Research Report

Expectations about dynamic visual objects facilitates early sensory processing of congruent sounds

Andrew Marin a,*, Viola S. Störmer b and Leslie J. Carver a

a University of California, San Diego (UCSD), Psychology Department, La Jolla, CA, USA
b Dartmouth College, Department of Psychological and Brain Sciences, Hanover, NH, USA

Abstract

The perception of a moving object can lead to the expectation of its sound, yet little is known about how visual expectations influence auditory processing. We examined how visual perception of an object moving continuously across the visual field influences early auditory processing of a sound that occurred congruently or incongruently with the object’s motion. In Experiment 1, electroencephalogram (EEG) activity was recorded from adults who passively viewed a ball that appeared either on the left or right boundary of a display and continuously traversed along the horizontal midline to make contact and elicit a bounce sound off the opposite boundary. Our main analysis focused on the audio-visual (AV) trials, a sound accompanied the visual input when the ball contacted the opposite boundary (AV-synchronous), or the sound occurred before contact (AV-asynchronous). We also included audio-only and visual-only trials. AV-synchronous sounds elicited an earlier and attenuated auditory response relative to AV-asynchronous or audio-only events. In Experiment 2, we examined the roles of expectancy and multisensory integration in influencing this response. In addition to the audio-only, AV-synchronous, and AV-asynchronous conditions, participants were shown a ball that became occluded prior to reaching the boundary of the display, but elicited an expected sound at the point of occluded collision. The auditory response during the AV-occluded condition resembled that of the AV-synchronous condition, suggesting that expectations induced by a moving object can influence early auditory processing. Broadly, the results suggest that dynamic visual stimuli can help generate expectations about the timing of sounds, which then facilitates the processing of auditory information that matches these expectations.

© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Introduction

In everyday life, dynamic visual objects often predict accompanying sounds. For example, observing two hands moving closer together precedes the onset of a clap, or a marble contacting another results in a sound precisely at the point of collision. These scenarios showcase how motion (i.e., directionality, speed, etc.) and physical cues (i.e., artificially defined object boundaries, collision, etc.) of dynamic visual objects in natural sensory environments elicit expected sounds at precise moments in time and space. Perceiving such events uniquely highlights how visual anticipation can directly interact with auditory processing—yet little is known about how auditory processing is influenced by preceding visual information about moving objects. We know that dynamic visual objects can elicit sounds in natural sensory environments, but does this visually driven anticipation facilitate early auditory processing?

Dynamically moving visual objects often generate sounds that can be predicted from the temporal expectancy laid forth by the object itself. Accurately inferring the source of sounds generated by a moving object involves matching temporally synchronous visual information with the sound. Such inferences may reflect a mechanism designed to exploit the temporal and spatial information of a moving object to make predictions about expected sounds in the environment. Questions regarding stimulus prediction and its brain bases have received considerable amounts of attention, particularly within the auditory domain (for review, Lange, 2013). Event-related potentials (ERPs), which reflect the averaged electroencephalogram (EEG) response time-locked to a particular event, have been utilized to examine how early auditory responses are shaped by predictable sensory information. One particular ERP response that is modulated by predictable sounds is the auditory evoked potential (i.e., the N1–P2 complex), which is an early sensory response elicited after a sound. Many studies have reported that the auditory response is attenuated when hearing temporally predictable sounds (Clementz et al., 2002; D’Andrea-Penna et al., 2020; Ford et al., 2007; Ford & Hillyard, 1981; Kononowicz & van Rijn, 2014; Lange, 2009; Menceloglu et al., 2020; Schafer et al., 1981). Auditory response suppression toward expected sounds has usually been interpreted within a general predictive coding framework (Friston, 2005; Lange, 2013), where the reduction of the auditory response is thought to arise due to top-down expectancies matching bottom-up sensory input. The synchronous match between bottom-up and top-down signals is thought to reduce the error signal of the predicted sound, which results in an overall reduction of the evoked ERP (Baldeweg, 2007; Lange, 2013). Yet, many of the studies mentioned here cued the expectation of the sound within the temporal domain by providing the perceiver foreknowledge about when a sound would occur. However, in natural sensory environments, sounds are more likely to be preceded by visual stimuli that are often moving across time and space.

The perception of simultaneity of discrete audio-visual (AV) events in time and space plays a large role in determining if the two sensory events will be perceptually bounded as one, or perceived as two separate events (Körding et al., 2007; for review, see Wallace & Stevenson, 2014). Discrete sensory events that remain in close temporal proximity to one another are more likely to be integrated as one, whereas sensory events that are further away in time and space are more likely to be perceived as two distinct events (Spence, 2007; Stevenson, Zemtsov, & Wallace, 2012; Stevenson, Fister, et al., 2012; van Wassenhove et al., 2007). Simple AV stimuli like pure auditory tones and geometrical visual shapes have been associated with an enlarged auditory neural response compared to the sum of unimodal presentations (Fort et al., 2002; Giard & Peronnet, 1999; Molholm et al., 2002). Enhancement of early neural responses while perceiving multisensory simultaneity has been theorized as a general principle of multisensory processing (Meredith et al., 1987). Other demonstrations of AV integration in natural environments involve the perception of speech sounds, human actions, and dynamic visual objects. The auditory-evoked potential has been found to occur earlier in time and elicit a smaller amplitude response when speech sounds are paired with synchronous mouth movements compared to auditory-only presentations, which has been interpreted as auditory processing being suppressed when paired with visual information (van Wassenhove et al., 2005). In the case of speech perception, visual information originating from mouth movements precedes paired auditory outputs by tens to a few hundreds of milliseconds (van Wassenhove et al., 2005), likely leading to strong expectations about when and what sound will appear. Other research has shown that such expectations can also be triggered by non-speech stimuli, for example human actions (i.e., a hand clap) and dynamic objects (i.e., a hammer tapping a cup), which have been also shown to attenuate the auditory response (Stekelenburg & Vroomen, 2007). Critically, decreases in the auditory response were not seen with objects that did not provide anticipatory visual motion information (Stekelenburg & Vroomen, 2007), suggesting that the amplitude reduction underlying the early auditory response occurs when visual information provides clear expectations about the onset of a sound. Together, these studies suggest that a smaller amplitude of the auditory response may arise in situations that allow a relatively long build-up of visual expectations (e.g., through the movement of hands or lips), in that the visual information allows one to predict the upcoming acoustic signal, and subsequently reduces uncertainty and lowers computational demands of auditory brain regions (Besle et al., 2004; Vroomen & Stekelenburg, 2010).

