentation in those trials in which an audiovisually synchronized talking face was present. The fact that the preference did not depend on the spatial location of the audiovisually synchronized talking face suggests that this face automatically captured attention. Second, participants gazed more at the eyes of an audiovisually synchronized talking face when such a face was present in the stimulus array than at the eyes of the competing but audiovisually desynchronized talking faces in that same array. By contrast, participants did not gaze more at the eyes of any of the four talking faces when an audiovisually synchronized talking face was not present in the stimulus array. Finally, participants gazed far more at the talker's mouth than eyes regardless of whether the audible and visible speech streams of one of the talking faces was temporally synchronized or not and they gazed more at the talker's mouth when none of the talking faces in the stimulus array were audiovisually synchronized than when one of the talking faces was synchronized.

The overall pattern of findings suggests that the marked preference for the audiovisually synchronized talking face was due to the detection of the temporal synchrony statistics of the audible and visible speech streams. This conclusion is justified by the fact that when an audiovisually synchronized talking face competed with three other audiovisually desynchronized talking faces for participants' attention, it was this face that attracted the bulk of their attention, but when this same talking face was audiovisually desynchronized it no longer did so. In addition, the preference for the audiovisually synchronized talking face was evident regardless of whether the array of four talking faces consisted of the same person's face and voice or of different people's faces and voices. It is interesting to note that the current findings are similar to the results from the *pip and pop effect* studies in which search for a target object embedded in a cluttered visual scene consisting of multiple objects is facilitated by a sound that is temporally synchronized with the actions of the target object (Van der Burg et al., 2008b). Of course, it should also be noted that the *pip and pop effect* reflects integration of abrupt, punctate events but that the effect found here reflects integration of a continuous event offering many points of intersensory congruence. Overall, the face-preference data from both experiments provide strong and convincing evidence of the power of temporal synchrony in audiovisual speech processing and, especially, in selective attention to and perceptual segregation of competing audiovisual speech inputs. The similarity of the current findings to those from studies of responsiveness to simple objects and sounds is a testament to the power and domain-generality of synchrony-based perceptual cues to direct attention and perceptual responsiveness.

The marked preference for the audiovisually synchronized talking face is interesting and might be interpreted as evidence that the individual identity cues available in Experiment 2 played no significant role in responsiveness. This interpretation is not, however, consistent with the fact that participants also preferred the audiovisually desynchronized version of the virtual target in Experiment 2. This finding suggests that identity cues did, indeed, play some role in responsiveness. This was probably due to the fact that participants were able to quickly associate each person's face and its dynamic "signature" with that person's voice over the course of the experiment. Although this is obviously a post hoc interpretation of the preference for the desynchronized target talking face in Experiment 2, it is consistent with findings that perceivers can rapidly learn the visual identity cues of talking faces (Jesse & Bartoli, 2018), that they can link the audible and visible identity cues of specific talkers (Kamachi et al., 2003; Lachs & Pisoni, 2004a, 2004b), and that the temporal relationship of auditory and visual speech cues not only signifies whether they constitute a unitary speech event but also their identity (Ten Oever, Sack, Wheat, Bien, & Van Atteveldt, 2013). Thus, the results from the asynchrony test trials in Experiment 2 suggest that participants take advantage of individual identity cues to segregate the multisensory clutter created by multiple talking faces. If so, it is also likely that differential identity cues, together with audiovisual temporal cues, play a role in the perceptual segregation of scenes composed of different talkers producing different utterances.