Another study further supports the notion that changes in early auditory processing only arise when visual information reliably predicts sound onset. Vroomen & Stekelenburg found early attenuation of the auditory evoked potential when viewing simple visual stimuli that provided expectations of an anticipated sound compared to auditory-alone presentations (2010). In their task, for AV expectation trials, two disks appeared to the extreme left and right of a vertically aligned rectangle that was presented at the center of a display. Here, each visual disk moved toward the rectangle and eventually collided with it, compressing it and eliciting a synchronous pure tone at the point of collision. Participants were also exposed to two other conditions where 1) the dynamic visual stimuli collided with the rectangle but did not contain an
expected sound (visual-only) and 2) audio-only trials that contained sound with no visual stimulus. These three trial types appeared in a single block, in random order. Importantly, in a new block of trials, participants were exposed to a new AV condition that did not provide visual expectations about when the sound would appear. In this condition, there were no visual disks, but the rectangle eventually compressed and made a sound upon doing so. Within this block, subjects were also presented with the same audio- and visual-only trials described previously, and each were presented in a random order. In a follow-up experiment, Vroomen & Stekelenburg presented the same two AV trials, as well as the audio- and visual-only conditions explained above. In addition, they 1) provided two new sensory conditions where the sound either happened before or after the collision event (AV-asynchronous; early and late) and 2) manipulated whether these sensory conditions appeared in a fixed or random order (2010). Here, the early amplitude response during fixed block ordering was reduced while perceiving synchronous AV expectations compared to audio-alone input, an effect not seen during mixed block ordering. This led the authors to suggest that auditory reduction arises when visual information reliably predicts AV onset across trials. Moreover, they found that the auditory response was not different in amplitude or latency between the synchronous and asynchronous AV inputs during mixed order presentations. Interestingly, a later component (i.e., the P2) showed a different pattern and was attenuated when early asynchronous and synchronous AV stimuli were fixed and varied from trial to trial. This suggests a possible dissociation between these two components of the auditory ERP. Taken together, these studies suggest that the neural effects of AV expectations depend on various factors, such as temporal synchrony, the amount of visual and auditory input, whether or not trial information was known beforehand, and the stimuli used.

One factor not considered in Vroomen and Stekelenburg (2010) was what would happen with a more naturalistic visual event such as a single object, moving in a uniform direction (i.e., a ball bouncing off a wall). Here, we fill this gap in the literature by better characterizing the neural correlates governing the anticipation of dynamic, temporally synchronous AV processing. Unidirectional dynamic visual stimuli might provide 1) the visual system more precise expectations about the collision event that elicits the anticipatory sound and/or 2) the auditory event itself may be more predictable when the accompanying visual stimulus is moving unidirectionally. Dynamic visual stimuli moving unidirectionally, in turn, may afford the visual system greater sensitivities toward small temporal AV asynchronies sooner in the auditory processing stream. Furthermore, it is currently unknown whether AV effects occur based on expectations alone, or whether visual objects and sounds need to be both present in order to affect sensory processing. Thus, the primary objective of our study was to examine how dynamic visual input—a single object moving continuously in one direction across the visual field— influences early auditory processing of a sound that is either congruent with the object’s motion, and thus likely perceived as being part of the visual object, or incongruent with the object’s motion. We were guided by the hypothesis that AV temporal synchrony would result in an attenuated and faster auditory response, compared to a unimodal auditory presentation—a response profile that would mimic the auditory effects seen in Stekelenburg and Vroomen (2007) and Vroomen and Stekelenburg (2010). Considering the null findings regarding the neural response toward the synchrony of dynamic AV input outlined above (i.e., 2007, 2010), we expected differences might appear because our stimuli were designed to constrain visual expectations in a single direction, perhaps affording the brain greater sensitivity toward small temporal asynchronies earlier in time. We also examined whether such auditory ERP effects only occur when a visual stimulus is presented at the same time as the sound, as predicted by multisensory integration accounts, or whether the expectation triggered by a moving visual stimulus is sufficient in influencing auditory processing. To test this hypothesis, we conducted a second experiment and examined auditory responses elicited by visual anticipatory information that becomes occluded prior to temporally congruent collision.

2. Methods

We report all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. A statement on how sample size was determined can be found in the methods section of Experiment 2.

2.1. Participants

Twenty-nine college-aged adults (M_age = 20.48 years, SD = 1.76; 14 female) were recruited via an online university subject pool and received course credit for participating. Prior to the experiment, each participant reported having normal or corrected-to-normal vision, normal hearing, and no history of neuropsychological, cognitive, or developmental disorders. All participants provided written informed consent in accordance with the tenets of the 1964 Declaration of Helsinki. An additional eight adults were tested but were excluded due to equipment malfunction (n = 3) and excessive (>10% of trials) EEG artifact (i.e., head motion, muscle artifact etc.; n = 5).

2.2. Audio-visual stimuli

The AV stimuli used in this experiment were the same as in Werchan et al. (2018). The stimuli were created using Adobe After Effects software, while stimulus delivery was controlled by E-prime software (Psychology Software Tools, 2016) and presented on a CRT monitor (13 width x 9.5 height; in inches), with a 60 Hz refresh rate. Participants viewed the stimuli at an average distance of 71 cm. The primary object of interest was a red ball that was one inch in diameter, subtending a visual angle of about 2.05°. The ball appeared within a black rectangle (7.75 width x 5.5 height in inches; visual angle width = 15.8°; subtended visual angle_height = 11.2°) that was overlaid on top of a neutral gray background (13 width x 9.5 height; in inches; visual angle width = 26.2°; visual angle_height = 19.3°). The inside of the black rectangle contained a grid of small white dots that emphasized the straight, horizontal motion of
the red ball. The ball’s horizontal movement was constrained to occur within the black rectangle at a rate of 2.5s per motion cycle (i.e., visual object starts and returns to its origin). The sounds were presented via two speakers presented to the left and right of the monitor. The sound itself was a 50 decibel (dB) complex tone that resembled a solid object colliding with a hard surface (a knocking sound) and had a duration of 200 msec.

2.3. Paradigm and procedure

Each participant was seated in a dark room and was shown a randomly presented stream of four AV sensory conditions: 1) visual-only, 2) audio-only, 3) AV-synchronous, and 4) AV-asynchronous, while high-density EEG (Electrical Geodesics, Inc.) was recorded. The experimental session took part during a single lab visit, and the EEG recording lasted approximately 45 min. At the start of each trial (see Fig. 1 for a single trial diagram), a geometric, achromatic fractal video was presented for 1000 msec, and served as visual input to promote participant attention and engagement during the passive viewing task.

A small fixation cross (.75 width × .75 in height; 1.5 of visual angle) then appeared for 1000 msec, followed by the random presentation of one of the four sensory conditions previously mentioned. There were a total of 416 experimental trials (104 trials per condition), split into four blocks. The experiment was split into blocks to provide breaks to the participant as needed. The primary part of each trial (where the ball moved across the screen, described in more detail below) lasted 2000 msec. Upon completion of each trial, a single letter (.75 width × .75 in height; 1.5 subtended visual angle), out of a possible of eight, appeared randomly in the center of the screen for 500 msec. At the start of each experimental block, the participant was instructed to identify and count, using a handheld clicker, a single target letter. This secondary task served as an attention check to keep each participant engaged during the passive viewing task. All participants were above a 95% accuracy rate in the secondary task so no subjects were removed due to poor attention.