The interpretation of the overall pattern of the results offered above provides an intriguing picture of the way we deal with the usual onslaught of different types of multisensory cues. Some such cues are inherently related to each other because they share a particular common perceptual dimension (e.g., the intersensory temporal statistics of fluent audiovisual speech) while others are specific to their modality of origin (e.g., color, pitch, smell, taste) and thus are not inherently related. The former are typically referred to as amodal cues and evidence indicates that adults are very good at detecting these types of cues and, because of this, adults are very good at perceiving their multisensory world as a coherent and unitary place (Marks, 1978; Stein & Meredith, 1993). The latter types of cues bear an arbitrary relationship to one another, but they can be associated whenever they co-occur. Overall then, even though the present results suggest that attention is primarily driven by amodal cues, they also suggest that learned associations of modality-specific cues contribute to responsiveness. Of course, the question that the current results do not address directly is whether amodal or modality-specific cues play a different role depending on the complexity of the information available and whether their relative importance varies as a function of the processing task required by a particular event. Only future studies will be able to answer this question. In the meantime, it is clear that modality-specific identity cues play a secondary and/or supportive role in perceptual segregation when they compete for attention with audiovisual temporal synchrony cues. Nonetheless, it is theoretically possible that modality-specific identity cues play a larger role in the processing of more complex audiovisual events. For example, modality-specific cues might play an especially important role when multiple talking faces consist of different people articulating different utterances. In this case, the differential semantic cues associated with different utterances are likely to contribute to perceptual segregation as well.

The eye and mouth gaze data provided interesting insights into the processes underlying participants' search behavior. The fact that they gazed more at the eyes of the audiovisually synchronized talking face than at the eyes of desynchronized talking faces is consistent with findings from other studies in which selective attention to talking faces has been tracked. When participants do not have to perform a speech processing task per se, they tend to attend more to the eyes (Buchan, Paré, & Munhall, 2007; Lewkowicz & Hansen-Tift, 2012; Võ et al., 2012). When, however, participants are engaged in speech processing and/or have to process speech presented in noise, they tend to attend more to the talker's mouth (Barenholtz et al., 2016; Birulés et al., 2020; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998; Võ et al., 2012). The mouth fixation data from the current study - showing that participants gazed longer at the mouth in the asynchrony test trials than in the synchrony test trials – also are consistent with previous findings. These data demonstrate that adults also attend more to a talker's mouth when they need to disambiguate an ambiguous temporal relationship between audible and visible speech streams. This sort of perceptual mechanism is useful and important whenever a perceiver is confronted with multiple audible speech utterances and must bind one of them with a particular person's face. To do so, attention to the mouth is essential to determine which of the competing audible speech streams belongs with a particular talker's face. Of course, in the present study, participants only had to bind one audible speech stream with one of several competing faces. This is a much easier task and probably explains why participants attended less to the talker's mouth in the synchrony condition. If this conclusion is correct, then participants would probably attend more to the talker's mouth producing audiovisually synchronized speech if they had to bind one of several distinct audible speech streams with one of several different talking faces.

The findings obtained here demonstrate that temporally synchronized talking faces are highly attractive and that they are preferred over desynchronized ones. This preference provides new insights into the ways that perceivers solve the multisensory Cocktail Party Problem. It shows that temporally coherent talking faces automatically attract maximum selective attention. This is highly adaptive for two reasons. First, it provides perceivers with access to redundant audiovisual speech cues that are perceptually more salient and thus easier to process than auditory speech cues (Grant & Seitz, 2000; MacLeod & Summerfield, 1987; Sumby & Pollack, 1954; Summerfield, 1979; Summerfield, 1992; van Wassenhove, Grant, & Poeppel, 2005). Second, it provides perceivers with a powerful way to de-clutter their multisensory world.