For the visual-only, AV-synchronous, and AV-asynchronous conditions, a single red ball randomly appeared on either the far left or right boundary of the black rectangle display. The ball then traversed horizontally to make contact with and bounce off the opposite boundary of the black rectangle. Participants were not explicitly told to maintain fixation but were encouraged to not track the exact motion of the ball and to take in the stimuli holistically. The EEG data was cleaned for any eye movements that occurred (see EEG data processing). The time between the start of the red ball’s motion to the time it reached the opposite boundary of the display was 1200 msec. For the AV-synchronous condition, the bounce-sound occurred at 1200 msec, exactly when the ball touched the opposite boundary of the grid. During the AV-asynchronous condition, the bounce-sound occurred at 750 msec post stimulus onset, which corresponded to the ball just having passed the vertical midline as it moved toward the opposite wall, and before it contacted the wall. The visual-only condition contained the ball moving and bouncing off the

Fig. 1 – Single trial schematic depicting the AV-synchronous condition. For each trial, a small achromatic fractal video was first presented for 1000 msec (ms) to promote participant engagement during the passive viewing task. A small fixation cross then appeared, followed by the random presentation of one of four sensory conditions for a total of 416 experimental trials, 104 trials per condition. The interval labeled “Event Stimuli” contains the primary part of the trial, where participants viewed a ball that moved across the screen and may or may not be presented with auditory input. A visual depiction and description of each sensory condition is outlined within Fig. 2. Lastly, upon completion of the event stimuli, each participant was instructed to identify a single target letter among 7 distractor letters, which served as an attention check.
opposite boundary of the black rectangle with no sound. The ball for the AV-synchronous, AV-asynchronous, and visual-only conditions remained stationary at the opposite boundary for 50 msec, so the time from the start of motion in one direction and the start of motion back to its origin was 1250 msec. At 1250 msec, the ball started to move toward its origin at the same speed and the stimulus subsequently terminated at 2000 msec, well before contact with the boundary of origin. The audio-only condition contained the black rectangle and no visual input provided by the red ball, but the bounce-sound occurred at 1200 msec after the start of the trial. The duration of the sound for the audio-only, AV-synchronous, and AV-asynchronous conditions was 200 msec. Each trial occurred with equal probability and was randomly generated within each experimental block (see Fig. 2 for a visual diagram of each sensory condition).

2.4. EEG data processing

Continuous EEG was recorded via a 128-channel HydroCel Geodesic Sensor Net (Electrical Geodesics, Inc.; EGI). Impedances were kept below 50 kOhms in all electrodes and the raw EEG data were referenced online to the vertex (Cz) and digitized at 500 Hz. EEG data were amplified according to the default settings of an EGI internal amplifier (model type: Net Amps 300). All data were processed off-line using MATLAB (Mathworks, Inc.) and EEGLAB/ERPLAB software (Delorme & Makeig, 2004; Lopez-Calderon & Luck, 2014). A video of each participant was obtained during the EEG recording to ensure they kept their eyes on the display during the EEG session.

The raw EEG data were first digitally filtered using a 0.05–50 Hz bandpass (Butterworth) and 60 Hz notch filters. Data were then manually inspected for individual bad channels present throughout at least 50% of the recording, as well as electromyographic (EMG) and other movement artifacts. EEG data with evidence of egregious EMG, movement, or muscle artifacts were rejected from the analysis. Data from bad channels were replaced using a spherical spline interpolation algorithm. The cleaned EEG data were then taken through an independent component analysis (ICA), where evidence of eye artifact (i.e., eye blinks and saccades) was removed from the data.

Fig. 2 – Depiction of the left-start sensory conditions (right-start not pictured). The moving spherical visual object is shown in red, the presentation of the sound is depicted as a bright yellow star, and the white arrows indicate the direction of motion of the visual object. The red rectangle drawn over a single frame of each condition reflects the point in time where we are event-locking the EEG data for each sensory condition. All ERPs were time-locked to the sound presentation, or in case of the visual-only condition, the moment when the ball bounced off the boundary. The audio-only condition contained no visual input at all. The visual-only condition contained the dynamic motion of the ball, with no audio input. For the synchronous condition, the bounce sound occurred when the ball first made contact with the boundary of the grid. For the asynchronous condition, the bounce sound corresponded to the ball just having moved past vertical midline. All conditions were equally likely to occur and were randomly intermixed across the experiment. The onset of each trial followed an interstimulus interval with a randomly presented jitter of 500–750 msec.
data set. To ensure that all ocular-related artifacts were eliminated by the ICA successfully, we scrolled through the entire raw EEG traces to look for any residual eye blinks or eye movements and if present, removed them by hand. This ICA procedure ensured that no eye movement artifacts were contained in the final data. Thus, while it is possible that some participants moved their eyes less in one condition than the other (e.g., audio-only vs. visual-present), this should not affect our ERP results. We also noticed ICA components in the data that resembled high-frequency harmonics and opted to remove them. The EEG data were then segmented into 1000 msec epochs (–200 to 800 msec relative to stimulus onset), and baseline corrected using mean voltage during the 200 msec prestimulus baseline period. ERPs were time-locked to the onset of the sound in all conditions except the visual-only condition in which the ERPs were time-locked to the exact moment the ball touched the boundary. Each segmented data set was again manually inspected for excessive artifacts. Once artifact rejection was completed, the EEG data were again filtered, this time using a 30 Hz lowpass (Butterworth) filter and then re-referenced to an average reference. Grand-averaged ERPs were then obtained for each participant by averaging all available epochs for each condition.

The total number of acceptable ERP segments per participant was on average 404.28 trials (SD = 8.08): (audio-only condition: M = 101.38, SD = 2.04; visual-only condition: M = 101.45, SD = 1.97; AV-synchronous condition: M = 100.66, SD = 3.05; AV-asynchronous condition: M = 100.79, SD = 3.02). There were no significant differences between the conditions in the amount of total useable segments included in the construction of each individual ERP response, F(3, 28) = 1.4, p = .25, η² = .05.

### 2.5. ERP regions & components of interest

To test whether early sensory processing was affected by the temporal synchrony of dynamic AV events, the auditory N1 and P2 components of the auditory evoked potential were evaluated. The N1 component is the first negative going peak of the auditory evoked potential and is thought to index the early sensory processing of auditory stimuli (Godey et al., 2001; Mayhew et al., 2010; Nätänen & Winkler, 1999; Picton et al., 1974; Ponton et al., 2002). The N1 was operationalized here as the minimum peak amplitude and latency occurring within 100–200 msec after sound onset. The auditory P2 component is the second positive going peak of the auditory evoked potential and its functional significance is much less clear compared to the preceding N1. One possible hypothesis posits that the auditory P2 may be involved in matching current sensory input with past perceptual representations (Freunberger et al., 2007; Luck & Hillyard, 1994). The P2 was operationalized here as the maximum peak amplitude and latency occurring within 200–300 msec after sound onset. Both the time window and the regions of interest were selected based on our hypotheses about the timing of each ERP component (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010) and from visual inspection using the grand averaged ERP across all participants and conditions. To quantify early processing of a sound across the entire auditory ERP, we calculated the N1–P2 peak-to-peak amplitude response, which reflects the amplitude change between the negative N1 trough and positive P2 peak. To obtain this value, we subtracted the amplitude of the P2 response from the amplitude of the N1 response for each subject. For our latency analyses, we planned to conduct individual N1 and P2 peak latency measures for both experiments. A six-channel frontal-central auditory region was constructed to evaluate differences in auditory activity between each sensory condition. The ERP data, stimuli, and scripts that support the findings of this study are available to download (Marin et al., 2021a, 2021b). Note that no part of the study’s procedures or analysis plan were formally pre-registered before the research was conducted.

### 3. Results

Fig. 3a presents the grand averaged (n = 29) ERP waveforms, split between sensory conditions.

As can be clearly seen in Fig. 3, all conditions that included a sound elicited auditory-evoked potentials, but – as expected – the visual-only condition did not elicit an auditory response, and was thus dropped from all subsequent analyses. We conducted two separate one-way within-subjects repeated measures ANOVAs with three levels (audio-only, AV-synchronous, AV-asynchronous) for the amplitude and latency responses, within frontal-central scalp regions. All statistical analyses presented below were conducted in R studio, using the ‘tidyverse’ and ‘emmeans’ plugin packages.

#### 3.1. N1–P2 peak-to-peak amplitude

As can be seen in Fig. 3a and b, the N1–P2 peak-to-peak amplitude differed between conditions. Statistical analysis of the N1–P2 peak-to-peak amplitude confirmed this observation and revealed a significant main effect of condition, F(2, 28) = 14.2, p < .001, η² = .34 (see Fig. 3b). Post-hoc tests revealed that the AV-synchronous condition (M = −7.52, SD = 2.83) exhibited a smaller N1–P2 peak-to-peak amplitude that significantly differed compared to the AV-asynchronous (M = −8.44, SD = 3.75; p = .006, d = .28) and the audio-only (M = −9.05, SD = 3.78; p < .001, d = .46) responses. The AV-asynchronous response was not different from the audio-only response (p = .1).

#### 3.2. N1 and P2 peak latency

To assess whether the timing of early auditory ERP was affected by the temporal synchrony of dynamic AV events, the N1 and P2 components were evaluated using analyses similar to the N1–P2 peak-to-peak. We predicted that the N1 and P2 responses toward the dynamic AV synchrony should elicit faster peak amplitudes compared AV asynchronous and audio only responses.

---

1 The visual-only condition was included in the design of the experiment to keep all conditions symmetrical and not bias participants’ expectations in any particular way. Because the visual-only condition contained continuous visual information but no sound, we planned not to look at the EEG data for this condition in any meaningful way.
Analysis of N1 minimum peak latency revealed a significant main effect of condition, $F(2, 28) = 33, p < .001, \eta^2_p = .54$ (see Fig. 3c). The N1 peaked earlier for the AV-synchronous condition ($M = 144.17$ msec, $SD = 12.7$) compared to AV-asynchronous ($M = 162.39$ msec, $SD = 17.2; p < .001, d = 1.21$) and audio-only ($M = 150.3$ msec, $SD = 12.4; p = .03, d = .49$) responses. Additionally, the N1 was slower for the AV-asynchronous response compared to the audio-only response ($p < .001, d = .81$).

Analysis of P2 maximum peak latency also revealed a significant main effect of condition, $F(2, 28) = 31.98, p < .001, \eta^2_p = .53$ (not depicted). Like the N1, the P2 peaked sooner for the AV-synchronous condition ($M = 234.32$ msec, $SD = 14.8$) compared to AV-asynchronous ($M = 255.33$ msec, $SD = 13.2; p < .001, d = 1.5$) and audio-only ($M = 242.12$ msec, $SD = 17.4; p = .01, d = .48$) responses. Additionally, the P2 was slower for the AV-asynchronous response compared to the audio-only response ($p < .001, d = .86$).

### 3.3. Summary and discussion of Experiment 1

We found that the auditory response was sensitive to the temporal relationship between dynamic AV input for Experiment 1. The auditory response was smaller in amplitude and occurred earlier in time when visual input was synchronously paired in time and space with an expected sound, compared to the response elicited from auditory-alone and asynchronous AV inputs. Additionally, the neural response toward asynchronous AV input was significantly delayed compared to AV-synchronous and auditory-alone presentations. Importantly, smaller auditory responses were seen even when the synchrony of the AV collision event varied unpredictably from trial to trial — a key distinction from Vroomen and Stekelenburg (2010). This pattern of results suggests that early sensitivity toward the temporal synchrony of anticipated sounds allows the brain to code for temporally congruent AV events, resulting in an
attenuated (or suppressed) early auditory response. Critically, Experiment 1 underscores the role of temporal synchrony in facilitating early auditory processing, providing further evidence that the auditory effects are relevant for non-predictable inanimate objects, not just expected human actions and AV speech perception (additional theoretical implications are included in the general discussion). Taken together, the results of Experiment 1 suggest that the continuous presentation of a moving object can alter the processing of incoming auditory information within the first 200 msec of processing.

4. Experiment 2

Experiment 1 showed that a moving visual object can alter early auditory processing of a subsequent sound that is perceived as part of the same object. One interpretation of the results of Experiment 1 is that the auditory effects occurred because the sound and visual object were present at the same time during the collision event — which would be consistent with a multisensory account of sensory facilitation. An alternative is, however, that the expectation of a sound induced by a single moving visual stimulus is sufficient to alter auditory processing, even if no visual object is present at the same time the sound occurs. Thus, in Experiment 2, we tested whether continuous visual input is necessary to generate the auditory effects found during temporally synchronous AV presentations, or whether the expectation about a moving object is sufficient to modulate early auditory processing. To do this, we added a new AV condition in which we showed the visual and motion cues provided by the ball and its motion, and then removed these cues via occlusion well before the object collided with an artificial boundary and subsequently elicited an expected bounce sound. Thus, the sound appeared at the moment the ball would collide with the boundary, only the ball was not visible to participants anymore. We compared this condition to the AV synchronous, AV-asynchronous, and audio-only conditions identical to those in Experiment 1. With this new AV-occluded condition, we hoped to elicit similar visual expectancies as in the AV-synchronous condition, but to eliminate the simultaneous presentation of visual object and sound during the collision itself, to tease apart effects of expectation alone, and multisensory integration.

We expected to replicate the auditory effects seen in Experiment 1 for AV-synchronous relative to AV-asynchronous and audio-only conditions. Of particular interest was the AV-occluded condition: If the AV-occluded auditory response was most similar to audio-only activity, it would imply that temporally concordant AV input is important to elicit the auditory effects, consistent with multisensory integration. Alternatively, if the AV-occluded response looks similar to the AV-synchronous response, it would suggest that visually-induced expectations alone are sufficient to alter early auditory processing. Lastly, if the AV-occluded condition resembled the AV-asynchronous response profile, this would suggest that both conditions elicit responses possibly related to detecting AV incongruencies.