Finally, the results obtained here raise an interesting theoretical question: Why might a temporally synchronized talking face be the preferred object? The answer is that multisensory integration is a fundamental feature of brain function (Calvert, Campbell, & Brammer, 2000; Ghazanfar & Schroeder, 2006; King & Calvert, 2001; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008), that integration appears early in development (Lewkowicz, 2000a; Lewkowicz & Ghazanfar, 2009; Murray et al., 2016), and that the development of multisensory integration is shaped by early experience (Lewkowicz & Röder, 2012). Indeed, the effects of early experience are especially key to the preference obtained here largely because our everyday social experiences consist of interactions with social partners whose visible and audible articulations have a common origin and thus are, by default, temporally and spatially coherent. Therefore, there is little doubt that early experience with coherent multisensory inputs contributes to the emergence of the unity assumption, a perceptual bias which compels us to treat our multisensory world as a coherent place even though it is specified by disparate multisensory inputs (Welch & Warren, 1980). Empirical evidence from studies of cats who are deprived of congruent and appropriate auditory and visual sensory inputs in early life supports this conclusion. It shows that they exhibit atypical responsiveness to integrated audiovisual inputs after such early experience (Xu, Yu, Rowland, & Stein, 2017; Xu, Yu, Stanford, Rowland, & Stein, 2015). This suggests that the years of exclusive exposure that humans have to congruent multisensory inputs during their everyday interactions with social partners and interlocutors imparts the unity assumption. If so, the sort of perceptual bias for audiovisually synchronized talking faces found here is not surprising. Of course, the functional advantage of such a bias is that it helps us overcome the multisensory Cocktail Party Problem and, in the process, enables us to quickly and efficiently identify and access the audiovisual communicative signals of specific talkers.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104743.

Acknowledgements

This work was supported by the National Science Foundation (Grant BCS 1749507) awarded to DJL.

References

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.

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.

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.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT 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.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*(11), 649–657.

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.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10*(6), 278–285.

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). https://doi.org/10.1111/desc.12381. n/a.

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.

Jesse, A., & Bartoli, M. (2018). Learning to recognize unfamiliar talkers: Listeners rapidly form representations of facial dynamic signatures. *Cognition, 176*, 195–208.

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.

King, A. J., & Calvert, G. A. (2001). Multisensory integration: Perceptual grouping by eye and ear. *Current Biology, 11*(8), R322–R325.

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. (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. (2000a). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*(2), 281–308.

Lewkowicz, D. J. (2000b). Infants' perception of the audible, visible and bimodal attributes of multimodal syllables. *Child Development, 71*(5), 1241–1257.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*(1), 66–77.

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, 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., & Röder, B. (2012). The effects of experience on the development of multisensory processing. In B. Stein (Ed.), *The new handbook of multisensory processing*. Cambridge, MA: MIT Press.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131–141.

Marks, L. (1978). *The unity of the senses*. New York: Academic Press.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology, 19*(22), R1024–R1027.

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.

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science, 283*(5406), 1272–1273.

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.

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.

Schroeder, C., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12*(3), 106–113.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*, 308.

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.

Shimojo, S., Watanabe, K., & Scheier, C. (2001). The resolution of ambiguous motion: Attentional modulation and development. In J. Braun, C. Koch, & J. Davis (Eds.), *Visual attention and cortical circuits* (pp. 242–264). MIT Press.

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology, 13*(13), R519–R521.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: The MIT Press.

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.

Stevenson, R. A., & Wallace, M. T. (2013). Multisensory temporal integration: Task and stimulus dependencies. *Experimental Brain Research, 227*(2), 249–261. https://doi.org/10.1007/s00221-013-3507-3.

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.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400–410.

Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & Van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology, 4*, 331.

Thelen, A., Matusz, P. J., & Murray, M. M. (2014). Multisensory context portends object memory. *Current Biology, 24*(16), R734–R735.

Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition, 138*, 148–160.

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. https://doi.org/10.1167/8.5.2.

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.

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72*(4), 871–884.

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

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186.

Watanabe, K., & Shimojo, S. (1998). Attentional modulation in perception of visual motion events. *Perception, 27*(9), 1041–1054.

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: Effects of auditory grouping on visual motion perception. *Psychological Science, 12*(2), 109–116.

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*.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15*(2), 77–84.

Xu, J., Yu, L., Rowland, B. A., & Stein, B. E. (2017). The normal environment delays the development of multisensory integration. *Scientific Reports, 7*(1), 1–13.

Xu, J., Yu, L., Stanford, T. R., Rowland, B. A., & Stein, B. E. (2015). What does a neuron learn from multisensory experience? *Journal of Neurophysiology, 113*(3), 883–889.

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.