5. Methods

5.1. Participants

Due to relatively large effect sizes in Experiment 1, we reduced the sample in Experiment 2 to match that used by Vroomen and Stekelenburg (2010). Nineteen college-aged adults (M_age = 20.51 years, SD = 1.46; 9 female) participated in Experiment 2. An additional five adults were tested but were excluded due to excessive EEG artifact based on the removal criteria outlined in the methods section of Experiment 1.

5.2. Audio-visual stimuli

The AV stimuli used in this experiment were the same as in Experiment 1, with the exception of the added AV-occluded condition described below. The new AV-occluded condition contained the same AV properties as the AV-synchronous condition. We did not include the visual-only condition in this experiment. Like in Experiment 1, the AV-asynchronous tone occurred 450 msec before contacting the opposite edge.

5.3. Paradigm and procedure

Each participant was seated in a dark room and was shown a randomly presented stream of four AV sensory conditions: 1) audio-only, 2) AV-synchronous, 3) AV-asynchronous, and 4) AV-occluded while high-density EEG was recorded. For the new AV-occluded condition (see Fig. 4 for a single trial stimulus presentation of the timing of events), a single red ball randomly appeared on either the far left or right boundary of a black rectangle display, at the horizontal midline of the monitor.

The ball in this condition traversed along the horizontal midline, but at approximately 600 msec, it began to move through an invisible slit midway in the display and became fully occluded before contacting the opposite boundary. For this condition, the bounce-sound occurred at 1200 msec, exactly when the occluded ball would contact the opposite boundary of the rectangle display. Thus, the timing of the sound was predictable based on when the object entered the occluding area, but the visual object itself was not visible when the bouncing sound was played. After auditory onset, the invisible ball started to move back toward its origin (still occluded at this point) and became fully visible half-way through the display (after another 600 msec), then the stimuli subsequently terminated at 2000 msec. Additionally, all participants again performed the secondary task in between trials and were above a 95% accurate in task, resulting in no subjects removed due to poor attention.

5.4. EEG data processing

The EEG/ERP pre-processing steps were the same as Experiment 1. The total number of acceptable ERP segments per participant was on average 404 trials (SD = 11.6): (audio-only condition: M = 101.74, SD = 2.28; AV-synchronous condition: M = 100.79, SD = 3.33; AV-asynchronous condition: M = 100.79, SD = 3.19; AV-occluded condition: M = 100.68, SD = 3.93).
There were no condition differences regarding the amount of total useable segments included in the construction of each individual ERP response, $F_{(3, 18)} = 1.7, p = .18, \eta^2_p = .09$.

6. Results

Fig. 5a displays the grand averaged ($n = 19$) auditory ERP waveforms, split between the four sensory conditions.

As seen in the grand averaged ERP for Experiment 2 (see Fig. 5a), ERP deflections between each condition were seen as early as the onset of the sound (0 msec) and by using N1–P2 peak-to-peak measures, we hope to account for these early ERP differences. The N1 and P2 peak latency analyses for Experiment 2 were conducted in the same manner as the latency analyses for Experiment 1 because the early visual amplitude drift seen between the conditions would not influence the interpretation of timing on the auditory ERP. We then conducted a within-subjects repeated measures ANOVAs with 4 levels (audio-only, AV-synchronous, AV-asynchronous, AV-occluded) for the N1–P2 peak-to-peak amplitude response and N1 and P2 peak latency responses within frontal-central scalp regions.

6.1. N1–P2 peak-to-peak amplitude

As shown in Fig. 5a, the auditory component differed in terms of amplitude and latency across the four conditions. Statistical analysis of N1–P2 peak-to-peak amplitude revealed a significant main effect of condition, $F_{(3, 18)} = 8.73, p < .001, \eta^2_p = .33$ (see Fig. 5b). Post-hoc tests revealed that the AV-synchronous response ($M = -5.79$, $SD = 2.09$) exhibited an attenuated N1–P2 peak-to-peak amplitude compared to the AV-asynchronous ($M = -7.18$, $SD = 2.77; p = .004, d = .57$) and audio-only ($M = -7.32$, $SD = 2.36; p = .001, d = .69$) responses, while the AV-asynchronous response was not statistically different from the audio-only response ($p = .98$), overall replicating Experiment 1. Importantly, the AV-occluded response ($M = -5.91$, $SD = 1.67$) was significantly smaller compared to the AV-asynchronous ($p = .01, d = .56$) and audio-only ($p = .003, d = .69$) responses. The AV-occluded response was not different compared to the AV-synchronous ($p = .99$) response.

6.2. N1 and P2 peak latency

Analysis of N1 minimum peak latency revealed a non-significant main effect of condition ($F_{(3, 18)} = 2.08, p = .11, \eta^2_p = .14$), diverging from Experiment 1. However, analysis of P2 maximum peak latency revealed a significant main effect of condition ($F_{(3, 18)} = 68.9, p < .001, \eta^2_p = .79$; see Fig. 5c). The P2 was significantly delayed for the AV-asynchronous response ($M = 244.8$ msec, $SD = 11.31$) compared to the AV-synchronous ($M = 221.49$ msec, $SD = 9.99$; $p < .001, d = 2.18$), audio-only ($M = 228.16$ msec, $SD = 10.78; p < .001, d = 1.51$), and AV-occluded ($M = 225.18$ msec, $SD = 10.68; p < .001, d = 1.78$) responses. The P2 for the AV-synchronous response also peaked sooner compared to the audio-only response ($p = .002, d = .64$). All other comparisons failed to reach statistical significance (all $p$'s $> .17$).

6.3. Summary and discussion of Experiment 2

The amplitude and latency effects seen in Experiment 1 were replicated in Experiment 2, providing further evidence that the early auditory response is sensitive to the temporal relationship between dynamic AV input. However, the effects of latency were not present at the N1, but were only seen at the P2 for Experiment 2. While we want to be careful in interpreting the N1 latency effects for Experiment 1, the perception of AV asynchrony, on average, delays the auditory response relative to each condition for both Experiments 1 and 2. Importantly, the partial replication of the effects of latency suggests they are overall less robust compared to the amplitude effects.

Of particular interest in Experiment 2 was the response pattern of the AV-occluded condition. We found that the auditory N1–P2 peak-to-peak amplitude response elicited during the AV-occluded condition was smaller compared to unimodal auditory and temporally asynchronous AV inputs, and closely mimicked the AV-synchronous response in amplitude and latency. These findings suggest that early auditory sensitivity toward the expectation of sounds can arise as the result of preceding visual input, without simultaneous audio and visual input. The AV-occluded P2 response also revealed a significant difference in speeded latency compared to the AV-asynchronous response. Importantly, the overall pattern of results suggests that the AV-occluded response closely resembled
the activity in the AV-synchronous condition. Overall, Experiment 2 demonstrated that visual expectation induced by a single moving object can facilitate early auditory processing, as most clearly indexed by the overall reduction of the auditory ERP.

7. General discussion

7.1. Early auditory processing is attenuated for synchronous audio-visual events

The goal of this study was to examine the electrophysiological correlates of dynamic AV temporal synchrony in the healthy adult brain. For Experiment 1, we were guided by the hypothesis that subtle manipulations of the temporal synchrony underlying dynamic AV events would result in unique patterns of neural responses underlying the auditory ERP, specifically the early auditory response. We found clear evidence that dynamic AV stimulation that differed in temporal onset synchrony subsequently altered the early sensory response (<200 msec) to sounds in fundamentally different ways. Specifically, the early auditory response to temporally synchronous AV events resulted in a pattern of reduced auditory processing (i.e., lower amplitude, faster peak latency) compared to discordant AV stimulation. A second experiment was conducted to assess early auditory responses toward
visualy occluded but temporally synchronous auditory input. We found that the N1–P2 peak-to-peak response toward AV-synchrony was similar to AV input that contained temporally synchronous, but visually occluded auditory input. These early sensitivities toward the temporal alignment of dynamic AV input demonstrate that general auditory processing is shaped by the temporal expectancies triggered by preceding dynamic visual input.

Our analyses for Experiment 1 revealed both an attenuated and accelerated auditory response when processing dynamic and temporally congruent AV inputs. These changes in the auditory evoked potential can be interpreted as very early auditory (<200 msec) processing being reduced during congruent AV conditions relative to incongruent AV (or audio-only) conditions, possibly indicating that participants coded the temporal synchrony of an expected sound generated by a moving stimulus very early in the auditory processing stream. Additionally, the AV-asynchronous response, where the sound occurred before it was expected, elicited a delayed auditory response compared to the response elicited from unimodal auditory events, likely reflecting a signature of detecting sensory conflict between the timing of the auditory and visual stimuli. The findings of early changes of the auditory evoked potential – amplitude reduction and shorter latency – for temporally congruent, dynamic AV stimuli provide more evidence for the idea that similar amplitude reductions (i.e., suppression) of the auditory response arises in scenarios in which preceding sensory input matches additional, yet expected sensory input (Clementz et al., 2002; D’Andrea-Penna et al., 2020; Ford et al., 2007; Ford & Hillyard, 1981; Kononowicz & van Rijn, 2014; Lange, 2009, 2013; Menceloglu et al., 2020; Schafer et al., 1981; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005; Vroomen & Stekelenburg, 2010). Note that no direct comparisons were made between visual-only presentations and the three other sensory conditions, due to the lack of an observed auditory ERP in the visual-only condition. Given that our visual stimulus was continuous rather than a discrete event, we were also not able to observe a clear visually-evoked response of the ball bouncing; however, our data clearly showed that the visual stimulus modulated auditory processing. We recognize that there are potentially important differences between the AV-synchronous and AV-asynchronous conditions. First, the visual position of the object during sound onset is different between the AV-synchronous and AV-asynchronous conditions, where the ball is closer to the origin for the AV-asynchronous condition, which makes a direct comparison of their baselines difficult. Second, the reversal of the object after the collision in the synchronous condition may have provided the observer additional visual information that was not seen during the AV-asynchronous condition. Because of these important, yet unavoidable sensory differences between the AV-synchronous and AV-asynchronous conditions, some degree of caution is needed when interpreting these results.

The present results deviate in some ways from the findings by Vroomen and Stekelenburg (2010). Specifically, Vroomen & Stekelenburg observed differences in the auditory response only when the visual input predicted the sound with high reliability (fixed blocks only; three condition comparison: audio-only, visual-only, and AV-synchronous expectations; their Experiment 1), but when the relation became less reliable (mixed vs. fixed blocks; six condition comparison: audio-only, visual-only, AV-synchronous expectation, AV-non expectation, and early and late AV-asynchronous sound onset; their Experiment 2), these early auditory modulations only appeared during fixed-block, AV-asynchronous expectation presentations. With regards to the later auditory response (~240 msec), they found an equally suppressed response toward synchronous AV and early asynchronous AV input regardless of the predictability of the audio-visual events (i.e., both in fixed and intermixed blocks). We, on the other hand, found that the auditory response was sensitive to AV asynchronies even when synchronous and asynchronous trial types were randomly intermixed, and the visual input thus did not reliably predict the timing of a sound across trials. The later auditory response appeared to show, on average, a general sensitivity to audio-visual inputs regardless of temporal synchrony, similar to Vroomen and Stekelenburg (2010). However, since we did not have clear a priori expectations with regards to the later auditory response in our tasks, we hesitate to strongly interpret these changes in the later auditory response.

Why do our results differ from those observed in the previous study? In the current set of experiments, a visual object provided expectations about upcoming auditory input in a single, uniform direction. In Vroomen and Stekelenburg (2010), two visual objects appeared to the left and right of a rectangle and moved toward it, eventually colliding with and bouncing off it. In our experiments, visual expectation, and therefore attention, was not divided between two objects, which may have reduced uncertainty, providing for a more accurate perceptual representation of the temporal relationships underlying the AV inputs. Additionally, the stimuli used in the current experiment were perhaps more reflective of natural sensory environments. For example, we opted to use a red sphere that appeared to move toward and bounce off a single, artificially defined barrier, eliciting a “knock” sound that resembled the sound of a ball bouncing of the wall. Vroomen and Stekelenburg used more simplified stimuli, including two white visual disks that elicited a pure auditory tone upon synchronous contact with an artificially defined barrier. Alternatively, the contrast between the results of our Experiment 1 and Vroomen and Stekelenburg (2010), where we found a difference in the neural response toward synchronous and asynchronous AV inputs during mixed trials while they did not, may have arisen due to differences in the temporal gap between discordant AV stimulation. In our experiments, the auditory onset for asynchronous AV input occurred 450 msec before visual collision. Vroomen and Stekelenburg presented auditory information 240 msec before visual collision (2010). The auditory response is thought to reflect an early sensory response that is modulated by low-level auditory characteristics like loudness and pitch (Hyde, 1997), and thus in principle could be sensitive to other low-level characteristics like systemically smaller asynchronous temporal gaps in dynamic AV input. We think it is unlikely that these small differences in the temporal onset of a sound paired with dynamic, yet discordant visual input drives the auditory response synchrony differences between the two studies.

We measured responses to AV asynchrony using a single offset in timing between the synchronous AV event and the
asynchronous one. However, differences in timing between when a sound is expected based on visual input and when it actually occurs could matter for how reliable the multisensory percept is. Additional studies will also be needed to assess the auditory response while perceiving systematically smaller temporal onsets, or even small delays between discordant AV input. Such research will help further characterize whether this mechanism relies more so on a general sensitivity toward the temporal expectation elicited by a moving object itself versus one that would rely on a unique multisensory interplay between the AV inputs.

### 7.2. Synchronous visual expectations about the timing of sounds reduces auditory responses

In Experiment 2, we asked whether expectation provided by continuous visual input preceding the sound was sufficient to elicit an attenuated auditory response even in the absence of the visual input continuing to the point of impact. The N1–P2 peak-to-peak response to audio-only and AV-asynchronous inputs were greater compared to sensory information that provided synchronous auditory stimulation but occluded visual information at the point of collision. Additionally, the AV-occluded condition was not statistically different in amplitude or latency from the AV-synchronous response. Thus, the reduced auditory amplitude toward occluded AV stimulation provides evidence for sensitivity toward the expectation of the impending sound in that the brain’s response resembled the perception of AV-synchronous input, even without a precise visual representation of the collision event. The human visual system displays a remarkable ability to represent the persistence of dynamic objects that undergo brief visual occlusion (see review, Scholl, 2007). Even six-month-old human infants are able to anticipate the exit trajectory of a briefly occluded visual object in motion (Johnson et al., 2003). Additionally, the ability to visually track and identify multiple target objects that undergo brief visual occlusion is unimpaired in normal sighted individuals (Scholl & Pylyshyn, 1999). In this case, tracking a briefly occluded visual object may help reduce the computational demands of auditory brain regions, allowing for the auditory system to better coordinate in time and space the physical properties of the occluded object (i.e., rate of motion, physical boundaries, etc.) with its expected sound. In other words, the suppressed auditory response elicited by occluded and synchronous AV inputs may have resulted from the brain generating successful predictions about when an expected sound would occur. We showed this by simply presenting a dynamic moving object for a relatively short time that contained accurate temporal and spatial information. These visual cues in essence allowed the perceiver to infer the source of the sound even without simultaneous visual input. It is worth noting that the visible condition contained more precise information about when the tone occurs, and thus the dissociation between expectation and integration ought to be interpreted carefully. Nonetheless, Experiment 2 underscores the importance of visual expectations in eliciting auditory suppression, in that the brief representation of a dynamic visual object’s spatial and temporal properties led to the expectation of an accompanying sound.

### 7.3. Summary

Early sensitivity toward the temporal synchrony of dynamic AV events is important for the successful identification of bimodal sensory signals that should be perceived as either unified or two separate sensory events. Sensitivity to the temporal relation between AV input allows the brain to either processes congruent AV events or detect asynchrony very early in the auditory processing stream. Thus, the reduction of the auditory response to AV temporal synchrony may manifest from the brain generating successful predictions about basic sensory events in the environment. In other words, the reduced auditory response seen here may have resulted from a perceptual match between top-down expectancies of a sound and correct bottom-up sensory input, like the ball’s motion and its synchronous relation to the timing of the sound itself. Such mechanisms may help lay the groundwork to further understand how neural activity is shaped in later processing stages that involve higher-level cognitive processes like attention and decision making. For example, one might direct less attention toward the low-level features of temporally synchronous AV events, while exerting more effort in extracting contextual information embedded within the AV signal. Conversely, AV input that is temporally asynchronous may evoke disturbances in fundamental mechanisms designed to bind bimodal sensory information as one. These highly specialized sensory processes afford the brain the ability to exhibit sensitivities toward relatively small temporal discrepancies between dynamic AV stimulation. Such mechanisms are crucial, as the successful integration of the expectation of an accompanying sound that arise via dynamic visual stimuli results in precise scene representations. Another interpretation of the data could be that the auditory response expectancy effects resulted from the pre-activation of the neural representation of the expected sound contained in the time prior to sound actually occurring (Blom et al., 2020; Kok et al., 2017). This is a very important distinction that can be addressed in future research. For example, sensory manipulations, like varying the speed of the object itself, the time spent behind occlusion, or a combination of both are needed to assess contextual occlusion influences over the visual ERP. The fact that the auditory response was not modified by occlusion but was different from unimodal input suggests that expectation plays some role in processing expected sounds based on the trajectory of a moving visual stimulus. The neural mechanism characterized in this study may be fundamental to proper AV processing more broadly, in that these early auditory responses are sensitive to temporal discrepancies between AV sensory input that differ in milliseconds.

In sum, our results provide evidence for a neural mechanism that is sensitive to the underlying temporal relation of dynamic AV input in healthy adults. Early sensitivities to temporally congruent and discordant events may help determine how the brain subsequently processes bimodal experiences in natural sensory environments. For synchronous events, the brain exhibited a reduced auditory response when the temporal predictions of an ensuing sound were in line with preceding visual input. In contrast, greater neural responses were seen in the presence of AV temporal incongruency. Early sensitivity toward the temporal synchrony of dynamic AV events, as reflected in the early auditory response...
modulations, may reflect a basic perceptual mechanism used to gage the plausibility of expected sensory events contained within our environment. Importantly, a moving visual object that provides accurate spatial and temporal expectations about when a sound is likely to appear are sufficient in attenuating early auditory processing. Broadly, this suggests that visual input – or the representation of visual input – leads to more efficient auditory processing within the first few hundred milliseconds of processing. These early influences likely have important consequences for other downstream processes.

Credit author statement

Andrew Marin: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original draft, Visualization, Project administration. Viola S. Stormer: Conceptualization, Methodology, Writing – Review & editing, Supervision. Leslie J. Carver: Conceptualization, Methodology, Resources, Writing – Review & editing, Supervision.

Open practices

The study in this article earned Open Data and open Materials badges for transparent practices. Data and materials for this study can be found at http://dx.doi.org/10.17632/k3j772tmwk.2.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The deidentified ERP data that support the findings of this study are available to download via an online data repository. The cleaned and segmented ERP data (.set/.fdt files), N1 and P2 peak amplitude and latency data sets, all programming scripts, and stimuli used for both experiments are available to download via a publicly available online data repository (https://dx.doi.org/10.17632/k3j772tmwk.4). Raw EEG data sets are also available for download via a separate online data repository (osf.io/d245g/).

Declaration of competing interest

None.

Acknowledgements

We would first like to thank Dr. Denise Werchan and Dr. Dima Amso for their willingness to share their stimuli with us. We would also like to acknowledge Zoë and Cameron Tauxe, and Dr. David Lewkowicz for the help with stimuli creation. Lastly, we would like to thank Jonathan Keefe and Elizabeth Baker for their helpful comments and edits when preparing this manuscript.

REFERENCES

Baldeweg, T. (2007). ERP repetition effects and mismatch negativity generation: A predictive coding perspective. Journal of Psychophysiology, 21(3–4), 204–213. https://doi.org/10.1027/0269-8803.21.3.204

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. European Journal of Neuroscience, 20(8), 2225–2234. https://doi.org/10.1111/j.1460-9568.2004.03870.x

Blom, T., Feuerriegel, D., Johnson, P., Bode, S., & Hogendoorn, H. (2020). Predictions drive neural representations of visual events ahead of incoming sensory information. Proceedings of the National Academy of Sciences, 117(13), 7510–7515. https://doi.org/10.1073/pnas.1917771117

Clementz, B. A., Barber, S. K., & Dzau, J. R. (2002). Knowledge of stimulus repetition affects the magnitude and spatial distribution of low-frequency event-related brain potentials. Audiology and Neuro-Otology, 7(5), 303–314. https://doi.org/10.1159/000064444

D’Andrea-Penna, G. M., Iversen, J. R., Chiba, A. A., Khalil, A. K., & Minces, V. H. (2020). One tap at a time: Correlating sensorimotor synchronization with brain signatures of temporal processing. Cerebral Cortex Communications, 1(1), tga036. https://doi.org/10.1093/ercem/tga036

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods, 134(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Ford, J. M., & Hillyard, S. A. (1981). Event-related potentials (ERPs) to interruptions of a steady rhythm. Psychophysiology, 18(3), 322–330. https://doi.org/10.1111/j.1469-8868.1981.tb03043.x

Ford, J. M., Roach, B. J., Faustman, W. O., & Mathalon, D. H. (2007). Synch before you speak: Auditory hallucinations in schizophrenia. The American Journal of Psychiatry, 164(3), 458–466. https://doi.org/10.1176/appi.ajp.2007.164.3.458

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. Cerebral Cortex, 12(10), 1031–1039. https://doi.org/10.1093/cercor/12.10.1031

Freunberger, R., Klimesch, W., Doppelmayr, M., & Höller, Y. (2007). Visual P2 component is related to theta phase-locking. Neuroscience Letters, 426(3), 181–186. https://doi.org/10.1016/j.neulet.2007.08.062

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. Journal of Cognitive Neuroscience, 11(5), 473–490. https://doi.org/10.1162/089892999563544

Godey, B., Schwartz, D., de Graaf, J. B., Chauvel, P., & Légeois-Chauvel, C. (2001). Neuronal source localization of auditory evoked fields and intracerebral evoked potentials: A comparison of data in the same patients. Clinical Neurophysiology, 112(10), 1850–1859. https://doi.org/10.1016/S1388-2457(01)00636-8

Hyde, M. (1997). The N1 response and its applications. Audiology and Neurotology, 2(5), 281–307. https://doi.org/10.1159/000259253
Johnson, S. P., Amso, D., & Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. Proceedings of the National Academy of Sciences, 100(18), 10568–10573. https://doi.org/10.1073/pnas.150655100

Kok, P., Mostert, P., & de Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. Proceedings of the National Academy of Sciences, 114(39), 10473–10478. https://doi.org/10.1073/pnas.1705652114

Kononowicz, T. W., & van Rijn, H. (2014). Decoupling interval

Meredith, M., Nemitz, J., & Stein, B. (1987). Determinants of

Mayhew, S. D., Dirckx, S. G., Niazy, R. K., Iannetti, G. D., &

Marin, A., Störmer, V. S., & Carver, L. J. (2021a). Expectations about dynamic visual objects facilitates early sensory processing of congruent sounds (Vol. 4). Mendeley Data. https://doi.org/10.17632/k377t2mwk.4

Marin, A., Störmer, V. S., & Carver, L. J. (2021b). Expectations about dynamic visual objects facilitates early sensory processing of congruent sounds (Raw Data). Retrieved from osf.io/d245g.

Mayhew, S. D., Dirckx, S. G., Niazy, R. K., Iannetti, G. D., & Wise, R. G. (2010). EEG signatures of auditory activity correlate with simultaneously recorded fMRI responses in humans. NeuroImage, 49(3), 849–864. https://doi.org/10.1016/j.neuroimage.2009.06.080

Mencelöglu, M., Graboweczyk, M., & Suzuki, S. (2020). Rhythm violation enhances auditory-evoked responses to the extent of overriding sensory adaptation in passive listening. Journal of Cognitive Neuroscience, 32(9), 1654–1671. https://doi.org/10.1162/jocn_a_01578

Meredith, M., Nemitz, J., & Stein, B. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. The Journal of Neuroscience, 7(10), 3215–3229. https://doi.org/10.1523/JNEUROSCI.07-10-03215.1987

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory audition—visual interactions during early sensory processing in humans: A high-density electrical mapping study. Cognitive Brain Research, 14(1), 115–128. https://doi.org/10.1016/S0926-6410(02)00066-6

Naätanen, R., & Winkler, R. (1999). The concept of auditory stimulus representation in cognitive neuroscience.

Psychological Bulletin, 125(6), 826. https://doi.org/10.1037/0033-2909.125.6.826

Picton, T. W., Hillyard, S. A., Krausz, H. I., & Galambos, R. (1974). Human auditory evoked potentials. I: Evaluation of components. Electroencephalography and Clinical Neurophysiology, 36, 179–190. https://doi.org/10.1016/0168-7523(74)90155-2

Ponton, C., Eggermont, J. J., Khosla, D., Kwong, B., & Don, M. (2002). Maturation of human central auditory system activity: Separating auditory evoked potentials by dipole source modeling. Clinical Neurophysiology, 113(5), 407–420. https://doi.org/10.1016/S1388-2457(01)00733-7

Psychology software Tools, Inc. [E-Prime 2.0.890]. Retrieved from https://www.pstnet.com, (2016).

Schafer, E. W. P., Amochea, A., & Russell, M. J. (1981). Knowledge of stimulus timing attenuates human evoked cortical potentials. Electroencephalography and Clinical Neurophysiology, 52(1), 9–17. https://doi.org/10.1016/0168-7523(81)90183-8

Scholl, B. J. (2007). Object persistence in philosophy and psychology. Mind & Language, 22(5), 563–591. https://doi.org/10.1111/j.1468-0017.2007.00321.x

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. Cognitive Psychology, 38(2), 259–290. https://doi.org/10.1006/cogp.1998.0698

Spence, C. (2007). Audiovisual multisensory integration. Acoustical Science and Technology, 28(2), 61–70. https://doi.org/10.1250/ast.28.61

Stecklenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. Journal of Cognitive Neuroscience, 19(12), 1964–1973. https://doi.org/10.1162/jocn.2007.19.12.1964

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., & Wallace, M. T. (2012). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. Experimental Brain Research, 219(1), 121–137. https://doi.org/10.1007/s00221-011-3072-1

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. Journal of Experimental Psychology. Human Perception and Performance, 38(6), 1517–1529. https://doi.org/10.1037/a0027339

Vroomen, J., & Steklenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. Journal of Cognitive Neuroscience, 22(7), 1583–1596. https://doi.org/10.1162/jocn.2009.21308

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, 102(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

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

Werchan, D. M., Baumgartner, H. A., Lewkowicz, D. J., & Amso, D. (2018). The origins of cortical multisensory dynamics: Evidence from human infants. Developmental Cognitive Neuroscience, 34, 75–81. https://doi.org/10.1016/j.dcn.2018.07.002